VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ
FACULTY OF BUSINESS AND MANAGEMENT

ÚSTAV INFORMATIKY
DEPARTMENT OF INFORMATICS

EXPERTNÍ SYSTÉM PRO ROZHODOVÁNÍ NA AKCIOVÝCH TRZÍCH S VYUŽITÍM SENTIMENTU INVESTORŮ

EXPERT SYSTEM FOR DECISION-MAKING ON STOCK MARKETS USING INVESTOR SENTIMENT

DIZERTAČNÍ PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE — **Ing. et Ing. Zuzana Janková**
AUTHOR

VEDOUCÍ PRÁCE — **prof. Ing. Petr Dostál, CSc.**
ADVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA PODNIKATELSKÁ

Zadání dizertační práce

Ústav:	Ústav informatiky
Studentka:	**Ing. et Ing. Zuzana Janková**
Vedoucí práce:	**prof. Ing. Petr Dostál, CSc.**
Akademický rok:	2020/21
Studijní program:	Ekonomika a management
Studijní obor:	Řízení a ekonomika podniku

Expertní systém pro rozhodování na akciových trzích s využitím sentimentu investorů

Charakteristika problematiky úkolu:

Úvod
Teoretické pozadí disertační práce
Kritický přezkum současného stavu
Cíle a metodologie disertační práce
Řešení a výsledky disertační práce
Diskuze disertační práce
Přínosy disertační práce
Závěr
Literatura
Přílohy

Cíle, kterých má být dosaženo:

Cíle: Hlavním cílem disertační práce je tvorba modelu expertního systému sloužící pro podporu investičního rozhodování na akciových trzích s využitím sentimentu investorů extrahovaného z textových zpráv.

Dílčí cíle disertační práce:
Cíl 1: Identifikovat z teoretického hlediska přístupy textové a sentimentální analýzy na akciových trzích a představit vhodný expertní systém sloužící k predikci.
Cíl 2: Zmapovat současný stav poznání z doposud publikovaných předních vědeckých článků a příspěvků v dané problematice.
Cíl 3: Klasifikovat skóre sentimentu z nestrukturovaných textových finančních zpráv a příspěvků zveřejněných online.
Cíl 4: Vytvořit expertní model predikce vývoje akciového trhu integrací extrahovaného skóre sentimentu investorů.

Cíl 5: Komparovat nově vytvořený expertní model se standardně užívanými modely predikce vývoje akciového trhu.

Základní literární prameny:

BUSTOS, O a A POMARES-QUIMBAYA. Stock market movement forecast: A Systematic review. Expert systems with applications. Elsevier, 2020, 156, 113464. ISSN 0957-4174.

DOSTÁL, Petr. Advanced decision making in business and public services. Brno: Akademické nakladatelství CERM, 2011, 167 s., grafy, tab. ISBN 978-80-7204-747-5.

LIU, Bing. Sentiment Analysis. Cambridge: Cambridge University Press, 2015. ISBN 9781139084789.

MENDEL, J.M, R.I JOHN a F LIU. Interval Type-2 Fuzzy Logic Systems Made Simple. IEEE transactions on fuzzy systems. PISCATAWAY: IEEE, 2006, 14(6), 808-821. ISSN 1063-6706.

ZADEH, Lofti. A. Fuzzy sets. Information and Control, 1965, 8, 338–353.

ZHAI, Cheng Xiang a Sean MASSUNG. Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. ACM, 2016.

Termín odevzdání dizertační práce je stanoven časovým plánem akademického roku 2020/21.

V Brně, dne 12. 6. 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA PODNIKATELSKÁ

L. S.

doc. Ing. et Ing. Stanislav Škapa, Ph.D.
předseda oborové rady

doc. Ing. Vojtěch Bartoš, Ph.D.
děkan

ABSTRAKT

Předložená disertační práce zkoumá potenciál využití skóre sentimentu extrahovaného z textových dat společně s historickými daty o akciovém indexu ke zlepšení výkonnosti predikce na akciovém trhu prostřednictvím vytvořeného modelu expertního systému. Vzhledem k tomu, že velké množství textových dokumentů souvisejících s financemi, které zveřejňují jak profesionální, tak amatérští investoři, nejen na online sociálních sítích, by mohly mít dopad na vývoj akciových trhů, je zásadním úkolem analyzovat finanční texty zveřejněné různými uživateli a zejména z nich extrahovat sentiment. V této práci je sentiment investorů získán zonline finančních zpráv a příspěvků zveřejněných na finanční sociální platformě StockTwits. Skóre sentimentu je stanoveno pomocí hybridního přístupu kombinující modely strojového učení s učitelem a neuronových sítí, přičemž ke klasifikaci polarity sentimentu je využito vícero lexikonů pozitivních a negativních slov. Je analyzován vliv skóre sentimentu na akciový trh prostřednictvím kauzality, kointegrace a koherence. V disertační práci je navržen model expertního systému založený na metodách fuzzy logiky. Fuzzy logika poskytuje pozoruhodné vlastnosti při práci s vágními, nepřesnými či nejasnými údaji a je schopna vypořádat se s chaotickým prostředím na akciových trzích. V nedávných vědeckých studiích na popularitě získává vyšší úroveň fuzzy logiky, která je označována jako type-2 fuzzy logika. Oproti klasické type-1 fuzzy logice, je tento vyšší typ schopen mezi zdvojené funkce členství integrovat určitou úroveň nejistoty. Tento typ expertního systému je ovšem v předmětné problematice predikce akciového trhu s využitím extrahovaného sentimentu investorů značně opomíjen. Z toho důvodu je v disertační práci zkoumán potenciál využít a výkonnost type-2 fuzzy logiky. Konkrétně je vytvořeno několik type-2 fuzzy modelů, které jsou trénovány na historických datech akciového indexu a skóre sentimentu investorů za období 2018-2020. Vytvořené modely jsou posouzeny k měření výkonu predikce bez sentimentu i s integrací sentimentu investorů. Následně je na základě vytvořeného expertního modelu stanovena investiční strategie a sledována jeho ziskovost. Výkonnost predikce fuzzy modelů je komparována s výkonností několika srovnávacích modelů, včetně SVM, k-NN, naivního Bayes a dalších. Z experimentů vyplynulo, že modely fuzzy logiky jsou schopny vhodným nastavením funkcí členství a nejistoty v nich obsažených vylepšit predikci a jsou schopny konkurovat klasickým modelům predikce, které jsou standardně využívané ve výzkumných studiích. Vytvořený model by měl sloužit jako nástroj pro podporu investičního rozhodování individuálním investorům.

KLÍČOVÁ SLOVA

Akciový trh; analýza sentimentu; expertní systém; fuzzy logika; rozhodovací model; sentiment investorů; textová analýza; type-1 fuzzy model; type-2 fuzzy model

ABSTRACT

The presented dissertation examines the potential of using the sentiment score extracted from textual data with historical stock index data to improve the performance of stock market prediction through the created model of the expert system. Given the large number of financial-related text documents published by both professional and amateur investors, not only on online social networks that could have an impact on real stock markets, but it is also crucial to analyze and in particular extract financial texts published by different users. investor sentiment. In this work, investor sentiment is obtained from online financial reports and contributions published on the financial social platform StockTwits. Sentiment scores are determined using a hybrid approach combining machine learning models with the teacher and neural networks, with multiple lexicons of positive and negative words used to classify sentiment polarity. The influence of sentiment score on the stock market through causality, cointegration and coherence is analyzed. The dissertation proposes a model of an expert system based on fuzzy logic methods. Fuzzy logic provides remarkable features when working with vague, inaccurate or unclear data and is able to deal with the chaotic environment of stock markets. In recent scientific studies, it has gained in popularity a higher level of fuzzy logic, which is referred to as type-2 fuzzy logic. Unlike the classic type-1 fuzzy logic, this higher type is able to integrate a certain level of uncertainty between the dual membership functions. However, this type of expert system is considerably neglected in the subject issue of stock market prediction using the extracted investor sentiment. For this reason, the dissertation examines the potential to use and the performance of type-2 fuzzy logic. Specifically, several type-2 fuzzy models are created. which are trained on historical stock index data and sentiment scores extracted from text data for the period 2018-2020. The created models are assessed to measure the prediction performance without sentiment and with the integration of investor sentiment. Subsequently, based on the created expert model, the investment strategy is determined, and its profitability is monitored. The prediction performance of fuzzy models is compared with the performance of several comparison models, including SVM, k-NN, naive Bayes and others. It has been observed from experiments that fuzzy logic models are able to improve prediction by appropriate setting of membership and uncertainty functions contained in them and are able to compete with classical expert prediction models, which are standardly used in research studies. The created model should serve as a tool to support investment decisions for individual investors.

KEYWORDS

Stock market; sentiment analysis; expert system; fuzzy logic; decision model; investor sentiment; text analysis; type-1 fuzzy model; type-2 fuzzy model

JANKOVÁ, Zuzana. *Expertní systém pro rozhodování na akciových trzích s využitím sentimentu investorů*. Brno: Vysoké učení technické v Brně, Fakulta podnikatelská, Ústav informatiky, 2021, 237 s. Dizertační práce. Vedoucí práce: prof. Ing. Petr Dostál, CSc.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Ing. et Ing. Zuzana Janková
VUT ID autora:	163485
Typ práce:	Dizertační práce
Akademický rok:	2021/2022
Téma závěrečné práce:	Expertní systém pro rozhodování na akciových trzích s využitím sentimentu investorů

Prohlašuji, že svou závěrečnou práci jsem vypracovala samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

„V každém okamžiku našeho života bychom měli být tak vděční, jak jen je to vůbec možné.“

— Seneca, římský filozof

Na tomto místě bych ráda poděkovala svému školiteli prof. Ing. Petru Dostálovi, CSc. za dlouhodobou spolupráci, cenné rady v průběhu celého doktorského studia a dále doc. Ing. Radku Doskočilovi, Ph.D., za připomínky při zpracování disertační práce.

Mé poděkování patří rovněž doc. RNDr. Bedřichu Půžovi, CSc., Mgr. Veronice Novotné, Ph.D., a ostatním kolegům Ústavu informatiky, Fakulty podnikatelské VUT v Brně, za vytvoření vhodných studijní a pracovních podmínkek pro zpracování této práce.

Děkuji také kolegům ze zahraničních pracovišť, jmenovitě Chengchi University na Taiwanu a Ekonomické univerzitě v Bratislavě, zejména doc. Ing. Miroslavu Hudcovi, Ph.D., za navázání vědecko-výzkumné spolupráce a v nespodlední řadě také Haldia Institute of Technology v Indii za sdílení poznatků v oblasti vědeckého bádání.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Finanční trhy zaujímají v moderní společnosti významné postavení. Úspěšné předpovídání cenových pohybů finančních instrumentů (tj. akcií, dluhových obligací a derivátů) může potenciálně zabránit škodlivým dopadům, které by mohla čekající finanční krize mít na každodenní život. Zejména akciové trhy vykazují vysokou míru chaotičnosti, nenadálých pohybů a jejich vývoj je předmětem zájmu mnoha výzkumníků. Jak dále uvádějí Hao a spol. (2021), je aspirací každého investora přesná předpověď tržního chování zaměřená na nejlepší rozhodnutí, pokud jde o nákup nebo prodej akcií usilujících o maximalizaci jeho zisků a snížení neočekávaných rizik. To je obtížný úkol, protože tržní chování je stochastické, nestálé a ovlivněné mnoha faktory, jako je globální ekonomika, politika, očekávání investorů a další. Tato problematika přitahovala pozornost mnoha akademiků a investorů. Většina studií předpovídání trendů cen akcií závisela na historických datech z akciových trhů, jako je cena a objem, jak popisuje Sezer a Ozbayoglu (2018). Ačkoli není obtížné získat historická data pro akciový trh, nevýhodou této metody je, že obvykle zanedbává kritické zpravodajské události. Ukázalo se, že některé události, jako jsou fúze a akvizice, kolísání zisku a změny vrcholového vedení ovlivňují ceny akcií dle Lien Minh a kol. (2018). Tyto kritické události, které obvykle zveřejňují poskytovatelé finančních zpráv, jako jsou Reuters nebo Bloomberg, jsou veřejným investorům vysílány jako zprávy. Ačkoli analýza těchto informací o trhu může zvýšit zisky, objem zpravodajských článků prudce vzrostl. S tak velkým objemem informací se organizace stále častěji spoléhají na vysokorychlostní počítačové zpracování, aby analyzovaly a vytvořily systémy podpory rozhodování pro předpovídání budoucích trendů a pomohly investorům učinit efektivnější rozhodnutí. Uvádí se, že mnoho systémů pro předpovídání cen akcií založených na článcích o finančních zprávách dokáže předpovídat cenové trendy (Matsubara a kol., 2018). Celkový proces toho, jak mohou zpravodajské události ovlivnit cenu akcií, je následující: událost se vysílá prostřednictvím finančních zpráv. Zprávy jsou interpretovány investory a poté přeneseny do záměrů nákupu/prodeje. Konečné rozhodnutí činí investoři. Nakonec je cena akcií ovlivněna volbou každého investora a odráží se v konečném cenovém trendu.

Behaviorální finanční teorie rozporuje základní zásady teorie efektivních trhů a ukazuje, že neefektivitu na akciových trzích lze vysvětlit tím, že budou splněny dva základní předpoklady. Prvním je, že investoři jsou pod vlivem sentimentu, což může vést ke špatnému určení ceny. Sentimentem se myslí názory investorů na budoucí peněžní toky a investiční riziko, které ovšem nejsou podloženy jednoznačnými fakty. Druhý předpoklad říká, že na trhu existuje limitovaná arbitráž – oprava špatného ocenění může být nákladná a riziková, takže se do ní arbitrážní spekulanti mnohdy vůbec nepouštějí. Toto vše má za následek, že ceny na akciovém trhu jsou

ovlivněny emocemi, náladami a názory účastníků trhu (Kaplanski a Levy, 2010). Analýza sentimentu je jednou z nejrychleji rostoucích oblastí výzkumu v informatice, což ztěžuje sledování všech aktivit v této oblasti. Kořeny analýzy sentimentu jsou ve studiích analýzy veřejného mínění na počátku 20. století a v analýze subjektivity textu provedené komunitou výpočetní lingvistiky v 90. letech. K propuknutí počítačové analýzy sentimentu však došlo pouze s dostupností subjektivních textů na webu. V posledních letech se analýza sentimentu přesunula z analýzy online recenzí produktů na sociální média z Twitteru a Facebooku dle Mäntylä a spol. (2018).

Vzhledem k posledním pokrokům v analýze sociálních sítí bylo možné zahrnout tento druh informací jako prediktivní vstup do prognostických modelů. Kromě toho se vyvinuly techniky zpracování textu, které nyní umožňují extrahovat složitější informace z textu. Nestrukturovaná data ve formě digitalizovaného textu rychle rostou, co se týká objemu, dostupnosti a významu. Zatímco tradiční pokusy o analýzu textu (tj. kvalitativní analýza) jsou omezené při zpracování velkého množství dat, dolování textu představuje soubor přístupů, které vědcům umožňují efektivně prozkoumat rozsáhlé sbírky textů dle Antons a kol. (2020). Studium sociálních sítí a vyhledávačů k předpovědi akciového trhu je relativně novým oborem. Kromě problémů při zpracování nestrukturovaných informací, jako jsou zprávy, je objem zpracovávaných informací enormní, což může překročit miliony záznamů a způsobit výpočetní problémy. Z této složitosti vyplývá, že ke zpracování je nutné použít specializované techniky a výkonné stroje. Na rozdíl od zpráv také komunikace na sociálních sítích není obvykle psána ve standardním formátu a může obsahovat pravopisné chyby a emotikony dle Porshnev a kol. (2015); Bustos a Pomares-Quimbaya (2020).

Na základě výše uvedených informací lze konstatovat, že ceny akcií jsou ovlivňovány jak publikovanými fundamentálními informacemi, tak myšlenkovými pochody v hlavách jednotlivých účastníků trhu (které obvykle nevznikají racionálně). Tyto dva vlivy se prolínají a působí současně. Bylo by proto zajímavé zkoumat, zda a jakou souvislost mají texty a pohyby akciového trhu. Texty mohou vyjadřovat jak fundamentální fakta (racionalita), tak emoce a názory lidí (iracionalita) a lze tak zkoumat, jaký vliv má (i)racionalita investorů na akciový trh. K tomuto účelu je ovšem nezbytné navrhnout vhodný expertní systém, který je schopen integrovat sentiment extrahovaný z textových dat vyjadřující pozitivní či negativní názory, postoje a emoce účastníků na trhu a tím usnadnit jejich investičního rozhodování. Cílem investování je především zisk, který motivuje investory ke vzdání se části svých peněžních prostředků a investovat je s vidinou toho, že se jim za určitý čas vrátí větší objem peněžních prostředků, než investovali. Tento prospěch plynoucí z investování má význam nejen pro samotné investory, ale potažmo i pro ekonomiku jako celek. Proto je pozornost zaměřena na návrh a aplikaci vhodného expertního modelu s cílem identifikovat ziskové příležitosti na akciovém trhu.

Relevance, aktuálnost a motivace

Zkoumanou výzkumnou oblastí jsou akciové trhy, které jsou předmětem zájmu mnoha výzkumníků s cílem získání užitečných vzorců a předpovědí jejich budoucích pohybů. Důvodem je, že kdo může porazit trh, může získat nadměrný zisk. Finanční analytici, kteří investují na akciových trzích, čelí problému obchodování s akciemi, které mají koupit a které prodat, aby získali větší zisky. Pokud dokáží předpovědět budoucí chování cen akcií, mohou podle toho okamžitě jednat a dosáhnout zisku. Čím přesnější systém předpovídá pohyb cen akcií, tím větší zisk lze z predikčního modelu získat. Prognóza vývoje cen akcií se těší velké oblibě pouze na základě technické a fundamentální analýzy dat. Číselná data časových řad však obsahují pouze událost a nikoli příčinu, proč k ní došlo. Textová data, jako jsou novinové články, mají bohatší informace, a proto využívání textových informací zejména kromě číselných dat časové řady zvyšuje kvalitu vstupu a očekávají se lepší předpovědi z tohoto druhu vstupu, nikoli pouze z číselných údajů. Bezpochyby je lidské chování vždy ovlivněno jejich prostředím. Jeden z nejvýznamnějších dopadů, které ovlivňují chování lidí, pochází z hromadných sdělovacích prostředků, konkrétněji ze zpravodajských článků. Na druhé straně jsou pohyby cen na akciových trzích důsledkem opatření investorů na to, jak vnímají události, které je obklopují, i na akciových trzích. Protože novinové články ovlivní rozhodnutí lidí a lidské rozhodnutí ovlivní ceny akcií, novinové články zase ovlivní akciový trh nepřímo. Na internetu je široce dostupné rostoucí množství důležitých a cenných zpravodajských článků v reálném čase, které se týkají akciových trhů. Extrakce cenných informací a zjištění vztahu mezi extrahovanými informacemi a akciové trhy je zásadní otázkou, protože pomáhá finančnímu analytikovi předpovědět chování akciového trhu a získat nadměrný zisk.

Motivací ke studiu a naléhavosti řešení předmětné problematiky je skutečnost, že lidé mají obvykle omezenou kapacitu zpracování informací, a proto se rozhodují pod časovým tlakem dle Gabaix a kol. (2003). Rychlá šíření informací umožňuje lidem okamžitě přijímat relevantní informace a provádět okamžité obchodní akce. V tomto rámci nedávné studie posunuly své zaměření na pozornost médií, neboť sdělovací prostředky hrají zásadní roli v procesu určování, kterým problémům bude věnována nízká nebo vysoká úroveň pozornosti veřejnosti. Pozornost investorů může také ovlivnit návratnost akciového trhu. Sociální média používají lidé nejen ke sdílení svých názorů, ale také k vyjadřování svých nálad. Z pohledu behaviorálního financování vědci prokázali, že akciový trh může být řízen emocemi účastníků trhu dle Sun a kol. (2020). Nicméně v současnosti jsou k dispozici stále rozporuplné názory a výstupy ohledně vlivu zveřejňovaných názorů a myšlenek na vývoj akciových trhů. Například Nguyen a kol. (2015) uvádí, že některé studie tvrdí, že sentiment sociálních médií má slabou nebo žádnou prediktivní sílu, zatímco jiné tvrdí, že so-

ciální média vykazují silné prediktivní schopnosti Nti a kol. (2020). Proto je použití sentimentu v sociálních médiích pro predikci cen na akciovém trhu stále otevřeným problémem výzkumu.

V současné době není výjimkou, že jeden příspěvek na sociálních sítí může zásadním způsobem rozkolísat nejen akciové trhy, ale také ostatní segmenty finančního trhu, což přispívá k aktuálnosti řešení předmětné problematiky. Kupříkladu jeden příspěvek na sociální síti Twitter, který zveřejnil Elon Musk, generální ředitel společnosti Tesla, znehodnotil prakticky celý trh s kryptoměnami o desítky procent. To se ovšem neděje jen s virtuálními měnami, ale také na akciových trzích. Síla, kterou může Twitter mít nad akciovým trhem, byla prokázána v minulosti. 30. března 2015 Elon Musk, tweetoval:

„Major new Tesla product line – not a car – will be unveiled at our Hawthorne Design Studio on Thurs 8pm, April 30".

Tím vzrostla hodnota společnosti na akciovém trhu o přibližně jednu miliardu USD za pouhých pár minut. Podobnou poznámku uvedl investor Carl Icahn ze dne 13. srpna 2013:

„We currently have a large position in APPLE. We believe the company to be extremely undervalued. Spoke to Tim Cook today. More to come".

To posílilo kapitalizaci o více než deset miliard USD. 23. dubna 2013 byl na twitterový účet americké tiskové organizace Association Press hacknut a hackeři zveřejnili:

"Breaking: Two Explosions in the White House and Barack Obama is injured".

Následoval pokles indexu S&P téměř o 1 %, což znamená, že investoři v daném časovém období z trhu odebrali více než 130 miliard USD. Důležitost Twitteru také dokazuje skutečnost, že společnosti začínají na Twitteru zveřejňovat informace o změnách cen akcií, i když tyto informace mohou být k dispozici i jinde, dle Simoes a kol. (2017). To přispívá k motivaci dalších přezkumu současných publikovaných výstupů.

Logická struktura disertační práce

Disertační práce je logicky strukturována do kapitol, jejichž schéma je zobrazeno na obrázku 1. **Úvod** – Obsahuje úvod do zkoumané problematiky včetně zasazení do širších souvislostí a zdůvodnění naléhavosti a důležitosti řešení problematiky.

Kapitola 1: Teoretické pozadí výzkumu – Tato kapitola prezentuje teoretické pozadí dolování textu a následnou analýzu sentimentu. Diskutuje také o hlavních přístupech a technikách klasifikace sentimentu včetně metrik evaluace klasifikátorů. V této

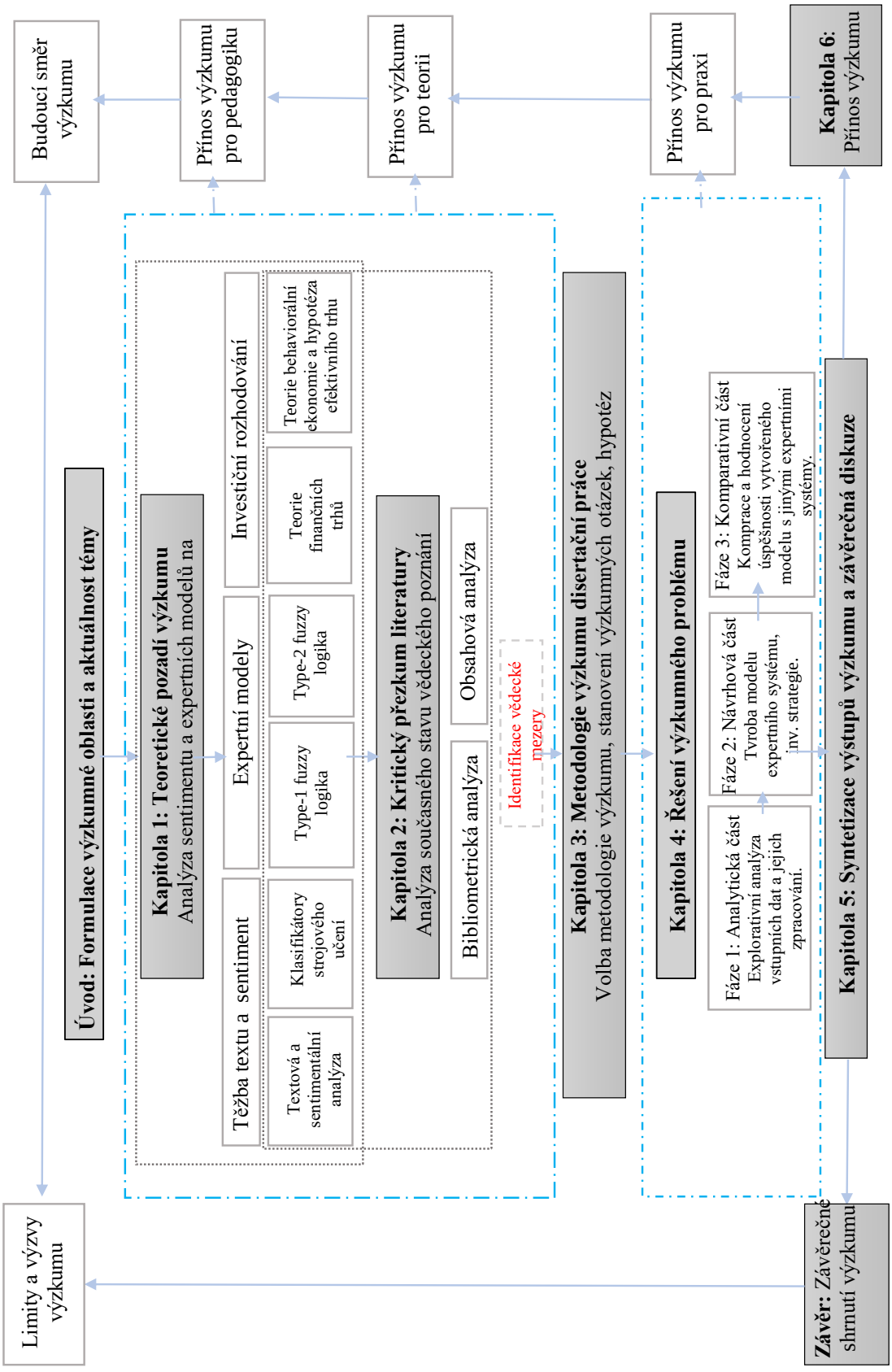
kapitole také představeno finanční pozadí výzkumu a expertních rozhodovacích modelů. Autorka práce považuje tuto část za nezbytnou, neboť doposud není k dispozici obdobný popis předmětné tématiky v českém jazyce a považuje tuto kapitola za příručku nejen pro odborníky v této oblasti, ale snahou bylo zpracovat tuto kapitolu co možná nejsrozumitelněji i pro širokou veřejnost.

Kapitola 2: Kritický přezkum současného stavu vědeckého poznání – V této kapitole je proveden kritický přezkum významných mezinárodních článků a příspěvků od renomovaných vědecko-výzkumných pracovníku získaných ze světových databází. Je zde využito grafické znázornění souvislostí a implikací jednotlivých výzkumných studií prostřednictvím bibliometrické analýzy. Následuje klíčová část sekundárního výzkumu, která syntetizuje současné poznatky v oblasti analýzy sentimentu na akciových trzích. Na základě této části jsou identifikována tzv. bílá místa či mezer ve výzkumu, které se pokusí tato disertační práce vyplnit. To je zásadní ke stanovení výzkumných hypotéz, které jsou v předkládané práci dále rozpracovány a řešeny.

Kapitola 3: Cíle a metodologie výzkumu disertační práce – Tato kapitola představuje formulaci cílů předkládané práce a výzkumných otázek a hypotéz řešených v průběhu výzkum, které vyplývají z kritického přezkumu. V rámci metodologie výzkumu je vycházeno z tzv. výzkumné cibule. Tato část definuje filozofii výzkumu, vědecké přístupy a metody včetně východisek řešení, staví na výsledcích sekundárního výzkumu a zahrnuje především informace o teoretických základech a definicích výzkumného rámce a zvoleném metodologickém přístupu výzkumu.

Kapitola 4: Řešení výzkumného problému – V této kapitole je provedena praktická aplikace zvolených metod na vybraném datovém vzorku. Zde je proveden explorační analýza vstupních dat z hlediska numerické a textové dimenze. Následně je provedena analýza sentimentu za využití vícero slovníků a následných klasifikačních technik. Podrobně je zkoumána výkonnost zvolených klasifikátorů. Po stanovení sentimentu investorů z textových dat je sledován jeho vliv na akciové trhy. Poté je v návrhové části vytvořen vlastní model založen na expertním systému. Je také generována obchodní strategie založená na sentimentu. V komparativní části je vytvořený model komparován s jinými expertními modely.

Kapitola 5: Diskuze, limity a náměty pro budoucí výzkum – Zde jsou zjištěné výsledky syntetizovány a verifikovány, dále rozebírány a diskutovány v kontextu již známých výstupů. Je upozorněno na zásadní limity a omezení předkládané studie s možnými náměti pro směřování budoucího výzkumu. **Kapitola 6**: Přínosy disertační práce – Tato předposlední kapitola uvádí přínosy této disertační práce ve třech dimenzích: vědecké, pedagogické a praktické. A na závěr jsou shrnuty a syntetizovány výstupy předkládané disertační práce.



Obr. 1: Konceptuální schéma disertační práce
Zdroj: vlastní zpracování

1 Teoretické pozadí disertační práce

Účelem této kapitoly je představit teoretické pozadí předmětné problematiky, ze které se bude vycházet v následujících částech zpracování disertační práce. Konkrétně podkapitola 1.1 představuje finanční pozadí výzkumu se zaměřením na akciové trhy, hypotézu efektivních trhů a behaviorální finance. V podkapitole 1.2 jsou uvedena teoretická východiska textové analýzy. Podkapitola 1.3 se zabývá analýzou sentimentu s využitím předdefinovaných slovníků a lexikonů včetně technik strojového učení, které se při analýze sentimentu využívají. Podkapitola 1.4 se zabývá modely expertních systémů, konkrétně fuzzy logikou včetně evaluačních metrika k vyhodnocení výkonnosti modelů.

1.1 Finanční pozadí

„Není nic bolestivějšího než přihlížet burzovnímu vzestupu a nebýt při tom. Bolí to dokonce víc než koupit akcie a prodělat.“

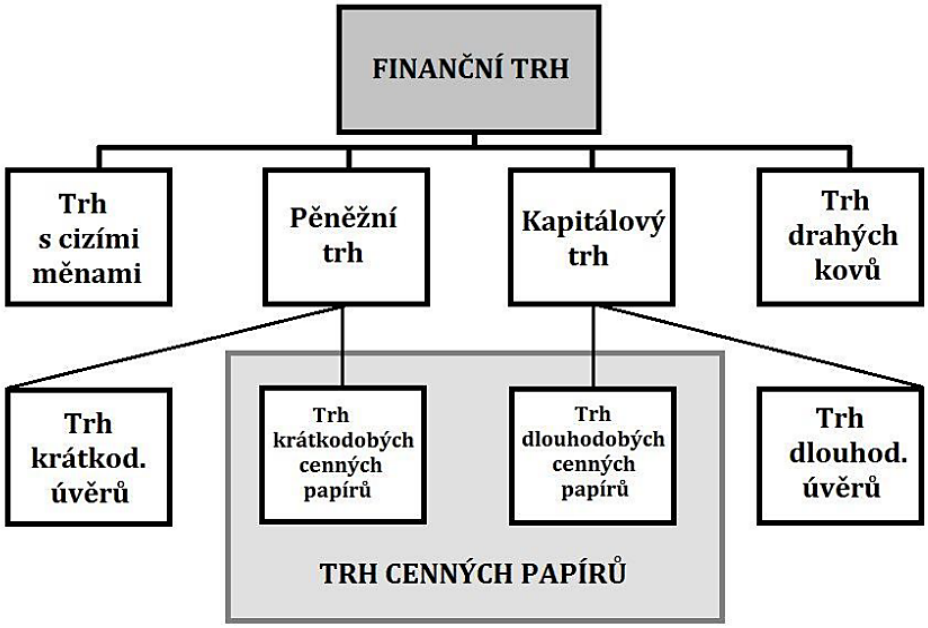
— Kostolany(2000)

Finanční trhy jsou jedním z nezastupitelných prvků soustavy tržní ekonomiky, organickou součástí její struktury. Na finančních trzích dochází k transferu disponibilních peněžních prostředků od přebytkových jednotek k jednotkám deficitním. Jinými slovy lze finanční trh popsat jako systém různých institucí a finančních instrumentů, pomocí nichž lze přesouvat kapitál od subjektů, kteří mají v daný moment nadbytek finančních prostředků, k těm, které v daný moment mají nedostatek finančních prostředků, a to na základě poptávky a nabídky. Finanční trh je tak nezbytnou podmínkou pro fungování každé tržní, resp. smíšené ekonomiky. Jeho zásadní význam spočívá ve schopnosti efektivně alokovat finanční prostředky na tržní bázi. Finanční trhy přímo ovlivňují bohatství jednotlivců, chování firem a domácností a přispívá k cykličnosti ekonomiky. Z makroekonomického hlediska, jak poznamenává Černohorský y(2020), jsou jedním z faktorů cykličnosti ekonomiky změny investic a jejich propojení skrz multiplikátor a akcelerátor s výstupem ekonomiky.

Finanční trhy vykazují určitou institucionální podobu, jsou vnitřně strukturované a tvořené řadou dílčích trhů a jejich segmentů, jak je znázorněno na obrázku 1.1. Finanční trhy je možné klasifikovat dle různých hledisek. Nejčastější forma členění vychází z jednotlivých druhů finančních instrumentů, které se na jednotlivých dílčích segmentech finančního trhu obchodují.

Na peněžním trhu jsou obchodovány finanční investiční nástroje s dobou splatnosti nejvýše jeden rok (krátkodobé úvěry nebo krátkodobé cenné papíry jako jsou směnky nebo pokladniční poukázky), na kapitálovém trhu jsou obchodovány finanční

investiční nástroje s delší dobou splatnosti (dlouhodobé úvěry nebo dlouhodobé cenné papíry jako jsou dluhopisy, hypoteční zástavní listy atp.) nebo majetkové cenné papíry (např. akcie nebo kmenové listy). Vedle toho však lze rozlišovat i další druhy finančních trhů, například trh devizový, jehož funkcí je umožnit nákup a prodej jednotlivých měn a dále trh drahých kovů. Hlavní funkcí kapitálového trhu je umožnit nepřímého dlouhodobého financování subjektů, které chtějí rozvíjet své obchodní aktivity, přičemž zdrojem jejich financování jsou peněžní prostředky investorů. Zde se jedná o primární trh, na kterém jsou nově emitované finanční investiční nástroje směňovány za peněžní prostředky investorů. Kapitálový trh však následně umožňuje také zpeněžení nakoupených finančních investičních nástrojů v okamžiku, kdy jejich držitelé pocítí potřebu likvidnějších peněžních prostředků. Kapitálový trh tak zajišťuje žádoucí likviditu finančních investičních nástrojů, když organizuje uzavírání smluvních vztahů, jejichž předmětem jsou již dříve emitované finanční investiční nástroje, tato součást kapitálového trhu je označena jako sekundární trh. Kohajda a Bakeš (2018).



Obr. 1.1: Základní členění finančního trhu
Zdroj: Rejnuš (2014)

Kapitálový trh lze členit dle druhu cenného papíru na trh akciový a trh dluhopisů. Trh dluhopisů reprezentuje všechny veřejně obchodované dluhopisy státních a veřejných institucí a korporátní dluhopisy. Akciovým trhem se zpravidla rozumí všechny veřejně obchodované, akcie a podílové listy. Předmětem zájmu této disertační práce je trh akciový, neboť se jedná o nepostradatelnou součást finančního systému, který podporuje ekonomický růst přitahováním zdrojů financování z odvětví s velkým ka-

pitálem k relativně menšímu kapitálu (Mamun a kol., 2018). A také důležitější pro rozvojové země, kde jsou omezené možnosti půjček a zejména dlouhodobých půjček v bankovním sektoru (Mauro, 2003). Rozvoj akciového trhu v ekonomice může zajistit lepší rozdělení zdrojů alespoň dvěma různými způsoby. První snižuje náklady na informování investorů o investičních projektech navržených společnostmi (King a Levine, 1993). Zadruhé, akciové trhy hrají rozhodující roli při ochraně zájmů investorů šířením a sdružováním rizik. To umožňuje investorům a firmám přijímat vhodná investiční rozhodnutí (Mamun a kol., 2018). Likvidita vytvořená na akciových trzích navíc umožňuje investorům nakupovat nebo prodávat akcie bez narušení jejich dlouhodobých investičních plánů a zároveň poskytuje společnostem dlouhodobý kapitál dle Tekin a Yener (2019).

Indexy akciového trhu jsou měřítkem cenové výkonnosti akciových portfolií, která jsou vytvořena tak, aby představovala akciový trh jako celek nebo konkrétní segment trhu dle Darškuvienė (2010). Známé indexy jsou například Dow Jones Industrial Average, Standard & Poor's 500 ve Spojených státech; FTSE 100 ve Velké Británii; Nikkei 225 v Japonsku; DAX v Německu; CAC 40 ve Francii; a Hang Seng v Hongkongu. Vnitrostátní akciové trhy mají alespoň jeden index a některé země s dobře rozvinutými akciovými trhy (zejména USA) mají četné indexy.

Základní kategorizace akciového indexu, dle Rejnuše (2014), je na cenové a hodnotově vážené. Cenově vážené akciové indexy mají váhu jednotlivých akcií stanovenou hodnotou jejich kurzu. Z toho je zřejmé, že jsou citlivé především na kurzové výkyvy společností s vysokou hodnotou kurzu bez ohledu na skutečnost, o jak významné, resp. o jak velké společnosti se jedná. Hodnotově vážené akciové indexy určují váhu akciových titulů nejen dle aktuální výše kurzu, ale také dle množství vydaných kusů (tržní kapitalizace), popřípadě dle jiného obdobného kritéria či kombinací souboru kritérií. Z toho vyplývá, že velké a likvidní společnosti mají na hodnotu hodnotově vážených indexů větší vliv než společnosti menší a méně likvidní.

Akciové indexy mají řadu využití:

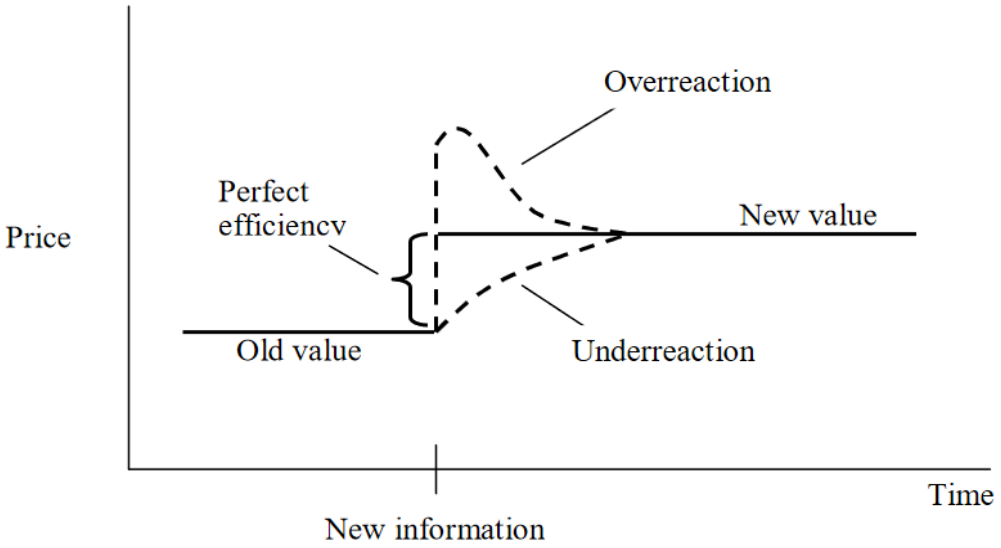
- měřit a sledovat pohyby trhu;
- poskytovat prostředky ke zjišťování změn v souhrnném bohatství v průběhu času;
- plnit roli barometrů ekonomiky; zejména pohyby na akciových trzích mají tendenci být předními ukazateli a poskytují náznaky pravděpodobných budoucích změn v úrovni aktivity v ekonomice jako celku;
- poskytovat rámec pro vytváření fondů, který má odrážet výkonnost akciového trhu;
- být používán modely kapitálového trhu, zejména modelem oceňování kapitálových aktiv (pro diskontní sazby u kapitálových projektů, odhad požadované míry návratnosti akcií aj.).

1.1.1 Hypotéza efektivních trhů

Hypotézu efektivního trhu *(angl. Efficient Market Hypothesis,* EMH) původně navrhl Fama (1965) a uvádí, že ceny cenných papírů vždy plně odrážejí všechny dostupné informace na efektivních kapitálových trzích. Jinými slovy, je nemožné vydělat zisky obchodováním na základě dostupných informací (Jensen, 1978). Tato silná hypotéza by však znamenala, že investoři potřebují pobídky k obchodování, dokud ceny ve skutečnosti neodrážejí všechny informace. Náklady na získání informací a transakční náklady by tedy musely být nulové, což zjevně neplatí (Grossman a Stiglitz, 1980). Reálnější se jeví méně omezená definice EMH, která uvádí, že ceny cenných papírů odrážejí pouze všechny informace, dokud mezní obchodní výhody (zisky) nepřekročí mezní obchodní náklady (Fama, 1991). Testy EMH se obvykle zabývají otázkou, jak rychle se informace odráží v cenách cenných papírů. Existují tři různé kategorie forem EMH, přičemž každá zvažuje jinou podmnožinu informací (Elton a kol., 2011; Fama, 1970):

- slabá forma efektivity trhu naznačuje, že aktuální ceny aktiv odrážejí všechny minulé ceny a cenové pohyby, tj. všechny informace související s obchodem. Jinými slovy, ke stanovení dnešní ceny byly použity všechny užitečné informace o předchozích cenách akcií. Investor tedy nemůže použít stejné informace k předpovědi zítřejší ceny, a přesto vydělat neobvyklý zisk. Empirické důkazy většiny světových akciových trhů naznačují, že tyto trhy jsou slabě účinné. Jinými slovy, není možné překonat („porazit“) trh pomocí informací o minulých cenách akcií.
- středně silná forma efektivity trhu naznačuje, že současné ceny akcií odrážejí všechny veřejně dostupné informace. Rozdíl mezi veřejnými informacemi a informacemi souvisejícími s trhem spočívá v tom, že veřejné informace zahrnují také oznámení o událostech společnosti, ekonomických a politických zprávách a událostech. Pokud tedy investoři používají investiční strategie založené na použití veřejně dostupných informací, nemohou vydělávat neobvyklé zisky. Neznamená to, že se ceny okamžitě mění, aby odrážely nové informace, ale spíše to, že informace jsou rychle zohledňovány v cenách akcií. Na ceny akcií mají vliv pouze neočekávané informace nebo překvapení.
- silná forma efektivity trhu předpokládá, že ceny aktiv odrážejí všechny veřejné a soukromé informace. Jinými slovy, trh (který zahrnuje všechny investory) ví všechno o všech cenných papírech, včetně informací, které nebyly zveřejněny. Silná forma znamená, že z obchodování nemůžete vydělat neobvyklé výnosy z interních informací, kde vnitřní informace jsou informace, které ještě nejsou veřejné (Darškuvienė, 2010).

Investoři dosud nedosáhli žádného kompromisu ohledně platnosti této hypotéz a jejího obsahu. Z praktické stránky se investoři zajímají o relevanci EMH pokaždé, když se snaží určit tržní trend pomocí technické analýzy, nebo podhodnocený nebo nadhodnocený titul pomocí fundamentální analýzy, nebo kdykoli se rozhodnou umístit fixní kapitál do pasivního podílového fondu nebo sledovat obecný index akcií. Pokud by investoři uznali neomylnost EMH, vyhnuli by se aktivní správě portfolia ve prospěch pasivního a nepodporovali by kontinuální fluktuaci a kompromisní proces, nezbytný pro udržení efektivního trhu. Podle EMH cena aktiva odráží všechny relevantní a dostupné informace v daném okamžiku: finanční, ekonomické, odvětvové, politické a sociální prostředí, a co je nejdůležitější, pocit investora k příslušnému aktivu dle Braşoveanu (2011) a Suciu (2015).



Obr. 1.2: Reakce trhu na nové informace
Zdroj: Darškuvienė (2010)

Panuje všeobecná shoda, že investoři nadměrně nebo nedostatečně reagují na informace. To neznamená, že trhy jsou neúčinné, pokud reakce nebude zaujatá (důsledně nadměrně nebo nedostatečně reaguje). V takovém případě bude investor, který rozpozná předpojatost, schopen dosáhnout neobvykle vysokého rizika upraveného výnosu. Tento koncept je znázorněn na obrázku 1.2. Předtím, než se informace stanou veřejnými, má aktivum starou hodnotu. Nové informace se dostávají na trh. Na efektivním trhu se cena okamžitě přizpůsobuje své nové rovnovážné úrovni. Pokud je trh neefektivní, může dojít k nadměrné reakci na nové informace. Pokud dojde k podhodnocení, je úprava ceny postupná. Při nadměrné reakci, tržní cena překročí novou rovnovážnou hodnotu. Pokud je trh neefektivní, mezi dobou odhalení zpráv a přizpůsobením se nové rovnovážné hodnotě by informovaní investoři byli schopni profitovat na úkor méně sofistikovaných investorů dle Darškuvienė (2010).

1.1.2 Behaviorální finance

"People in standard finance are rational. People in behavioral finance are normal."

—Statman (2014)

Hypotéza efektivního trhu je založena na myšlence, že většina investorů je při zpracování informací racionální. Behaviorální finance naopak studují, jak lidé nedosahují tohoto ideálu ve svých rozhodnutích a jak jsou trhy do určité míry neefektivní. Behaviorální věda patří mezi aplikované znalosti, které vstoupily do věd ve formě specializovaného oboru zvaného psychologie. Základní infrastrukturou těchto znalostí jsou lidské a přirozené, behaviorální a charakteristické složitosti, které je třeba chápat ve vědeckém rámci nazvaném psychologie. Behaviorální finance se snaží ukázat využití rozhodovacích procesů psychologie při rozpoznávání a předpovídání finančních trhů. Většina předchozích studií o validaci aktiv se zaměřila na behaviorální reakci investorů. Tyto studie prokázaly, že změny behaviorální reakce investorů mohou způsobit změny cen aktiv. Reakce investorů na chování byla zavedena jako významný faktor v procesu stanovení tržních cen. Behaviorální finance studují interpretaci jednotlivců od informací po vědomá investiční rozhodnutí. Nárůst behaviorálních financí za poslední tři desetiletí byl citelný v celé oblasti financí a ekonomiky. Mnoho vědců je nyní připraveno uvážit důsledky racionálních nebo iracionálních aspektů lidského úsudku, které jsou relevantní pro konkrétní aplikaci dle Hirshleifer (2015).

Behaviorální finance mají dvě hlavní základny; jedním je omezení v arbitráži, které uvádí, že racionální investoři nemohou snadno využít příležitosti arbitráže, protože to vyžaduje přijetí některých rizik. Zadruhé je to psychologie pro studium chování a úsudku investorů a chyb, kterých se jednotlivci dopouštějí (Raei a Fallahpour (2004). Zjištění naznačují, že investoři se ne vždy chovají racionálně. Behaviorální ekonomové se odvolávají na výsledky testu kognitivních psychologů, kteří studují různé předsudky chování, aby získali informace o iracionálním chování investorů (Darabi a kol., 2016). Vzhledem k roli změn cen akcií při rozhodování investorů, pozadí výzkumu přimělo vědce zkoumat fenomény přehnané reakce a nedostatečné reakce na nové informace a prezentovat některé modely s využitím údajů týkajících se akcií cen.

Rozpor mezi behaviorálními financemi a EMH

Vznikající rozpor mezi hypotézou efektivního trhu a realitou podpořil hlubší vhled zaměřený na psychologii jako důležitý faktor ve finanční teorii. Bylo formulováno behaviorální financování – nové odvětví teorie, kombinující znalosti psychologie, sociologie a dalších společenských věd díky integraci různých vědeckých poznatků behaviorální financování lépe vysvětluje anomálie trhu a finanční chování jednotlivců.

Efektivní trh je spojen s teorií racionálních očekávání, včetně vyhodnocení všech informací o majetku. Pokud však existuje mnoho iracionálních investorů a jejich finanční chování nekoreluje a jejich transakce se navzájem zneplatňují a mají dopad na ceny. Studie behaviorálních financí zkoumá účinky sociálních, kognitivních a emocionálních faktorů na ekonomická rozhodnutí jednotlivců a institucí a důsledky pro přínos a alokaci zdrojů. Důraz je kladen na nerovnovážné procesy, akce různých agentů s omezenou racionalitou, kteří se mohou poučit ze zkušeností a interakcí dle Rezaei a Elmi (2018).

Psychologické základy

Protože lidé potřebují rychle činit soudy a rozhodnutí s využitím omezených kognitivních zdrojů, nutně používají tzv. „heuristiky". Veškeré myšlení staví na kognitivních algoritmech, které fungují automaticky pod úrovní vědomí. Termín „heuristika" zahrnuje jak vrozené, tak automatické procesy a naučená nebo vědomě vybraná pravidla. Heuristika často funguje dobře v některých doménách a pro některé typy problémů, ale v jiných je jeho využití nevhodné. Heuristické zjednodušení implikuje více chyb pro rozhodovací problémy. Kahneman (2011) popisuje lidské myšlení jako převážně intuitivní a silně ovlivněné asociacemi, které jsou vyvolány prezentací rozhodovacího problému. Lidé mají přílišnou jistotu, že jejich intuitivní způsob myšlení o problému je správný; informace, které ihned nepřijdou na mysl, bývají zcela opomíjeny. Pocity poskytují váhu hodnoty přiřazenou možným výsledkům, aby motivovaly k rozhodnutím a činům. Afektivní reakce mohou také usnadnit rychlé využití naléhavých informací. Například riskantní investiční příležitost může vyvolat strach a tím i užitečné váhání. Pocity však často zkratují užitečnou analýzu, například při opuštění akciového trhu v náhlé panice nebo nákupu akcie na základě nadšení spíše než kritického hodnocení dle Hirshleifer (2015). Dle Triverse (1991) lidé přeceňují své osobní přednosti. Takový sebeklam přichází za cenu chyb vyplývajících z příliš sebevědomých přesvědčení. Tři výše uvedené prvky – heuristické zjednodušení, afektivní zkrat a sebeklam vysvětlují většinu psychologických předsudků studovaných v behaviorálním financování.

Omezená pozornost a kognitivní zpracování

Kvůli omezené pozornosti a síle zpracování mají lidé tendenci zanedbávat relevantní informační signály a strategické rysy rozhodovacího prostředí. To se projevuje řadou konkrétnějších účinků, které je třeba popsat dle Hirshleifer (2015).

Neschopnost zpracovat signály a vlastnosti rozhodovacího prostředí

Lidé mají tendenci zanedbávat slabé signály a přehnaně reagovat na hlavní nebo nedávné zprávy. Nemají tendenci si takové chyby uvědomovat, a proto je často neo-

pravují. Lidé také zanedbávají důležité rysy svých rozhodovacích prostředí, například strategické motivy jednání ostatních.

Finanční teorie založené na pocitech

Finanční teoretizace o pocitech byla většinou neformální (viz Mehra a Sah, 2002), což je vzhledem k jejich psychologickému významu překvapivé. Základním tématem je, že výkyvy nálady ovlivňují optimismus, toleranci vůči riziku a tržní ceny. Kvůli nesprávnému přiřazení nálady k dlouhodobým vyhlídkám mohou výkyvy nálady ovlivnit ceny, jak dokumentují Edmans a kol. (2007).

Sentiment, měnící se optimismus a tolerance vůči riziku

Sentiment investorů je kolísavý obecný postoj k investičním kategoriím, jako jsou růstové akcie nebo dlouhodobé dluhopisy. Může to souviset s posuny v hodnocení očekávaných výnosů nebo rizika. Vlny iracionálního nadšení pro určité investiční charakteristiky nebo jejich odmítání se odvíjejí od posunů ve výběžku emočních nebo kognitivních spouštěčů v ekonomickém prostředí. Takové posuny lze zvětšit pomocí sebeztužujících sociálních procesů vyvolaných zkreslením médií. Pokud sentiment vyvolá nesprávnou cenu, měl by předpovídat budoucí výnosy.

Behaviorální finance vyšetřují jemné aspekty a interakce v lidském mozku, čelí nejistotě při přijímání ekonomických rozhodnutí. Nejběžnější lidské vlastnosti (strach, hněv, chamtivost) kladou značný důraz na naše rozhodování o penězích. Intelekt (uchopení situace), rozum (dlouhodobé důsledky provedené akce) a emoce (s ohledem na postup) spolu souvisejí; jsou prameny za lidským rozhodnutím. Lidské chování je obecně reaktivní, nikoli proaktivní; proto je obtížné předpovídat na základě pravidel. Kromě výše uvedené prezentace chyb chování na finančním trhu lze vytvořit popis jejich mechanismu, který je podobný mechanismu kyvadla neustále oscilujícího mezi optimistickými investory a pesimistickými investory. Optimističtí investoři, stejně jako sebevědomí, jsou ochotnější investovat riskantně. Provádějí iracionální transakce a jejich iracionální reakce mohou vést k abnormálním výnosům a objemu obchodů (Oprean a Tanasescu, 2014). Vědci naznačují, že nálada výrazně ovlivňuje úsudek a rozhodování a následně mění chování investorů. Nálada nebo psychologický stav investorů při rozhodování mohou ovlivnit jejich preference, hodnocení rizik a racionální kogitace a nakonec i jejich investiční rozhodnutí. Finanční rozhodnutí by se proto měla lišit podle nálady investora. Ve srovnání s rozsáhlými empirickými důkazy, že nálada ovlivňuje finanční trhy, jsou však její účinky na rovnovážné ceny aktiv relativně nepodložené. Ačkoli některé studie v posledním desetiletí modelovaly psychologii investorů, většina se zaměřila na psychologickou zaujatost (např. přehnané sebevědomí, reprezentativní heuristiku, přehnanou a nedostatečnou reakci, averzi k nejednoznačnosti a důvěrnost) nebo na obavy z budoucích pocitů

(např. ztráta a averze ke zklamání) spíše než zkoumání dopadu na akciové trhy.

Vliv sentimentu investorů na rovnovážné ceny aktiv a výnosy tak zůstává otevřenou výzvou dle Shu (2010), Rezaei a Elmi (2018). Proto se proces formování názoru a jeho dopadové faktory stávají předmětem zájmu při průzkumu akciového trhu. Je jasně vidět, že ve srovnání s institucionálními investory s více informacemi a vyspělejším investičním chováním je pro jednotlivé investory obtížnější být racionální, když čelí masivním informacím týkajícím se akcií. Pravděpodobně tedy mají sklon řídit se názorem ostatních a mají zjevnější proces šíření sentimentu. To zvyšuje nepopiratelnou důležitou prozkoumat proces tvorby sentimentu, který by pomohl pochopit, jak tento důležitý sentiment generuje a ovlivňuje obchodní strategie investorů.

1.2 Těžba textu

„Ačkoli se sdělovací prostředky – noviny, časopisy a zpravodajská média spolu se svými novými partnery na internetu prezentují jako nezávislí pozorovatelé tržních událostí, jsou sami nedílnou součástí těchto tržních událostí.“

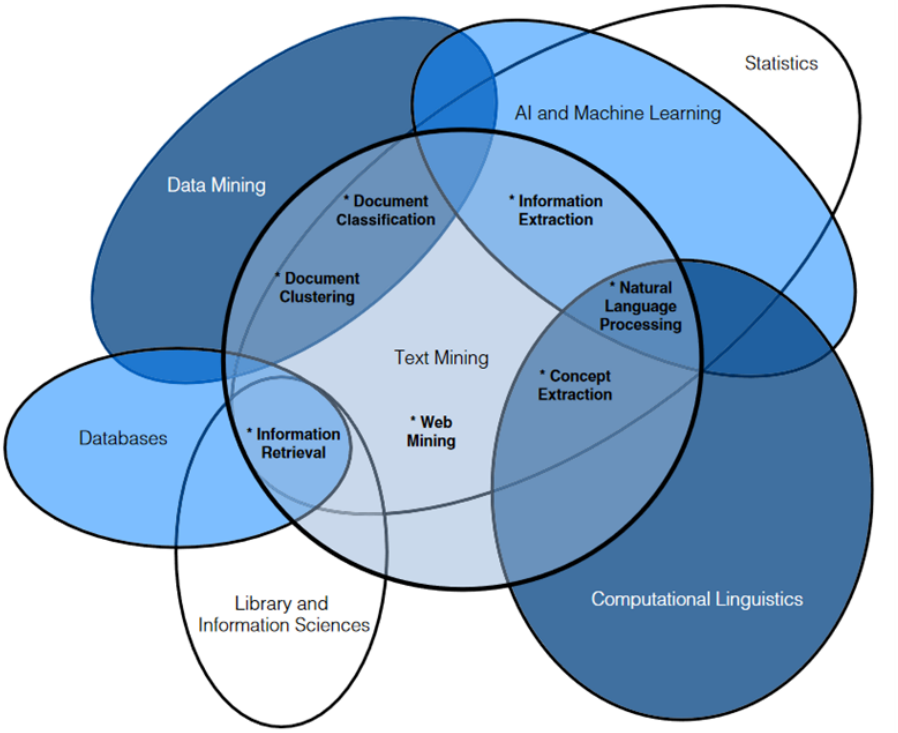
—Shiller (2000)

Text (přirozený jazyk) je nejpřirozenějším způsobem kódování lidských znalostí. Výsledkem je, že většina lidských znalostí je zakódována ve formě textových dat. Text je zdaleka nejběžnějším typem informací, s nimiž se lidé setkávají. Většina informací, které člověk denně produkuje a spotřebovává, je ve skutečnosti v textové podobě. Explozivní růst online textových informací vytvořil silnou poptávku po inteligentních softwarových nástrojích, které by poskytly související služby, které lidem pomáhají spravovat a využívat velká textová data, jak popisuje Zhai a Massung (2016).

Růst textových dat znemožňuje lidem shromáždit veškerá tato data včas. Jelikož textová data kódují většinu našich nashromážděných znalostí, nelze je obecně zahodit, což vede například k akumulaci velkého množství údajů z literatury, která je nyní nad možnosti jakéhokoli jednotlivce dokonce překonat. Rychlý růst online textových informací také znamená, že nikdo nemůže denně strávit všechny nové informace, které byly vytvořeny. Je tedy naléhavě nutné vyvinout inteligentní systémy pro vyhledávání textu, které by lidem pomohly rychle a přesně získat přístup k potřebným relevantním informacím, což by vedlo k rozvinutí těžby a analýzy textu.

Těžba textu a analýza textu jsou široce zastřešující termíny popisující řadu technologií pro analýzu a zpracování polostrukturovaných a nestrukturovaných textových dat. Sjednocujícím tématem každé z těchto technologií je potřeba „převést text na čísla“, aby bylo možné na databáze velkých dokumentů použít výkonné algoritmy.

Převod textu do strukturovaného číselného formátu a použití analytických algoritmů vyžaduje znalost používání a kombinování technik manipulace s textem, od jednotlivých slov přes dokumenty až po celé databáze dokumentů. K dnešnímu dni těžba textu odolávala komplexnější definici, protože se vynořuje skupina příbuzných, ale odlišných oborů. Obrázek 1.3 ukazuje šest dalších hlavních oborů, které se protínají s těžbou textu.



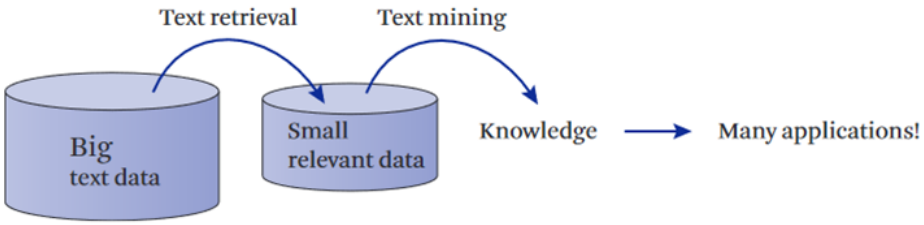
Obr. 1.3: Vennův diagram těžby a analýzy textu
Zdroj: Miner a kol. (2012)

V zásadě všechny formy těžby textu, se zabývají otázkou, jak zpracovat exponenciálně rostoucí množství textových dat, aby se získaly smysluplné a užitečné informace. Zatímco množství textu generovaného v elektronické podobě exploduje, množství informací, které lze zpracovat jednou osobou nebo skupinou jednotlivců, zůstává konstantní. Je zřejmé, že je třeba vyvinout způsob, jak se vypořádat s touto povahou textových dat.

Kvůli šíři těžby textu není cílem plně pokrýt všechny uvedené obory v této disertační práci. V důsledku toho práce pokrývá do hloubky pouze ty oblasti, které jsou nezbytné k naplnění cíle práce. Disertační práce odkazuje na další zdroje v oblastech, které vyžadují dodatečné odborné znalosti, nebo jsou použitelných v jiných oblastech než na akciových trzích. Záměrem této části disertační práce je poskytnout systematický úvod do vybraných přístupů.

1.2.1 Koncepční základy těžby textu

Před zahájením těžby a analýzy textu je nejprve nutné porozumět jejím koncepčním základům a poté pochopit, jak začít využívat sílu těchto dat k vytváření rozhodnutí. Většina textových dat, se kterými se lze denně setkat, je nestrukturovaná. Na rozdíl od strukturovaných dat, která odpovídají dobře definovaným schématům a jsou relativně snadno zpracovatelné pro počítače, má text méně explicitní strukturu, takže vyžaduje, aby počítačové zpracování porozumělo obsahu kódovanému v textu. Současná technologie zpracování přirozeného jazyka dosud nedosáhla bodu, který by umožňoval počítači přesně rozumět textu přirozeného jazyka, ale má širokou škálu statistických a heuristických přístupů k řízení a analýza textových dat byla vyvinuta v posledních několika desetiletích, dle Zhai a Massung (2016). Obvykle jsou velmi robustní a lze je použít k analýze a správě textových dat.



Obr. 1.4: Koncepční základy textové analýzy
Zdroj: Zhai a Massung (2016)

Vyhledávání a těžba textu koncepčně odpovídají dvěma přirozeným krokům v procesu analýzy jakýchkoli „velkých textových dat“, jak je znázorněno na obrázku 1.4. I když mohou být nezpracovaná textová data velká, konkrétní aplikace často vyžaduje jen malé množství nejdůležitějších textových dat, takže koncepčně by prvním krokem v každé aplikaci měla být identifikace příslušných textových dat pro konkrétní aplikaci nebo rozhodování a vyhnout se zbytečnému zpracování velkého množství nerelevantních textových dat.

Účelem algoritmů těžby textu je poskytnout určité porozumění tomu, jak je text zpracováván, aniž by ho člověk musel číst. Počítač však může zkoumat pouze přímo jednotlivé znaky v každém slově, a jak jsou tato slova uspořádána. Počítač nemůže vědět, jaké informace sděluje text, může pochopit pouze strukturu textu.

Syntaxe se týká struktury jazyka a způsobu skládání jednotlivých slov tak, aby vytvářely dobře formulované věty a odstavce. Jazyk se řídí specifickými pravidly gramatiky a jazykovými konvencemi, což vede ke statistickým vzorům, které se často objevují ve velkém množství textu. Tuto strukturu počítač relativně snadno zpracovává. Samotná syntaxe však není dostatečná pro úplné pochopení významu. Jinými slovy, účelem syntaktické analýzy je zjistit, jak jsou slova ve větě vzájemně spjata,

a odhalit tak syntaktickou strukturu věty, jak uvádí Miner a kol. (2012). Sémantika dává přednost významu jednotlivých slov. Účelem sémantické analýzy je určit význam věty. To obvykle zahrnuje výpočet významu celé věty nebo větší jednotky na základě významů slov a jejich syntaktické struktury dle Zhai a Massung (2016). Je zřejmé, že úplný sémantický význam textu může být obtížné určit automaticky bez rozsáhlého pochopení používaného jazyka. Naštěstí lze k získání praktické hodnoty z textu použít samotnou syntaxi bez úplného sémantického porozumění, a to kvůli úzké vazbě mezi syntaxí a sémantikou. Mnoho úloh dolování textu, jako je klasifikace dokumentů a vyhledávání informací, se týká hodnocení nebo hledání konkrétních typů dokumentů ve velké databázi dokumentů. Hlavním předpokladem těchto algoritmů je, že syntaktická podobnost (podobná slova) implikuje sémantickou podobnost (podobný význam). Ačkoli se tyto přístupy spoléhají pouze na syntaktické informace, fungují, protože dokumenty, které sdílejí mnoho klíčových slov, jsou často na stejné téma. V jiných případech je cíl opravdu o sémantice. Například extrakce konceptu je o automatické identifikaci slov a frází, které mají stejný význam dle Miner a kol. (2012).

1.2.2 Zdroje textových dat

Prvním krokem v těžbě textu je shromáždění dat. Textový vstup může mít několik zdrojů a typů obsahu. Obvykle se textový vstup označuje jako korpus. Texty mohou obsahovat především faktické (např. finanční výsledky) nebo názorové (např. pozitivní nebo negativní vztah k firmě či produktu) informace, případně obojí. V druhém případě tak může zkoumat, zda pohyb ceny akcie souvisí s tím, jaké emoce, nálady a názory mají lidé, obchodující na akciových trzích, resp. zda a jak se jimi nechají ovlivnit. Mezi těmito dvěma typy dokumentů nebude explicitně rozlišováno, i když hrubé rozlišení poskytnou samotné zdroje textů, neboť lze očekávat, že novinové články budou obsahovat faktické, zatímco statusy na sociálních sítích spíše názorové informace. S rostoucí snadností archivace lidské řeči a projevu bude objem textových dat v průběhu času jen narůstat. Tento trend je podporován rostoucí všudypřítomnost digitalizací.

Získávání dat je výchozím bodem každého projektu datové vědy a vstupy lze získat buď ze soukromých zdrojů, jako jsou interní údaje společnosti nebo veřejné zdroje, a právě zde se otevírá svět možností. Veřejné zdroje zahrnují oficiální data z vládních institucí, deníků nebo obecněji jakákoli data, která lze najít na webu bez ohledu na to, zda se jedná o volně dostupná data, nebo je nutné si je zakoupit. Při získávání dat z webu lze využít několik způsobů. Jedná se buď o tzv. API *(angl. Application Programming Interface)*, což je rozhraní pro programový přístup k dané aplikaci. Nebo lze získat data pomocí web scrapingu, což je proces získávání textu

z webové stránky zpracováním HTML kódu dané stránky (Russell, 2013). K tomuto účelu existuje pro každý "běžný" programovací jazyk mnoho nástrojů (knihoven).

1.2.3 Předzpracování textu

Sjednocujícím tématem spojujícím různé techniky dolování textu je myšlenka přeměny nestrukturovaných dat ve formě textu na strukturovaná data ve formě čísel, aby bylo možné použít matematické a statistické algoritmy (Miner a kol., 2012). Při těžbě dat obecně a konkrétně při těžbě textu má fáze předběžného zpracování významný dopad na celkové výsledky (Uysal a Gunal, 2014). Zásadní problém vyplývá ze značné variability lidského jazyka. V textu jsou použity slovesa v různých časech (budoucí, přítomný, minulý), slova v jednotném či množném čísle a v neposlední řadě také jinak skloňované formě. V automatizované analýze textu to zvětšuje rozměrnost termínů v dokumentu. Proto byly vyvinuty techniky předzpracování, které snižují takovou variabilitu při zachování významu slova (Rüdiger a spol., 2017). Typicky se jedná o následující základní kroky předzpracování.

Výběr rozsahu dokumentu

U mnoha úloh těžby textu je snadné určit rozsah textu. Například e-maily nebo protokoly hovorů se přirozeně překládají do jednoho vektoru pro každou zprávu. U delších dokumentů se však musí rozhodnout, zda použít celý dokument, nebo rozdělit dokument na části, odstavce nebo věty. Výběr správného rozsahu závisí na cílech úkolu dolování textu: pro úkoly klasifikace nebo shlukování je správným rozsahem často celý dokument; pro analýzu sentimentu, shrnutí dokumentu jsou vhodnější kratší části textu, například odstavce nebo oddíly dle Miner a kol. (2012).

Tokenziace

Dalším krokem předzpracování je rozdělení jednotek textu na jednotlivá slova nebo tokeny. Jinými slovy, tokenizace je úkol „rozřezat" posloupnost znaků na kousky, které se nazývají tokeny, a případně odstranit určité znaky například interpunkci. Tento proces může mít mnoho podob, v závislosti na analyzovaném jazyce. Pro angličtinu je přímou strategií efektivní strategie tokenizace použití mezer a interpunkce jako oddělovačů tokenů. Tuto strategii lze snadno implementovat, ale existují případy, kdy nemusí odpovídat požadovanému chování. V případě akronymů a zkratek by kombinace použití neuspořádaného vektorového prostoru a oddělení tokenů při interpunkci dala různé komponenty zkratky do různých tokenů dle Weiss (2005).

Odebrání stop slov

U mnoha úloh těžby textu je užitečné odstranit slova, která se objevují téměř v každém dokumentu, aby se ušetřil úložný prostor a urychlilo se zpracování. Tato běžná slova se nazývají „stopwords“ a proces odstraňování těchto slov se nazývá „stopping“. Mezi často používaná slova v angličtině patří „a“, „of“, „the“, „such“, „then“, „by“ a „and“. Stopwords je běžně obsaženou funkcí téměř v každém softwarovém balíčku pro dolování textu. Odstranění stop je možné bez ztráty informací, protože u velké většiny úkolů a algoritmů těžby textu mají tato slova malý dopad na konečné výsledky algoritmu. Miner a kol. (2012)

Stemming

Stemming je proces normalizace souvisejících slovních tokenů do jediné formy. Proces obvykle zahrnuje identifikaci a odstranění předpon, přípon a nevhodných pluralizací. U mnoha úkolů těžby textu, včetně klasifikace a shlukování, vede stemming k vylepšení přesnosti zmenšením počtu dimenzí používaných algoritmy a seskupením slov podle konceptu. Toto snížení rozměrnosti zlepšuje fungování algoritmů. Lemmatizace je pokročilejší forma stemming, která se pokouší seskupit slova na základě jejich hlavního konceptu nebo lemmatu. Lemmatizace používá k určení lemmatu jak kontext obklopující slovo, tak další gramatické informace dle Miner a kol. (2012).

Detekce konce vět

Detekce hranice věty je proces rozdělení celých dokumentů na jednotlivé gramatické věty. U textu je snadné najít každý výskyt interpunkce jako „.“; "?"; nebo „!“ v textu. Některá interpunkce se však vyskytují jako součást zkratek nebo akronymů (před tokenizací). Tyto podmínky naznačují několik jednoduchých heuristických pravidel, která mohou správně identifikovat většinu hranic vět. K dosažení téměř dokonalé přesnosti se používají techniky statistické klasifikace. Miner a kol. (2012)

Normalizace velkých a malých písmen

Většina textů je psána smíšeně, což znamená, že text obsahuje jak velká, tak malá písmena. Velká písmena pomáhají čtenářům rozlišovat například mezi vlastními jmény a názvy a mohou být užitečná také pro automatizované algoritmy. Za každých okolností by se však s velkým písmenem na začátku vět nemělo zacházet jinak než se stejným slovem, které se v jiném dokumentu vyskytuje malými písmeny. Normalizace případu převede celý dokument buď na zcela malá písmena, nebo na úplně velká písmena. Miner a kol. (2012). Tyto kroky jsou prováděny téměř vždy a v některých dílech jsou jedinými kroky.

1.2.4 Transformace textu na datové struktury

Počítače rozumí pouze číslům. Reprezentace textu pomocí čísel je výzva. Zároveň je to příležitost vymyslet a vyzkoušet přístupy k reprezentaci textu tak, aby bylo možné v procesu zachytit maximum informací. V této podkapitole je pozornost zaměřena na transformaci textových dat na matematické datové struktury, které poskytnou informace o tom, jak text skutečně reprezentovat pomocí čísel dle Weiss a kol. (2015).

Jedna z nejdůležitějších inovací při zpracování textu je zobecněný model vektorového prostoru. Dokumenty jsou reprezentovány řetězcem hodnot (vektorem), ve kterém každý prvek v řetězci představuje jedinečné slovo nebo skupinu slov obsaženou mezi dokumenty. Vektory jsou jednorozměrná řada čísel, ve kterých lze každé slovo identifikovat podle příslušných indexů.

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{1.1}$$

V tomto příkladu rovnice 1.1 má vektor x tři prvky a tyto tři prvky ukládají informace o vektoru. Vektory jako objekt v prostoru, kde každý prvek vektoru představuje projekci tohoto vektoru podél dané osy. Jakmile je objekt abstrahován jako vektor, musí splňovat všechny vlastnosti vektoru a můžeme s ním provádět libovolné vektorové operace. Vektor může být buď binární (slovo se v dokumentu vyskytuje nebo ne), nebo může obsahovat četnost výskytu, případně váhy, založené na důležitosti slov v celé kolekci dokumentů, využívající váhovací metody TF-IDF. Lze setkat s různými přístupy k reprezentaci textových dat, které jsou diskutovány níže.

Bag of Words

Velmi intuitivním přístupem k reprezentaci dokumentu je použití frekvence slov v konkrétním dokumentu. To je přesně to, co se děje v rámci přístupu Bag of Words (BoW). Slovní zásoba je předpokladem této metodiky BoW. Jakmile je slovní zásoba k dispozici, může být každá věta představována jako vektor. Délka tohoto vektoru by se rovnala velikosti slovní zásoby. Každá položka ve vektoru by odpovídala výrazu ve slovníku a číslem v tomto konkrétním záznamu by byla četnost výrazu v podhodnocování věty. Dolní limit pro toto číslo by byl 0, což naznačuje, že slovník se v dané větě nevyskytuje. (Kedia a Rasu, 2020) Výsledkem je vektor v prostoru funkcí, který odpovídá tokenům nalezeným v textu. S tímto předpokladem by skóre dotazu, bylo založeno na skóre vypočítaném pro každé jednotlivé slovo dotazu. To znamená, že skóre bude záviset na skóre každého slova dle Zhai a Massung (2016).

Existují však určitá omezení. Model se spoléhá pouze na počet výrazů v dokumentu. Tyto modely neberou v úvahu sémantiku nebo významy spojené s tokenem nebo frázemi v dokumentu. Ignoruje možnost zachycení prvků ze sousedství fráze, které mohou naznačovat kontext, ve kterém se slovo nebo fráze používá. Proto zcela ignoruje související kontext. Model BoW může také být extrémně obrovský, pokud jde o slovník pro velký textový korpus. To může vést k vektorům obrovských velikostí představujících každý dokument, což by mohlo způsobit zhoršení výkonu modelu. (Kedia a Rasu, 2020)

N-gramy

Bag-of-word modely jsou jednoduchý a často velmi efektivní přístup. Důležité podrobnosti o původním dokumentu, jako jsou fráze, slovní spojení, kontext a věty, se však ztratí. Alternativně může přidání víceslovných výrazů pomoci identifikovat kontext textu.

Metoda N-gram založená na modelu vektorového prostoru zvažuje, že vztah sousedních slov může mít pravděpodobně vliv na klasifikaci. Parametr N zde představuje počet slov souvisejících s určitým předcházejícím slovem. To znamená, že vznik N-tého slova je spojen s N-1 slovy před ním. Běžně se lze setkat s N = 1, 2, 3. Například nastavení N = 2 znamená, že pro každý dokument je vygenerována sekvence dvou slov. Metodou N-gram založenou na modelu vektorového prostoru je každý text vyjádřen jako vektor v jazykovém prostoru a každá vlastnost získává váhu podle své důležitosti v textu. N-gram je souvislá posloupnost N položek, kterými jsou obvykle slova z dané posloupnosti textu.

Ukázalo se, že extrakce založená na N-gramech má robustní výkon při extrakci funkcí z textu kvůli různým aspektům. Nejprve automatické zachycení nejčastějších kořenů v datech. Zadruhé, dobrá reprezentace, kterou poskytuje N-gram, nevyžaduje použití konkrétního slovníku. Zatřetí, jeho tolerance vůči deformacím a pravopisným chybám. (Khedr a kol., 2017)

Skip-gram

Skip-gramy jsou technikou k překonání řídkosti dat. K-skip-n-gramy jsou tvořeny „přeskakováním“ tokenů a povolením sousedních sekvencí slov. Například aplikace skip-gramů na model trigramu pomocí věty „I hit the tennis ball“, přeskočení tokenu „tennis“, také vytvoří funkci „hit the ball“, což může být stejně důležitá vlastnost. Formálně jsou k-skip-n-gramy pro větu $w_1 \ldots w_m$ definovány takto:

$$\left(w_{i_1}, w_{i_2}, \cdots, w_{i_n} | \sum_{j=1} ni_j - i_{j-1} < k \right) \tag{1.2}$$

kde k je definováno jako vzdálenost přeskočení pro konstrukci N-gramu. Guthrie a kol. (2006) ve svém výzkumu ukazuje, že skip-gramy mohou přesně modelovat kontext při zachování dezinformací na minimu. Výsledky dále naznačují, že skip-gramy jsou obzvláště užitečné, když je testovací korpus podobný trénovanému korpusu. (Purgstaller, 2018)

Váhová metoda TF-IDF

Přirozenou myšlenkou je zvážit více výskytů termínu v dokumentu *(angl. Term Frequency,* TF) na rozdíl od binární reprezentace. Ke zvážení rozdílu mezi dokumentem, kde se dotazovaný termín vyskytl vícekrát, a tím, kde se dotazovaný termín vyskytl jen jednou, je nezbytné vzít v úvahu termínovou frekvenci – počet výrazu v dokumentu. Nejjednodušší způsob, jak vyjádřit TF slova w v dokumentu d, je:

$$TF_{w,d} = count\,(w, d) \tag{1.3}$$

S binární reprezentací lze zachytit pouze přítomnost nebo nepřítomnost termínu, nicméně ignoruje skutečný počet výskytů termínu. To znamená, že prvky vektoru dotazu i vektoru dokumentu nebudou nuly a jedničky, ale místo toho to budou počty a slovo v dotazu nebo dokumentu dle Zhai a Massung (2016).

Inverzní frekvence dokumentu *(angl. Inverse Document Frequency,* IDF) je velmi důležitý signál používaný v moderních vyhledávacích funkcích. Frekvence dokumentu je počet dokumentů, které obsahují konkrétní výraz. Způsob, jak to začlenit do vektoru, je upravit počet frekvencí vynásobením IDF příslušného slova. Nyní lze penalizovat běžná slova, která mají obecně nízkou hodnotu IDF, a odměna informativní slova, která mají vyšší IDF. IDF lze definovat jako:

$$IDF_w = \frac{M + 1}{df(w)} \tag{1.4}$$

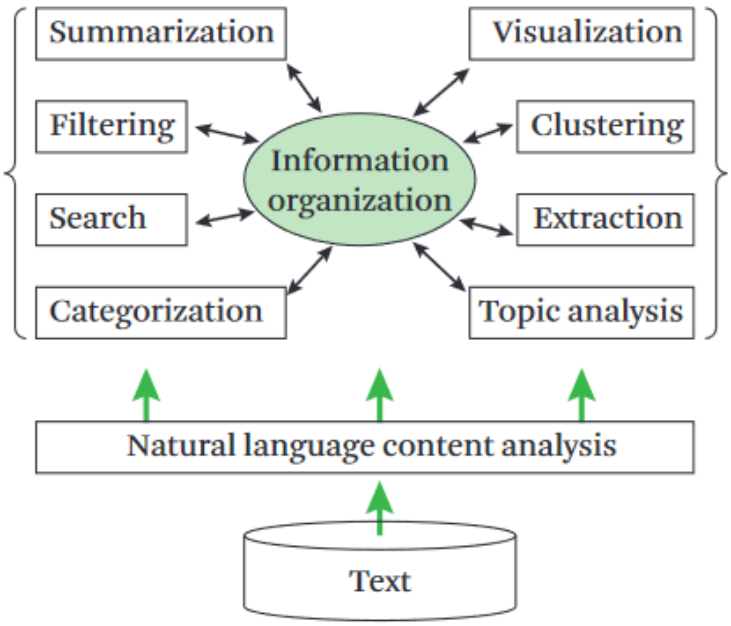
kde M je celkový počet dokumentů ve sbírce a $df(.)$ počítá frekvenci dokumentů (celkový počet dokumentů obsahujících w). Díky vážení IDF nyní lze upravit váhu TF (četnost termínů) vynásobením váhou IDF. IDF tedy bude rozlišovat tato dvě slova podle toho, jak jsou informativní. Váhové schéma TF-IDF je následně definováno jako:

$$w_{w,d} = TF_{w,d} \cdot IDF_w \tag{1.5}$$

1.2.5 Základní aplikace vytěženého textu

Význam textových dat pro náš život lze snadno zjistit ze skutečnosti, že denně zpracováváme velké množství textových dat. Vzhledem k širokému pokrytí znalostí

v textových datech a naší spoléhání se na textová data pro komunikaci je možné si představit aplikace pro analýzu textu nepřímo v jakékoli doméně. Níže je ilustrativně uvedeno jen několik konkrétních příkladů, které mohou poskytnout kontexty aplikace pro porozumění technikám analýzy textu, které jsou stručně popsány. Text mining lze využít pro tyto základní úlohy, jak vyobrazuje obrázek 1.5 dle Zhai a Massung (2016).



Obr. 1.5: Základní aplikace vytěženého textu
Zdroj: Zhai a Massung, (2016)

Vyhledávání informací: Vezme dotaz uživatele a vrátí relevantní dokumenty. Oblast webových vyhledávačů patří mezi nejužitečnější vyhledávače, které uživatelům umožňují efektivně pracovat s velkým množstvím textových dat.

Filtrování informací: Monitorujte příchozí stream, rozhodněte, které položky jsou relevantní (nebo nerelevantní) pro zájem uživatele, a poté uživateli doporučí relevantní položky (nebo odfiltrujte nerelevantní položky).

Klasifikace dokumentů: Klasifikuje textový objekt do jedné nebo několika předdefinovaných kategorií, kde se kategorie mohou lišit v závislosti na aplikacích. Komponenta kategorizace může anotovat textové objekty všemi druhy významných kategorií, čímž obohacuje textová data reprezentace, což dále umožňuje efektivnější a hlubší analýzu textu. Kategorie lze také použít pro organizaci textových dat a usnadnění přístupu k textu. Specifické příklady textového kategorizačního systému je kupříkladu klasifikace sentimentu dokumentů, vět či slov na pozitivní, negativní nebo

neutrální polaritu, která je detailně rozebrána v následující subkapitole 1.3.

Sumarizace informací: Vezme jeden nebo více textových dokumentů a vygenerujte stručné shrnutí základního obsahu. Souhrn snižuje lidské úsilí při trávení textových informací a může také zlepšit efektivitu při těžbě textu.

Analýza témat: Vezme sadu dokumentů a extrahujte a analyzujte v nich témata. Témata přímo usnadňují pochopení textových dat uživateli a podporují procházení textových dat. V kombinaci s doprovodnými netextovými daty, jako je čas, místo, autoři a další metadata, může analýza tématu generovat libovolné zajímavé vzory, jako jsou časové trendy témat, prostorová distribuce témat a profily témat autorů.

Extrakce informací: Extrahujte z textu entity, vztahy entit nebo jiné „funkce se znalostmi“. Složka extrakce informací umožňuje konstrukci grafů vztahů mezi entitami.

Shlukování dokumentů: Objevte skupiny podobných textových objektů (např. výrazy, věty, dokumenty, ...). Využívá empirická data k vytváření smysluplných struktur, které mohou být užitečné pro procházení textových objektů a získání rychlého porozumění velké sadě textových dat. Je také užitečné pro objevování odlehlých hodnot identifikací položek, které netvoří přirozené shluky s jinými položkami.

Vizualizace: Vizuální zobrazení vzorů v textových datech. Vizualizační komponenta je důležitá pro zapojení lidí do procesu objevování zajímavých vzorů, protože lidé dokážou velmi dobře rozpoznat vizuální vzorce.

1.3 Analýza sentimentu

„Jednoho dne budeme moci měřit sílu slov.“

—Maya Angelou (2011)

Média neuvádějí pouze stav trhu, ale aktivně utvářejí dopad na dynamiku trhu na základě novinek, které vydávají (Wisniewski a Lambe, 2013). Interpretace stejných informací lidmi se velmi liší. Účastníci trhu mají kognitivní předsudky, jako je nadměrná důvěra, přehnaná reakce, zkreslení informací a různé další předvídatelné lidské chyby v uvažování a zpracování informací (Friesen a Weller, 2006). Behaviorální finance a teorie sentimentu investorů pevně prokázaly, že chování investorů lze ovlivnit tím, zda se budou cítit optimisticky (býčí) nebo pesimisticky (medvědí) ohledně budoucích tržních hodnot (Bollen a Huina, 2011).

Analýza sentimentu je řada metod, technik a nástrojů pro detekci a extrakci subjektivních informací, jako jsou názory a postoje, z jazyka Liu (2009). Analýza sentimentu byla tradičně o polaritě názorů, tj. o tom, zda má někdo na něco pozitivní, neutrální nebo negativní názor Dave a spol. (2003). Vypuknutí moderní

analýzy sentimentu nastalo až v polovině roku 2000 a zaměřilo se na recenze produktů dostupné na webu, např. Dave a spol. (2003). Od té doby se použití analýzy sentimentu dostalo do řady dalších oblastí, jako je predikce finančních trhů Nassirtoussi a kol. (2014). Analýza Mäntylä a spol. (2018) ukazuje, že ačkoli porozumění názorům bylo vždy důležité, to, co z analýzy názorů učinilo trendové výzkumné téma, byla možnost automaticky shromažďovat a analyzovat velké korpusy názorů pomocí nástrojů pro dolování textu.

1.3.1 Sentiment, názor a jeho intenzita

Analýza sentimentu je objevující se součástí analýzy textových dat, která z dokumentů získává subjektivní pojmy a názory. Spojuje techniky jazykového zpracování a výpočetní lingvistiku společně za účelem identifikace nebo extrakce subjektivních informací. Jedním z potenciálních faktorů pro využití zpravodajských článků na akciovém trhu je jejich sentiment nebo tón. Pozitivní tón může motivovat investory ke koupi akcií, zatímco negativní tón může mít opačný účinek. Výzkum analýzy sentimentu začal identifikací názorově (nebo náladově) styčných slov, která jsou například slova jako skvělý, nádherný, radost, špatný, bída, smutek, vztek. U takových slov se poté určovala sémantická orientace neboli polarita – pozitivní nebo negativní (Liu, 2007). Později se určilo několik základních lingvistických pravidel, pomocí kterých se určil sentiment slov z rozsáhlého korpusu (Hatzivassiloglou a McKeown, 1997). Tato metoda byla nadále rozvíjena. Významným pokrokem v této oblasti bylo třídění názorů (sentimentu), tzv. klasifikace sentimentu, které je nyní vlastně primární úlohou analýzy sentimentu.

Definice 1 (sentiment)
Sentiment je základní pocit, postoj, hodnocení nebo emoce spojené s názorem. Je reprezentován jako trojitá dimenze dle Liu (2015):

$$(y, o, i) \tag{1.6}$$

kde y je typ sentimentu, který lze dále členit na racionální a emocionální. Orientace sentimentu o může být pozitivní, negativní nebo neutrální. Neutrál obvykle znamená nepřítomnost sentimentu nebo žádný sentiment nebo názor. Orientace sentimentu se také nazývá polarita, sémantická orientace nebo valence ve výzkumné literatuře. Intenzita sentimentu i může mít různé úrovně síly nebo intenzity.

Definice 1.1 (racionální sentiment)
Racionální nálady pocházejí z racionálního uvažování, hmatatelných přesvědčení a utilitárních postojů. Nevyjadřují žádné emoce.

Definice 1.2 (emoční sentiment)
Emoční sentimenty jsou z nemateriálních a emocionálních reakcí na entity, které jdou hluboko do psychologického stavu mysli lidí. Názvy vyjadřující emocionální sentiment se nazývá také emocionální názory. Emoční sentiment je silnější než racionální sentiment a v praxi je obvykle důležitější. Každá z těchto širokých kategorií může být dále rozdělena do menších kategorií (Russell, 1980). Lze je definovat jako: (i) valence: příjemnost podnětu; (ii) vzrušení: intenzita emocí vyvolaná podnětem; (iii) dominance: stupeň kontroly vyvolaný podnětem. Sentiment lze považovat za zvláštní případ tohoto druhého pohledu na emoce jako body v prostoru. Zejména dimenze valence, která měří, jak příjemné nebo nepříjemné je slovo, se často používá přímo jako míra sentimentu dle Liu (2015).

Definice 2 (názor)
Výzkumníci v oblasti zpracování přirozeného jazyka *(angl. Natural Language Processing,* NLP) na pojem názor pohlížejí z různých hledisek. Například je na něj pohlíženo jako na objekt, tj. to, co je komentováno, a jeho komponenty a atributy (Liu, 2007), nebo jako na stavy zaujetí vůči cíli (Wiebe a kol., 2005). Nejvíce názor zřejmě dekomponoval Liu (2015), a ten ho definoval jako dimenzi:

$$(o_i, a_{ij}, s_{ijkl}, h_k, t_l) \tag{1.7}$$

kde o_i je objekt, tzn. název tématu, kterého se názor týká, potažmo někdy i cíl názoru; a_{ij} je aspekt názoru neboli prvek objektu o_i, cíl názoru (někdy shodný s objektem o_i); s_{ijkl} je sentiment aspektu (prvku) daného objektu; h_k je nositel názoru – osoba, organizace, která názor vyjadřuje; t_l je čas, kdy je názor vyjádřen. Části informací definovaných v této pětici musí navzájem korespondovat. To znamená, že sentiment s_{ijkl} musí být dán nositelem h_k o aspektu a_{ij} objektu oi v čase t_l. Jakýkoliv nesoulad je chybou. Pokud je objekt cílem názoru, tak je shodný i s jeho aspektem.

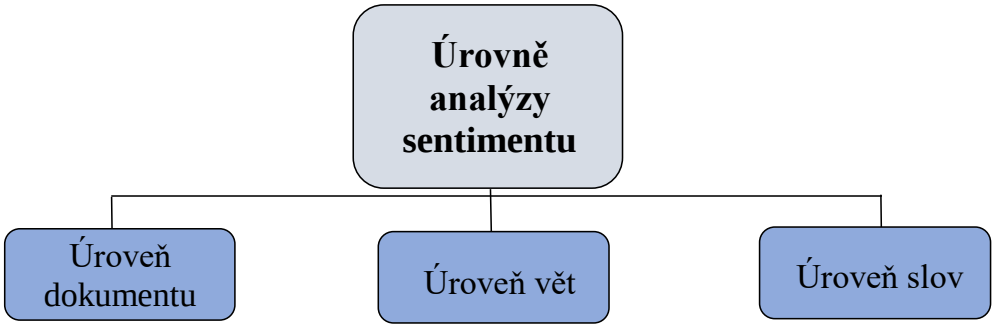
Sentiment objektu/aspektu může být pozitivní, negativní, případně neutrální, nebo vyjádřen různým stupněm intenzity, např. 1-5 hvězdiček (často u recenzí) nebo slovním hodnocením (nejlepší, špatný atp.). Běžný názor se dělí na další dva podtypy (Liu, 2015): (i) přímý názor, který je vyjádřen přímo na objekt nebo aspekt objektu; (ii) nepřímý názor, který je vyjádřen nepřímo na objekt nebo aspekt objektu a má vliv na některé další objekty.

Dalším typem je tzv. komparativní názor, který vyjadřuje vztah založený na podobnostech či odlišnostech více než jednoho objektu či aspektu názoru a který již vyžaduje jinou definici a jiné analytické techniky.

Ačkoliv analýza sentimentu objevuje sentiment především u subjektivních vět nebo tvrzení, tak to ovšem neznamená, že každá subjektivní věta či tvrzení obsahuje sentiment. Naproti tomu objektivní věty nebo tvrzení mohou sentiment obsahovat.

1.3.2 Úrovně analýzy sentimentu

Analýza sentimentu obvykle funguje na třech různých úrovních členitosti, a to na úrovni dokumentu, úrovni věty, úrovni aspektů, jak je znázorněno na obrázku 1.6. V této části práce je podrobně popsána každá z těchto úrovní.



Obr. 1.6: Úrovně analýzy sentimentu
Zdroj:vlastní zpracování

Sentiment na úrovni dokumentu

Úroveň dokumentu pracuje s celými dokumenty jako základní informační jednotkou. Jedná se tedy o nejabstraktnější úroveň analýzy sentimentu, a proto není vhodná pro přesná hodnocení (Moraes a kol., 2013). Výsledkem na této úrovni jsou často obecné informace o dokumentech. Polarita je nakonec shrnuta na celém dokumentu jako pozitivní nebo negativní. Většina raných studií v (Turney, 2002; Pang a kol., 2002) se zaměřila na úroveň dokumentu a spoléhala na datové sady, jako jsou recenze filmů a produktů.

Sentimentu na úrovni vět

Klasifikace sentimentu na úrovni vět zkoumá názory, jak již název vypovídá, na úrovni jednotlivých vět v dokumentu, přičemž za základní informační jednotku je zde považována právě jedna věta. U jednotlivých vět určuje, zda vyjadřují pozitivní, negativní, případně neutrální (žádný) názor či sentiment. Větná úroveň klasifikace se nejčastěji skládá ze dvou od sebe oddělených úloh klasifikace. V prvním kroku je třeba určit, zda zkoumaná věta vyjadřuje nebo nevyjadřuje názor či sentiment. Tato úroveň analýzy je úzce spjata s klasifikací subjektivity, která rozlišuje objektivní (vyjadřující faktické informace) a subjektivní věty (vyjadřující subjektivní mínění). Ve druhém kroku, když je věta klasifikována jako subjektivní, se určuje, zda je pozitivní nebo negativní. Jelikož výsledek analýzy sentimentu na úrovni dokumentu

je obecný a nepřináší přesné informace a je potřeba hlubší analýzy, mnoho studií začalo používat věty v dokumentu jako přístup k analýze názorů (Boiy a Moens, 2009; Abdul-Mageed a kol., 2014; Chenlo a Losada, 2014; Appel a kol., 2016).

Sentimentu na úrovni slov (aspektů)

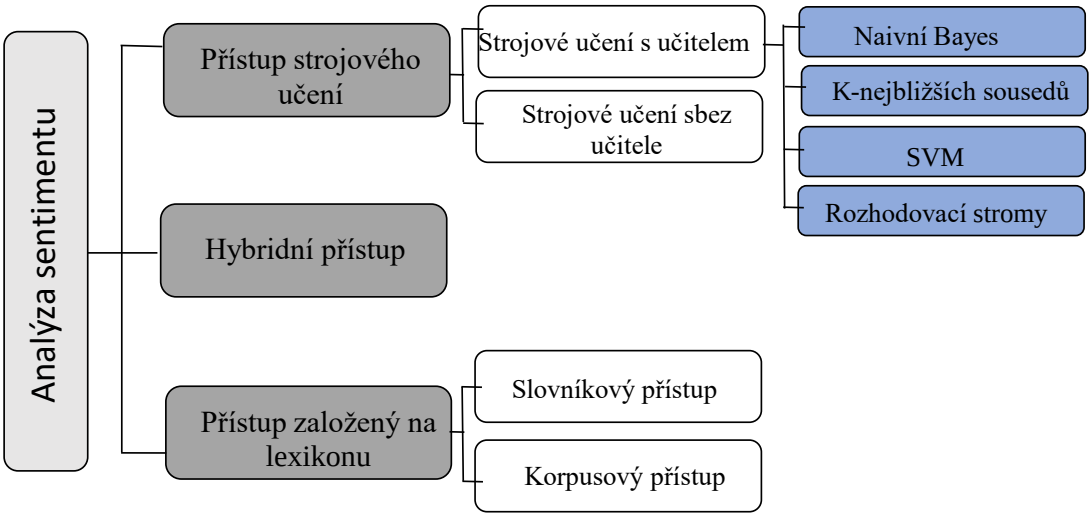
Klasifikace textových nálad na úrovni dokumentu a věty je v mnoha případech zásadní, ale nepředstavuje všechny požadované podrobnosti. Například pozitivní postoj k určité osobnosti neznamená, že autorův názor je kladný na všechny aspekty entity. Podobně negativní sentiment nepředstavuje negativní autorský názor na všechny aspekty entity (Liu a Zhang, 2012). Pro jemnozrnné srovnání dvou nebo více produktů podobných kategorií se musí zjistit výhody a nevýhody různých komponent a funkcí (aspektů). Klasifikace na úrovni dokumentu nebo věty tento typ informací neposkytuje a pro získání těchto podrobností se musí provést dolování názorů na úrovni aspektů.

Aspektová či slovní úroveň, známá také pod názvem *feature-based sentiment analysis*, poskytuje mnohem celistvější a zároveň podrobnější pohled na problematiku analýzy sentimentu a její klasifikaci. Úkolem je určit, co konkrétně se autorovi dokumentu líbí či nelíbí. Základem je myšlenka, že názor se skládá ze sentimentu (pozitivního nebo negativního) a cíle. Cíle jsou obvykle představovány entitami a jejich různými aspekty (vlastnostmi). Úkolem je objevit mínění ohledně nalezených entit a jejich aspektů, z čehož lze poté vytvořit shrnutí. Stojí za povšimnutí, že tato úroveň vytěžování názorů poskytuje podrobnější analýzu cílové entity.

1.3.3 Klasifikace sentimentu

Nejdůležitějším a nejkritičtějším krokem vytěžování názorů je výběr vhodné techniky pro klasifikaci sentimentů. Klasifikace sentimentu, nazývaná také jako stanovení polarity, se zabývá určováním polarity objektu (dokumentu, věty atd.), Ať už objektu vyjadřuje pozitivní, negativní nebo neutrální sentiment vůči subjektu. Jako takový byl aplikován na sociální sítě, recenze produktů, fóra, blogy, novinové články atd. Metody klasifikace, které jsou nabízeny v literatuře, mohou spadat do tří skupin: přístup založený na strojovém učení, přístup založený na lexikonu a hybridní přístup. Základní členění přístupů k analýze sentimentu a dolování názorů je znázorněn na obrázku 1.7.

Podle Kumara a Sebastiana (2012) může analýza prováděná na úrovni slov sledovat přístup lexikonu, které lze rozlišit na jednu ze dvou metod: slovníkovou nebo korpusovou. Metody založené na slovnících používají předem připravené slovníky obsahující termíny a skóre sentimentu k přiřazení skóre sentimentu k textovým datům



Obr. 1.7: Znázornění přístupů k analýze sentimentu
Zdroj:vlastní zpracování

porovnáním termínů ve vzorku s konkrétním slovníkem. Korpusová analýza sentimentu na úrovni slov staví na termínech v korpusu s identifikovatelnými sentimenty k vytvoření asociační sítě pro synonymní termíny, které lze použít ke klasifikaci termínů s neznámým sentimentem (Kumar a Sebastian, 2012). Přístup lexikonu přiřazuje polaritu slovům z dříve vytvořeného slovníku. Tento slovník definuje slovo a jeho polaritu. Pokud lexikon obsahuje stejné slovo nebo frázi, které se objevují v textu, je vrácena jeho hodnota polarity. (Anandarajan a kol., 2019)

Další přístup, který lze využít k analýze sentimentu je strojové učení. Tento přístup je založen na učení a vytváří automatický klasifikátor sentimentu pro sadu dokumentů, která byla dříve opatřena poznámkami. Odtud se proškolí klasifikátor, který lze použít na nová data (Ignatow a Mihalcea, 2016).

Ačkoli jsou všechny tyto přístupy kategorizovány samostatně, jsou vzájemně propleteny, například strojové učení s učitelem lze trénovat s lingvistickými rysy odvozenými od lexikonů polarity. Většina přístupů by tedy mohla být považována za hybridní.

Lexikálně založený přístup

Přístup k analýze sentimentu založený na lexikonu používá dříve přiřazená slova a slovní fráze se stanovenou hodnotou sentimentu k novému textu. Každé slovo nebo fráze, které odpovídají stanovenému slovu nebo frázi v lexikonu, má tuto hodnotu. U celého textu jsou hodnoty poté sečteny (Silge a Robinson, 2017). Pro použití při analýze sentimentu existuje řada lexikonů se skórováním. Štítky obvykle obsahují

indikátor pro pozitivní a negativní nebo skóre, které označuje sílu polarity. Slova a fráze, které vyjadřují pozitivní nebo negativní sentiment, slouží tak k analýze sentimentu. Ve výzkumné literatuře se taková slova nazývají také sentimentální slova, polární slova nebo názorová slova. Pozitivní sentimentální slova se používají k vyjádření některých požadovaných stavů nebo vlastností, zatímco negativní sentimentální slova se používají k vyjádření některých nežádoucích stavů nebo vlastností. Kromě jednotlivých slov existují také sentimentální fráze a idiomy. Souhrnně se jim říká lexikon sentimentu (nebo názorový lexikon). Existují tři hlavní přístupy ke kompilaci sentimentálních slov: manuální přístup, přístup založený na slovníku a přístup založený na korpusu. Manuální přístup je náročný nejen na práci ale i čas a je proto obvykle používán jako kontrola automatizovaných přístupů, protože automatizované přístupy dělají chyby. (Liu, 2015)

Slovníkový přístup

Použití slovníku k sestavování sentimentálních slov je zřejmým přístupem, protože většina slovníků uvádí pro každé slovo synonyma a antonyma. Jednoduchou technikou je tedy použít několik úvodních názorových slov na základě synonymní a antonymové struktury slovníku. Konkrétně tato metoda funguje následovně: Malá sada názorových slov (semen) se známými pozitivními nebo negativními orientacemi (nebo polaritami) se nejprve sbírá ručně, což je velmi snadné. Algoritmus poté tuto sadu rozroste hledáním synonym a antonym v online slovníku. Nově nalezená slova byla přidána do seznamu semen a začíná další iterace. Iterativní proces končí, když již nelze najít žádná nová slova (Hu a Liu, 2004; Valitutti a kol., 2004). Po dokončení procesu se k vyčištění seznamu použije ruční kontrola (odstranění chyb). Seznam lze také vyčistit přiřazením síly sentimentu každému slovu pomocí pravděpodobnostní metody. (Liu, 2015) Populární slovníky sentimentu jsou komparovány z hlediska počtu termínů, typu a formátu skóre sentimentu v tabulce 1.1. Tyto slovníky se významně liší, protože některé sentimentální lexikony definují skóre sentimentu s různým číselným rozsahem, zatímco ostatní definují jednu nebo více kategorií sentimentu, jako je pozitivní, negativní, neutrální.

Kromě výše uvedených obecných slovníků, se zejména ve finanční oblasti využívají také specializované lexikony jako je InvestorWords a Loughran-McDonald. Slovník Loughran-McDonald obsahuje všechna slova použitá ve finančním kontextu. InvestorWords je finanční glosář obsahující přibližně 5 000 definic a 15 000 odkazů mezi souvisejícími pojmy a k vytvoření sémantické sítě souvisejících slov k jejich vyhodnocení dle Yekrangi a Abdolvand (2020). Tyto slovníky využívají upravenou slovní zásobu typickou, jako je znázorněno v tabulce 1.2.

Obecně výhodou použití přístupu založeného na slovníku je, že lze snadno a rychle najít velké množství sentimentálních slov s jejich orientací. Přestože výsledný seznam

Název slovníku	Počet slov	Typ sentimentu	Skóre sentimentu
Opinion Finder	6 886	kategorické	pozitivní, neutrální, negativní
General Inquirer	11 789	kategorické	pozitivní, neutrální, negativní
SentiWordNet	147 306	číselné	-1 až +1
SenticNet	15 143	číselné	-1 až +1
SentiSense	4 404	kategorické	14 emočních kategorií
VADER	7 502	číselné	-1 až +1
Stanford CoreNLP	-	kategorické	velmi negativní, negativní, neutrální, pozitivní, velmi pozitivní
WordNet-Affect	4 552	kategorické	pozitivní, negativní, neurčitý, neutrální
AFINN	2 477	číselné	-5 až +5
SO-CAL	6 306	číselné	-5 až +5
TextBlob	-	číselné	-1 až +1

Tab. 1.1: Přehled obecných slovníků sentimentu
Zdroj: Cho a kol. (2014)

Panel A: Pozitivní slova
buy; buying; bought; long; bull; bullich; good; acceptable; excellent; exceptional; favorable; great; positive; awesome
Panel B: Negativní slova
sell; selling; sold; short; bear; bearish; awful; crummy; dreadful; poor; sad; unacceptable; blah; junky; abominable; faulty; inadequate; unsatisfactory

Tab. 1.2: Ukázka polarity slov finančního lexikonu
Zdroj: McGurk a kol. (2020)

může obsahovat mnoho chyb, lze jej vyčistit manuální kontrolou. Ruční vyčištění je časově náročné, ale je to jen jednorázové úsilí, které pro rodilého mluvčího vyžaduje jen několik dní. Hlavní nevýhodou přístupu založeného na slovníku je, že orientace sentimentu takto shromážděných slov je obecná nebo nezávislá na doméně a kontextu. (Liu, 2015)

Přístup strojového učení

Přístup strojového učení *(angl. Machine Learning,* ML) uplatňuje prominentní algoritmy ML a využívá jazykové funkce. Metody klasifikace textu využívající přístup ML lze rozdělit na metody učení s učitelem a bez učitele. Metody s učitelem používají k analýze sentimentu klasifikátor na základě datové sady s označenými sentimenty. Metody bez učitele se používají, když je obtížné najít tyto označené funkce trénovacích dokumentů (Kaur a Gupta, 2013). Cílem tohoto přístupu je vyvinout

algoritmus pro optimalizaci výkonu systému pomocí minulých zkušeností. Poskytuje řešení problému klasifikace a zahrnuje dva základní kroky: učení modelu z korpusu tréninkových dat a následná klasifikace dat na základě vytvořeného modelu na testovacích datech korpusu (Massung a kol., 2013).

Strojové učení bez učitele

Na rozdíl od strojového učení s učitelem nemá tento přístup žádný explicitní cílený výstup spojený se vstupem. Štítek třídy pro libovolnou instanci není znám, takže učení bez učitele se chystá naučit pozorováním. Na rozdíl od učení s učitelem, je tento přístup nezávisí na doméně a tématu tréninkových dat. Překonává tak obtíže se shromažďováním a vytvářením označených tréninkových dat. V metodách bez učitele se uvažuje o sadě tréninkových vzorků, pro které je zadána pouze vstupní hodnota, a přesné informace o výstupu nejsou k dispozici. Nevyžaduje velké množství údajů o tréninku s anotací člověka k získání přijatelných výsledků. Jednou z hojně využívaných technik je shlukování. Proces shromažďování objektů podobných charakteristik do skupiny. Objekty v jednom klastru jsou odlišné od objektů v jiných klastrech.

Strojové učení s učitelem

Pro přístup strojového učení s učitelem, jsou vyžadovány dva typy datových souborů: tréninková a testovací. Automatický klasifikátor se naučí klasifikační faktory dokumentu z tréninkové sady a přesnost klasifikace lze vyhodnotit pomocí testovací sady. Algoritmy ML, jako je SVM, Naive Bayes, rozhodovací stromy či k-nejbližších sousedů jsou úspěšně používány v mnoha výzkumných pracích a při klasifikaci sentimentu fungovaly dobře. Tradiční klasifikace textu klasifikuje hlavně dokumenty různých témat, v takových klasifikacích jsou klíčovými rysy slova související s tématem. V klasifikaci sentimentu jsou však důležitější slova sentimentu nebo názoru, která naznačují pozitivní nebo negativní názory.

Naivní Bayes

Modely naivního Bayese *(angl. Naive Bayes,* NB) jsou široce používány, protože se snadno implementují a snadno se používají (Allahyari a spol., 2017). Model se opírá o Bayesovo pravidlo, které se používá k odhadu podmíněné pravděpodobnosti.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1.8}$$

Naivní Bayes předpokládá, že podmínky jsou podmíněně nezávislé vzhledem k třídě, což zjednodušuje potřebné výpočty. Naivní Bayesovo klasifikační pravidlo říká, že nejpravděpodobnější klasifikace $\hat{C}$ dokumentu D se rovná:

$$\hat{C} = \arg\max P(D|C) \cdot P(C) \tag{1.9}$$

kde $P(D|C)$ je podmíněná pravděpodobnost dokumentu vzhledem k jeho třídě a $P(C)$ je pravděpodobnost klasifikace. Jelikož se dokumenty skládají z termínů, lze určit pravděpodobnost dokumentu danou klasifikací vynásobením každé z pravděpodobností výrazu, který patří do dané třídy. Tento předpoklad činí odhad modelu efektivním, ale nemusí být nutně přesný v aplikacích využívajících data z reálného světa (Zhang, 2004).

K– nejbližších sousedů
Algoritmus k-nejbližšího souseda *(angl. k-Nearest Neighbors,* k-NN) se používá k testování stupně podobnosti mezi dokumenty a k tréninkovými daty a k uložení určitého množství klasifikačních dat, čímž se stanoví kategorie testovacích dokumentů. Tato metoda je algoritmus okamžitého učení, který kategorizuje objekty na základě nejbližšího prostoru funkcí v tréninkové sadě (Han a kol., 2001). Tréninkové sady jsou mapovány do vícerozměrného prostoru funkcí. Prostor funkcí je rozdělen do oblastí na základě kategorie tréninkové sady. Bod v prostoru funkcí je přiřazen konkrétní kategorii, pokud se jedná o nejčastější kategorii mezi nejbližšími tréninkovými daty. Při výpočtu vzdálenosti mezi vektory se obvykle používá euklidovská vzdálenost. Klíčovým prvkem této metody je dostupnost míry podobnosti pro identifikaci sousedů konkrétního dokumentu. Fáze tréninku spočívá pouze v uložení vektorů funkcí a kategorií tréninkové sady. Ve fázi klasifikace se vypočítají vzdálenosti od nového vektoru představujícího vstupní dokument ke všem uloženým vektorům a vybere se k nejbližších vzorků. Anotovaná kategorie dokumentu je predikována na základě nejbližšího bodu, který byl přiřazen konkrétní kategorii Baharudin a kol., (2010).

Support Vector Machines
Metoda podpůrných vektorů *(angl. Support Vectore Machine,* SVM) zkonstruuje nadrovinu v prostoru funkcí představovaném vektorem w, který odděluje mezi dvěma třídami, které jsou zastoupeny v pozitivních a negativních trénovacích vektorech dokumentů s výpočtem maximálního rozpětí. Pro nalezení nejlepší poloroviny je nezbytné vyřešit dvojí optimalizace. Nechť y_i se rovná (+1), pokud $\vec{d_i}$ dokument leží ve třídě (+) a y_i se rovná (-1), pokud $\vec{d_i}$ dokument leží ve třídě (-) dle Joshi a Itkat (2014).

$$\vec{w} = \sum_{i=1}^{n} a_i \cdot Y_i \vec{d_i} a_i \geq 0 \tag{1.10}$$

kde a_i ukazuje, že váhový vektor nadroviny, který vyplynul z řešení předchozích optimalizačních problémů, je konstruován jako lineární kombinace $\vec{d_i}$. Tyto vektory

se nazývají podpůrné vektory, protože pouze přispívají k $\vec{w}$ jsou vektory dokumentů. Joshi a Itkat (2014) SVM je přesný, efektivní a rychlý trénink a bylo prokázáno, že překonává jiné metody, včetně naivního Bayese a další metod. (Anandarajan a kol., 2019)

Rozhodovací stromy
Klasifikace rozhodovacího stromu *(angl. Decision Tree,* DT) je neparametrický přístup, který využívá rekurzivní dělení k oddělení tříd v datové sadě (Rokach a Maimon, 2005). Rozhodovací strom znovu sestavuje manuální kategorizaci tréninkových dokumentů vytvořením dobře definovaných pravdivých/nepravdivých dotazů ve formě stromové struktury. Ve stromové struktuře rozhodnutí představují listy odpovídající kategorii dokumentů a větve představují spojení prvků, které k těmto kategoriím vedou. Dobře organizovaný rozhodovací strom může snadno klasifikovat dokument tak, že jej umístíte do kořenového uzlu stromu a nechá se běžet strukturou dotazu, dokud nedosáhne určitého listu, který představuje cíl klasifikace dokumentu. Naučené stromy lze také reprezentovat jako sady pravidel if-then (Mitchell, 1997).

Hybridní přístup

Někteří vědci kombinovali přístupy založené na ML a lexikonu pod dohledem, aby zlepšili výkon klasifikace sentimentu. Hybridní přístup kombinuje výhody obou technik. Koncept kombinování klasifikátorů je navržen jako nový směr pro zlepšení výkonu jednotlivých klasifikátorů. V poslední době bylo navrženo mnoho metod pro vytvoření souboru klasifikátorů. Mechanismy, které se používají k sestavení souboru klasifikátorů, dle Ikonomakis a kol. (2005); Baharudin a kol., (2010), zahrnují: i) použití různých podmnožin tréninkových dat s jedinou metodou učení, ii) použití různých parametrů tréninku s jedinou metodou tréninku (např. použití různých počátečních vah pro každá neurální síť v souboru) a iii) používání různých metod učení.

Srovnání vybraných klasifikátorů

Je patrné, že ke klasifikaci sentimentu lze využít mnoho algoritmů strojového učení, které byly popsány výše. V tabulce 1.3 jsou popsány výhody a nevýhody těchto algoritmů.

Metoda	Výhody	Nevýhody
K– nejbližších sousedů	• Intuitivní algoritmus. • Fáze školení je velmi rychlá a její náklady jsou nulové. • Jednoduchá a snadná implementace. • Dokáže si poradit s hlučnými daty.	• Uživatel musí určit počet klastrů, což nebude vždy snadné. • Vysoká relativní výpočetní složitost. • Je velmi citlivý na irelevantní vlastnosti. • K uložení všech příkladů školení potřebuje obrovskou paměť. • Pro velké datové sady může být kNN velmi neefektivní.
Naivní Bayes	• V malých datových sadách funguje dobře, pokud platí podm. nezávislý předpoklad. • Jsou snadno implementovatelné. • Rychlé klasifikace. • Není citlivý na irelevantní funkce.	• Předpoklad nezávislosti mezi funkcemi. • Relativně nízký výkon klasifikace ve srovnání s jinými diskriminačními algoritmy.
SVM	• Může poskytovat nelineární řešení. • Poměrně robustní proti přetížení a dokáže se vypořádat s daty, která obsahují chyby. • Dobře se přizpůsobuje vysoce dimenzionálním datům. • Přesnost predikce je velmi vysoká a poskytuje dobrý výkon zobecnění. • Poskytuje jedinečné řešení, protože optimalizační problém je konvexní, což znamená, že má jedinečnou minimální hodnotu.	• K dosažení dobrého výkonu vyžadují znalosti o použitém jádře. • Jsou náročné na paměť. • Obtížná interpretace. • Nedostatek transparentnosti ve výsledcích, protože se jedná o neparametrickou metodu.
Rozhodovací stromy	• Zvládá kategorické funkce. • Obsahuje pouze několik parametrů k vyladění. • V souborech dat s velkým počtem funkcí fungují dobře. • Snadný převod na soubor pravidel, která jsou čtenáři často srozumitelná. • Může snadno zpracovávat odlehlé a chybějící hodnoty. • Snadná replikace pomocí jednoduchých matematických algoritmů.	• Interpretovatelnost souboru může být zpochybněna. • Výpočetní náklady mohou být vysoké. • Má problém s overfittingem. • Rozhodovací stromy mohou mít u některých konceptů výrazně složitější zastoupení kvůli problému s replikací.

Tab. 1.3: Přehled výhod a nevýhod přístupů strojového učení s učitelem
Zdroj: vlastní zpracování dle Juarez-Orozco a kol. (2018), Mohamed (2017)

1.4 Expertní systémy

Během posledního desetiletí vzrůstá zájem o výsledky výzkumu umělé inteligence. Zejména oblast systémů založených na znalostech, jedna z prvních oblastí umělé inteligence, která byla komerčně plodná, získala velkou pozornost. Fráze systém založený na znalostech je obecně používán k označení informačních systémů, ve kterých je aplikováno určité symbolické znázornění lidského poznání, obvykle způsobem podobným lidskému uvažování. Z těchto systémů založených na znalostech byly v současnosti nejúspěšnější expertní systémy. Expertní systémy (angl. Expert System, ES) jsou systémy, které jsou schopny nabídnout řešení konkrétních problémů v dané oblasti nebo které jsou schopny poskytovat rady, a to způsobem a na úrovni srovnatelné s úrovní odborníků v daném oboru. Budování expertních systémů pro konkrétní aplikační domény se stalo dokonce samostatným předmětem známého znalostního inženýrství dle Lucas (1991). Problémy v oblastech, pro které jsou vyvíjeny expertní systémy, jsou ty, které pro jejich řešení vyžadují značné lidské znalosti. Příklady takových problémových domén jsou lékařská diagnóza nemoci, finanční poradenství, design produktů atd. Většina současných odborných systémů je schopna řešit pouze omezené problémové oblasti. Expertní systémy jsou většinou upřednostňovány, protože vytvářejí přijatelná řešení i pro některé špatně strukturované problémy, které nemají efektivní algoritmické řešení (Giarratano a Riley, 2004). Tento základní pohled na vývoj současných expertních systémů je formulován v následující rovnici, která se někdy nazývá paradigma designu expertního systému:

$$\text{expertní systém} = \text{znalost} + \text{inference}$$

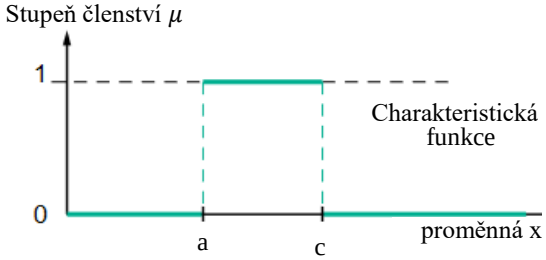
Expertní systém tedy obvykle zahrnuje následující dvě základní součásti: znalostní základnu zachycující znalosti specifické pro doménu a inferenční nástroj skládající se z algoritmů pro manipulaci se znalostmi zastoupenými ve znalostní bázi. Znalosti specifické pro doménu jsou stanoveny ve znalostní bázi pomocí speciálního formalismu reprezentace znalostí. V prostředí expertního systému nebo nástroji pro vytváření expertních systémů je předdefinován jeden nebo více formalismů reprezentace znalostí pro kódování znalostí domény. Dále je k dispozici odpovídající inferenční nástroj, který je schopen manipulovat se znalostmi představovanými v takovém formalismu. Vývoj konkrétního expertního systému se provádí konzultováním různých zdrojů znalostí, jako jsou učebnice, a databáze či konzultace s odborníky. Vybudování expertního systému je úkol vyžadující vysoké dovednosti; osoba provádějící tento úkol se nazývá znalostní inženýr. Proces sběru a strukturování znalostí v problémové doméně se nazývá získávání znalostí Lucas (1991).

1.4.1 Fuzzy logika

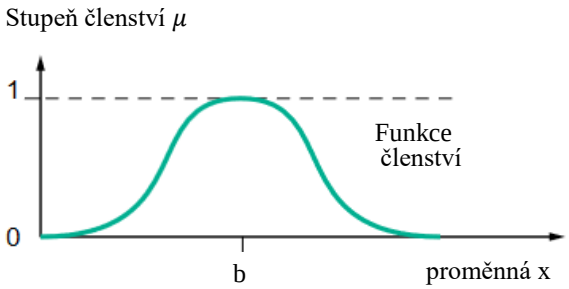
„Roste-li složitost systému, klesá naše schopnost formulovat přesné a významné soudy o jeho chování, až je dosaženo hranice, za níž jsou přesnost a relevantnost prakticky vzájemně se vylučující charakteristiky.“

—Zadeh (1965)

Fuzzy logika *(angl. Fuzzy Logic,* FL) vyvinutá profesorem Lotfim Zadehem (1965), je schopna popsat neurčitost a nejednoznačnost obsaženou v datech ze skutečného světa. Profesor Zadeh uvažoval o inteligenci v lidském uvažování, aby připravil půdu pro založení fuzzy množin a fuzzy logiky. Motivovala ho skutečnost, že lidé častěji komunikují prostřednictvím termínů přirozeného jazyka nebo jazykových výrazů, které nelze vždy kvantifikovat číselnými hodnotami. Typické lidské výrazy zahrnují buď jednu, nebo několik jazykových frází jako malé, velké, nízké, střední, vysoké atd. Tyto fráze, i když výstižně popisují lidské pocity, nelze číselně kvantifikovat. Zadeh vytvořil termín „fuzzy“ (stál za něčím, co je vágní, nejasné a nepřesné), aby replikoval pojem neměřitelného lidského porozumění a logiky. Fuzzy množiny tedy tvoří páteř efektivnějších a robustnějších systémů, které jsou imunní vůči všem druhům nejistot a nepřesností převládajících v reálném světě. Fuzzy systémy proto fungují v jazykovém rámci a jejich síla spočívá v jejich schopnosti zpracovávat jazykové informace a provádět přibližné uvažování prostřednictvím přiřazení neměřitelných logických vlastností dle Ross (2004) a Ross a kol. (2003). Fuzzy logika poskytuje příležitost pro modelování podmínek, které jsou ze své podstaty nepřesně definovány. Fuzzy techniky ve formě přibližného uvažování poskytují podporu rozhodování a expertní systémy s výkonnými schopnostmi uvažování.



Obr. 1.8: Charakteristická funkce
Zdroj: Chevrie a Guely (1998)



Obr. 1.9: Funkce členství
Zdroj: Chevrie a Guely (1998)

Konvenční logika (ostré množiny) není schopna zvládnout různou míru nepřesnosti, nejistoty a neurčitosti, které vykazují pozorování v reálném světě. Důvodem je

skutečnost, že v ostrých množinách jsou informace častěji vyjádřeny v kvantifikačních propozicích, které nejsou vždy schopny popsat základní nejistotu v pozorováních v reálném životě. Přesněji řečeno, ostrá logika je bivalentní logika, která s sebou nese pouze jeden ze dvou možných výsledků (pravda nebo nepravda) jakéhokoli pozorování. (Kosko, 1997). Fuzzy množina je rozšířením ostré množiny. Zatímco ostré sady umožňují pouze úplné členství nebo vůbec žádné členství k základním prvkům, fuzzy sady umožňují částečné členství. V ostré množině je členství nebo nečlenství prvku x v množině A popsáno charakteristickou funkcí $\chi_A(x)$, kde $\chi_A(x) = 1$, pokud $x \in A$ a $\chi_A(x) = 0$, pokud $x \notin A$. Teorie fuzzy množiny rozšiřuje tento koncept definováním částečných členů, což umožňuje existenci všech prvků ve vesmíru diskurzu U s různým stupněm členství. Fuzzy množina $\tilde{\mathrm{A}}$ ve vesmíru diskurzu U je charakterizována funkcí členství $m_{\tilde{\mathrm{A}}}(x)\hat{I}$ [0,1]. Čím blíže je hodnota členství prvku k jednotě, tím silnější je zadržení prvku ve fuzzy množině. Podobně nižší hodnota členství znamená slabší zadržení prvku v množině. Bhattacharyya a Dutta (2012)

Fuzzy množina je definována „funkcí členství“, která odpovídá pojmu „charakteristická funkce“ v klasické logice (Zadeh, 1965). Funkce členství jsou obecně definovány jako jednorozměrné křivky. Jsou-li definovány ve vyšších dimenzích, stávají se z těchto křivek povrchy nebo hyperplochy. Charakteristická funkce (viz obrázek 1.8) udává „0“ pro proměnnou x mimo rozsah $[a, b]$ a „1“ pro proměnnou x v tomto rozsahu. Fuzzy množina je definována „funkcí členství“, která se liší od charakteristické funkce v tom, že může nabývat jakékoli hodnoty v rozsahu [0, 1]. Každé možné hodnotě proměnné x je přiřazen „stupeň členství“ vyobrazená fuzzy množina (viz obrázek 1.9) s hodnotou mezi 0 a 1 dle Chevrie a Guely (1998).

V literatuře existují dva hlavní přístupy k fuzzy logice:

- fuzzy množiny typu 1: členské funkce jsou zcela jisté.
- fuzzy množiny typu 2: členské funkce jsou sami fuzzy.

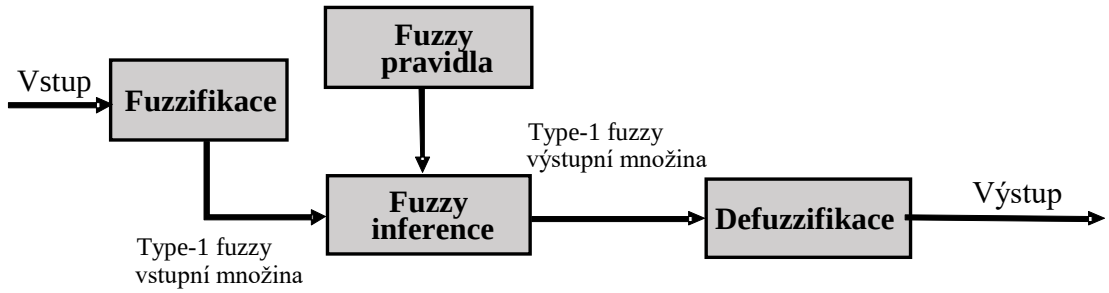
Následující podkapitoly jsou věnovány pochopení podstaty teoretických konceptů fuzzy logiky obou výše uvedených typů.

Type-1 fuzzy logika

Jak uvádí Novák (2000), fuzzy logika umožňuje obsáhnout nepřesnost a poměrně jednoduchým způsobem pracovat s významy slov přirozeného jazyka. Důvod proč fuzzy logika funguje, je poměrně překvapivý, neboť využívá vágně charakterizované expertní znalosti. Jde o vztah mezi relevancí a přesností informace, tedy princip, který hlavní představitel fuzzy logiky Lotfi A. Zadeh (1965), nazval principem inkompatibility. Jura (2005) dodává, že člověk se v řadě situací rozhoduje na základě nepřesných či neurčitých informací, které získává z vnějšího okolí, nicméně i přesto je výsledek jeho činnosti dosažený na základě těchto vágních údajů dostačující. Chang

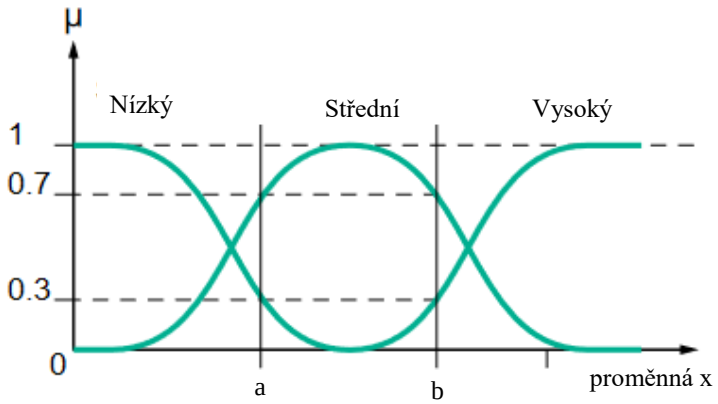
a spol. (2011) navazují a uvádějí, že prognóza akciového trhu může být úspěšná pouze s použitím nástrojů a technik, které mohou překonat problém nejistoty a nelinearity cen. Potenciál fuzzy logiky při zlepšování prognostických modelů lze nalézt v různých aplikacích díky své známé schopnosti překlenout propast mezi číselnými údaji (kvantitativní informace) a jazykovým výrazem (kvalitativní informace).

Type-1 fuzzy logika (T1FLS) se skládá ze čtyř hlavních prvků, jak je znázorněno na obrázku 1.10 a stručně popsána následovně:



Obr. 1.10: Struktura systému type-1 fuzzy logiky
Zdroj: Janková a Dostál (2019)

Fuzzifikace je rozhraní, které mapuje ostré číslo na fuzzy doménu definovanou fuzzy množinou. Fuzzifikace umožňuje převést skutečnou hodnotu na fuzzy. Skládá se ze stanovení stupně členství hodnoty k fuzzy množině. Lze definovat řadu fuzzy množin, například množiny „malá“, „střední“ a „vysoká“, přičemž každý popis je vysvětlen pomocí funkce členství, jako je znázorněno na obrázku 1.11. Proměnná, stejně jako pojmy (například: střední, vysoký) definované funkcemi členství, jsou známé jako lingvistická proměnná, respektive jako lingvistické termíny.



Obr. 1.11: Členská funkce a lingvistický termín
Zdroj: Chevrie a Guely (1998)

Typické členské funkce zahrnují trojúhelníkové, lichoběžníkové, generalizované zvonové tvary, Gaussovy křivky, polynomické křivky a sigmoidní členské funkce. Tyto členské funkce se často označují jako běžné členské funkce, protože každému prvku ve vesmíru diskurzu přiřazují konkrétní hodnotu členství. Bhattacharyya a Dutta (2012). Giarratano a Riley (2004) navrhli zobecněné členské funkce pro specifikaci nejasností v situacích, kdy běžné členské funkce nedokážou přesně určit hodnoty členství kvůli nedostatku odpovídajících informací. Typickou generalizovanou funkcí členství je funkce členství v intervalech, která je konstruována na základě horní a dolní hranice stupňů členství, typicky u type-2 fuzzy logiky, pro každý prvek vesmíru. Odpovídající fuzzy množiny se označují jako intervalové fuzzy množiny. Členství prvku x ve fuzzy množině s hodnotami intervalu tedy představuje interval členství (α_1, α_2). Kromě toho lze s intervalem (α_1, α_2) zacházet jako s fuzzy množinou. Výsledná funkce členství se pak označuje jako fuzzy množina typu 2. Je možné poznamenat, že volba konkrétní funkce členství závisí na uvažovaném chování systému.

Účelem **báze fuzzy pravidel** je formalizovat a implementovat metodu rozumu lidského úsudku. Jako takový jej lze zařadit do pole umělé inteligence. Nejběžněji používaným nástrojem v aplikacích fuzzy logiky je základna fuzzy pravidel. Fuzzy databáze je vytvořena z pravidel, která se běžně používají paralelně, ale která mohou být také zřetězena v některých aplikacích. Pravidlo je typu: JESTLIŽE „predikát“ PAK „závěr“. Základny fuzzy pravidel, stejně jako konvenční systémy odborníků, spoléhají na znalostní základnu odvozenou z lidských znalostí. Chevrie a Guely (1998)

Chování fuzzy systému je charakterizováno množinou jazykových pravidel, která tvoří základ pravidel. Typické jazykové pravidlo má následující formu:

JESTLIŽE (je splněna sada podmínek), PAK (lze vyvodit sadu důsledků)

Předpokladem pravidla je podmínka ve vstupní doméně V a důsledkem je akce, která se má provést ve výstupní doméně W. Protože předpoklady a důsledky těchto pravidel jsou spojeny s fuzzy pojmy, jsou pravidla vyjádřena jako fuzzy pravidla, například:

$$JESTLIŽE\ x_1\ je\ A_1\ AND\ x_2\ je\ A_2,\ POTOM\ y\ je\ B \tag{1.11}$$

kde x_1 a x_2 jsou skalární vstupy, A_1 a A_2 jsou vstupní jazykové termíny reprezentované fuzzy množinami a B je výstupní lingvistický termín reprezentovaný fuzzy množinou. Obecnou základnu pravidel pro n-rozměrné fuzzy systémy, jejichž fuzzy pravidla jsou v následujícím tvaru: Leondes (1998).

Každé pravidlo má K_l n-dimenzionální vstupní fuzzy termíny, jejichž projekce do každé dimenzionální je vstupní lingvistické termíny. W-rozměrné vstupní fuzzy členy

jsou reprezentovány n-dimenzionálními fuzzy množinami A^l. Řadu vstupních fuzzy členů lze kombinovat pomocí AND a poté OR za vzniku předpokladu fuzzy pravidla. Je třeba poznamenat, že zde použité vstupní fuzzy termíny a výstupní fuzzy termíny jsou fuzzy množiny s vícerozměrnými funkcemi členství. Fuzzy pravidla daná dříve jsou velmi obecná, takže lze zahrnout fuzzy pravidla konvenčních fuzzy systémů dle Leondes (1998).

$$\begin{aligned} JESTLIŽE\ (x\ je\ A_1^l\ AND\ x\ je\ A_2^l\ AND, \cdots, AND\ x\ je\ A_{l1}^l) \\ OR\ (x\ je\ A_{l1+1}^l\ AND\ x\ je\ A_{l1+2}^l\ AND, \cdots, AND\ x\ je\ A_{l2}^l) \\ \vdots \\ OR\ (x\ je\ A_{l_{K-1}+1}^l\ AND\ x\ je\ A_{l_{K-1}+2}^l\ AND, \cdots, AND\ x\ je\ A_{lk1}^l) \\ POTOM\ y\ je\ B^l,\ k_1 \in 1, 2, \ldots, K_l,\ l = 1, 2, \cdots, L \end{aligned} \tag{1.12}$$

Existují různé modely **inference** pro analýzu FIS modelu (Passino a Yurkovich, 1998). Mezi nejčastěji užívané patří Mamdani a Takagi Sugeno Kang. Tyto grafické metody jsou uživatelsky přívětivější než běžné matematické nástroje dostupné na trhu, a dokonce umožňují začátečníkům experimentovat a navrhovat robustní fuzzy odvozovací systémy počínaje od nuly. Mamdani (Mamdani, 1977) je nejvíce přijímaný inferenční model, vstupy a výstupy jsou v grafickém zobrazení tohoto modelu reprezentováni jako členské funkce. Vzhledem k tomu, že má intuitivní povahu, má široké přijetí. Kromě toho je navržen tak, aby přijímal lidské vstupy. Model Takagi-Sugeno Kang (Passino a Yurkovich, 1998) je v několika ohledech podobný modelu Mamdani, pokud jde o proces fuzzy inference, proces fuzzifikace vstupu a použití fuzzy operátorů. Zřetelný rozdíl mezi těmito dvěma modely spočívá ve skutečnosti, že výstupní členské funkce Sugeno jsou buď lineární nebo konstantní. Proto je výpočetně efektivní a funguje dobře s lineárními, optimalizačními a adaptivními technikami dle Bhattacharyya a Dutta (2012).

Vzhledem k inherentní síle fuzzy množin, které představují neurčitost a nejednoznačnost informací ze skutečného světa, je nutné konstatovat, že konvenční výpočetní technika se při provádění rozhodovacího procesu spoléhá v zásadě na přesné ostré hodnoty. Fuzzy proměnná proto musí být převedena na její ostrý protějšek, aby bylo možné dosáhnout přesně kvantifikovaného rozhodnutí. Proces převodu proměnných z fuzzy do ostré domény se označuje jako **defuzzifikace** (Ross, 2004).

Type-2 fuzzy logika

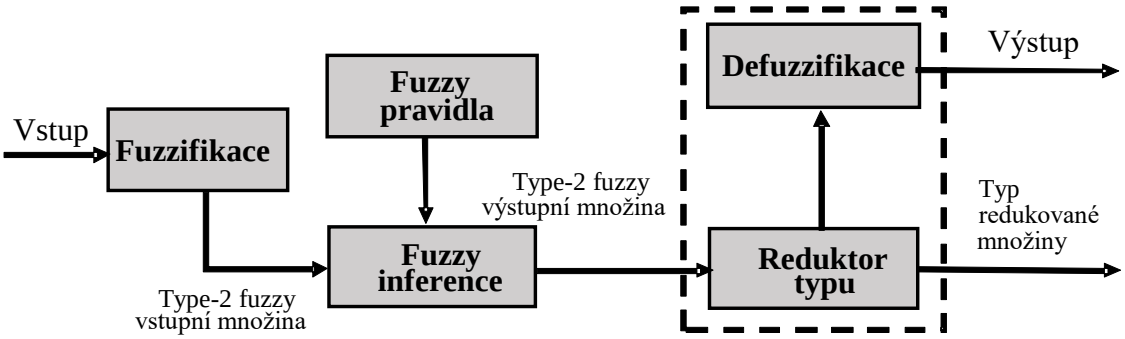
Type-2 fuzzy logiky (T2FLS) byla zavedeny Zadehem v roce 1975 jako rozšíření fuzzy množin typu 1. Karnik a Mendel (2001) vyvinuli teorii fuzzy množin typu 2. Teoretické základy fuzzy systému intervalu typu 2 a jeho konstrukční principy jsou

popsány v Mendel a spol. (2006). T2FLS se zdají být slibnější metodou než jejich protějšky typu 1 pro řešení nejistot, jako jsou hlučná data a měnící se prostředí. Ve studii Khanesar a kol. (2010) jsou simulovány účinky šumu měření typu 1 a typu 2 a identifikátorech za účelem provedení srovnávací analýzy. Byl učiněn závěr, že použití T2FLS v aplikacích v reálném světě, které vylučují šum měření a nejistoty modelování, může být lepší volbou než T1FLS. Pokud má systém velké množství nejistot, nemusí být T1FLS schopny dosáhnout požadovaná úroveň výkonu s rozumnou složitostí struktury. V takových případech se doporučuje použití T2FLS jako preferovaného přístupu v literatuře v mnoha oblastech, jako je předpovídání časových řad, resp. jako podpora pro rozhodování viz Janková a Dostál (2019).

Jak uvádí Melin a Castillo (2014) a Castillo a spol. (2007), není rozumné využívat přesné funkce příslušnosti fuzzy logiky pro něco nejistého. V tomto případě je nezbytné využít k řešení jiný typ fuzzy logiky, která je schopna zvládnout tyto nejistoty. Castillo a spol. (2013) uvádí, že T2FLS jsou v podstatě „fuzzy fuzzy“ množiny. Chen a spol. (2018) upozorňují, že T2FLS mají vyšší aproximační schopnost než neuronové sítě. Nicméně výzkumníci museli nějakou dobu počkat, aby se teorie mohla dále vyvíjet a rozvíjet. Pokrok T2FLS primárně omezovalo hardwarové vybavení.

Nedávno získal systém type-2 fuzzy logiky popularitu v široké škále aplikací zejména díky své schopnosti zvládnout vyšší stupeň nejistoty. Janková a Dostál (2019) upozorňují na skutečnost, že znalosti, které jsou využívány při konstrukci pravidel v systému type-1 fuzzy logiky (T1FLS), jsou nejisté. Existují tři způsoby, jak se taková nejistota v pravidlech může objevit: (1) slova, která jsou používána v antecedentech a konsekventech pravidel mohou pro různé lidi znamenat různé věci; (2) konsekventy získané při hlasování skupiny odborníků se často liší pro stejná pravidla; (3) šum v trénovacích datech. Antecedentní nebo konsekventní nejistota se transformuje do antecedentní a konsekventní funkce příslušnosti. Systémy type-1 fuzzy logiky nejsou schopny přímo zapracovat tyto nejistoty, naproti tomu systémy type-2 fuzzy logiky mohou tuto nejistotu zvládnout.

Struktura T2FLS je velice podobná struktuře T1FLS. Naměřené reálné proměnné jsou nejprve transformovány v bloku fuzzifikace na jazykové proměnné, přičemž jazykové proměnné vycházejí ze základních lingvistických proměnných. Dostál (2011) konstatuje, že se obvykle využívá tři až sedm atributů této základní proměnné. Stupeň příslušnosti atributů dané proměnné v množině je znázorňován matematickou funkcí. V T2FLS jsou k dispozici tři typy fuzzifikace. Jestliže jsou naměřená data perfektní, modelují se jako ostrá množina, data se šumem a data se stacionárním šumem se modelují jako type-1 fuzzy množiny, s nestacionárním šumem se modelují jako type-2 fuzzy množiny. Přičemž posledně jmenovaný typ fuzzifikace nelze provést v T1FLS, jak popsiuje Janková a Dostál (2021a).



Obr. 1.12: Struktura systému type-2 fuzzy logiky
Zdroj: Janková a Dostál (2019)

Jak uvádí Medasani a spol. (1998), všechny existující T2FLS funkce příslušnosti nejsou nic jiného než modifikované verze konvenčních T1FLS funkcí. Jinými slovy, základ původní funkce příslušnosti type-1 je rozmazaný, jestliže si odborník není jistý hodnotami funkce příslušnosti kolem určitého konkrétního bodu. Existuje řada fuzzy funkcí příslušnosti typu 2, tj. trojúhelníkové, Gaussovy, lichoběžníkové, sigmoidální apod. Kayacan a spol. (2018) dodává, že v aplikacích teorie fuzzy množin jsou funkce příslušnosti voleny na základě subjektivního vnímání nejasných nebo nepřesných kategorií. Krom toho neexistují žádná kritéria, která by hodnotila vhodnost či korektnost zvolené funkce příslušnosti.

Po fuzzifikaci následuje fuzzy inference, která definuje chování systému za pomoci pravidel IF-THEN a na jazykové úrovni vyhodnocující stav příslušnosti či pravdivosti proměnné. Každá kombinace atributů proměnných vstupujících do systému a vyskytujících se v podmínce vyjadřuje jedno pravidlo. Následně je potřeba pro každé pravidlo stanovit stupeň podpory neboli váhu pravidla v daném systému. Výsledkem fuzzy inference, jak uvádí Dostál (2011), je jazyková proměnná. Ve většině aplikací je ovšem zapotřebí konečný výstup jako konkrétní číslo, a ne fuzzy množina. V důsledku toho musí být výstupní fuzzy množina konvertována na číslo.

Fuzzy pravidla definují spojení mezi vstupními a výstupními fuzzy proměnnými. Pravidla T2FLS mohou nabízet alternativu, jestliže je potřeba modelovat nejistotu problému a jsou lepší pro nevyužití přesných stupňů členství, například když jsou tréninková data ovlivněna šumem.

V T1FLS je proces konverze na konkrétní číslo nazvaný defuzzifikace. Existuje mnoho způsobů, jak výsledku docílit, např. výpočet těžiště funkce příslušnosti pro danou množinu, výpočet váženého průměru těžiště jednotlivých funkcí členství atd. Nicméně celá záležitost je mnohem komplikovanější pro T2FLS, protože přejít z type-2 fuzzy množiny na konkrétní číslo vyžaduje dva kroky. První krok je pojmenován

jako reduktor typu, ve kterém je type-2 fuzzy množina snížena na type-1 fuzzy množinu. Existuje tolik reduktorů typů jako je metod defuzifikace T1FLS. Nejčastěji se využívá algoritmus vyvinutý Karnikem a Mendelem (2001) a Mendelem (2001), který je iterativní a rychlý. Reduktor typu generuje fuzzy množinu T1FLS, která se následně transformuje na číselný výstup pomocí defuzzifikace. V případě využití metody středu součtů (cos) reduktor typu je jeho matematické vyjádření, dle Taskin a Kumbasar (2015), následující:

$$Y_{\cos(x)} = [y_t, y_r] = \bigcup_{f^n \in F^n(x)} \frac{\sum_{n=1}^{N} y^n f^n}{\sum_{n=1}^{N} f^n} \tag{1.13}$$

kde y_t a y_r definovány jako:

$$y_t = \frac{\sum_{n=1}^{L} y^n f^n + \sum_{n=L+1}^{N} y^n f^n}{\sum_{n=1}^{L} f^n + \sum_{n=L+1}^{N} f^n} \tag{1.14}$$

$$y_r = \frac{\sum_{n=1}^{R} y^n f^n + \sum_{n=R+1}^{N} y^n f^n}{\sum_{n=1}^{R} f^n + \sum_{n=R+1}^{N} f^n} \tag{1.15}$$

přičemž R a L jsou body, které lze nalézt pomocí iterativního algoritmu KM.

Druhý krok zpracování výstupu, který následuje za reduktorem typu, se stále nazývá defuzzifikace. Z obrázku 1.12 je patrné, že mohou existovat dva číselné výstupy u T2FLS označené jako „ostrý“ výstup a typ redukované množiny. Posledně jmenovaná vyjadřuje míru nejistoty, které mají T2FLS kvůli nejistým vstupním měřením dle Mendel (2007); Zarandi a spol. (2009). Z reduktoru typu jsou získány průměrné hodnoty y_t, y_r a výpočet výstupu procesu defuzzifikace je následující:

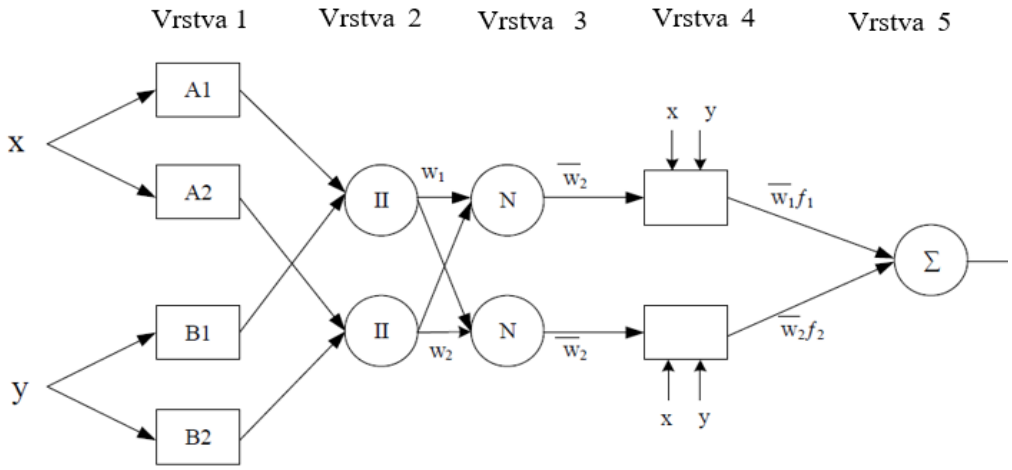
$$y = \frac{y_t - y_r}{2} \tag{1.16}$$

1.4.2 Hybridní model

Díky své konstrukci může adaptivní neuro-fuzzy inferenční systém *(angl. Adaptive Neuro-Fuzzy Inference System,* ANFIS) těžit z výhod obou technik umělé inteligence, ale také překonat jejich nedostatky. Ve srovnání s neuronovými sítěmi je model ANFIS pro uživatele transparentnější a způsobuje méně chyb v zapamatování. V důsledku toho existuje několik výhod ANFIS, včetně adaptační schopnosti, nelineární schopnosti a schopnosti rychlého učení. Neuronové sítě se mohou snadno naučit z dat. Je však obtížné interpretovat získané vědomosti jako význam spojený s každým neuronem a každou hmotností, kterou je docela složité pochopit. Naproti tomu, jak popisuje Mathur a spol. (2016), fuzzy logika sama o sobě se nemůže z dat poučit. Fuzzy modely jsou však snadno srozumitelné, protože používají spíše lingvistické termíny než číselné a strukturu pravidel IF-THEN. Jazykové proměnné jsou

definovány jako proměnné, jejichž hodnoty jsou slova nebo věty v přirozeném jazyce s přidruženými stupni členství. Fuzzy množina, do které lingvistické proměnné patří, je rozšířením „ostré“ množiny, kde prvek může mít plné nebo žádné členství. Fuzzy sady však umožňují i částečné členství, což znamená, že prvek může částečně patřit do více než jedné sady.

V ANFIS je ostrý vstupní signál převeden pomocí funkce příslušnosti na fuzzy vstupy. Fuzzy vstup spolu s funkcí příslušnosti jsou následně přiváděny do bloku neuronové sítě. Blok neuronové sítě sestává z báze pravidel, která je připojena k fuzzy inferenci. Back propagation algoritmus se používá k výcviku inferenčního bloku pro správný výběr základny pravidel. Po zaškolení mohou být generována a vypuštěna vhodná pravidla z bloku neuronové sítě, čímž se získá optimální výstup. Jazykový výstup z bloku neuronové sítě je pak pomocí dešifrovací jednotky převeden na ostrý výstup. Struktura neuro-fuzzy modelu sestává z různých adaptivních vrstev. Každá z těchto vrstev má uzly s přidruženou sítí přenosových funkcí, prostřednictvím kterých se zpracovávají fuzzy vstupy. Výstup z těchto uzlů se pak spojí a získá jediný ostrý výstup, protože konfigurace ANFIS umožňuje pouze jeden výstup modelu. Tento ostrý výstup je zpětnou vazbou jako vstup do modelu a porovnává se s nastavenou hodnotou. Pokud dojde k nějaké odchylce, takto generovaný chybový signál se stane vstupem do ANFIS, čímž se udržuje stabilita v systému dle Boyacioglu a Avci (2010).



Obr. 1.13: Základní struktura modelu ANFIS
Zdroj: Salleh a kol. (2018)

ANFIS je fuzzy Sugeno model zařazený do rámce adaptivních systémů pro usnadnění učení a adaptace. Takový rámec způsobuje, že modelování s ANFIS je systematičtější a méně závislé na odborných znalostech skupiny odborníků či expertů. Pro představení architektury ANFIS jsou zvažována dvě fuzzy pravidla typu IF-THEN

založena na Sugeno modelu prního řádu. Architektura ANFIS pro implementaci těchto dvou pravidel je znázorněna na obrázku 1.13, ve kterém kružnice označuje pevný uzel, zatímco čtverec označuje adaptivní uzel.

Mathur a spol. (2016) popisují model ANFIS. V první vrstvě jsou všechny uzly adaptivní. Výstupy vrstvy 1 jsou fuzzy členskou třídou vstupů, které jsou dány:

$$O_i^1 = \mu_{A_i}(x) i = 1, 2 \tag{1.17}$$

$$O_i^1 = \mu_{B-2}(y) i = 3, 4 \tag{1.18}$$

kde $\mu_{A_i}(x)$, $\mu_{B-2}(y)i$ mohou přijmout jakoukoli fuzzy funkcí členství. Například pokud je použita fuzzy funkce členství ve tvaru zvonku, je zápis $\mu_{A_i}(x)$ následující:

$$\mu_{A_i}(x) = \frac{1}{1 + [(\frac{x-c_i}{a_i})^2]^{b_i}} \tag{1.19}$$

kde a_i, b_i a c_i jsou parametry členské fuzzy funkce, odpovídajícím způsobem řídí funkce ve tvaru zvonku. Ve druhé vrstvě jsou uzly pevné. Jsou označeny písmenem M, což znamená, že fungují jako jednoduchý multiplikátor. Výstupy této vrstvy lze reprezentovat jako:

$$O_i^2 = w_i = \mu_{A_i}(x)\mu_{B_i}(y) i = 1, 2 \tag{1.20}$$

Ve třetí vrstvě jsou uzly také pevné. Jsou označeny písmenem N, což znamená, že hrají normalizační úlohu předchozí vrstvy. Výstupy této vrstvy lze reprezentovat jako:

$$O_i^3 = w_i = \frac{w_i}{w_1 + w_2} i = 1, 2 \tag{1.21}$$

Ve čtvrté vrstvě jsou uzly adaptivní. Výstupem každého uzlu v této vrstvě je jednoduše součin normalizované váhy a polynomu prvního řádu (pro Sugeno model prvního řádu). Výstupy této vrstvy jsou tedy dány:

$$O_i^4 = w_i f_i = w_i(p_i x + q_i y + r_i) i = 1, 2 \tag{1.22}$$

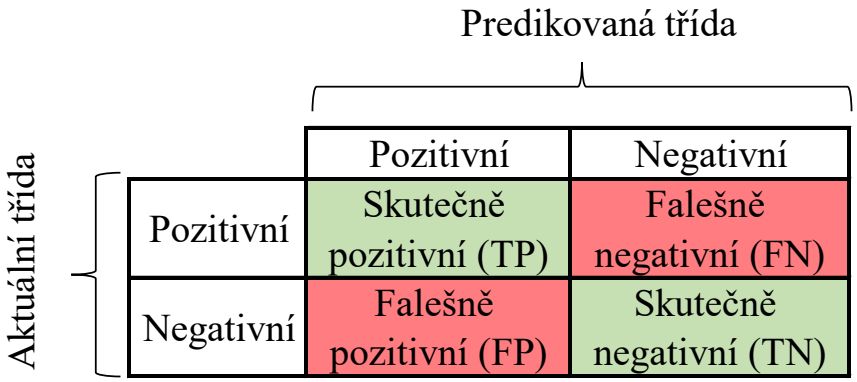
V páté vrstvě je pouze jeden pevný uzel označený S. Tento uzel provádí součet všech předchozích signálů. Celkový výkon modelu je dán:

$$O_i^5 = \sum_{i=1}^{2} w_i f_i = \frac{\sum_{i=1}^{2} w_i f_i}{w_1 + w_2} \tag{1.23}$$

Lze pozorovat, že v této architektuře ANFIS jsou dvě adaptivní vrstvy, a to první vrstva a čtvrtá vrstva. V první vrstvě jsou tři modifikovatelné parametry $\{a_i, b_i, c_i\}$, které se vztahují k vstupním členským funkcím. Ve čtvrté vrstvě jsou také tři modifikovatelné parametry $\{p_i, q_i, r_i\}$ vztahující se k polynmu prvního řádu.

1.4.3 Evaluační metriky

K vyhodnocení metody pro správnou klasifikaci sentimentu, potažmo i pro zhodnocení výkonnosti modelů predikce, je obligatorně využívána matice záměn (angl. confusion matrix) podobné té, která je zobrazena na obrázku 1.14. Matice záměn je tabulka pro vizualizaci toho, jak algoritmus funguje s ohledem na štítky sentimentu, pomocí dvou dimenzí (systémový výstup a štítky sentimentu) a každé označení buněk množinou možných výsledků. Například v případě detekce sentimentu jsou skutečnými pozitivními dokumenty, které skutečně obsahují pozitivní sentiment, o nichž systém správně řekl, že náleží do této kategorie. Falešně negativní jsou dokumenty, které ve skutečnosti vyznívají pozitivně, ale náš systém je nesprávně označen jako dokumenty s negativním sentimentem. V pravém dolním rohu tabulky je uvedena přesnost rovnice, která se ptá, jaké procento korpusu, je systémem označeno správně. Ačkoli se přesnost může zdát přirozená, obvykle se nepoužívá pro úkoly klasifikace textu. Je to proto, že přesnost nefunguje dobře, když jsou třídy nevyvážené. Jurafsky a Martin (2009)



Obr. 1.14: Matice záměn
Zdroj: Jurafsky a Martin (2009)

Proto je místo Accuracy obecně přihlíženo ke dvěma dalším metrikám: Precision a Recall. Precision měří procento položek, které upřesňuje detekovaný systém (tj. systém označený jako pozitivní), které jsou ve skutečnosti pozitivní (tj. jsou pozitivní podle označení štítků). Precision je definována jako:

$$Precision = \frac{TP}{TP + FP} \tag{1.24}$$

Recall měří procento položek skutečně přítomných ve vstupu, které byly systémem správně identifikovány. Recall je definováno jako:

$$Recall = \frac{TP}{TP + FN} \tag{1.25}$$

Precision a Recall tedy na rozdíl od Accuracy zdůrazňují skutečná pozitiva.

Existuje mnoho způsobů, jak definovat jedinou metriku, která zahrnuje aspekty precision a recall. Nejjednodušší z těchto kombinací je F-míra (van Rijsbergen, 1975), definovaná jako:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{1.26}$$

Parametr β váží rozdílně důležitost Precision a Recall, pravděpodobně na základě potřeb aplikace. Hodnoty $\beta > 1$ upřednostňují recall, zatímco hodnoty $\beta < 1$ upřednostňují precision. Když $\beta = 1$, precision a recall jsou stejně vyvážené; toto je nejčastěji používaná metrika a nazývá se $F_\beta = 1$ nebo jen F_1:

$$F_1 = \frac{2PR}{2P + R} \tag{1.27}$$

F-míra vychází z váženého harmonického průměru precision a recall. Harmonický průměr množiny čísel je převrácená hodnota aritmetického průměru převrácených čísel:

$$\bar{x_h}(a_1, a_2, a_3, a_4, \cdots, a_n) = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \cdots + \frac{1}{a_n}} \tag{1.28}$$

a proto F-míra je následující:

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha)\frac{1}{R}} \tag{1.29}$$

popřípadě

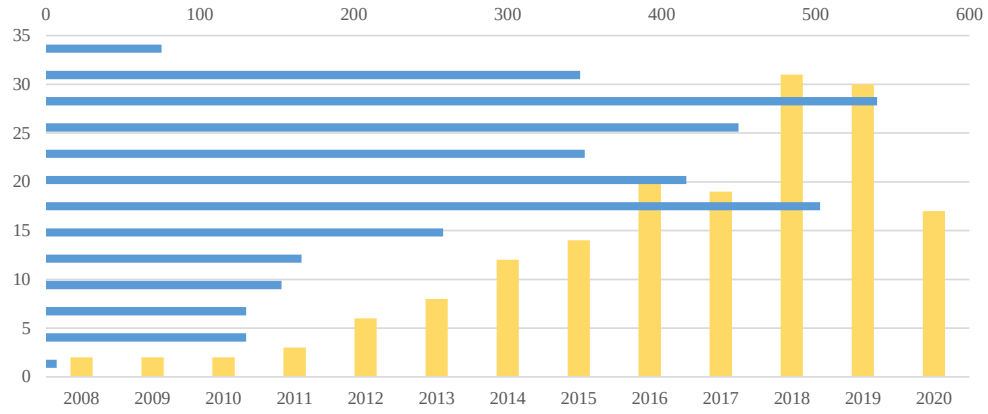
$$F = \frac{(\beta + 1)^2 PR}{\beta^2 P + R} \tag{1.30}$$

Harmonický průměr se používá, protože se jedná o konzervativní metriku; harmonický průměr dvou hodnot se blíží minimu dvou hodnot než aritmetický průměr. Váží tedy nižší z těchto dvou čísel. Jurafsky a Martin (2009)

2 Kritický přezkum současného stavu vědeckého poznání

Cílem této části práce je vytvořit kritický přehled literatury zvláště s ohledem na nejnovější mezinárodní poznatky výzkumných článků ve zvolené tématice. To vyžaduje zkoumat a kriticky rozebrat metody využité při analýze sentimentu z textových dat se zvláštním ohledem na možnost generalizace a přenositelnosti výsledků zkoumání. Z toho důvodu je využit také analytický přístup při práci s literaturou a kritický přístup při její organizace zejména kvůli úplnosti, koherenci a konzistenci.

Informační příprava výzkumu hraje důležitou roli v procesu výzkumné práce. Poskytuje nejen znalostní základnu, ale umožňuje výzkumníkovi poučit se z pracích jiných odborníků. Tato část je zaměřena na klíčové autory a výzkumné problémy, kterými se v rámci zkoumaného tématu zabývali. Se zaměřením na klíčové výzkumné metody a nástroje, které byly využity a klíčové výsledky a závěry, ke kterým autoři dospěli. Je nezbytné odhalit, které otázky v této oblasti zůstaly otevřené a v další části práce se zaměřit právě na tyto bílá místa ve výzkumu.



Obr. 2.1: Vývoj počtu publikací a citací
Zdroj: vlastní zpracování

Pro poskytnutí strukturovaného přehledu o stavu a vývoje aplikace textového dolování a sentimentální analýzy na akciových trzích, je proveden systematický přehled literatury, který byl proveden následovně. Jako první krok systematického přezkumu byly vybrány dvě databáze primárních vědeckých prací, Web of Science a Scopus. Jako druhý krok byl sestaven komplexní seznam klíčových slov, která autoři pravidelně používají k označení aplikace techniky predikce akciového trhu za použití analýzy sentimentu a těžby textu. Tento seznam obsahoval výrazy, které se týkaly různých alternativních zápisů metodické oblasti (např. „text mining“; „sentiment

analysis“), některé z nejběžnějších přístupů predikce za využití předmětných prostředků expertního systému (např. „expert system“; „fuzzy logic“) a konkrétní oblast výzkumu (např. „stock“; „stock market“). Byly vzaty v úvahu pouze články na základě těchto klíčových slov.

Počet publikací a citací v posledních letech podrobně vysvětluje rostoucí výzvy a využití analýzy sentimentu v predikci vývoje akciového trhu. Počet publikací a citací za jednotlivé roky v předmětné oblasti výzkumu je uveden na obrázku 2.1. Svislé sloupce představují počet publikací za rok a vodorovné sloupce vyobrazují trend citací zkoumaných publikací. Počet publikací dosáhl v posledních letech signifikantního nárůstu. Počet citací za rok do roku 2020 stoupá. Je patrné, že výzkum analýzy sentimentu v kontextu predikce vývoje akciového trhu za posledních 12 let zaznamenal velký pokrok a zdaleka není výzkum ještě u konce, o čemž svědčí výrazný nárůst publikací za poslední tři roky.

Podle růstové křivky výzkumu lze identifikovat tři fáze následovně:

(1) Raná fáze (2008–2011): v této fázi je počet dokumentů publikovaných každý rok velice omezený a pohybuje se okolo 2 článků za rok. V tomto období stojí za zmínění výzkum Chena a Zimbra (2010), kteří pokládají základy umělé inteligence v těžbě názorů. Autoři dospěli k závěru, že pozitivní sentiment snižuje objem obchodování, možná proto, že spokojení akcionáři drží své akcie, zatímco negativní sentiment vyvolává obchodní aktivitu. Tato domněnka byla v následujících letech několikrát jinými autory přezkoumávána s rozporuplnými výsledky.

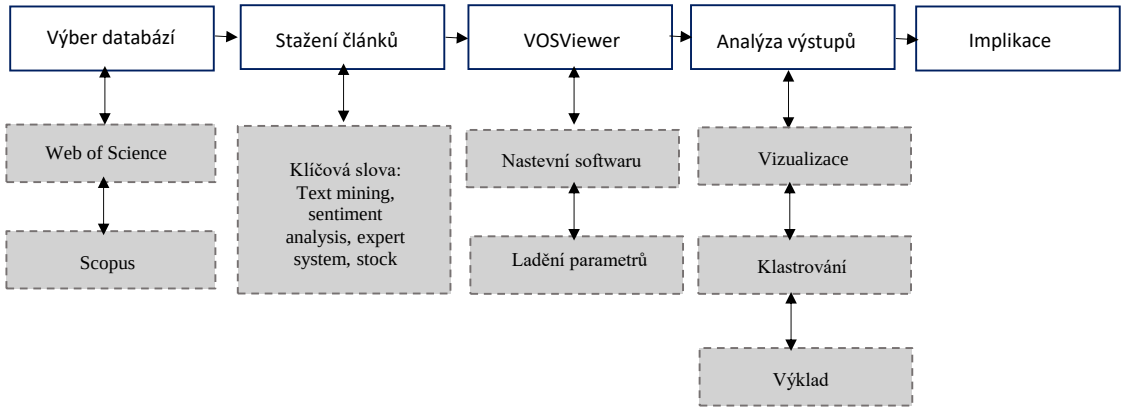
(2) Vývojová fáze (2012–2017): v této fázi se výzkum tohoto zaměření postupně zvyšuje a rychle roste. Počet ročně publikovaných prací se zvýšil na průměrně 13 publikací ročně. Lze shledat postupně zvyšující se zájem a povědomí vědecké komunity o způsoby těžby textu a analýzy sentimentu a jejího potenciálního využití na akciových trzích. Během tohoto období lze nalézt vysoce citovaná díla jako je například Smailovic a kol. (2014), jejichž dílo zaznamenalo 132 citací. Jejich článek analyzuje, zda sentiment vyjádřený ve zdrojích Twitteru, které pojednávají o vybraných společnostech a jejich produktech, může naznačovat jejich změny cen akcií. Jedná se o jednu z prvních studií zaměřených na sociální sítě a rovněž lze tuto studii považovat za odstartování popularity sociálních sítí v predikci vývoje nejen akciového trhu. Na tuto studii dále navazuje Pagolu a kol. (2017), kteří aplikovali analýzu sentimentu a dohlíželi na principy strojového učení na tweety extrahované z Twitteru a analyzovali korelaci mezi pohyby akciového trhu společnosti a sentimenty ve tweetech.

(3) Expanzivní fáze (2018–202x): obsahuje poslední roky, kdy došlo opravdu k expanzivnímu rozšíření a publikování výstupů v předmětné problematice a zdaleka se nejedná o konečné slovo v této oblasti. V této fázi dosáhl počet publi-

kovaných článků 26 ročně a stále roste. Zcela zřejmé je kombinování různých technik expertních a hybridních systémů nejen v textové analýze pro extrakci sentimentu, ale také pro predikci vývoje akciového trhu jako je například SVM, LSTM a další. Mäntylä a kol. (2018) si všimli nejrychleji se rozvíjejících oblastí výzkumu v této oblasti, z toho důvodu představili počítačově podporovanou literární rešerši, kde využívají jak těžbu textu, tak kvalitativní kódování a analyzují 6996 prací ze Scopus. Obdobně Xing a kol. (2018) poskytují přehledový článek a objasňuje rozsah výzkumu zpracování přirozeného jazyka a jeho implikací a strukturovaných technik a aplikací ze související práce.

2.1 Bibliometrická analýza

Bibliometrická analýza je vědecká a často používaná statistická metoda pro analýzu publikovaných studií a slouží k prohledávání citační vztahy a vlivy na jiné obory a publikace. Tato metoda je užita k vyhledání citačních sítí mezi nejpoužívanějšími klíčovými slovy ve studiích zkoumajících sentimentální analýzu na akciovém trhu. K využití této metody je možné použít data z různých databází. Výběr dat na základě dostupnosti se zúžil na dva nejdůležitější pro tuto studii – Web of Science a Scopus. Získaná data byla poté zpracována programem VOSViewer verze 1.6.15, softwarem schopným zobrazit velké bibliometrické mapy s interpretací ukázaných shluků (van Eck a Waltman, 2009).



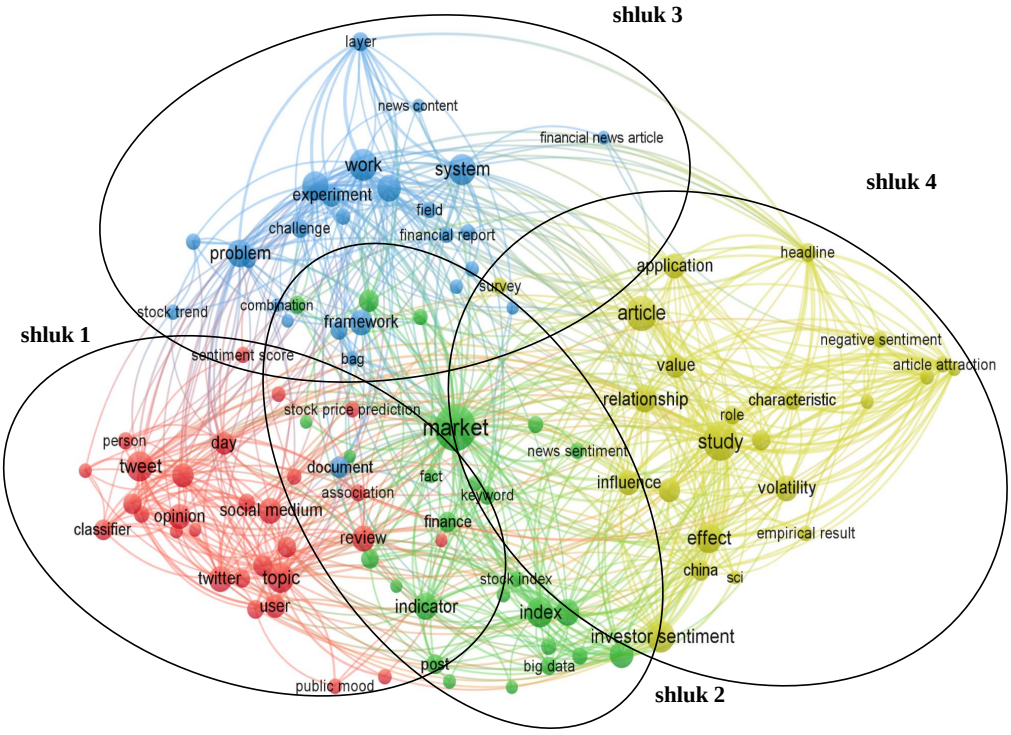
Obr. 2.2: Diagram architektury využití VOSViewer
Zdroj: vlastní zpracování

Skryté potenciální vztahy lze intuitivně pozorovat z mapy vědecké distribuce článků, která může položit dobrý základ pro vědecký výzkum. Obrázek 2.2 ukazuje diagram architektury procesu sběru dat a použití VOSViewer, který představuje výběr konkrétních parametrů během používání VOSViewer. Na základě dat dokumentu

jsou typy uzlů rozhraní VOSViewer nastaveny na „klíčové slovo“ „citovaný časopisu“ a „citovaný dokumentu“. Znalostní mapy výzkumu jsou zobrazeny intuitivně. Tyto analýzy mohou poskytnout vhled do předmětné problematiky a identifikovat oblasti, které jsou již výzkumem pokryté.

2.1.1 Analýza výskytu klíčových slov

Analýza výskytu klíčových slov je zaměřena na všechna klíčová slova s metodou úplného počítání, což znamená, že se počítá každý výskyt výrazu v dokumentu. Byly vybrány pouze bibliometrické sítě, kde se klíčová slova objevila alespoň 5krát, které vyústily v 173 klíčových slov zobrazených na konečné mapě. Konečná mapa ukazuje řadu barevných a různě velkých bublinkových polí, která jsou spojena na kratší nebo delší vzdálenost. Barva představuje část určité skupiny – shluk. Toto mapování skupin identifikuje témata s vysoce propojenými aspekty. Velikost bublinových polí představuje frekvenci používání – čím častěji je klíčové slovo použito ve vzorku, tím větší je jeho bublina. Vzdálenosti mezi klíčovými slovy představují spojení, kratší vzdálenost znamená silnější asociaci mezi klíčovými slovy, a nakonec spojení termínů ve studiích symbolizují řádky. Po bibliometrické analýze s využitím softwaru VOSviewer bylo přijato kvalitativní šetření literatury.



Obr. 2.3: Bibliometrická analýza klíčových slov
Zdroj: VOSviewer (2021)

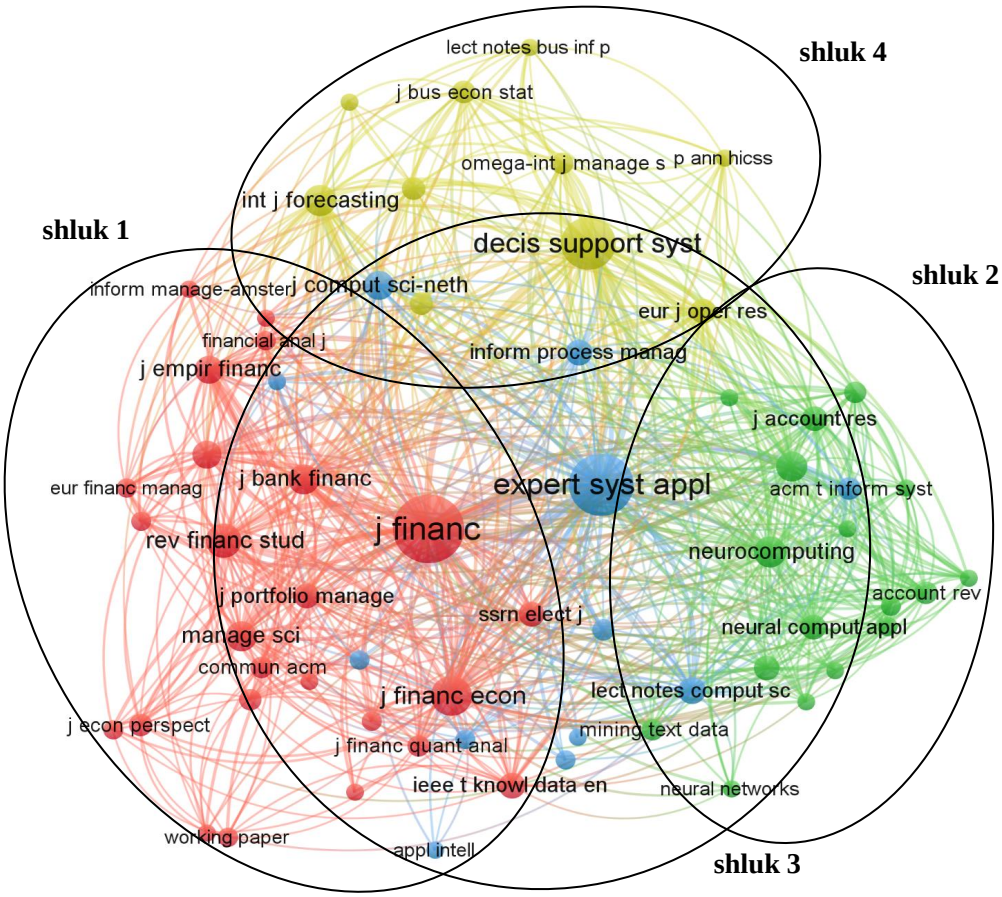
Obrázek 2.3 ukazuje společný výskyt klíčových slov. Celkový počet položek je seřazen do 4 klastrů s 700 odkazy. Jádrem prvního klastru byl pojem „social medium“ s odkazy na „twitter“, „tweet“, „topis“ a „opinion“. Zdá se, že tento klastr obsahuje nejvíce termínů souvisejících se sentimentální analýzy a analýzy názoru na sociálních sítí. Nejčastějším klíčovým slovem druhého klastru byly „market“, poté „index“, „finance“ a „indicator“, tedy pojmy související s akciovým trhem. Třetí skupina zahrnuje vazby mezi slovy „system“, „experiment“, „problem“ a „stock trend“. Indikuje tak problém s extrakcí textových dat a dolováním sentimentu, ke kterému je zapotřebí jisté nástroje, systémy, algoritmy. A konečně, čtvrtý klastr je ten největší a zahrnuje používané klíčové slovo vše „article“, „study“, „investor sentiment“, tedy propojuje klastr 1 a klastr 2 a odkazuje tak na získávání dat z textových dokumentů. Velikost bublin je největší u oblasti zajmu, což je akciový trh, textových dat a sentimentem investora. Vztah mezi bublinami výrazů nejsou příliš vzdálené. Zejména klastr 2 vykazuje nejkratší odkazy a má docela blízké vazby na shluk 1, ale také odkazy na ostatní shluky. Naopak největší vzdálenost lze pozorovat mezi klastry 1 a 4. Klastr 1 odkazuje na sociální média jako je Twitter, kdežto klastr 4 odkazuje spíše na tradiční média, jako jsou články a studie.

2.1.2 Analýza kocitací časopisů

Tato část analyzuje kocitační síť časopisů o sentimentální analýze na akciových trzích. McCain (1990) navrhl kocitační analýzu časopisu, která má přispět ke studiu tematiky organizace vědeckých oborů. Četnost citování dvou zdrojů označuje podobnosti mezi rozsahem časopisu a jeho výzkumnými tématy. Obrázek 2.4 ukazuje zdroje, které byly citovány nejméně dvakrát. Každý uzel představoval zdroj a velikost každého uzlu představoval počet přijatých citací. Spojení mezi dvěma položkami naznačovalo kocitační vztah. Uzly byly seskupeny podle podobnosti; to znamená, že zdroje ve stejném klastru (barvě) a ty, které jsou si bližší, byly si navzájem více podobné.

Journal of Finance byl citován celkem 90 s největší celkovou sílou odkazu 2002. Je třeba poznamenat, že rozsah tohoto časopisu zahrnuje studie z oblasti financí související s akciovým trhem, takže tento výsledek lze očekávat.*Expert system with Application* měl největší celkovou sílu odkazu 1503, tento časopis publikuje vysoce kvalitní články z oblasti expertních systémů, které jsou nezbytnou součástí při dolování textu a sentimentální analýzu. Další časopis s vysokou sílou dosahu (1269) je *Decision Support System*. Tento časopis zaměřený na systémy slouží jako podpora pro rozhodování. Jejich citace je prakticky povinná, pokud chtějí vědci zahrnout do své analýzy nejnovější poznatky z akciových trhů a algoritmů expertních systémů. Následující nejcitovanější časopisy byly *Journal of Financial Economics* (24 citací

a celková síla odkazu 627), *Review of Financial Studies* (16 citací a celková síla odkazu 452), *Neurocomputing* (13 citací a celková síla odkazu 432), *International Journal of Forecasting* (13 citací a celková síla odkazu 412), *Journal of Banking and Finance* (12 citací a celková síla odkazu 368).

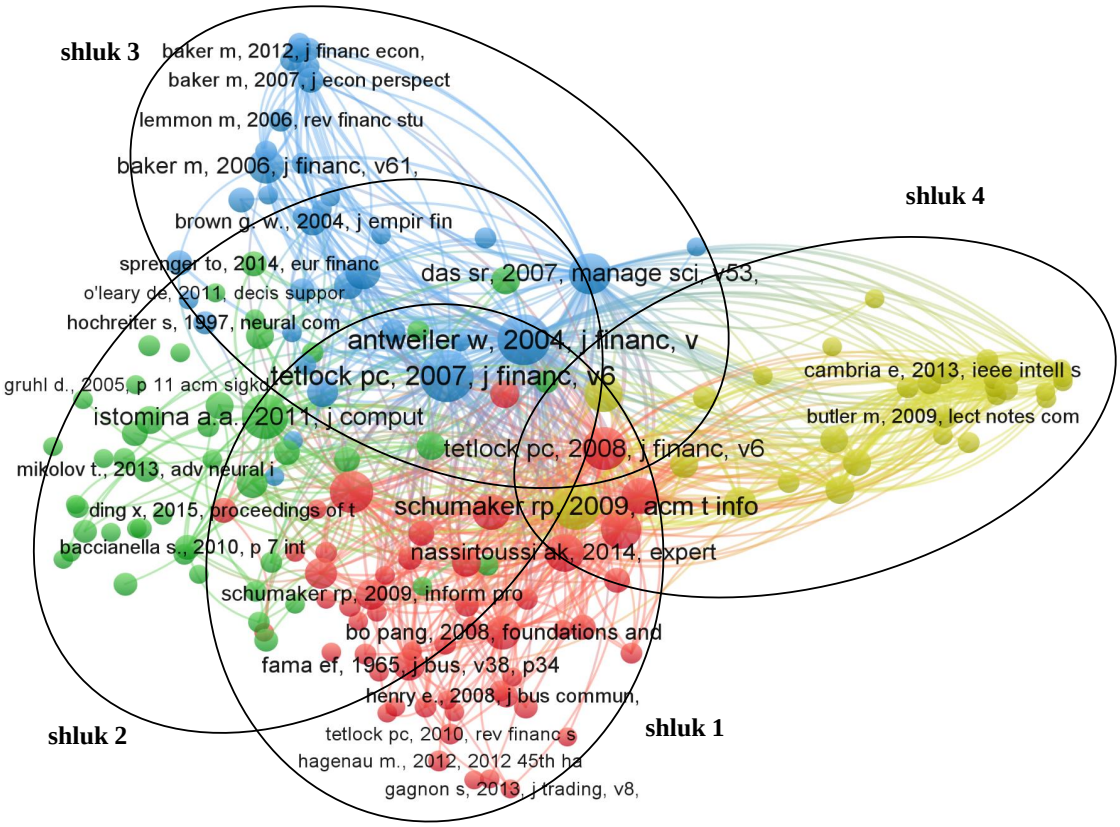


Obr. 2.4: Bibliometrická analýza kocitací časopisů
Zdroj: VOSviewer (2021)

Software VOS Viewer klasifikoval deníky do čtyř klastrů. Červený klastr obsahoval 42 zdrojů a zahrnoval časopisy, které se týkají zejména financí, finančním managementem a investováním. Modrá skupina se v zásadě zaměřovala na relevantní výzkum v oblasti expertních systémů a algoritmů učení a zahrnovala 28 zdrojů. Žlutý klastr (16) byl nejmenší a obsahoval zejména časopisy zabývající se rozhodování, predikcí a operačním výzkumem. A konečně zelený klastr (26) se tematicky nejvíce lišil od ostatních, protože se skládal hlavně z časopisů, které se zaměřují na neurovědu, což je přední expertní systém využívá při textové a sentimentální analýze na akciovém trhu.

2.1.3 Analýza kocitací dokumentů

Další kocitační studie analyzovala kocitační síť dokumentů o zkoumaném problému. Small (1973) předpokládá, že často citované práce představují klíčové pojmy, metody nebo experimenty v oboru. Tato analýza tedy umožní zjistit, které dokumenty definují intelektuální strukturu v otázkách těžby textu a analýze sentimentu. Z analyzovaných článků splňovalo 167 prahovou hodnotu minimálně 3 citací na článek (obrázek 2.5). Každý uzel představoval jeden odkaz a jeho velikost udávala počet citací na dokument. Odkaz představoval spolucitaci. Počet identických citujících položek definoval sílu kocitací mezi dvěma citovanými články. Kocitace byla tedy četnost, s jakou byly citovány dva články z dřívější literatury, a to spolu s pozdější literaturou.



Obr. 2.5: Bibliometrická analýza kocitací dokumentů
Zdroj: VOSviewer (2021)

Uzly se stejnou barvou patřily do stejného klastru. Metoda normalizace síly asociace používaná softwarem VOS Viewer identifikovala čtyři skupiny. Červený (51) a modrý (36) klastr obsahují především vlivné články o finanční teorii od autorů Markowitze či Famy, jejichž východiska jsou nezbytné pro jakoukoli analýzu na finančních, potažmo akciových trzích. Žlutý klastr obsahuje 30 dokumentů. Zelený

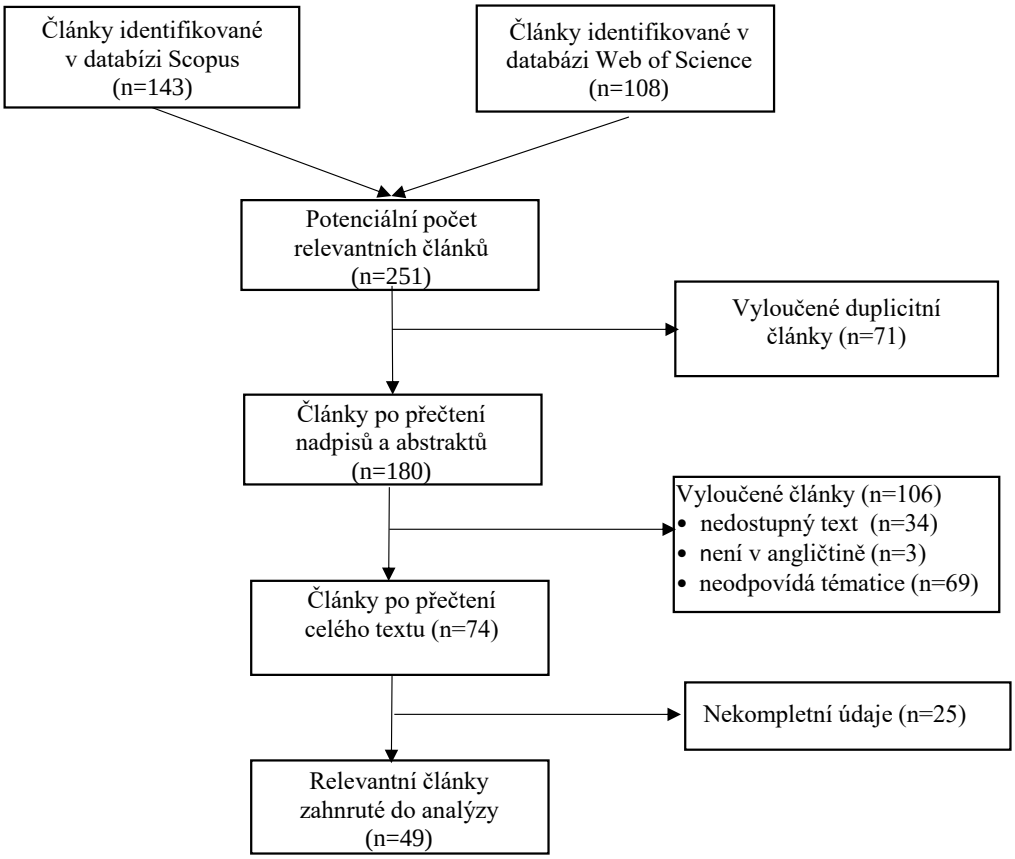
klastr obsahuje 46 citovaných dokumentů.

Mnohé z těchto citovaných studií zkoumalo prediktivní sílu a účinky sentimentu investora na výnosy akcií s nekonzistentními výsledky. Například Antweiler a Frank (2004) zjišťují, že online zprávy pomáhají předpovídat volatilitu trhu, zatímco jejich účinek na výnosy akcií je statisticky významný, ale ekonomicky malý na základě údajů o vysoké frekvenci za rok 2000. Rovněž naznačují, že neshody mezi zveřejněnými zprávami jsou spojené se zvýšeným objemem obchodování. Tetlock (2007) zjišťuje, že vysoký mediální pesimismus předpovídá tlak na snižování tržních cen s následným návratem k základům a že neobvykle vysoký nebo nízký pesimismus předpovídá vysoký objem obchodování na trhu. Kim a Kim (2014) však nezískali žádné důkazy o tom, že sentiment investorů předpovídá budoucí návratnost akcií buď na souhrnné úrovni, nebo na úrovni jednotlivých firem, nebo že sentiment investorů ze zveřejňování zpráv na internetu má sílu předpovědět volatilitu nebo objem obchodování. Bu a kol. (2018) zjistili, že sentiment investorů získaný z internetových diskusních fór na burze nemá sílu předpovídat denní výnosy, volatilitu nebo objem obchodování na čínském akciovém trhu. Sentiment investorů má však významný pozitivní vliv na současnou cenu akcií. Kromě toho Bu a kol. (2018) odhalili, že sentiment investorů získaný z textu vytvořený pomocí zpráv před otevřením trhu by mohl předpovědět otevřenou cenu akciového trhu. Baker a Wurgler (2006) naznačují, že sentiment je vírou o budoucích peněžních tocích a investičními riziky a není odůvodněn dostupnými fakty.

2.2 Obsahová analýza

Následuje obsahová analýzy vědeckých článků a příspěvků. K tomu je nezbytné vyselektovat pouze relevantní studie úzce souvisejících s předmětnou problematikou řešenou v disertační práci. Proces výběru je uveden na obrázku 2.6. Po spuštění vyhledávání v databázích se vyhledávací rovnici shodovalo 251 článků, přičemž z databáze Scopus bylo identifikováno 143 článků a z databáze Web of Science 108 článků. Z tohoto vyhledávání bylo vyloučeno 71 duplicitních článků a příspěvků. Po přečtení titulů a jejich abstraktů bylo z této recenze vyloučeno dalších 106 článků, protože jejich primárním účelem není předpovídání směru akciového trhu (69), nejsou zveřejněny v anglickém jazyce (3) a některé články byly nedostupné (34). Po úplném přečtení těchto článků bylo vyřazeno 25 prací, protože neuváděli nezbytné údaje. Kritický přezkum je proveden u 49 vědeckých článků a příspěvků, které byly identifikovány jako relevantní pro výzkum disertační práce. Hlavním účelem je pomoci vyvinout dobré porozumění a vhled do relevantního předchozího výzkumu a trendů, které se objevily. Současně posoudit silné a slabé stránky předchozího zkoumaného

výzkumu a identifikovat ty klíčové oblasti, které jsou nedosostatečně rozpracovány a kterým je potřeba věnovat pozornost.



Obr. 2.6: Diagram procesů selekce relevantních článků
Zdroj: vlastní zpracování

2.2.1 Zdroje textových a numerických dat

Práce provedené na predikci akciového trhu lze klasifikovat v závislosti na typu vstupů, které používají. Velká část recenzovaných článků využívá vstupy strukturovaného typu, pro které již existují techniky zpracování, a jejich význam byl důkladně prostudován. Nejnovější umožňují použití nestrukturovaných informací, které se obtížněji zpracovávají a získávají užitečné informace.

Textová datová základna

Použití nestrukturovaných informací je další výzvou při předpovídání akciového trhu. Tyto informace musí být předzpracovány a převedeny na kategorické nebo číselné informace, aby mohly být použity jako vstup do modelu. Textové nestrukturované vstupy vyžadují techniky těžby textu, které extrahují segmenty zpráv nebo názory

ze sociálních sítí a mohou generovat číselné reprezentace, které mohou předpovídat ceny akcií, jak popisuje Bustos a Pomares-Quimbaya (2020). Vzhledem k tomu, že velké množství testových dokumentů souvisejících s financemi, které zveřejňují profesionální i amatérští investoři, nejen na online sociálních sítích, které by mohly mít dopad na skutečné finanční trhy, je zásadním úkolem analyzovat finanční texty zveřejněné různými uživateli.

Textový vstup může mít několik zdrojů a typů obsahu, jak je uvedeno v tabulce 2.1. Část použitých studií (25%) využívá za textová data hlavní finanční weby jako The Wall Street Journal (Hwang a Kim, 2019, Khedr a kol., 2017), Reuters (Hwang a Kim, 2019, Khedr a kol., 2017), Dow Jones (Siering, 2012) stejně jako Yahoo! Finance (Derakhshan a Beigy, 2019, Chen a Chen, 2019, Al-Ramahi a kol., 2015, Nguyen a kol., 2015), Google Finance (Jammalamadaka a kol., 2019, Chen a Chen, 2019), NYSE (Hao a kol., 2021, Jammalamadaka a kol., 2019) a NASDAQ (Khedr a kol., 2017). Drtivá většina těchto studií používá finanční zprávy, protože se má za to, že mají menší šum ve srovnání s obecnými zprávami. Zde se extrahuje text zprávy nebo nadpis zprávy. Titulky zpráv se občas používají a tvrdí se, že jsou přímější, a tudíž méně hlučné kvůli podrobnému textu (Huang a kol., 2010). Kromě předních finančních webů, se u analyzovaných zdrojů objevily jako vstupní textované údaje také obecné finanční zprávy (Kim a kol., 2018, Hajek, 2018, Kraus a Feuerriegel, 2017) nebo finanční zprávy konkrétních zkoumaných společností jako například v práci Feuerriegel a Gordon (2018). Další část vědců zkoumalo méně formální zdroje textových informací, např. Li a kol. (2020), Ren a kol. (2019), Long a kol. (2018), se ve své práci podívali na text, který je zveřejňován a diskutován na online diskusních fórech Eastmoney, což je největší a nejreprezentativnější burza zpráv v Číně. A jedná se o důležitý a slavný finanční a ekonomický web, který má nejvíce měsíčních jedinečných návštěvníků ze všech webových stránek s finančními informacemi v Číně.

Nicméně v poslední době zkoumali textový obsah ze sociálních médií. Tento textový zdroj dat tvoří 39% z analyzovaných zdrojů. Jedna skupina vědců se zaměřuje pouze na Twitter a využívají jej pro predikci trhu a analýzu veřejné nálady efektivněji (Sakhare a kol., 2020). Gross-Klussmann a kol. (2019) uvádějí, že směrová sentimentální opatření na Twitteru významně odrážejí současné směry návratnosti indexu akciového trhu. Zejména expertní uživatelé jsou hlavní hnací silou vzájemné závislosti sentimentu Twitteru a finančních trhů. Ve srovnání se sadou doplňkových uživatelů vykazují kandidáti odborní uživatelé vyšší počty sledovatelů, což naznačuje větší vliv na síť a vydává zprávy blíže finančním a ekonomickým tématům. Obdobně Eliacik a Erdogan (2018) inspirovaní výzkumem analýzy sociálních sítí a analýzami sentimentu zjistili, že důvěra lidí v komunitu má důležité místo při určování polarity sentimentu komunity v daném tématu. Při zkoumání studií v literatuře je vidět, že

důvěryhodní uživatelé v komunitě jsou ve skutečnosti vlivní uživatelé.

Jiní autoři se zaměřili na platformu StockTwits (Owen a Oktariani, 2020, Batra a Daudpota, 2018, Al Nasseri a kol., 2015, Al Nasseri a kol., 2014), což je největší sociální síť pro investory a obchodníky s více než pěti miliony členů komunity a miliony měsíčních návštěvníků. Jako rozhodující hlas „sociálního financování“ je Stocktwits nejlepším způsobem, jak zjistit, co se právě děje na trzích a akcích, na kterých vám záleží, a přetváří finanční média pro další generaci investorů. Al Nasseri a kol. (2015) potvrzují, že příspěvky StockTwits obsahují cenné informace a předcházejí obchodním aktivitám na kapitálových trzích. Změny průměrných výskytů různých sémantických výrazů v příspěvcích StockTwits informovaly rozhodnutí o tom, zda koupit nebo prodat akcie. Zjištění tohoto výzkumného článku mohou přinést slibný pohled na potenciální poskytnutí mechanismu podpory investic pro analytiky, investory a jejich kolegy.

Jiní výzkumníci a řešitelé se zaměřili na diskuzní a akciová fóra, mikroblogy, online recenze aj. Tato skupina tvoří 14% analyzovaných textových dat z relevantních studií. Jedná se o minoritní skupinu, které zejména v posledních letech ztrácí na atraktivnosti a posunuje se do pozadí současných výzkumů. Například Sun a kol. (2020) a Zhao a kol. (2016), kteří považují za vhodné zdroje finančních zpráv Sina Weibo. Vzhledem k tomu, že v roce 2009 byl v Číně zablokován Twitter, Sina Weibo umožňuje uživatelům zveřejňovat krátké tweety nebo zprávy. Sina Weibo poskytuje platformu pro zkoumání vztahu mezi skupinami uživatelů s různými identitami na základě pozornosti uživatelů, úrovní nálad a sociální interakce z pohledu zúčastněných stran. Navíc díky pokročilé vyhledávací schopnosti Sina Weibo můžeme přizpůsobit naše data vyhledáním konkrétnějšího výrazu v kombinaci s klíčovými slovy tématu, datem, typy dokumentu a zdrojem dokumentu (Sun a kol., 2020). Na tento zdroj textového korpusu se zaměřují výhradně výzkumníci zkoumající čínský akciový trh.

Celkem 22% autorů využilo kombinaci finančních zpráv a sociálních médií, popřípadě diskuzních fór. Nti a kol. (2020) zjistili přesnost (49,4–52,95%) na základě Google trends, (55,5–60,05%) na Twitteru, (41,52–41,77%) na základě příspěvku na fóru, (50,43–55,81%) na základě webových zpráv a (70,66 –77,12%) na základě kombinované datové sady. Lze si povšimnout, že různou kombinací textových dat lze dosáhnout vyšší přesnost. Dále navrhovaný model v Khedr a kol. (2017) zlepšil přesnost predikce budoucího trendu na akciovém trhu tím, že autoři zohlednili různé typy denních zpráv s různými hodnotami číselných atributů během dne. Uvažovány byly tři kategorie údajů o novinkách: zprávy týkající se trhu, zprávy o společnostech a finanční zprávy, které finanční experti zveřejnili o akciích.

V tabulce 2.1 je kromě typu a zdroje textového korpusu uveden také počet položek zkoumaného textového zdroje. Lze pozorovat, že tento počet se pohybuje od

Autor	Typ korpusu	Zdroj korpusu	Počet položek
Hao a kol. (2021)	Finanční zprávy	NYSE	2,474
Owen a Oktariani (2020)	Sociální sítě	StockTwits	1,454
Sakhare a kol. (2020)	Sociální sítě	Twitter	n/a
Li a kol. (2020)	Akciové fórům	Eastmoney	18 milionů
Sun a kol. (2020)	Mikroblog	Sina Weibo	22,504
Nti a kol. (2020)	Sociální sítě	Twitter, Google trends, fóra	4028
Bouktif a kol. (2020)	Sociální sítě	Twitter	170 000 pro akcii
Bouktif a kol. (2019)	Sociální sítě	Twitter	n/a
Birbeck a Cliff (2019)	Sociální sítě	Twitter	1,474,747
Moro a kol. (2019)	Sociální sítě	Twitter	10 milionů
Jammalamadaka a kol. (2019)	Finanční zprávy	Yahoo Finance, Google Finance a Twitter	n/a
Gross-Klussmann a kol. (2019)	Sociální sítě	Twitter	102 milionů
Derakhshan a Beigy (2019)	Finanční zprávy	Yahoo Finance, Sahamyab	787,547
Chen a Chen (2019)	Finanční zprávy	ChinaTimes, cnYES, Yahoo, Google	n/a
Ren a kol. (2019)	Akciové fórům	Sina, Eastmoney	1,930,592
Hwang a Kim (2019)	Finanční zprávy	CNBC, Reuers, Wall Street Journal aj.	n/a
Chen a Shih (2019)	Finanční zprávy	Wantgoo	1,312
Batra a Daudpota (2018)	Sociální sítě	StockTwits	300,000
Eliacik a Erdogan (2018)	Sociální sítě	Twitter	1,241,234
Kim a kol. (2018)	Finanční zprávy	8-K	n/a
Feuerriegel a Gordon (2018)	Finanční zprávy	DGAP, EQS Group	80,813
Long a kol. (2018)	Akciové fórům	Eastmoney	5,990
Hajek (2018)	Finanční zprávy	10-K	n/a
Shi a kol. (2018)	Akciové fórům	Eastmoney	5,163,210
Pagolu a kol. (2017)	Sociální sítě	Twitter	250,000
Ab. Rahman a kol. (2017)	Finanční zprávy	Edge Markets	14,992
Khedr a kol. (2017)	Finanční zprávy	NASDAQ, Reuters, Wall Street Journal aj.	n/a
Domeniconi a kol. (2017)	Sociální sítě	Twitter	10 milionů
Kraus a Feuerriegel (2017)	Finanční zprávy	8-K	34,782
Xie a Jiang (2017)	Finanční zprávy	CNBC, Reuers, Wall Street Journal aj.	2,302,692
Oliveira a kol. (2017)	Sociální sítě	Twitter	31 milionů
Urolagin (2017)	Sociální sítě	Twitter	n/a
Simoes a kol. (2017)	Sociální sítě	Twitter	n/a
Alostad a Davulcu (2017)	Finanční zprávy	NASDAQ, Twitter	53 641+780 139
Zhao a kol. (2016)	Mikroblog	Sina Weibo	6.1 milionů
Das a Das (2016)	Recenze	TripAdvisor	15,763
Eliacik a Erdogan (2016)	Sociální sítě	Twitter	1,148,181
Hajek a Bohacova (2016)	Reporty	Zprávy bank	n/a
Al-Ramahi a kol. (2015)	Finanční zprávy	Yahoo Finance	200
Nguyen a kol. (2015)	Finanční zprávy	Yahoo Finance	n/a
Al Nasseri a kol. (2015)	Sociální sítě	StockTwits	n/a
Smailović a kol. (2014)	Sociální sítě	Twitter	152,570
Al Nasseri a kol. (2014)	Sociální sítě	StockTwits	2,892
Meesad a Li (2014)	Sociální sítě	Twitter	4,622
Tirea a Negru (2013)	Finanční zprávy	M.bursa, M.antena3.	6217
Oliveira a kol. (2013)	Sociální sítě	Twitter	n/a
Nann a kol. (2013)	Sociální sítě	Twitter, Yahoo Finance	2,971,381
Siering (2012)	Finanční zprávy	Dow Jones News	11,518
O'Hare a kol. (2009)	Finanční zprávy	n/a	232

Tab. 2.1: Vstupní údaje o textových datech
Zdroj:vlastní zpracování

stovek (Al-Ramahi a kol., 2015, O'Hare a kol., 2009) až po několik milionů (Li a kol., 2020, Moro a kol., 2019, Gross-Klussmann a kol., 2019, Domeniconi a kol., 2017, Oliveira a kol., 2017, Zhao a kol., 2016) položek textového zdroje. Nicméně některé zdroje neuvádějí velikost korpusu, který ve svých pracích zkoumali. Alostad a Davulcu (2017) uvádějí, že vyšší objem Tweetů vede ke statisticky významnému zvýšení přesnosti. Nicméně Kraus a Feuerriegel (2017) upozorňují na běžné překážky vyplývající ze skutečnosti, že velké předem sestavené datové sady často nejsou snadno dostupné. Začlenění těchto velkých korpusů je však zásadní pro vytváření výkonných modelů. Vzhledem k datovým omezením ve většině studií zůstává debata o dlouhodobých výhodách signálů sociálních médií pro finanční rozhodování doposud neprůkazná.

Numerická datová základna

Dalším zdrojem vstupních dat pro systémy pochází z číselných hodnot na finančních trzích ve formě akcií nebo akciových indexů nebo fondů. Tato data se používají většinou za účelem trénování expertních systémů a pro účely predikce. V tabulce 2.2 jsou uvedeny zásadní podrobnosti o těchto tržních datech. Minulý výzkum byl většinou zaměřen na predikci akciového trhu, a to buď ve formě indexu akciového trhu (39%), samostatných akcií (8%), skupiny indexů či akcií (29%). Zarážející je, že celkem v 24% relevantních prací nebyl konkrétně uveden název akcie či indexu, který byl zkoumán.

Z hlediska akciových indexů je často zkoumán americký Dow Jones Industrial Average (Domeniconi a kol., 2017, Alostad a Davulcu, 2017, Al Nasseri a kol., 2015 a 2014) a S&P 500 (Nann a kol., 2013, O'Hare a kol., 2009). Kromě amerických akciových indexů byly hojně zkoumány také čínské indexy například CSI 300 (Li a kol., 2020, Moro a kol., 2019, Jammalamadaka a kol., 2019, Long a kol., 2018, Shi a kol., 2018), jako proxy pro celý čínský akciový trh, protože je to nejvíce důležitý index akciového trhu v Číně, který popisuje pohyby akcií na Šanghajské burze cenných papírů a Shenzhenské burze cenných papírů. Kromě tohoto akciového indexu je zkoumán na území Číny také SSE 50 Index (Ren a kol., 2019) a Shanghai Composite Index (Zhao a kol., 2016). V minoritním zastoupení lze nalézt také studie zaměřené na tchajwanský index TAIEX (Chen a Chen, 2019), dále turecký index BIST100 (Eliacik a Erdogan, 2018, Eliacik a Erdogan, 2016), popřípadě i německý index CDAX (Kraus a Feuerriegel, 2017, Siering, 2012). Kromě zkoumání samostatných akciových indexů je také velké zastoupení výzkumníků zabývajících se skupinou akciových indexů (Gross-Klussmann a kol., 2019, Feuerriegel a Gordon, 2018, Oliveira a kol., 2017, Simoes a kol., 2017). Dále jsou zkoumány samostatné akcie jako Boeing (Owen a Oktariani, 2020), Apple (Batra a Daudpota, 2018), Microsoft (Pa-

golu a kol., 2017), TripAdvisor (Das a Das, 2016) nebo skupina společností (Hao a kol., 2021, Bouktif a kol., 2020, Birbeck a Cliff, 2019, Kim a kol., 2018, Ab. Raman a kol.(2017), Smailović a kol., 2014, Oliveira a kol., 2013). Většina výzkumníků v této oblasti použila pouze jednu akcii nebo v mnoha z nich byl počet testovacích vzorků nedostatečný, což se zdá být pro dosažení závěru o vlivu sentimentální analýzy na akciových trzích nedostatečné. Prozatím neexistuje dlouhodobý výzkum, který by ukazoval pozoruhodné výsledky u souboru několika akcií.

Většina studií se soustředila na americký akciový trh. Ten byl analyzován celkem v 55% případů a stal se tak jasně dominujícím trhem, na kterém byl zdokumentován vliv analýzy sentimentu extrahovaný z textových dat. V 15% pracích se pozornost zaměřila na čínský akciový trh a v 7% na taiwanský trh. Z hlediska evropského trhu se ve výstupech objevuje akciový trh Německa (7%). Kombinací více akciových trhů využili Gross-Klussmann a kol. (2019) a Derakhshan a Beigy (2019). Zatímco důkazy o ostatních trzích jsou značně omezené. Na základě předešlého argumentu bylo zjištěno, že minimální počet studií se soustředil na zkoumání dopadů sentimentu investorů na akciové trhy v rozvojových ekonomikách. Zde se nabízí prostor pro budoucí výzkum a rozšíření stávajících poznatků na ostatní, doposud neprozkoumané trh.

Dále je porovnáván prediktivní časový rámec pro každou práci. Časový rámec od vydání zprávy po sledování dopadu na trh se může lišit od sekund do dnů, týdnů nebo měsíců. Li a kol. (2020) dále dodávají, že denní sentiment investorů může adekvátně předpovědět otevřené tržní ceny následného obchodního dne, zatímco prediktivní informace pro denní uzavírací cenu jsou slabé. Navíc zjistili, že začlenění hodinového sentimentu přináší minimální zlepšení pro předpovídání otevřených cen, protože denní vstupní proměnné mají dostatek informací. Většina zkoumaných studií se zabývá denní údaji o akciových titulech či indexech. Zkoumaná časová perioda se pohybuje od několika měsíců (Owen a Oktariani, 2020, Hwang a Kim, 2019, Chen a Shih, 2019 a další) až po desítky let (Bouktif a kol., 2020, Feuerriegel a Gordon, 2018, Kim a kol., 2018).

Pozornost byla zaměřena také na to, zda analyzované studie berou jako vstupní údaje pouze cen akciových titulů či akciových indexů nebo i jiné ukazatelé vycházející z technické či fundamentální analýzy při zohlednění sentimentu pocházejícího z textových dat více viz Janková (2020). Jak lze pozorovat z tabulky 2.2, některé studie do modelu začleňují také ukazatelé pocházející z těchto dvou analýz. Konkrétní ukazatele, které nejčastěji vstupují do modelů pro predikci, je uvedeno v příspěvku Jankové (2020), kde je tato problematika detailně rozebírána. Nicméně Jammalamadaka a kol. (2019) uvádějí, že přítomnosti moderních i klasických prediktorů ve fundamentálním i technickém smyslu má polarita zpráv stále komplementární účinek.

Autor	Zdroj	Název akcie/indexu	Země	Časový rámec		FA	TA
Hao a kol. (2021)	Akcie	30 akcií	Tchaj-wan	2017	2018		
Owen a Oktariani (2020)	Akcie	Boeing	USA	-	2019		x
Sakhare a kol. (2020)	Akcie	n/a	n/a	n/a	n/a		
Li a kol. (2020)	Index	CSI 300 index	Čína	2009	2014		
Sun a kol. (2020)	Akcie	n/a	Čína	-	2015	x	
Nti a kol. (2020)	Akcie	GCB, MTNGH, TOTAL	Ghana	2010	2019		
Bouktif a kol. (2020)	Akcie	Amazon, Apple, Microsoft aj.	USA	2008	2018		
Bouktif a kol. (2019)	Index	CSI 300 index	USA	2007	2017		x
Birbeck a Cliff (2019)	Akcie	Apple, Tesla, Twitter aj.	USA	2015	2016		x
Moro a kol. (2019)	Index	CSI 300 index	USA	2008	2018		
Jammalamadaka a kol. (2019)	Index	CSI 300 index	USA	2016	2018	x	x
Gross-Klussmann a kol. (2019)	Index	ASX200, HangSeng, EURO STOXX 50, Nikkei225, SP 100	USA, EU	2010	2018		
Derakhshan a Beigy (2019)	Akcie	n/a	Irán, USA	2012	2013		
Chen a Chen (2019)	Index	TAIEX	Tchaj-wan	2016	2017		
Ren a kol. (2019)	Index	SSE 50 Index	Čína	2014	2016		
Hwang a Kim (2019)	Akcie	n/a	S. Korea	-	2019		
Chen a Shih (2019)	Akcie	n/a	Tchaj-wan	-	2018		x
Batra a Daudpota (2018)	Apple	Apple	USA	2010	2017		
Eliacik a Erdogan (2018)	Index	BIST100	Turecko	2015	2016		
Kim a kol. (2018)	Akcie	Citigroup, Wells fargo and Co, Goldman sachs, JP Morgan	USA	2002	2012		
Feuerriegel a Gordon (2018)	Index	DAX, CDAX, STOXX Europe 600	Německo	1996	2016		
Long a kol. (2018)	Index	CSI 300	Čína	n/a	n/a		
Hajek (2018)	Akcie	n/a	USA	n/a	n/a	x	
Shi a kol. (2018)	Index	CSI 300	Čína	2011	2015		
Pagolu a kol. (2017)	Akcie	Microsoft	USA	2015	2016		
Ab. Rahman a kol. (2017)	Akcie	Axiata Group, CIMB Group Holdings aj.	Malajsie	2014	2017		
Khedr a kol. (2017)	Akcie	Yahoo, Microsoft, Facebook	USA	n/a	n/a		
Domeniconi a kol. (2017)	Index	DJIA	USA	n/a	n/a		
Kraus a Feuerriegel (2017)	Index	CDAX	Německo	2010	2013		
Xie a Jiang (2017)	Akcie	n/a	Čína	2008	2015		
Oliveira a kol. (2017)	Indexy	SP 500, Nasdaq 100, Russell 2000, DJIA	USA	2012	2015		x
Urolagin (2017)	Akcie	n/a	n/a	-	2017		
Simoes a kol. (2017)	Indexy	SP 500, NASDAQ 100, DJIA	USA	n/a	n/a		
Alostad a Davulcu (2017)	Index	DJIA	USA	2010	2014		
Zhao a kol. (2016)	Index	Shanghai Composite Index	Čína	-	2015		
Das a Das (2016)	Akcie	TripAdvisor	USA	-	2015		
Eliacik a Erdogan (2016)	Index	BIST 100	Turecko	2014	2015		
Hajek a Bohacova (2016)	Akcie	Akcie bank	USA	n/a	n/a	x	
Al-Ramahi a kol. (2015)	Akcie	n/a	USA	2012	2014		
Nguyen a kol. (2015)	Akcie	n/a	USA	2012	2013		
Al Nasseri a kol. (2015)	Index	DJIA	USA	2012	2013		
Smailović a kol. (2014)	Akcie	Apple, Amazon, Baidu aj.	USA	-	2011		
Al Nasseri a kol. (2014)	Index	DJIA	USA	2012	2013		
Meesad a Li (2014)	Akcie	n/a	n/a	2013	2014		
Tirea a Negru (2013)	Akcie	n/a	Rumunsko	2012	2013	x	x
Oliveira a kol. (2013)	Akcie	AMD, Amazon, Dell aj.	USA	2012	2013		
Nann a kol. (2013)	Index	SP 500	USA	-	2011		x
Siering (2012)	Index	DAX	Německo	2010	2011		
O'Hare a kol. (2009)	Index	SP500	USA	-	2009		

Tab. 2.2: Vstupní údaje o trhu, časový rámec
Zdroj:vlastní zpracování

2.2.2 Způsob předzpracování datových zdrojů

Jakmile jsou vstupní data k dispozici, musí být připravena, aby mohla být vložena do algoritmu pro další vyhodnocování. To pro textová data znamená transformovat nestrukturovaný text do reprezentativního formátu, který je strukturovaný a může být zpracován strojem. Při těžbě dat obecně a konkrétně při těžbě textu má fáze předběžného zpracování významný dopad na celkové výsledky (Uysal a Gunal, 2014).

Rozhodnutí o prvcích, kterými má být text reprezentován, je zásadní, protože od nesprávného reprezentačního vstupu nelze očekávat nic jiného než nesmyslný výstup. V tabulce 2.3 je uveden typ reprezentace textové struktury na čísla. Ve většině literatury byly nejzákladnější techniky použity při řešení problémů s predikcí trhu založených na dolování textu a rovněž v souladu s našimi poznatky. Bag of Words (BOW) je populární přístup k reprezentaci textových dat a byl použit v různých souvisejících studiích. Tato technika je detailně vysvětlena v sekci 1.2.4. Jak je uvedeno v tabulce **??**, přibližně 58% prací ji aplikuje ve svém výzkumu (Owen a Oktariani, 2020, Sakhare a kol., 2020, Li a kol., 2020, Sun a kol., 2020, Birbeck a Cliff, 2019, Moro a kol., 2019 a mnoho dalších) se spoléhají na tuto základní techniku výběru prvků, ve které je pořadí a společný výskyt slov zcela ignorovány. Hájek (2018) ukazuje, že kvalita predikce se významně zvýšila při použití korelačního výběru funkcí BoW.

Druhý přístup je technika N-gramů, kterou lze nalézt samostatně u 19% autorů. N-gram je souvislá posloupnost N položek, kterými jsou obvykle slova z dané posloupnosti textu. Sekvence slov a syntaktické struktury však mohly být v zásadě pokročilejší. Příkladem takového nastavení může být použití syntaktických N-gramů (viz práce Bouktif a kol., 2019, Pagolu a kol., 2017, Ab. Rahman a kol., 2017, Khedr a kol., 2017, Urolagin, 2017, Eliacik a Erdogan, 2016). Jiné studie využívají (10%) latentní Dirichletovu alokaci (LDA), ta slouží k získání funkcí tématu za účelem objevování abstraktních témat v článcích. Důvodem je, že například novinové články ovlivňující cenu akcií mohou mít konkrétní abstraktní témata, která lze zachytit a použít ke klasifikaci trendů cen akcií (Hao a kol., 2021, Gross-Klussmann a kol., 2019, Derakhshan a Beigy, 2019, Das a Das, 2016 a Nguyen a kol., 2015). Některé studie zkoumali kombinace uvedených technik, těch je celkem 13%. Například lze nalézt kombinace bag of words a N-gram (Eliacik a Erdogan, 2018; Long a kol., 2018; Hajek a Bohacova, 2016) nebo kombinaci všech tří výše zmíněných (Bouktif a kol., 2020).

Omezený počet funkcí je nesmírně důležitý, protože zvýšení počtu funkcí, ke kterým může snadno dojít při výběru funkcí v textu, může způsobit, že problém s klasifikací nebo shlukováním bude extrémně těžké vyřešit snížením účinnosti většiny algoritmů učení, situace je obecně známá jako kletba dimenzionality. V tabulce

Autor	Převod	Způsob předzpracování	Váha
Hao a kol. (2021)	LDA	Stemming, odstranění stop slov	TF-IDF
Owen a Oktariani (2020)	BoW	Odstranění zkratek, převod na malá písmena	Binární
Sakhare a kol. (2020)	BoW	Odstranění HTML, hashtagů, stop slova, tokenizace	TF-IDF
Li a kol. (2020)	BoW	nespecifikováno	TF
Sun a kol. (2020)	BoW	Odstranění HTML, hashtagů, stemming, stop slova	TF-IDF
Nti a kol. (2020)	n/a	Stemming, odstranění hastagů, HTML	Binární
Bouktif a kol. (2020)	BoW, N-gram, LDA	Malá písmena, stemming a lemmatizace, odstranění URL, hashtagů, PCA	Binární
Bouktif a kol. (2019)	N-gram	Odfiltrování neanglických tweetů, malá písmena, stop slova	Binární
Birbeck a Cliff (2019)	BoW	Malá písmena, stemming, odstranění numerických tokenů, URL, stop slov	TF-IDF
Moro a kol. (2019)	BoW	Stemming, odstranění stop slov	TF-IDF
Jammalamadaka a kol. (2019)	BoW	nespecifikováno	Binární
Gross-Klussmann a kol. (2019)	LDA	Stemming, stop slov, hastagů, malá písmena	TF-IDF
Derakhshan a Beigy (2019)	LDA-POS	Odstranění stop slov, lemmatizace	Binární
Chen a Chen (2019)	N-gram	Odstranění duplicitních dat, chybějících hodnot, HTML, PCA	TF-IDF
Ren a kol. (2019)	BoW	Odstranění interpunkčních znamének	Binární
Hwang a Kim (2019)	BoW	Stemminh, převedením na malá písmena	TF-IDF
Chen a Shih (2019)	BoW	Stemming, převod na malá písmena, odstranění čísel, adres URL, stop slova	TF-IDF
Batra a Daudpota (2018)	BoW	Odstranění adres URL, tagy, symboly, stop slova	Binární
Eliacik a Erdogan (2018)	BoW, N-gram	Tokenizace, stemming	Binární
Kim a kol. (2018)	BoW	Stemming, malá písmena, odstranění čísel, URL, stop slov, PCA	Binární
Feuerriegel a Gordon (2018)	BoW	Odebrání čísel, interpunkčních znamének, PCA	TF-IDF
Long a kol. (2018)	N-gram, BoW	Odstranění interpunkce	TF-IDF
Hajek (2018)	BoW, N-gram	Stemming, malá písmena, odstranění URL, stop slov	TF-IDF
Shi a kol. (2018)	BoW	Odstranění interpunkce, statistika Chi2	Binární
Pagolu a kol. (2017)	N-gram	Stemming, převod na malá písmena	Binární
Ab. Rahman a kol. (2017)	N-gram	Chi2	TF-IDF
Khedr a kol. (2017)	N-gram	Tokenizace, odstranění stop slov, slučování synonym	TF-IDF
Domeniconi a kol. (2017)	BoW	Stemming, tokenizace, odstranění stop slov	TF-IDF
Kraus a Feuerriegel (2017)	BoW	Stemming, tokenizace, odstranění stop slov	TF-IDF
Xie a Jiang (2017)	BoW	Stemming, tokenizace, odstranění stop slov	Binární
Oliveira a kol. (2017)	N-gram	Stemming, tokenizace, odstranění stop slov	Binární
Urolagin (2017)	N-gram	Stemming, tokenizace, odstranění stop slov	Binární
Simoes a kol. (2017)	BoW	Tokenizace, stemming	TF-IDF
Alostad a Davulcu (2017)	N-gram	Stemming, tokenizace, odstranění stop slov, Chi2	Binární
Zhao a kol. (2016)	BoW, LDA	nespecifikováno	TF-IDF
Das a Das (2016)	LSA, LDA	Stemming	Binární
Eliacik a Erdogan (2016)	N-gram	Tokenizace, odstranění stop slov	Binární
Hajek a Bohacova (2016)	BoW, N-gram	Stemming, tokenizace, odstranění stop slov	TF-IDF
Al-Ramahi a kol. (2015)	BoW	Chi-square	Binární
Nguyen a kol. (2015)	LDA	Lematizace	TF-IDF
Al Nasseri a kol. (2015)	BoW	Stemming, převod na malá písmena, stop slova	IG
Smailović a kol. (2014)	BoW	Stemming, malá písmena, odstranění URL, stop slov	Binární
Al Nasseri a kol. (2014)	BoW	Odstranění mezer, čísel, stop slov	Binární
Meesad a Li (2014)	BoW	Tokenizace, stemming, stop slova	TF-IDF
Tirea a Negru (2013)	BoW	Tokenizace, stemming, stop slova	Binární
Oliveira a kol. (2013)	BoW	Tokenizace a odstranění mezer	Binární
Nann a kol. (2013)	BoW	Stemming, malá písmena, odstranění čísel, URL	Binární
Siering (2012)	BoW	Odtsranění stop slov	TF-IDF
O'Hare a kol. (2009)	BoW	Tokenizace, stop slova, slučování synonym a stemming	Binární

Tab. 2.3: Proces předzpracování textových dat
Zdroj:vlastní zpracování

2.3 pod redukcí rozměrů je zdůrazněn přístup zvolený u každé z prací. Souborem činností, které obvykle tvoří snížení dimenzí, jsou: stemming, tokenizace, převod na malá písmena, odstranění interpunkce a odstranění čísel, adresy webových stránek a stop slova, které byly detailně rozepsány v sekci 1.2.3. Tyto kroky jsou prováděny téměř vždy a v drtivé většině jsou jedinými kroky, které jsou prováděny ke snížení dimenzí analyzovaných zdrojů (Hao a kol., 2021, Owen a Oktariani, 2020, Sakhare a kol., 2020, Sun a kol., 2020, Nti a kol., 2020, Bouktif a kol., 2019 a další). Kromě běžných kroků ke snížení dimenzí prvku navíc autoři Bouktif a kol. (2020), Chen a Chen (2019), Kim a kol. (2018) a Feuerriegel a Gordon (2018) využili také metodu PCA. Další pomocnou metodou je statistika Chí-kvadrát. Tuto statistiku lze nalézt v práci Shi a kol. (2018), Ab. Rahman a kol. (2017) a Alostad a Davulcu (2017).

Poté, co je zredukován počet funkcí, musí být každá funkce představována číselnou hodnotou, aby mohla být zpracována algoritmy strojového učení. Proto se pro sloupec používá nadpis „reprezentace prvků“, přičemž se u všech recenzovaných prací porovnává typ číselné hodnoty, která je spojena s každým prvkem. Tato přiřazená číselná hodnota funguje jako skóre nebo váha. Existuje několik typů, které jsou velmi populární, a to: Information Gain (IG), Document Frequency (DF), Accuracy Balanced (Acc2) and Term Frequency-Inverse Document Frequency (TF – IDF).

Nejzákladnější váhovací technikou je logická nebo binární reprezentace, přičemž dvě hodnoty jako 0 a 1 představují nepřítomnost nebo přítomnost prvku. Binární reprezentace byla zvolena u poloviny analyzovaných studiích (Owen a Oktariani, 2020, Nti a kol., 2020, Bouktif a kol. 2020 a 2019, Jammalamadaka a kol., 2019, Derakhshan a Beigy, 2019, Ren a kol., 2019 a další). Druhá část autorů využívá další nejběžnější technikou je Term Frequency-Inverse Document Frequency nebo TF-IDF (Hao a kol., 2021, Sakhare a kol., 2020, Sun a kol., 2020, Birbeck a Cliff, 2019, Moro a kol., 2019 a další). Hodnota TF-IDF se zvyšuje úměrně s počtem výskytů slova v dokumentu, ale je vyvážena četností slova v korpusu, aby se vyvážila obecná popularita některých slov. Výjimečně lze v relevantních zdrojích nalézt techniku Information Gain (IG) u autorů Al Nasseri a kol. (2015).

2.2.3 Modely klasifikace a predikce

Po dokončení předběžného zpracování a převedení textu na řadu funkcí s numerickým vyjádřením lze zapojit algoritmy strojového učení. V následující části je uvedeno stručné shrnutí těchto algoritmů, které se v analyzovaných zdrojích využívají. Přehledně jsou využité algoritmy zobrazeny v tabulce 2.4. Kategorizace recenzovaných studií na základě použitého typu algoritmu je provedena do 8 tříd.

Metody strojového učení a umělé neuronové sítě jsou vhodné pro předpovídání vysoce volatilních finančních časových řad s nelinearitou, časovou korelací a silným

Autor	SVM	LSTM	ANN	Naivní Bayes	Rozh. stromy	Regr. analýza	Log. regrese	Ostatní
Hao a kol. (2021)	x							x
Owen a Oktariani (2020)		x	x					x
Sakhare a kol. (2020)		x			x	x		x
Li a kol. (2020)	x	x		x			x	
Sun a kol. (2020)	x			x				
Nti a kol. (2020)			x					
Bouktif a kol. (2020)	x			x				x
Bouktif a kol. (2019)	x		x				x	x
Birbeck a Cliff (2019)	x			x			x	x
Moro a kol. (2019)	x				x			
Jammalamadaka a kol. (2019)		x	x					x
Gross-Klussmann a kol. (2019)	x							x
Derakhshan a Beigy (2019)	x							
Chen a Chen (2019)								
Ren a kol. (2019)	x							
Hwang a Kim (2019)			x			x		x
Chen a Shih (2019)	x							
Batra a Daudpota (2018)	x							
Eliacik a Erdogan (2018)	x			x				
Kim a kol. (2018)	x			x			x	x
Feuerriegel a Gordon (2018)	x							
Long a kol. (2018)	x							
Hajek (2018)	x				x			x
Shi a kol. (2018)	x							
Pagolu a kol. (2017)							x	x
Ab. Rahman a kol. (2017)	x							
Khedr a kol. (2017)	x			x				x
Domeniconi a kol. (2017)	x				x			
Kraus a Feuerriegel (2017)	x	x	x		x	x		
Xie a Jiang (2017)	x							
Oliveira a kol. (2017)	x		x			x		x
Urolagin (2017)	x			x				
Simoes a kol. (2017)	x							x
Alostad a Davulcu (2017)	x							
Zhao a kol. (2016)	x					x		
Das a Das (2016)	x			x				x
Eliacik a Erdogan (2016)	x			x				
Hajek a Bohacova (2016)	x			x	x			x
Al-Ramahi a kol. (2015)						x		
Nguyen a kol. (2015)	x							
Al Nasseri a kol. (2015)	x							
Smailović a kol. (2014)	x							
Al Nasseri a kol. (2014)				x				x
Meesad a Li (2014)	x							
Tirea a Negru (2013)								x
Oliveira a kol. (2013)						x		
Nann a kol. (2013)				x				
Siering (2012)	x							
O'Hare a kol. (2009)	x			x				

Tab. 2.4: Typ použitého algoritmu
Zdroj:vlastní zpracování

šumem, jak uvádí Li a kol. (2020). Pokud je prediktivní síla investorského sentimentu silná, lze očekávat, že všechny modely mohou odhalit správné výsledky i přes své různé prediktivní schopnosti. Pokud je však prediktivní síla slabá, některé modely nemusí být schopny hlásit správné výsledky. Když nemohou fungovat žádné modely, sentiment investorů nemá žádnou prediktivní sílu nebo je prediktivní síla příliš slabá na to, aby se dala v praxi použít.

Z tabulky je patrné, že drtivá většina relevantních studií a výzkumu aplikuje pro klasifikaci sentimentu metodu SVM. Jedná se o více než 75%, což podtrhuje algoritmus učení s učitelem jako hlavní proud nástrojů klasifikace sentimentu. Xie a Jiang (2017) dokonce uvádí, že SVM vykazuje fantastickou míru přizpůsobení při předpovídání fluktuace akcií. SVM je v analyzovaných studiích aplikována buď samostatně (Hao a kol., 2021, Ren a kol., 2019, Chen a Shih, 2019 a další) nebo v kombinaci s jinými algoritmy (Li a kol., 2020, Sun a kol., 2020, Bouktif a kol., 2020 a 2019, Birbeck a Cliff, 2019, Moro a kol., 2019 a další). Například nejnovější studie Hao a kol. (2021) využívá kombinaci fuzzy logiky a SVM. Dle autorů je fuzzy SVM je robustnější ve srovnání s jinými metodami, když data obsahují některé odlehlé hodnoty. Kromě toho stupeň členství odhadovaný hranicí fuzzy rozhodnutí poskytuje lepší vhled do míry důvěry v předpovídané výstupy. Dle výše zmíněných prací se ukázalo se, že SVM překonává většinu tradičních učebních strojů v oblasti predikce akciových trhů.

Umělé neuronové sítě (NN) a rekurentní neuronové sítě (RNN) se obvykle používají pro předpovídání cen aktiv, zejména cen akcií. Ve srovnání s jinými typy neuronových sítí je nejdůležitější výhodou RNN jeho konkrétní síťová struktura, která uvažuje perzistenci informací. Využití RNN lze nalézt v práci Kraus a Feuerriegel (2017). RNN však nemůže dokonale řešit problém dlouhodobé závislosti. K řešení tohoto problému je navržen určitý typ RNN, konkrétně LSTM. V poslední době byly v diskusi o předvídatelnosti pozornosti investorů a sentimentu investora na akciovém trhu použity metody LSTM, zejména na základě opatření sentimentu investora nebo pozornosti investora konstruovaných technikami dolování textu. Výhody LSTM využívají výzkumy kupříkladu Owen a Oktariani (2020), Sakhare a kol. (2020), Kraus a Feuerriegel (2017). Navíc Li a kol. (2020) ve svém výzkumu odhalují nadřazenost modelu LSTM při detekci prediktivní síly sentimentu investora, což znamená, že metody hlubokého učení jsou užitečné pro zkoumání předvídatelnosti sentimentu investora. Autoři považují LSTM za velmi vhodný model k popisu charakteristik akciového trhu. Jejich výsledky odhalují rozdíly mezi modely a poskytují důkazy o nadřazenosti modelu LSTM nad SVM, Naïve Bayes a logistické regrese, zvláště když vstupy modelu obsahují prediktivní informace.

Jiní autoři dávají přednost klasifikačnímu algoritmu Naïve Bayes kvůli jeho nízké složitosti. Jammalamadaka a kol. (2019) začlenil online dolování textu do pokroči-

lého multivariačního modelu Bayesovské časové řady strojového učení, což otevírá dveře současného uplatnění jak textové těžby, tak strojového učení v moderním kvantitativním finančním výzkumu. Dle autorů MBSTS se sentimentálními prediktory překonává modely jako ARIMA, RNN či LSTM. Odlišuje se od přístupů jako k-Nearest Neighbors (k-NN), umělých neuronových sítí (ANN) nebo Support Vector Machine (SVM) v tom, že staví na pravděpodobnostech (prvku patřícího do určité kategorie), zatímco druhý zmíněné přístupy interpretují matici prvků dokumentu v prostoru.

V sekci ostatní jsou využity algoritmy jako náhodný les (Sakhare a kol., 2020, Bouktif a kol., 2019, Kim a kol.,2018, Pagolu a kol., 2017 a Oliveira a kol., 2017), K-means (Gross-Klussmann a kol., 2019, Hajek, 2018 a Khedr a kol., 2017) a další, které nebyly příliš zastoupeny nebo nejsou v této oblasti zájmu rozšířené.

Techniky navržené v literatuře poskytly přijatelné výsledky a zajímavé informace o sentimentální analýze a vztahu mezi akciovým trhem a sentimentální analýzou. Průzkum o těžbě textu pro predikci akciových trhů dospěl k závěru, že SVM a Naïve Bayes jsou výzkumníky silně upřednostňováni, zatímco NN jsou v této fázi v oblasti prediktivní těžby textů na akciových trzích výrazně nedostatečně prozkoumány, přestože NN ukázaly slibný potenciál pro textovou klasifikaci a analýzu sentimentu.

2.2.4 Analýza sentimentu

Dále jsou rozebírány konkrétní slovníky či lexikony, které výzkumníci využívají ke stanovení skóre sentimentu. Uvedený přehled relevantních studií je zaznamenán v tabulce 2.5. Bouktif a kol. (2020) využívají VADER (Valence Aware Dictionary and sEntiment Reasoner), což je API pythonové knihovny pro analýzu textu na sociálních médiích. Tento přístup bez učitele nepotřebuje žádná označená data, protože model je sestaven z generalizovatelného, valenčního zlatého standardního seznamu lexikálních funkcí spolu s přidruženými intenzitami sentimentu. Polarita sentimentu se získá použitím bez učitele.

Bouktif a kol. (2019) použili TextBlob pro analýzu sentimentu, která nabízí snadno použitelné API pro běžné úkoly NLP. Dává sentimentální polaritu a skóre subjektivity. Skóre polarity jsou v rozsahu -1,0 a 1,0 a skóre subjektivity v rozmezí 0,0 a 1,0, kde 0,0 je velmi objektivní a 1,0 je velmi subjektivní.

Jammalamadaka a kol. (2019) přijímají lexikonový přístup ke klasifikaci sentimentu z hlediska polarit (pozitivní, negativní, neutrální) zpravodajských článků a emoce (úzkost, klid, nechuť, strach, laskavost, láska, radost, smutek, neznámé) příspěvků na Twitteru. Přístupy založené na lexikonu pro klasifikaci sentimentu mají tu výhodu, že se vyhnou těžkopádnému kroku označování tréninkových dat, a jsou založeny na poznatku, že sentimentů vyjádřených částí textu lze dosáhnout na zá-

kladě polarit a emocí slov. Zdrojem lexikonu použitým v tomto článku je SenticNet, který poskytuje sadu sémantiky a polarit spojených se 100 000 koncepty přirozeného jazyka. Jako znalostní základna je SenticNet schopen identifikovat polaritu a afektivní informace, včetně získávání názorů na složité koncepty, jako je dosahování cílů, slavení zvláštních příležitostí, ztráta nálady atd. Skóre polarity v rozsahu -1 až 1 pro tyto běžné smyslové koncepty poskytuje tento lexikon. Obdobně Gross-Klussmann a kol. (2019) čerpá ze slovníkového přístupu k extrakci sentimentu, kde se polarita slova měří podle pozitivních a negativních slovních slovníků. Gross-Klussmann a kol. (2019) využívají na seznamech pozitivních a negativních slov předložených Hu a Liu (2004) a dále uvažují psychologické slovníky Harvard General Inquirer IV-4 (HGIV-4) založené na Stoneovi a kol. (1966). Slovník HGIV-4 pokrývá širokou škálu témat mimo seznamy pozitivních a negativních slov. V průběhu studie využívají seznamy slov z HGIV-4 na téma „ekonomika“, aby zjistili podobnost textových dat s kontextem.

Feuerriegel a Gordon (2018), Chen a Shih (2019) a další integrují finanční specifický slovník Loughran-McDonald, který se vyvinul jako kvazi-standard ve výzkumu souvisejícím s financemi. Tento slovník byl speciálně konstruován tak, aby mohl z finančních zpráv extrahovat kvalitativní materiály, aby získal numerické skóre. Hajek a Bohacova (2016) dále dodávají, že použití slovníků specifických pro finance v přístupu analýze sentimentu na akciových trzích ukázalo významně vyšší přesnost predikce ve srovnání s použitím obecných slovníků. Obecné slovníky jsou obzvláště nevhodné pro analýzu sentimentu finančních informací, což způsobuje vysoké procento nesprávné klasifikace sentimentu.

Hajek (2018) kombinuje výše uvedený finanční slovník se slovníkem Diction 7.0 používá řadu 35 slovníků k výpočtu pěti obecných sémantických rysů, konkrétně aktivity, optimismu, jistoty, realismu a shodnosti.Word2Vec, populárního slovníku sentimentu, který lze použít k výpočtu vztahu mezi slovy z hlediska podobnosti a váhy, což umožňuje předem transformovat slova na vícerozměrné vektory, které se pak používají k reprezentaci slovo. Tento slovník je integrován v Python a využívá jej Chen a Chen (2019), Hwang a Kim (2019), Kim a kol. (2018) a Pagolu a kol. (2017). Ren a kol. (2019) využívají HowNet, což je online znalostní báze, která odhaluje koncepční vztahy a vztahy mezi pojmy, jak jsou obsaženy v lexikonech čínštiny a jejich anglických ekvivalentů.

Li a kol. (2020) a Long a kol. (2018) klasifikují zprávy do tří kategorií: pozitivní, negativní a neutrální, tj. podle očekávání nebo přesvědčení vyjádřených ve zprávách. Pozitivní sentiment ve zprávě znamená, že se očekává, že cena akcií uvedená ve zprávě v blízké budoucnosti vzroste, nebo naznačuje tendenci ke koupi této akcie či býčí trend na trhu. Negativní sentiment ve zprávě znamená, že se očekává, že cena akcií uvedená ve zprávě v blízké budoucnosti poklesne, nebo naznačuje ten-

Autor	Druh lexikonu	Ukazatelé výkonnosti	Výkon
Hao a kol. (2021)	Chinese LIWC (CLIWC)	Accuracy	87 %
Owen a Oktariani (2020)	AFINN, BingLiu, NRC Hashtag, General Inquirer, SentiWordNet	MAPE, AMAPE	n/a
Sakhare a kol. (2020)	Natural Language Toolkit	Accurancy	75 %
Li a kol. (2020)	n/a	Precision, Recall, F-skóre	85 %
Sun a kol. (2020)	Chinese Tencent NLP platform2	n/a	n/a
Nti a kol. (2020)	Natural Language Toolkit	Accuracy, RMSE, MAPE, Sensitivity, Specitivity	75 %
Bouktif a kol. (2020)	VADER	Accuracy	60 %
Bouktif a kol. (2019)	TextBlob	Accuracy, F-skóre	60 %
Birbeck a Cliff (2019)	SentiWords	Precision, Recall, F-skóre	85 %
Moro a kol. (2019)	n/a	Accuracy, RMSE	89 %
Jammalamadaka a kol. (2019)	SenticNet	MAFE	80 %
Gross-Klussmann a kol. (2019)	Harvard General IV-4	Accuracy	60 %
Derakhshan a Beigy (2019)	SentiWordNet, Stanford CoreNLP	Accuracy, F-skóre	70 %
Chen a Chen (2019)	Word2Vec	Accuracy	55 %
Ren a kol. (2019)	HowNet	Accuracy	89 %
Hwang a Kim (2019)	word2vec	n/a	n/a
Chen a Shih (2019)	Harvard-IV-4, Loughran-McDonald	Accuracy	56 %
Batra a Daudpota (2018)	Natural Language Toolkit	Accuracy	76 %
Eliacik a Erdogan (2018)	Turkish Accounting and Auditing Standards Institution+Harvey's	Accuracy, F-skóre	87 %
Kim a kol. (2018)	word2vec	Accuracy	68 %
Feuerriegel a Gordon (2018)	Loughran-McDonald	RMSE	n/a
Long a kol. (2018)	n/a	Accuracy	70 %
Hajek (2018)	Loughran-McDonald	Accuracy, RMSE	70 %
Shi a kol. (2018)	n/a	Accuracy	60 %
Pagolu a kol. (2017)	Word2vec	Accuracy, F-skóre	71 %
Ab. Rahman a kol. (2017)	NLTK Corpus	Accuracy, Costing, Gamma	56 %
Khedr a kol. (2017)	n/a	Kappa, Accuracy	89 %
Domeniconi a kol. (2017)	n/a	Accuracy	80 %
Kraus a Feuerriegel (2017)	n/a	Accuracy, RMSE, MAE, MAPE	58 %
Xie a Jiang (2017)	Domain-Specific	Accurancy, MSE	60 %
Oliveira a kol. (2017)	Stanford CoreNLP	NMAE, MAE	n/a
Urolagin (2017)	AFINN's word dictionary	Accuracy	80 %
Simoes a kol. (2017)	n/a	Accuracy	82 %
Alostad a Davulcu (2017)	Loughran-McDonald	Accuracy	82 %
Zhao a kol. (2016)	SentiRank, FinLex	Accuracy, F-skóre	75 %
Das a Das (2016)	n/a	F-skóre	90 %
Eliacik a Erdogan (2016)	Turkish finanční slovník	Accuracy, F-skóre	87 %
Hajek a Bohacova (2016)	Harvard IV-4	Accuracy, RMSE, ROC, F-skóre	75 %
Al-Ramahi a kol. (2015)	Loughran-McDonald	Accuracy	75 %
Nguyen a kol. (2015)	SentiWordNet	Accuracy	65 %
Al Nasseri a kol. (2015)	SentiWordNet, Harvard IV	ROC, F-skóre	73 %
Smailović a kol. (2014)	n/a	F-skóre	n/a
Al Nasseri a kol. (2014)	Harvard IV-4, SentiWordNet	Accurancy	n/a
Meesad a Li (2014)	SentiWordNet	Accurancy, F-skóre	85 %
Tirea a Negru (2013)	nespecifikováno	n/a	n/a
Oliveira a kol. (2013)	Harvard General, MPQA SentiWordNet, Emoticons, ALL	R2	n/a
Nann a kol. (2013)	n/a	Prediction	n/a
Siering (2012)	Harvard-IV-4, FIN	Accuracy	70 %
O'Hare a kol. (2009)	n/a	Kappa	n/a

Tab. 2.5: Detekce sentimentu a ukazatelé výkonnosti
Zdroj:vlastní zpracování

denci tuto akci prodat či indikaci medvědího trendu. Neutrální sentiment znamená, že se očekává, že cena akcií zůstane v blízké budoucnosti beze změny bez zjevného očekávání a nemá tendenci obchodovat. Kim a kol. (2018) uvádí, že ve finančních dokumentech mohou stejná slova vyjadřovat různé nálady v různých sektorech; pokud se dokumenty z více sektorů učí současně, výkon se může zhoršit. Výsledky experimentu ukazují, že vytvoření slovníku sentimentu speciálně navrženého pro finanční doménu extrahovat sentimentální slova, která jsou specifická pro finanční sektor, zlepšuje predikční výkon o 25,4%. Siering (2012) také zjistili, že kombinace dolování textu a analýzy sentimentu zlepšuje výsledky prognóz. Vyšší přesnosti lze navíc dosáhnout použitím seznamů slov souvisejících s financemi pro analýzu sentimentu namísto obecného slovníku.

Nejnižší výkonnost deklarují Ab. Rahman a kol. (2017) a dodávají, že tento výkon lze připsat nejistotě akciových trhů a náročné finanční oblasti. Přesnost je však konzistentní s některými přesnostmi získanými jinými výzkumy v podobných výzkumných zájmech.

Lze si povšimnout, že přesnost predikce vývoje akciového trhu, resp. jednotlivých akcií se se zahrnutím sentimentu u relevantních studií pohybuje v rozmezí 56-90%. Přesnost se značně liší dle využitého lexikonu pozitivních a negativních slov a zvoleného klasifikátoru, stejně tak jako vstupech do modelu predikce. Vysokou přesnost vykazují zejména predikce čínského (Hao a kol., 2021; Ren a kol., 2019) a tureckého (Eliacik a Erdogan, 2016; 2018) akciového trhu, kteřé využívají speciální slovníky uzpůsobené jejich jazyku. Některé studie nevyužívají nebo neuvádějí využitý slovníkový přístup nebo vynechávají údaj o přesnosti predikce vývoje akciového trhu.

2.2.5 Hlavní zjištění, limity a výzvy současných studií

Na základě výše provedené bibliometrické a obsahové analýzy jsou nyní zjištěné poznatky syntetizovány a je upozorněno na hlavní limity a výzvy současných studiích včetně z nich vyplývajících vědeckých mezer neboli bílých míst, které jsou nedostatečně rozpracovány či zcela opomíjeny ve zveřejněných publikací. Je nezbytné konstatovat, tak jak je níže uvedeno, že výstupy jednotlivých autorů poskytují mnohdy protichůdné výstupy a aktuálně neexistuje shoda ohledně kalkulace, vlivu a implementace sentimentu v kontextu akciových trhů **(VO1)**. Identifikované vědecké mezery jsou stěžejní pro správné stanovení výzkumných otázek a hypotéz, které jsou následně v disertační práci řešeny a přispívají tak rozšíření či prohloubení současného stavu vědění v předmětné problematice **(VO2)**.

Počáteční výzkum predikce akciového trhu byl zcela založen na náhodných procházkách a numerické predikci, ale se zavedením behaviorálních financí byla při předvídání pohybu akcií zohledněna také víra a nálada lidí, přičemž tato oblast zís-

kává na popularitě jak upozornují Batra a Daudpota (2018). Feuerriegel a Gordon (2018) odhalují několik výzev v současném stavu poznání. Mezi ně patří věrohodnost finančních zpráv nebo, jinak řečeno, kvalita informací spojená se slovními projevy. I když se finanční zprávy jeví jako významná hnací síla oceňování akcií, stále existuje velká část nevysvětlitelných odchylek. Tuto složku lze snížit lepší technikou předpovídání, i když zbytková složka může být dokonce nepředvídatelná, zvláště když jsou signály nejasné nebo hlučné. Oliveira a kol. (2017) dodává, že různé textové zdroje mají odlišné charakteristiky, které se mohou doplňovat a umožňují lepší předpovědi. Například blogy mají úplnější názorový obsah, obsah mikroblogování má větší objektivitu, interaktivitu a frekvenci zveřejňování a vyhledávání Google představují větší počet uživatelů. Dynamická kombinace různých zdrojů dat na webu může vést k informačnějším finančním ukazatelům. Z toho důvodu se jako adekvátní jeví při práci s textovými daty uvažovat v disertační práci nejen příspěvky zveřejňované na sociální platformě obsahující množství dat se šumem, ale také finanční zprávy z renomovaných serverů obsahující především faktické a kvalitní informace.

Z rozsáhlého přezkumu literatury také vyplývá, že drtivá většina výzkumníků při analýze sentimentu využívá výhradně jeden anotovaný lexikon jako na například Hao a kol. (2021); Bouktif a kol. (2020, 2019); Birbeck a Cliff (2019) a mnoho dalších. Koncentrace pozornosti výhradně na jeden slovník může značným způsobem zkreslit výsledné skóre sentimentu. Nesprávně stanovené skóre sentimentu následně může zkreslit vliv na akciové trhy, potažmo způsobit značné ztráty investorům při implementaci do investiční strategie, neboť v konečném důsledku může indikovat nesprávné nákupní a prodejní signály. Disertační práce se z toho důvodu zaměřuje na vícero slovníků pozitivních a negativních slov **(VO3)**. Rovněž z literární rešerše někteří autoři tvrdí (Feuerriegel a Gordon, 2018; Hajek, 2018; Alostad a Davulcu, 2017), že speciální finanční slovníky inklinují na akciových trzích k přesnější kalkulaci sentimentu investorů. Tato domněnka je v práci také revidována prostřednictvím stanovené výzkumné hypotézy **(H1)**. V nepolední řadě se při analýze sentimentu autoři opírají o různé klasifikátory, které hrají ústřední roli pro správnou klasifikaci sentimentu. Obdobně je postupováno i v této práci a je zkoumán vliv volby binárního klasifikátoru na přesnost kalkulace skóre sentimentu **(VO4)**.

Dále z výstupů kritického přezkumu vyplývají nekonzistentní výstupy týkající se samotného vlivu sentimentu na akciové trhy. Shi a kol. (2018) zjistili, že sentiment investorů má krátkodobé pozitivní účinky a střednědobé reverzní účinky na akciový trh. Rovněž byl nalezen asymetrický efekt, že vysoký sentiment investorů získává zjevnější vliv na výnosy akcií. Hwang a Kim (2019) zjistili, že ceny akcií mají větší dopad na sentiment článků. Jinými slovy, ceny akcií reagovaly na sociální problémy dříve než články. Kromě toho nemusí být snadné předpovídat ceny akcií pomocí zpravodajských článků. Predikce ceny akcií je ještě obtížnější, pokud jsou zahrnuty

minulé a budoucí články. Nguyen a kol. (2015) upozorňuje, že sentiment nemusí být faktorem, který způsobí pohyb ceny akcií. Avšak ikdyž sentiment může být jedním z faktorů, které ovlivňují pohyb cen, extrahované sentimenty z diskusních fór aj. neodrážejí cenu kvůli chaotickému, chybovému komentáři nebo předpovědi chyby člověka, když zveřejňují zprávy. Z toho lze usuzovat, že doposud neexistuje jednotné stanovisko ohledně vlivu sentimentu na akciového trhu **(VO5)**. Je nezbytné se touto nejednoznačností dále zabývat a před integrací samotné investiční strategii prozkoumat, zda sentiment investorů je schopen předpovídat budoucí vývoj akciových trhů **(H2, H3)**.

Empirické výsledky Chen a Chen (2019) naznačují, že techniky analýzy velkých dat k hodnocení emocionálního obsahu komentářů k aktuálním akciím nebo finančním problémům mohou účinně předpovídat pohyb cen akcií. Lze dojít k závěru, že integrace sentimentu jak z dokumentů, tak konkrétních témat ze sociálních médií by mohla pomoci zlepšit predikci akciového trhu dle Nguyen a kol. (2015). Al Nasseri a kol. (2015) uvádí, že v praxi by to mohlo být použito ke stanovení přesného času, kdy mají být akcie drženy, přidávány (kupovány) nebo odebírány (prodávány) z portfolia, čímž se pro investora získá maximální návratnost investice. To by mohlo ušetřit čas a úsilí a povede to k informovanějšímu investičnímu rozhodnutí na akciovém trhu. Je třeba věnovat pozornost volbě vhodného expertního modelu, který by byl schopen s vysokou přesností predikovat vývoj akciového trhu s integrací sentimentu. Veškeré relevantní studie, které jsou detailně v disertační práci zkoumány zcela opomíjí expertní systém založený na fuzzy logice. Pouze nejnovější studie Hao a kol. (2021) implementuje fuzzy logiku prvního řádu, přičemž výsledky jeho studie jsou více než uspokojivé. Nabízí se příležitost prozkoumat v tomto kontextu fuzzy logiku nejen prvního, ale také druhého typu, která je v této oblasti zcela zanedbávána a opomíjena, přitom fuzzy logika poskytuje na akciových trzích pozoruhodné výsledky **(VO6, VO7)**.

Simoes a kol. (2017) na základě výsledků provedené studie konstatují, že je možné implementovat ziskovou obchodní strategii pomocí textových údajů ze sociálních medií. Nicméně Gross-Klussmann a kol. (2019) stále tvrdí, že dosud neexistuje shoda ohledně implementace sentimentálních signálů do investičních strategií. Nguyen a kol. (2015) uvádí, že jednoduchým předpokladem účinnosti funkce sentimentu je, že analýza sentimentu nemusí poskytnout žádné další informace, pokud lze pohyb akcií dobře předpovědět pouze historickou cenou. Pokud je přesnost modelu pouze s cenou vysoká, jsou na skladě trendy a historické opakování. V takových případech mohou k předpovědi stačit pouze historické ceny a integrace sentimentu nemusí přesnost příliš zlepšit. Z toho důvodu se disertační práce zaměřuje na přezkum tohoto problému a navrhuje vhodný expertní model integrující skóre sentimentu s cílem vytvořit ziskovou obchodní strategii **(VO8)**. V neposlední řadě je nezbytné poskytnout

komparaci přesnosti predikce vytvořeného expertního modelu s jinými, standardněji využívanými modely predikce a poskytnou rozsáhlou srovnávací pasáž, která ve výše uvedených studiích chybí **(VO9)**.

Existují ovšem další výzvy, které nejsou touto disertační pokryté z časových důvodů a omezeného rozsahu práce. Nicméně na tyto výzvy je možné navázat v dalším výzkumu, potažmo poskytnou inspiraci dalším výzkumníkům. Ab. Rahman a kol. (2017) doporučuje rozšířit stávající výzkumy integrací se správou portfolia a sledovat tak skutečné výnosy na základě predikce. Je pravděpodobné, že na dosažené výsledky mohla mít vliv řada omezení. Jako kupříkladu Hájek (2018), který se potýká s omezeným vzorkem zvoleného akciového trhu, které nelze generalizovat na všechny trhy, neboť nedávné empirické důkazy naznačují, že akciové trhy jiných regionů vykazují specifické chování. Domeniconi a kol. (2017) doporučuje dále prozkoumat možné korelace mezi různými tržními indexy a možnostmi akcií, které rozšíří analýzu na další zdroje nestrukturovaných textových proudů. Dalším nedostatek výše zmíněných prací je zkoumání pouze omezeného časového období. V neposlední řadě je velký nedostatek, jak poznamenává Pagolu a kol. (2017), že autoři využívají data z jednoho zdroje pro analýzu sentimentu lidí, taková data mohou být ovšem zkreslená, protože ne všichni lidé, kteří obchodují s akciemi, sdílejí své názory na předmětné sociální platformě či webové stránce.

3 Cíle a metodologie disertační práce

3.1 Cíle a výzkumné otázky disertační práce

3.1.1 Cíle disertační práce

Tento výzkum je zaměřen na využití expertních systémů k identifikaci možných vztahů mezi textovými informacemi potažmo z nich extrahovaných myšlenek a názorů investorů v souvislosti s budoucím vývojem akciového trhu.

Hlavním cílem disertační práce je tvorba modelu expertního systému sloužící pro podporu investičního rozhodování na akciových trzích s využitím sentimentu investorů extrahovaného z textových zpráv.

Tento obecně pojatý cíl je nyní podrobněji specifikován. Pro dosažení hlavního cíle je potřeba splnění dílčích cílů práce.

Dílčí cíle disertační práce:

- **Cíl 1**: Identifikovat z teoretického hlediska přístupy textové a sentimentální analýzy na akciových trzích a představit vhodný expertní systém sloužící k predikci.
- **Cíl 2**: Zmapovat současný stav poznání z doposud publikovaných předních vědeckých článků a příspěvků v dané problematice.
- **Cíl 3**: Klasifikovat skóre sentimentu z nestrukturovaných textových finančních zpráv a příspěvků zveřejněných online.
- **Cíl 4**: Vytvořit expertní model predikce vývoje akciového trhu integrací extrahovaného skóre sentimentu investorů.
- **Cíl 5**: Komparovat nově vytvořený expertní model se standardně užívanými modely predikce vývoje akciového trhu.

3.1.2 Formulace hypotéz a výzkumných otázek

Základní ideou empirického výzkumu, jak popisuje Punch (2015), je používat data pro zodpovídání otázek nebo navrhování a testování jistých představ a myšlenek. Nejprve je tak nezbytné soustředit se na to, co chce výzkum nalézt (tedy na otázky) a teprve pak pozornost zaměřit na to, jak bude výzkum prováděn (na metody). Schéma propojeí cílů disertační práce a výzkumných otázek s hypotézami je znázorněn na obrázku 3.1.

Hlavní cíl	Dílčí cíle	Výzkumné otázky	Výzkumné hypotézy	Sekundární výzkum	Empirický výzkum
Tvorba modelu expertního systému sloužící pro podporu investičního rozhodování na akciových trzích s využitím sentimentu investorů extrahovaného z textových zpráv.	**DC1**: Identifikovat z teoretického hlediska přístupy textové a sentimentální analýzy na akciových trzích.			x	
	DC2: Zmapovat současný stav poznání z doposud publikovaných předních vědeckých článků a příspěvků v dané problematice.	**VO1**: Jsou zjištěné poznatky vlivu sentimentu na akciové trhy jednoznačné a konzistentní a poskytují standardní modely přijatelné výstupy?		x	
		VO2: Jaká existují bílá místa v současném světovém výzkumu v oblasti predikce vývoje akciového trhu s ohledem na sentiment investorů?		x	
	DC3: Klasifikovat skóre sentimentu z nestrukturovaných textových finančních zpráv a příspěvků zveřejněných online.	**VO3**: Jaký vliv na skóre sentimentu má zvolený lexikon pozitivních a negativních slov?	**H1**: Speciální finanční slovník generuje přesnější skóre sentimentu jako obecný slovník.		x
		VO4: Jak souvisí volba binárního klasifikátoru s přesností kalkulace skóre sentimentu s ohledem na zvolený lexikon?			
	DC4: Vytvořit expertní model predikce vývoje akciového trhu integrací extrahovaného skóre sentimentu investorů.	**VO5**: Jaký vliv má skóre sentimentu na pohyb akciových trhů?	**H2**: Mezi sentimentem investorů a výnosem akciového trhu existuje kointegrace.		x
		VO6: Jaká míra neurčitosti ve fuzzy funkcích členství je vhodná pro predikci vývoje akciového trhu?			
		VO7: Který typ fuzzy funkce členství nejlépe odpovídá charakteru vstupních dat?	**H3**: Sentiment investorů způsobuje Grangerovu kauzalitu výnosů akciového trhu.		x
		VO8: Jak ovlivňuje investiční strategii integrace sentimentu?			
	DC5: Komparovat nově vytvořený expertní model se standardně užívanými modely predikce vývoje akciového trhu.	**VO9**: Který model poskytuje přesnější výsledky predikce vývoje akciového trhu?			x

Obr. 3.1: Schéma cílů a otázek disertace
Zdroj: vlastní zpracování

Výzkumné otázky

Výzkumné otázky vycházejí z výše obecně uvedených cílů a přetvářejí je do specifičtější a konkrétnější podoby. V této práci je na základě sekundárního výzkumu tvrzeno, že použití sentimentální analýzy k předpovědi pohybu akciových trhů není dostatečně vyspělé a je zapotřebí dalšího průzkumu a zejména vytvoření vhodného modelu expertního systému integrující tento sentiment a sloužící pro podporu investičního rozhodování. K tomu poslouží zodpovězení níže položených výzkumných otázek. Jedinečnost tohoto výzkumu tak spočívá v řešení výzkumných otázek a testování výzkumných hypotéz, které přispívají k prohloubení a rozšíření teorie.

Nejprve je nezbytné nadefinovat výzkumné otázky vyplývající ze sekundárního výzkumu a odkazující se na teoretické pozadí výzkumu:

- **VO1**: Jsou zjištěné poznatky vlivu sentimentu na akciové trhy jednoznačné a konzistentní a poskytují standardní modely přijatelné výstupy?
- **VO2**: Jaká existuje vědecká mezera v současném světovém výzkumu v oblasti predikce vývoje akciového trhu s ohledem na sentiment investorů?

Jedná se o zásadní výzkumné otázky, neboť pouze v případě, že prozatím publikované výstupní zjištění nejsou konzistentní a jednoznačné, je možné se výzkumným problémem dále zabývat. V opačném případě by nemělo smysl ve výzkumu dále pokračovat. Navíc pro relevantní výzkum, který prohloubí či rozšíří současnou teorií v předmětné oblasti, je nezbytné identifikovat tzv. bílá místa výzkumu, která jsou prozatím ostatními výzkumníky opomíjena a nedostatečně či vůbec zpracována.

Po zodpovězení těchto výzkumných otázek, je pozornost zaměřena na již identifikované vědecké mezery ve výzkumu a následující výzkumné otázky vycházejí tedy z těchto zjištění, která byla identifikována v kapitole 2.2.5.

Důležitým předpokladem sestavení klasifikátoru je označení vstupních dat na pozitivní nebo negativní. Automatické označování vyžaduje sladění textových datových zdrojů s příslušnými slovníky. Ve výsledku tyto faktory ovlivňují charakteristiky souboru tréninkových dat. Určují nejen kvalitu označení na pozitivní a negativní slova, ale také množství označení slov v jednolitých kategoriích. To zase ovlivňuje očekávanou kvalitu předpovědi. Vyvstává následující výzkumná otázka:

- **VO3**: Jaký vliv na skóre sentimentu má lexikon pozitivních a negativních slov?

Předpovídat ceny akcií pomocí technik těžby textu vyžaduje dosažení nejvyšší možné výkonnosti predikce. To zahrnuje výzvu volbu správného binárního klasifikátoru. Úspěch takového klasifikátoru do značné míry závisí na kvalitě vstupních dat. Je nutné učinit správnou volbu klasifikátoru vhodnou pro specifický cíl. Je tedy třeba si položit důležitou výzkumnou otázku:

- **VO4**: Jak souvisí volba binárního klasifikátoru s přesností kalkulace skóre sentimentu s ohledem na zvolený lexikon?

Al Nasseri a kol. (2015) zjistili, že příspěvky investorů zveřejňované na sociální platformě určují následné pohyby na akciovém trhu. Což je v rozporu s výzkumem Hwang a Kim (2019), kteří tvrdí, že, ceny akcií reagovaly na sociální problémy dříve než samotné články. Na základě této úvahy a protichůdných zjištění v recenzované literatuře je třeba znovu posoudit směr působení mezi sentimentem investorů a vývoje akciového trhu.

- **VO5**: Jaký vliv má skóre sentimentu na pohyb akciových trhů?

Ambiciózní cíl předpovídat ceny akcií pomocí sentimentu investorů vyžaduje dosažení nejvyšší možné výkonnosti predikce. To zahrnuje výzvu volbu vhodného expertního modelu. Jelikož různé expertní modely vykazovaly v minulosti rozdílný výkon predikce vývoje akciových trhů, je nutné navrhnout a vytvořit vlastní expertní model umožňující snadnou integraci vytvořeného sentimentu. Z kritického přezkumu vyplývá jako vhodný nástroj pro predikci akciového trhu fuzzy logika. Je tedy třeba si položit klíčové otázky:

- **VO6**: Jaká míra neurčitosti ve fuzzy funkcích členství je vhodná pro predikci vývoje akciového trhu?
- **VO7**: Který typ fuzzy funkce členství nejlépe odpovídá charakteru vstupních dat?
- **VO8**: Jak ovlivňuje investiční strategii integrace sentimentu?
- **VO9**: Který model poskytuje přesnější výsledky predikce vývoje akciového trhu?

Hypotézy

Hypotézy mají ve výzkumu důležitou roli, pokud mohou být dedukovány z teorie nebo jsou pomocí teorie vysvětleny, takže výzkum, který je testuje, skutečně testuje teorii, která za hypotézami stojí. Což představuje tradiční hypoteticko-deduktivní model výzkumu. Hypotézy jsou tak specifické predikce o tom, co se očekává, že se nalezne v datech. Hypotézy výzkumu vyplývají jednak z nekonzistentních zjištění předchozích studií a jednak ze zjištěných mezer, pokud jde o vztahy mezi sentimentem investorů a vývoj akciových trhů.

Na základě úvah v literatuře je zřejmé, že sociální média hrají důležitou roli při poskytování platformy pro vyjádření názorů. V dostupném výzkumu byl však studován většinově výhradně jeden slovník. To vyžaduje přezkum se zaměřením na vícero slovníků, ať již v obecné nebo specializované rovině, aby byla prozkoumána přesnost binární klasifikace zpráv na pozitivní a negativní. Slovníky totiž hrají klíčovou

roli při určování sentimentu. Volba nevhodného slovníku může generovat nesprávné skóre sentimentu a tím pádem poskytovat mylné nákupní a prodejní signály. Proto je uvedena následující hypotéza:

- **H1**:Speciální finanční slovníky poskytují přesnější klasifikaci polarity slov příspěvků zveřejňovaných na finanční sociální platformě než slovníky obecné.

Akademická literatura prokázala nekonzistentní zjištění při testování vztahu extrahovaného sentimentu z online zpráv či sociálních platforem a jeho vlivem na akciové trhy. Z toho důvodu je znovu testován a analyzován vztah mezi sentimentem investorů z textových dat zveřejňovaných na online finančních portálech a finanční sociální platformě.

- **H2**: Mezi sentimentem investorů a výnosem akciového trhu existuje kointegrace.
- **H3**: Sentiment investorů způsobuje Grangerovu kauzalitu výnosů akciového trhu.

Výzkumné otázky, hypotézy a metody je nezbytné ve výzkumu vzájemné propojit. To je totiž součástí interní validity výzkumu.

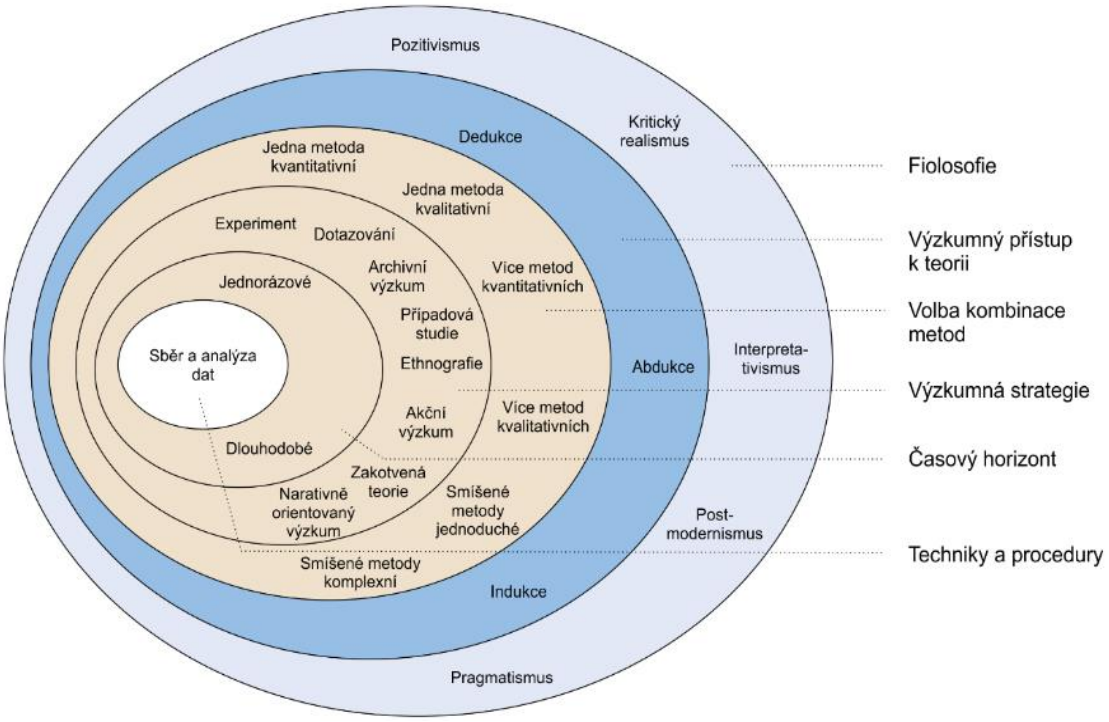
3.2 Metodologie výzkumu

V této kapitole je stručně uvedeno zařazení výzkumu disertační práce v rámce výzkumné filozofie a přístupu k budování teorie včetně představení výzkumného designu. Nicméně před samotným rozborem jednotlivých metod a postupů je nezbytné vymezit z terminologického hlediska základní pojmy, se kterými bude dále pracováno.

Metodologie výzkumu je vědní nauka, která zahrnuje plánování a realizaci výzkumu a také vyhodnocení výzkumných dat. Zahrnuje v sobě výzkumné metody, které využívá při organizaci výzkumu. **Metodika** výzkumu je soubor technik a postupů, které výzkumník využívá v rámci řešení výzkumného úkolu. **Metody** výzkumu jsou pak vědomé a cílené postupy, které používá výzkumník při realizaci svého výzkumu při řešení výzkumného problému pro dosažení teoretického nebo praktického cíle (Linderová a kol., 2016). Je nutné rozlišovat pojmy „metodologie“ a „metoda“ a to tak, že metodologie, jakožto vědní disciplína, využívá metod, jako prostředků, při řešení výzkumného úkolu. **Technika** nebo také **postup** je souhrn pracovních postupů, prostředků ale i zásad dané metody. Technika umožňuje dosahování požadovaných výsledků a cílů a umožňuje vykonávat danou metodu efektivně (Molnár, 2012).

Výzkum jako vícestupňový proces, který je nutné dodržet, aby mohl být proveden a dokončen výzkumný projekt. Přesný počet stupňů se liší, avšak na počátku

každého výzkumného projektu je potřeba si ujasnit předmětnou tematickou oblast zájmu. Demonstrovat, jak myšlenky souvisejí s výzkumem, který již byl proveden v tematické oblasti, a vytvořit jasný design výzkumu. Krom toho je nezbytné ujasnit si způsob shromažďování a analyzování dat. V rámci toho je také nezbytné zvážit platnost a spolehlivost údajů, které jsou využity ve výzkumu, spolu s přidruženými problémy s jejich přístupem. Vhodnost a přiměřenost analytických technik, které jsou využity, je stejně tak důležité zhodnotit. Nakonec výzkumného procesu následuje zveřejnění poznatků o výzkumném projektu v co možná nejjasnější a nejpřesnější podobě Saunders a kol. (2009). Zároveň výzkum musí prokázat interní validitu. To znamená, že použité metody musí odpovídat definovaným otázkám a logika výzkumu musí být jasná a interně konzistentní.



Obr. 3.2: Saundersova „výzkumná cibule“
Zdroj: Saunders a kol. (2012)

Jednotlivé vědecké výzkumné metody a postupy jsou popsány a zdůvodněny v kontextu jejich využití v této práci. Autorka práce metodicky vychází z publikace Saunderse (2012), která komplexně pokrývá potřebné oblasti realizovaného výzkumu. Je systematicky postupováno dle Saundersovi „výzkumné cibule“, který poslouží jako vhodný nástroj k rozebrání jednotlivých vrstev, které je nezbytné v průběhu výzkumného procesu zohlednit k poskytnutí kvalitního vědeckého výstupu. Nyní následuje diskuse o hlavních vrstvách „výzkumné cibule“, která zobrazena na obrázku 3.2.

3.2.1 Filozofie výzkumu

Za každým výzkumem je nějaká filosofie, i kdyby nebyla explicitně vyřčená. Filosofie totiž odpovídá na základní otázky, které se promítají do toho, jak výzkumník problém řeší a jak pracuje s informacemi. Základní přehled filosofických přístupů je také důležitý pro hodnocení zdrojů a informací – je jedním z pokročilých klíčů pro rozeznání kvality výzkumu. V Saundersově výzkumné cibuli jsou vysvětleny různé filozofie. Nejvýznamnější z nich jsou: pozitivismus, realismus, postmodernismus a pragmatismus.

Pozitivizmus je jeden z nejvlivnějších filozofických směrů 19. a 20. stol. Vychází z pozitivních (daných) poznatků a pozitivních faktů empirických věd; usiluje být filozofií a metodologií vědy. Pozitivizmus odmítá jakékoliv metafyzické otázky, tedy úvahy mimo oblast vědeckých pravd vyvozených (a experimentálně potvrzených) ze vztahů a zákonů prokazatelnými zkušeností (empirií), jak uvádí Olecké a Ivanové (2010).

Pozitivismus je založen na teoretické víře, že existuje objektivní realita, kterou může výzkumník znát správným používáním správných metod. Pozitivizmus lze považovat za paradigma kvantitativního výzkumu. Kvantitativní výzkum se opírá o základy takového vědění, které je získáváno empiricky a ověřováno. Vědecká fakta zde mají postavení neosobní, jsou empiricky zaručena a rigorózně testována. Logický postup užívaný v kvantitativním šetření je deduktivní. Cílem výzkumu je odpovědět na konkrétní výzkumnou otázku pomocí kvantitativních dat, je vysoce strukturovaný a využívá velký vzorek zejména textových dat.

Využití: Filozofií přijatou pro tento výzkum je pozitivismus.

3.2.2 Výzkumné vědecké přístupy

Filozofie výzkumu obsahuje důležité předpoklady o způsobu, jakým nahlížíte na zkoumanou problematiku. Tyto předpoklady podpoří výzkumnou strategii a metody jako součást této strategie. To, do jaké míry je jasně stanovená teorie na začátku výzkumu, vyvolává důležitou otázku týkající se koncepce výzkumného projektu. To je to, zda by výzkum měl využívat deduktivní přístup, při kterém se rozvíjí teorie a hypotézy a navrhuje se strategii výzkumu k testování hypotézy, nebo induktivní přístup, ve kterém jsou sesbíraná data a ve výsledku se rozvíjí teorii analýzy dat, popřípadě jinou metodu dle Saunders a kol. (2009).

Logické metody

Mezi přístupy k vypracování dizertační práce patří logické metody. Princip těchto metod je založen na logickém myšlení. Mezi logické metody patří následující dvojice

(Kučera a Radvan, 2000):

- analýzy –syntéza
- indukce –dedukce
- abstrakce –konkretizace
- formalizace –kvantifikace
- komparace –analogie.

Analýzy –syntéza

Ochrana (2019) definuje **analýzu** jako klíčovou obecně vědní metodu. Využívá se zejména ve fázi poznávání vědeckého problému při jeho detailním zkoumání. V rámci analýzy se určitý celek rozkládá na nižší entity. Těmito entitami mohou být prvky, vztahy mezi prvky či procesní vztahy. Analýza od sebe odděluje jednotlivé jevy a zkoumá je jako relativně izolované. Tím je umožněno proniknout až k samotné podstatě zkoumaného problému. V rámci metodologie vědy je rozlišeno několik druhů analýz. V této práci je uplatněna především systémová analýza. Systémová analýza je jednou ze systémových disciplín (obecná teorie systémů, systémová analýza, operační analýza a systémové inženýrství). Systémovou analýzou se rozumí metodická disciplínu, která směřuje k poznání systému postupnou dekompozicí systému na podsystémy až na prvky se známou funkcí a vazebností, nebo jako disciplínu, která je zaměřena na analýzu řídících a informačních systémů s využitím výpočetní techniky.

Využití: Metody analýzy jsou využívané především v analytické části práce při formulaci výzkumného problému. Data získaná literární rešerší byla analyzována na základě bibliometrické a obsahové analýzy, na jejím základě byly vyhodnoceny poznatky zkoumané problematiky.

Syntéza má v porovnání s analýzou opačný cíl použití včetně odlišného postupu. Syntézu a analýzu je možné v určitém smyslu považovat za dvě vzájemně se doplňující metody. Analýza je typická pro úvodní fázi vědeckého bádání ve chvíli, kdy chceme zkoumaný jev rozkladem poznat blíže. Syntéza poté navazuje na analýzu v další fázi vědeckého zkoumání, kdy je vytvářen nový vědecký obraz zkoumaného jevu. Syntéza funguje na spojování dvou nebo více prvků do jednoho celku, jak uvádí Sedláková (2014).

Využití: Metoda syntézy byla využita při sestavování získaných informací do vzájemně propojených logických celků. Syntéza je využita např. při formulaci závěrů disertační práce na základě shrnutí dílčích poznatků a jejich zobecňováním.

Indukce –dedukce

Indukce je myšlenkový postup od jednotlivého a zvláštního k obecnému. Je to tedy proces zobecňování. Z pozorování ojedinělých jevů, z analýzy dílčích poznatků o projevech či procesech se vyvozují obecné závěry. Induktivní postupy umožňují dospívat od dílčích, částečných pozorování k zákonitostem, které platí obecně, pro všechny entity uvažovaného druhu. V souvislosti s indukcí se stává důležitá, jak popisuje Janíček a kol. (2013), problematika verifikovatelnosti představ a hypotéz a problematika věrohodnosti úsudku s úmyslem zobecnit to, co bylo pozorováno a naměřeno v dílčích případech na celé universum všech případů.

Využití: S induktivním přístupem je v práci spojené zevšeobecňování poznatků z realizovaného výzkumu.

Postup indukce je spojen s postupem **dedukce**, který spočívá v tom, že vyvozuje z obecného poznatku jednotlivé jevy. Je to tedy proces přechodu od myšlenkové operace, která z jedné nebo více premis vyvozuje výrok, který je jejich logickým důsledkem. Je to způsob myšlení, při němž se z obecnějších závěrů, tvrzení přechází k méně obecným, dle Janíčka a kol. (2013).

Využití: Dedukce je v disertační práci uplatněna při stanovení cílů, výzkumných otázek a ověřování platnosti výzkumných hypotéz.

Abstrakce –konkretizace

Abstrakce je proces, v němž se berou do úvahy pouze skutečnosti podstatné a odlučují se nepodstatné stránky zkoumaného jevu v poznávacím procesu. Cílem je uvažovat pouze hlavní znaky entity. Abstrakce je nezastupitelný a nevyhnutelný stupeň v procesu poznávání objektivního světa. Využívá se ve smyslu teoretické zevšeobecňování neboli abstrahování, vytváření systému podstatných veličin při řešení problémů modelováním dle Janíčka a kol. (2013). Jak uvádí Olecká a Ivanová (2010), abstrakce je důležitým prostředkem myšlení, který umožňuje vyčlenit z celistvé, spojené a nerozlišené reality některou její část, vlastnost, vztah, proces apod.

Využití: Prostřednictvím abstrakce se selektují podstatné a konzistentní informace z různých bibliografických zdrojů, aby se mohla definovat teoretická základna pro praktickou část dizertační práce. Rovněž je abstrakce využita při identifikaci vstupů textových dat pro textovou analýzu.

Konkretizace je opakem abstrakce. Využívá se v případě, kdy dává určité entitě názorný, předmětný charakter, něčemu konkrétnímu výraz nebo při vyhledávání konkrétního prvku z určité třídy entit. Konkretizace se realizuje pozorováním, měřením a analogií. (Janíček a kol., 2013)

Formalizace –kvantifikace

Formalizací se označuje proces oddělení formy od obsahu konkrétní entity, která se tím stává abstraktní. Proces formalizace slouží k zefektivnění operací s informacemi o charakteristikách jednotlivých entit pomocí výrazů běžného jazyka prostřednictvím různých symbolů či znaků, tím pádem je entita zakódovaná, dle Janíček a kol. (2013).

Využití: V disertační práce je formalizace systematicky využívána při přebírání termínů, které výzkumníci již definovali ve svých předcházejících výzkumech.

Kvantifikace je proces přiřazování číselných hodnot formalizovaným veličinám v klasické matematice, resp. lingvistickým veličinám ve fuzzy matematice použitím kvantifikátoru, který určuje míru charakteristiky vzhledem k vybranému srovnávacímu objektu.

Využití: Kvantifikace je využita v části výzkumu, ve které jsou textové zdroje kvantifikovány a převedeny do numerické podoby.

Komparace –analogie

Komparace umožňuje stanovení rozdílů mezi porovnávanými jevy nebo objekty. Pozorovací kritérium může být stanovené věcně, časově anebo prostorově. Existují dva základní způsoby porovnávání, dle Hendl (2005):

- porovnání pojetí problému, názorů, premis pro vytvoření, ověření či zdůvodnění vlastního stanoviska nebo úvah;
- porovnání jako nástroj měřené, zjišťování, objektivizace a hodnocení dosažených výsledků.

Využití: V práci je komparace využita jak v teoretické části kritického přezkumu při vzájemné porovnávání výstupů jednotlivých autorů při řešení předmětné problematiky, tak v praktické části při porovnávání výkonnosti či přesnosti predikce jednotlivých expertních modelů a klasifikátorů.

Analogie vychází z metod komparace. Je určitým druhem úsudku, který je založen na vyhledávání analogií a usuzování z určitých již známých souvislostí, dle Gavetti a kol. (2005).

Využití: Využívá se při identifikací rozdílů a podobností jednotlivých klasifikátorů a modelů.

Metoda zpětné vazby

Pomocí této metody bude zajištěna reflexe každého výzkumného kroku tak, aby se výzkum neodchýlil od původního cíle a jeho východisek. Tuto metodu lze chá-

pat jako součást ostatních použitých metod, přičemž v procesu modelování zaujímá podstatnou roli.

Využití: Metody zpětné vazby je využíváno jak ve zhodnocení celkového modelu, tak v dílčích etapách jeho tvorby. Byla zjišťována zpětná vazba závislostí vstupních proměnných na výstupní proměnnou, ale především i verifikace výsledků z vytvořeného modelu.

3.2.3 Výzkumné vědecké metody

Další vrstvou ve výzkumné cibuli jsou metody výzkumu. Tyto metody jsou: monometodická, smíšená a multimetodická metoda. Tyto metody často souvisejí s odpovídajícími postupy analýzy dat, ať už jsou kvalitativní nebo kvantitativní. Jak uvádí Saunders (2012), při výběru svých výzkumných metod proto použije buď jedinou techniku sběru dat a odpovídající analytické postupy (monometodická metoda), nebo použije více než jednu techniku sběru dat a analytické postupy k zodpovězení své výzkumné otázky (multimetodická metoda).

Využití: Tato disertační práce využívá monometodickou metodu výzkumu. Všechna data se shromažďují stejným způsobem, a to z volně dostupných online databází, ať již se jedná o data pocházející z akciového trhu či textová data z online finančních zpráv a příspěvků na sociálních sítí.

Kvalitativní výzkum

Kvalitativní výzkum lze chápat jako nenumerické šetření, které umožňuje popsat reálný stav a který využívá jiných než matematických metod. Kvalitativní výzkum pracuje nejčastěji s textem a zodpovídají otázku „proč?“. Cílem kvalitativního výzkumu je formulování nových zjištění, hypotéz a teorií a pracuje nejčastěji s induktivním přístupem.

Kvantitativní výzkum

Kvantitativní výzkum umožňuje reprezentativní šetření populace, které lze zobecnit na populaci, dle Olecké a Ivanové (2010). Cílem kvantitativního výzkumu je testování hypotéz, nikoliv jejich formulování. V rámci kvantitativního výzkumu se nejčastěji využívá metod dedukce. Má četné výhody jako například relativně rychlý sběr data jejich rychlou analýzu, dále poskytuje přesná numerická data a jistotu, že výsledky jsou nezávislé na výzkumníkovi. Užitečný je zejména při zkoumání velkých skupin. Kvantitativní výzkum má však i své nevýhody. Tím, že se výzkumník soustřeďuje pouze na určitou teorii a její testování, může opomenout důležité fenomény. Způsob

získávání dat je navíc omezen na standardizované postupy, což zapříčiňuje poměrně nízkou validitu výsledků.

Využití: Disertační práce je založena na kvantitativním výzkumu, neboť používá statistické metody a pracuje tak s měřitelnými daty, pomocí nichž lze určitý jev kvantifikovat, uspořádat nebo porovnat.

3.2.4 Výzkumná strategie

Každou strategii lze použít pro průzkumný, popisný a vysvětlující výzkum (Yin 2003). Některé z nich jasně patří k deduktivnímu přístupu, jiné k induktivnímu přístupu. Kromě toho je třeba zdůraznit, že žádná výzkumná strategie není ve své podstatě lepší nebo horší než jakákoli jiná. Nejdůležitější proto není přiřazení ke konkrétní strategii, ale to, zda umožní odpovědět na konkrétní výzkumnou otázku a cíle dle Saunders a kol. (2009).

Existují různé strategie, které vědci přijímají pro konkrétní výzkumnou studii. V Saundersově výzkumu cibule jsou vysvětleny různé výzkumné strategie. Hlavní strategie jsou: experiment, průzkum, akční výzkum, případová studie, zakotvená teorie a archivní výzkum. Experiment je forma výzkumu, který vděčí za mnoho přírodních věd, i když se silně projevuje v mnoha společenských vědách. Účelem experimentu je studium příčinných souvislostí, tedy zda změna v jedné nezávislé proměnné způsobí změnu v jiné závislé proměnné (Hakim 2000). Experimenty proto mají tendenci být použity v průzkumném a vysvětlujícím výzkumu k zodpovězení otázek „jak“ a „proč“. Nicméně experimentální strategie nevyhnutelně nebude pro mnoho otázek výzkumu podnikání a managementu proveditelná, jak dále popisuje Saunders a kol. (2009). Jak uvádí Saunders (2012), nejjednodušší experimenty se zabývají otázkou, zda existuje souvislost mezi dvěma proměnnými. Složitější experimenty rovněž zohledňují velikost změny a relativní důležitost dvou nebo více nezávislých proměnných.

Využití: Pro tuto studii je experimentální výzkum vhodnou strategií výzkumu. Vazba mezi dvěma proměnnými je přesně to, co se tato studie snaží zjistit (sentiment a kurz akciového indexu). K tomu je třeba provést experiment expertních model. Tato disertační práce tak využívá experimentální výzkum k zodpovězení konkrétních výzkumných otázek.

3.2.5 Časový horizont výzkumu

Časové horizonty odkazují na časový limit, kterým je výzkum omezen. Existují dva typy časových horizontů – longitudinální a průřezový. V longitudinální studii výzkumník sleduje jevy po delší dobu, zatímco v průřezové studii je čas omezený.

Využití: Vzhledem k tomu, že časový rámec tohoto výzkumu je omezený, je využit průřezový časový horizont.

3.2.6 Metody a techniky sběru dat

Nejdůležitějšími prvky výzkumné studie jsou sběr a analýza dat. Data shromážděná a analyzovaná systematickým způsobem umožní odpovědět na výzkumnou otázku. Jsou rozlišovány různé metody a techniky sběru dat s ohledem na charakter samotných dat a charakter výzkumu. Rozlišují se sekundární data **(sekundární výzkum)**, která jsou dostupná z různých zdrojů a databází a byla získána za jiným účelem, než je studovaný problém, a primární data **(primární výzkum)**, která jsou nově získaná data pro konkrétní účel řešeného problému. Techniky sběru těchto dat pak lze rozdělit podle povahy výzkumu (Linderová a kol., 2016; Molnár, 2012):

Primární data

Primární data se týkají informací, které jsou generovány poprvé nebo které jsou generovány za účelem splnění konkrétních požadavků daného šetření. Primární data jsou shromažďována přímo od respondentů nebo subjektů experimentu. Hlavní nevýhodou použití primárních dat je skutečnost, že jejich shromažďování může být časově náročné a může být obtížné získat velké množství dat. Příklady zdrojů primárních dat zahrnují: průzkumy, dotazníky, plány rozhovorů a rozhovory, focus groups, případové studie, experimenty a pozorování.

Sekundární data

Sekundární údaje zahrnují jak kvantitativní, tak kvalitativní údaje a používají se hlavně v deskriptivním a vysvětlujícím výzkumu. Data, která používáte, možná používají data, u nichž došlo k malému, pokud vůbec nějakému zpracování, nebo kompilovaná data, získala nějakou formu výběru nebo shrnutí (Kervin, 1999). V rámci výzkumu v oblasti podnikání a řízení se tato data používají nejčastěji jako součást případové studie nebo strategie průzkumu. Neexistuje však žádný důvod nezahrnout sekundární údaje do jiných výzkumných strategií, včetně archivního výzkumu, akčního výzkumu a experimentálního výzkumu. Saunders a kol. (2009) Sekundární data, která jsou v rámci disertační práce upřednostňována se dělí na interní a externí.

- Interní sekundární data – jsou ty informace, které subjekt shromáždil za jiným účelem v minulosti a které zle použít i pro řešení aktuálního problému. Příkladem jsou účetní výkazy, databáze zákazníků nebo evidence pracovníků.

- Externí sekundární data – mohou být například výsledky výzkumu prováděných výzkumnými institucemi. Příkladem jsou výzkumy agentur, údaje Českého statistického úřadu nebo údaje dostupné z médií (Liška, 2004).

U mnoha výzkumných otázek a cílů je hlavní výhodou použití sekundárních dat obrovská úspora zdrojů, zejména vašeho času a peněz (Ghauri a Grønhaug, 2005). Obecně je mnohem levnější použít sekundární data, než si je sbírat. V důsledku toho lze analyzovat mnohem větší soubory dat, než jako jsou ty shromážděné předcházejícími průzkumy. Navíc je pravděpodobné, že se bude jednat o kvalitnější data, než která by mohla být získána sběrem vlastních dat (Stewart a Kamins, 1993). Nicméně sekundární data jsou bohužel shromážděna pro konkrétní účel, který se ve většině případů liší od výzkumníkových výzkumných otázek nebo cílů (Denscombe, 2007). Navíc pokud byly údaje shromažďovány z komerčních důvodů, může být získání přístupu nákladné. Saunders a kol. (2009)

Využití: Sběr sekundárních dat vychází z rešerše teoretických poznatků obsažených v odborných článcích, knižních publikacích a internetových zdrojů. Informace získané rozborem literatury a ze sekundárních údajů byly využité jako základ pro výzkum a jeho jednotlivé části. Pro provedení výzkumu jsou zapotřebí dva typy údajů: ceny akciového indexu, které jsou interními sekundárními údaji získanými z finančních databází; online články z finančních zpráv, které jsou shromažďovány z externích zdrojů.

3.2.7 Metody analýzy dat

Logika kvantitativního výzkumu je založen na sumarizaci, syntéze a redukci informací, které data zprostředkovávají. V rámci analýzy dat jsou učiněny závěry, dle položených výzkumných otázek a testovaných hypotéz. K tomu je nejdříve nezbytné popsat numerická i textová data prostřednictvím popisné statistiky a následně využít statistické metody k jejich zpracování, přičemž v rámci modelování je využit expertní systém fuzzy logiky, který byl již popsán v teoretické části práci.

Popisná statistika

Jde o disciplínu popisující a sumarizující informace obsažené ve velkém množství dat pomocí tabulek, grafů, funkcionálních a číselných charakteristik. Činí tak pomocí základních matematických operací. Cílem popisné statistiky je zpřehlednit informace skryté v datových souborech (Budíková a kol., 2010)

Statistické metody

Matematická statistika je věda, která buduje metody pro analýzu dat a využívá přitom princip statistické indukce. Její součástí je teorie odhadu, testování statistických hypotéz, statistická predikce. (Budíková a kol., 2010). Ze všech možných testů byly v disertační práci použity následující (Klímek a kol., 2006):

- Engle-Grangerův test kointegrace
- Test Grangerovy kauzality
- Wilcoxonův test

Testování statistických hypotéz je jednoduchý rozhodovací postup, při němž se na základě výsledků získaných náhodným výběrem vyslovíme buď pro testovanou (nulovou) hypotézu nebo alternativní hypotézu. Statistická hypotéza je určité tvrzení o parametrech (parametrické testy) pozorované náhodné veličiny pocházejícího ze základního souboru nebo o tvaru rozdělení znaku základního souboru (neparametrické testy) na základě pozorované náhodné veličiny. Testování statistických hypotéz je podrobně rozpracovaná teorie v řadě statistických učebnic, kde se postupuje podle následujících 6 kroků.

1. Volba nulové (testované) hypotézy H_0 a alternativní hypotézy H_1 (která popírá nulovou hypotézu).
2. Na základě vstupních informací učiníme rozhodnutí, kterým testem se bude testovat.
3. Sestrojení kritického oboru – W. Na základě zvoleného testu, alternativní hypotézy a hladině významnosti α (většinou 5% – chyba prvního druhu – zamítnutí hypotézy H_0, ačkoliv je správná) sestrojení intervalu W, který povede k zmítnutí hypotézy H_0.
4. Sestrojení testového kritéria T. Testovací (testové) kritérium T je statistika, jejichž rozdělení pravděpodobnosti známe a která vhodným způsobem souvisí s n, α, β. (β – chyba druhého druhu, přijmeme hypotézu H_0, ačkoliv platí H_1).
5. Rozhodnutí o platnosti hypotézy:
   - $T \notin W$ přijímám H_0, zamítám H_1
   - $T \in W$ zamítnu H_0 a přijmu H_1

Engle-Grangerův test kointegrace

Engle-Grangerův test se využívá ke zjištění vzájemné kointegrace časových řad. Tento test se opírá o předpoklad nestacionarity časových řad a je založen na testování odhadnutých reziduí z kointegrační regrese na přítomnost jednotkového kořene. Kointegrace je provedena pomocí metody nejmenších čtverců dle Hendl (2012):

$$Y_t = \beta_0 + \beta_1 X_1 + \cdots + \beta_t X_t + \varepsilon_t \tag{3.1}$$

kde X_1 a Y_t jsou časové řady, β_t jsou koeficienty a ε_t je reziduum. Následně je nezbytné testovat náhodné složky pomocí ADF testu, který umožňuje zjistit přítomnost jednotkového kořene prostřednictvím regrese obsahující zpožděné hodnoty těchto reziduí ve formě prvních diferencí, jak popisuje Arlt a Arltová (2007). S využitím rovnice lze získat odhad reziduí, které jsou využity v ADF testu:

$$X_t = \emptyset X_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta X_{t-1} + \varepsilon_t \tag{3.2}$$

Z rovnice je testováno, že parametr $\emptyset = 0$, jinými slovy, že obsahuje jednotkový kořen, X_t je závisle proměnná, p je zpoždění a ε_t je reziduum. Vyhodnocení testu je totožné s ADF testem dle nejnižší hodnoty Akaikova kritéria. Následně je testována nulová hypotéze, zda nejsou časové řady kointegrované. Je využito následující hypotéz:

H_0: časové řady nejsou kointegrované
H_1: časové řady jsou kointegrované

V případě, že p-hodnota testovaných reziduí je vyšší než zvolená hladina významnosti, která je obvykle volena jako 5% či 1%, tak nulová hypotéza není zamítnuta a časové řady nejsou kointegrované. V případě, že časové řady nejsou párově kointegrované, znamená to, že neexistuje dlouhodobý vztah a jejich regrese indikuje vztah zdánlivý dle Arlt a Arltová (2007).

Test Grangerovy kauzality

Časové řady jsou následně testovány na možné vzájemné příčinné vazby na základě Grangerovy kauzality. Základní myšlenka spočívá v tom, že pokud působí časová řada X na časovou řadu Y, pak řada X by měla vylepšit předpověď u řady Y, jak udává Arlt a Arltová (2007).

H_0: proměnná X neovlivňuje proměnnou Y v Grangerově smyslu
H_1: proměnná X ovlivňuje proměnnou Y v Grangerově smyslu

Přičemž základní modely, jak popisuje Hušek (2007), mají následující formu:

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-1} + u_t \tag{3.3}$$

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{i-r} + \sum_{i=1}^{p} \beta_i X_{t-i} u_t \tag{3.4}$$

kde α_i a β_i jsou koeficienty proměnných, X_t a Y_t jsou časové řady proměnných, p je zpoždění a u_t je náhodná složka. První rovnice odhaduje závisle proměnnou na

základě vlastních zpožděných hodnot, zatímco druhá rovnice integruje i zpoždění druhé proměnné. Tento test je realizován prostřednictvím VAR modelu. Následné vyhodnocení je založeno na komparaci vygenerované hodnoty testovací statistiky a kritických hodnot při zvolené hladině významnosti. V případě, že je dosažená hodnota testovací statistiky nižší než kritická hodnota, není zamítnuta nulová hypotéza a změny nezávisle proměnné nevysvětlují změny v závisle proměnné.

Wilcoxonův test

Wilcoxonův je neparametrickým protějškem párového testu, který srovnává rozdíly ve výkonnosti dvou modelů nad každou datovou sadou. Test v podstatě porovnává kladné a záporné rozdíly. Rozdíly jsou seřazeny na základě jejich absolutních hodnot, v případě shody jsou vypočítány průměrné hodnosti. Nechť je rozdíl mezi skóre výkonu dvou modelů na i-té z datových sad. Nechť R^+ je součet řad pro datové sady, kde druhý model překonává první a naopak R^-, jak poznamenává Trawiński a kol. (2012):

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{3.5}$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \tag{3.6}$$

Pořadí, kde $d_i = 0$, jsou rozděleny na polovinu a přidány k součtu. Pokud T označuje menší součet, tj. $T = min(R^+, R^-)$, z-statistika:

$$z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{3.7}$$

Pro větší počet souborů dat, například větší než 25, bude přibližně normálně distribuováno. Kroky Wilcoxonova testu jsou následující: Nechť Y označuje pozorovanou hodnotu, M_1 označuje předpovídanou hodnota prvního modelu a M_2 označují prediktivní hodnotu druhého modelu.

H_0: dva modely generují stejně přesné výsledky
H_1: jeden model je přesnější než druhý

Poté je rozhodnutí, zda zamítnout nulovou hypotézu, či nikoli, založeno na výsledné p-hodnotě. Pokud je p-hodnota větší, než zvolená hladina významnosti nulovou hypotézu není zamítnuta. V tomto případě je využito jednostranného intervalu spolehlivosti.

Modelování

V rámci vědeckého výzkumu je často nezbytné provést simulaci sledovaného systému. K simulaci v laboratorních podmínkách se využívá tzv. modelů, které představují ekvivalent reálného sledovaného systému. Modelování dokáže převést sledovaný systém do počítačové podoby, která bude sloužit k testování požadovaných situací. Model je tedy určitým přiblížením skutečnosti, přičemž se může od reálného systému lišit svou složitostí. Model nemusí zobrazovat komplexní souhrn všech množin vazeb a prvků reálného systému z důvodu jeho často velké náročnosti. Samotné modelování se využívá téměř ve všech vědeckých disciplínách, a to primárně z důvodu nízké finanční i časové náročnosti oproti sledování a vyhodnocování výsledku pouze na reálných systémech (Pravica a Suprr, 2011)

Vytvořený model musí splňovat několik následujících kritérií:

- konzistentnost s reálnými daty,
- umožnění ověření hypotéz,
- jednoduchost s preferencí co nejmenšího množství parametrů,
- řešení by mělo být optimální,
- umožnění racionálno rozhodování.

Následující kroky jsou definovány jako postup při tvorbě modelu:

- identifikace systému,
- sestavení modelu,
- zkoumání chování modelu,
- srovnání řešení se skutečností,
- interpretace získaných výsledků.

Modelování ekonomických procesů se v dnešní době provádí výhradně za pomocí počítačové techniky a sofistikovaných nástrojů pro modelování. Práce s modely v kombinaci s počítačovou technikou vyžadují přesné definování systému pomocí matematického vyjádření. Matematické vyjádření modelu definuje požadované vlastnosti sledovaného systému a následně pomocí počítačové techniky provádí simulaci, z níž lze odvodit určité závěry výzkumné činnosti.

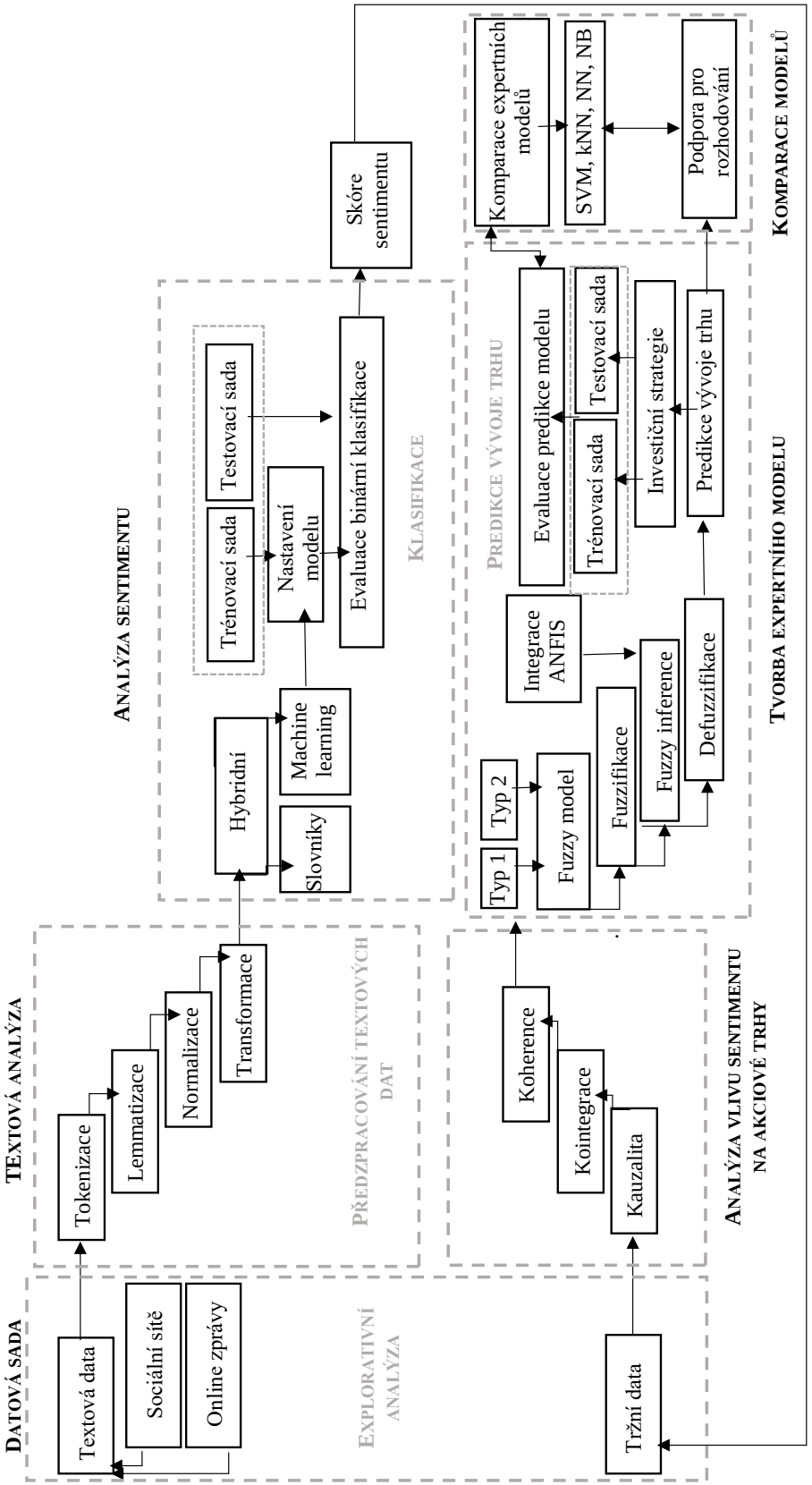
V disertační práci je, při tvorbě rozhodovacího modelu, využito techniky fuzzy modelování, které byly detailně popsány v kapitole 1.4. Technika fuzzy modelování je založena na principu fuzzy množin a využívá algoritmu fuzzy logiky. Technika fuzzy modelování byla zvolena proto, že řešená problematika, je charakterizována velkou měrou výskytu vágních (fuzzy) pojmů. Fuzzy modelování spočívalo v aplikaci všech základních etap modelování, přes výstavbu modelu po jeho naprogramování až po experimentování s modelem. Sestavený model má charakter dynamického modelu.

4 Řešení a výsledky disertační práce

Disertační práce je zaměřena na tvorba modelu expertního systému sloužící pro podporu investičního rozhodování na akciových trzích s využitím sentimentu investorů extrahovaného z textových zpráv. Většina výzkumů v této oblasti používá strukturovaná (kvantitativní) data k predikci vývoje akciového trhu. Nestrukturovaná data (jako text) však mohou poskytnout další doplňující informace s dodatečnými obtížně kvantifikovatelnými znalostmi (Groth a Muntermann, 2011). Proto se práce zaměřuje na propojení textových dat s cenami akciového indexu. Texty obsahují objektivní nebo subjektivní informace (Dařena a kol., 2018). Teorie behaviíuorálního financování říká, že emoce mohou hluboce ovlivnit chování a rozhodování jednotlivců i celých lidských společností (Kearney, 2014). To znamená, že ceny na akciových trzích jsou (víceméně) ovlivňovány emocemi, náladami a názory účastníků trhu (Bollen a kol., 2011). Tyto atributy jsou často obsaženy v textových dokumentech. Spolu s objektivními informacemi mohou pomoci určit, jaké je veřejné mínění o dané společnosti nebo akciovém trhu jako celku.

Jak je patrné ze schématu zpracování empirické části disertační práce na obrázku 4.1, je nezbytné k naplnění cíle práce postupovat v několika na sebe navzájem navazujících krocích. Nejprve je nezbytné shromáždit relevantní datové zdroje. Jedná se zejména o proud finančních online zpráv a příspěvků na sociální síti, které pojednávají o informacích týkajících se akciového trhu. Sesbírané textová data jsou poté předzpracována dle postupu prezentovaného v kapitole 1.2, aby mohly postoupit ke klasifikaci na pozitivní a negativní zprávy prostřednictvím zvolených lexikonů dle postupu uvedené v kapitole 1.3. Vzhledem ke stanoveným výzkumným otázkám a hypotézám, které vyplývají z kritického přezkumu v kapitole 2, je sledována výkonnost finančně, ale i obecně anotovaných slovníků pozitivních a negativních slov. Stejně tak je sledována výkonnost vícero binárních klasifikátorů. Tento krok hraje ústřední roli ke správnému stanovení skóre sentimentu, se kterým je v dalších částech empirického výzkumu disertační práce pracováno.

Na druhé straně jsou shromážděny údaje finanční časové řady, konkrétně ceny akciového indexu pro každý sledovaný den. Jsou stanoveny výnosy akciového indexu, aby byly splněny stacionární podmínky požadované testem kauzality Grangera popsaného v sekci 3.2. Vzhledem k denním změnám časové řady pravděpodobnosti sentimentu a denním výnosům v časové řadě se provádí Grangerova analýza kauzality (s ohledem na zpožděné hodnoty časové řady), aby se otestovalo, zda je sentiment investorů užitečný pro předpovědi pohybu cen na akciovém trhu a jak významné jsou výsledky. Rovněž je využita vlnková koherence k identifikaci vzájemné závislosti sentimentu investorů a vývoje akciového trhu a identifikaci jejich vzájemného působení.



Obr. 4.1: Schéma zpracování empirické části disertace
Zdroj: vlastní zpracování

Následně je vytvořena model expertního systému, konkrétně je pozornost zaměřena na fuzzy logiku, která vykazuje v kontextu akciového trhu řadu výhod, jak uvádí kapitola 1.4. Je vytvořeno několik reprezentativních modelů fuzzy logiky nižšího a vyššího stupně integrující různou úroveň neurčitosti mezi fuzzy funkcemi členství včetně rozličných tverů těchto funkcí s cílem verifikovat jejich výkonnost. Po návrhu fuzzy modelu je vytvořena investiční strategie sledující ziskovost expertního modelu integrující sentiment investorů. Na závěr je fuzzy model komparován s jinými, běžně užívanými modely pro predikci vývoje akciového trhu na základě evaluačních ukazatelů dle podkapitoly 1.4.3.

4.1 Explorační analýza vstupních dat

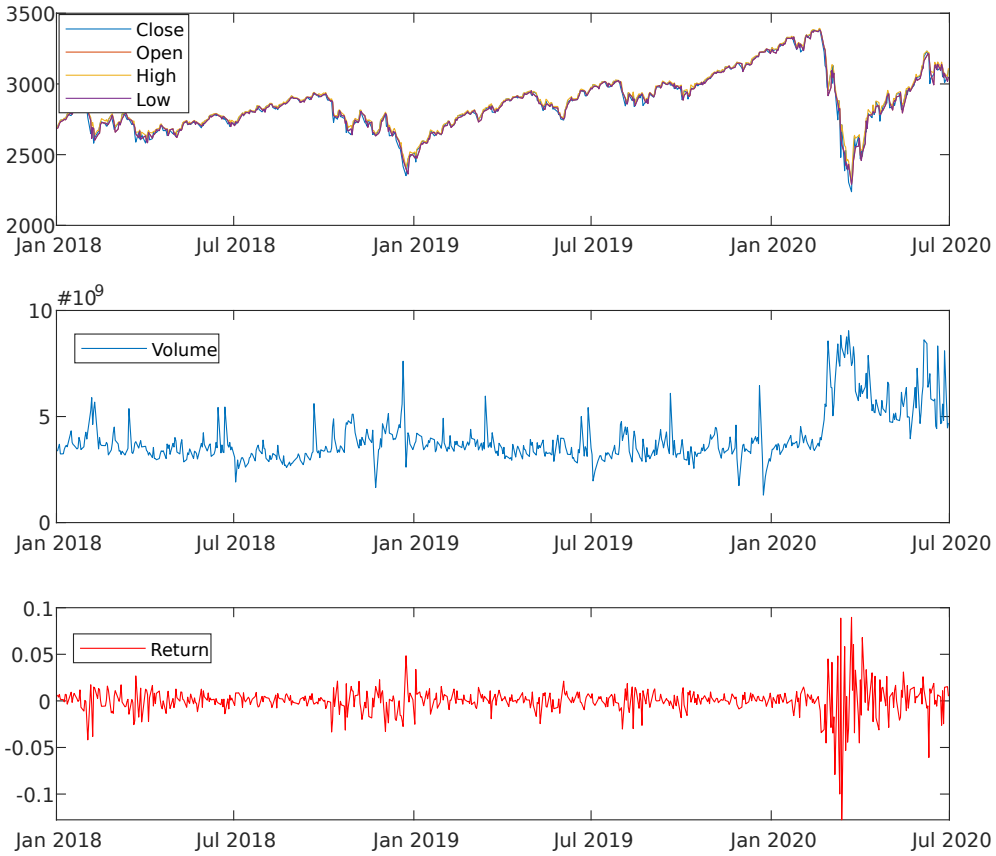
V této části práce je nejdůležitější definovat typ použitých tržních a textových zdrojů; zdroj, ze kterého byla tato data čerpána, množství analyzovaných dat a v neposlední řadě také software použitý ke spuštění předzpracování stejně jako následné analýzy. V disertační práci je využit software MATLAB 2021a a jeho dostupné balíčky a funkce relevantní k řešení předmětné problematiky. Je využito celouniverzitní licence, kterou univerzita VUT v Brně nabízí pro vědecké a výzkumné účely. Ve srovnání s některým proprietárním softwarem to umožňuje úplnou introspekci toho, jak se s daty zachází a jak se používají analytické algoritmy. Kromě toho umožňuje, aby byl výzkum plně transparentní vůči kolegům, vědcům a recenzentů.

4.1.1 Tržní data

Údaje o cenách a objemech předního akciového indexu jsou převzaty z webu Finance Yahoo!. K tomuto účelu je využit Datafeed Toolbox™, který poskytuje přístup k finančním údajům. Díky tomuto toolboxu je možné navázat spojení přímo z MATLABu k načtení nejen historických, ale také vnitrodenních datových toků a dat v reálném čase. Jelikož jsou využita textová data v anglickém jazyce, tak je pozornost zaměřena na americký akciový trh, který je zastoupen předním akciovým indexem Standard & Poor's 500 (zkráceně S&P 500). Tento americký akciový index je považován za indikátor vývoje ekonomiky nejen v USA, ale je také dokázáno, že se od něj odvíjí vývoj dalších akciových indexů potažmo ekonomik na jiných kontinentech. Jak již název indexu napovídá, obsahuje akcie 500 nejkvalitnějších společností kótovaných na burze v USA.

Prostřednictvím tickerů jsou z finančního webu staženy uzavírací, otevírací ceny, ale také nejnižší a nejvyšší ceny akciového indexu za jednotlivé dny sledovaného období 22.̇2.̇2018 až 17.̇7.̇2020. Vývoj akciového indexu je zachycen na obrázku 4.2, přičemž v horní části je vyobrazen vývoj cen, zatímco ve střední části obrázku je

zobrazen vývoj objemu obchodování a ve spodní části je graficky zachycen vývoj výnosu akciového indexu. Pokud jde o finanční údaje, může být vizualizace časové řady vysoce odhalující, pokud jde o identifikaci určitých vzorců. Akciový index vykazuje pozitivní trend ve z hlediska vývoje ceny. Zdá se, že jejich vývoj je v souladu s ekonomickým cyklem a růstem HDP USA. Vytvořená obchodní strategie bude pronásledovat příležitosti v krátkých časových intervalech a předvídat vzestupy a pády na základě sentimentu, který projevuje široká investorská veřejnost v online prostoru.



Obr. 4.2: Vývoj akciového indexu S&P 500
Zdroj: vlastní zpracování

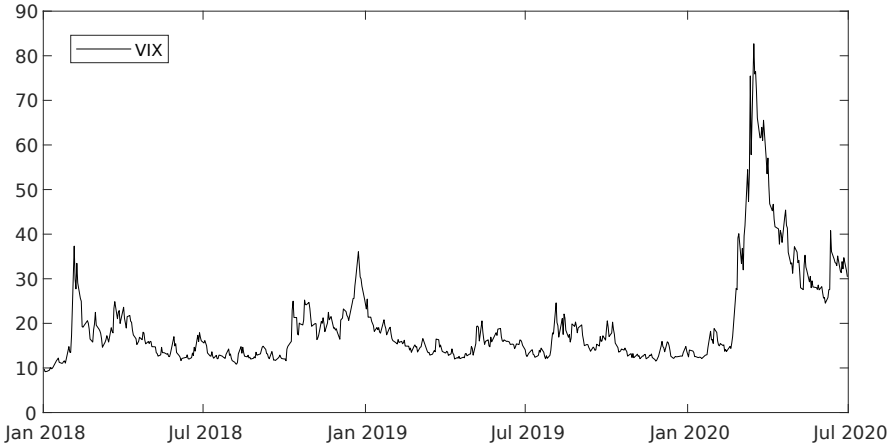
Je vhodné kromě ceny sledovat také vývoj objemu obchodování, neboť ceny na akciových trzích velmi kolísají v závislosti na nabídce a poptávce cenných papírů. Časovou řadu objemu obvykle charakterizuje několik vrcholů spojených s masivními nákupními nebo prodejními akcemi. Tyto vrcholy se obvykle objevují po příznivých nebo nevýhodných zprávách, podle toho, zda jde o nákup nebo prodej. Navíc v panických stavech může docházet k nadměrnému prodeji držených titulů investorskou veřejností, zatímco v nadměrném optimistické náladě může docházet k nadměrným nákupům akciových titulů. Tyto nezvyklé aktivity spojené s přehnanou pesimis-

tickou či optimistickou náladou projevovanou na trhu může zachytit právě tento ukazatel. Negativní nebo kontroverzní informace o určitém titulu či obecně o vývoji ekonomiky zvyšují nejistotu a vytváří vyšší variabilitu cen akcií, protože budoucí scénář je nejasný, a proto činí investice riskantnější.

Investoři obecně spekulují s variabilitou na základě různých typů analýz, ale ve všech studií je dominantní zřetele věnován právě výnosům, neboť jde o hodnotovou referenci pro investice do cenných papírů a velikost se má maximalizovat. Výnosy generované z akciového trhu jsou generovány ve formě zisku prostřednictvím obchodování nebo dividend vyplacených společností. Ve skutečnosti investoři nevěnují pozornost povaze dosaženého výnosu, protože upravená konečná cena odráží obojí, a výsledkem je pouze dosažení nejvyššího možného zisku bez ohledu na to, zda jej lze dosáhnout prostřednictvím dividend nebo obchodováním s akciemi. Výnos akciového indexu je počítá podle Sun a kol. (2020) dle následujícího vzorce:

$$R_{i,t} = \ln(P_{i,t}) - \ln(P_{i,t-1}) \tag{4.1}$$

kde $P_{i,t}$ je uzavírací cena akcie i v časovém období t a $P_{i,t-1}$ je uzavírací cena den před $t-1$.



Obr. 4.3: Vývoj indexu VIX za sledované období
Zdroj: vlastní zpracování

K finančním výpočtům v prostřední MATLAB slouží Financial Toolbox™, který poskytuje funkce pro matematické modelování a statistickou analýzu finančních dat. Ke konverzi ceny na výnos lze jednoduše využít funkci `price2ret` nebo `tick2ret`. Ve sledovaném období vykazuje výnos indexu S&P 500 stabilní tendenci, nicméně je patrné značné rozkolísání zejména v období vypuknutí koronavirové krize na území USA směrem k vysokým výnosům, ale také směrem ke značným ztrátám.

Kromě akciového indexu S&P 500 je důraz také kladen na vývoj indexu VIX, který je označován jako index strachu a poskytuje měřítko tržního rizika a sentimentu investorů. V podstatě by tento index měl v sobě integrovat očekávání trhu na 30denní volatilitu do budoucnosti. Index VIX je odvozen z cenových vstupů indexových opcí S&P 500. Vývoj indexu je zachycen na obrázku 4.3. Z vývoje je patrný značný nárůst indexu VIX v období již zmíněné rozšíření pandemické situaci na území USA. Nárůst indexu označuje narůstající strach, nejistotu a obavy investorské veřejnosti. Naopak pokles indexu VIX indikuje klesající strach s výhledem pozitivního vývoje trhu. Tento index může sehrát důležitou roli při predikci vývoje akciového trhu.

	Průměr	**SE**	**Min**	**Medián**	**Max**	**Skewness**	**Kurtosis**	**J-B test**
SP 500 cena	2875.96	201.76	2237.4	2864.36	3386.15	0.26	0.04	240.52*
SP 500 výnos	0.00	0.02	-0.12	0.00	0.09	-0.58	14.93	3828.94*
VIX cena	19.74	10.68	10.85	15.93	82.69	2.74	9.06	1781.41*
VIX výnos	0.00	0.09	-0.27	-0.01	0.39	1.14	3.11	138.56*

* označuje p-hodnotu nižší než 0.01

Tab. 4.1: Deskriptivní statistiky indexů
Zdroj:vlastní zpracování

V tabulce 4.1 jsou uvedeny deskriptivní statistiky akciového indexu S&P 500 a indexu VIX. Adjustované uzavírací ceny indexu S&P 500 vykazují pozitivní trendy a výnosnost je značně stabilní (směrodatná odchylky kolem 0.02%) a není zvlášť vzdálená od 0. Maximální výnos za sledované období dosahoval 9.4%, naopak nejnižší výnos, resp. nejvyšší ztráta se pohybuje okolo -12%. Co se týče výnosu indexu strachu, tak maximální změna je 39.2% označující prudký vzestup strachu na trhu, zatímco nejvyšší negativní změna je -26.6%, což označuje uklidnění na trhu s převahou pozitivismu. Skewness je míra asymetrie nebo přesněji nedostatek symetrie. V tomto případě je výnos akciového indexu vychýlen směrem doleva, znamená to negativní skewness a tím pádem atribut distribuce má jeden ocas těžší než druhý. Dále ukazatel Kurtosis indikuje, že akciový index nemá hodnotu velice blízkou 3, což je teoretická hodnota pro Gaussovo rozdělení pravděpodobnosti. To naznačuje, že pravděpodobnostního rozdělení této časové řady se nezdá být normálně distribuováno. Pro potvrzení tohoto předpokladu je proveden Jarque-Bera test s nulovou hypotézou, že příslušné rozdělení pravděpodobnosti je Gaussovo (chí-kvadrát s 2 stupni volnosti). Vykázané hodnoty vedou k odmítnutí nulové hypotézy. Tento nedotčený standard je v souladu se známými „stylizovanými fakty“ tržních výnosů. Na základě hodnot lze konstatovat, že akciový index má asymetrické rozdělení pravděpodobnosti výnosů. Souhrnně lze říci, že indikátor americké ekonomiky představují relativně stabilní ceny s pozitivními trendy a s distribucí výnosů, které jsou z velké

části normální, soustředěné kolem nuly, bez velmi velkých odchylek.

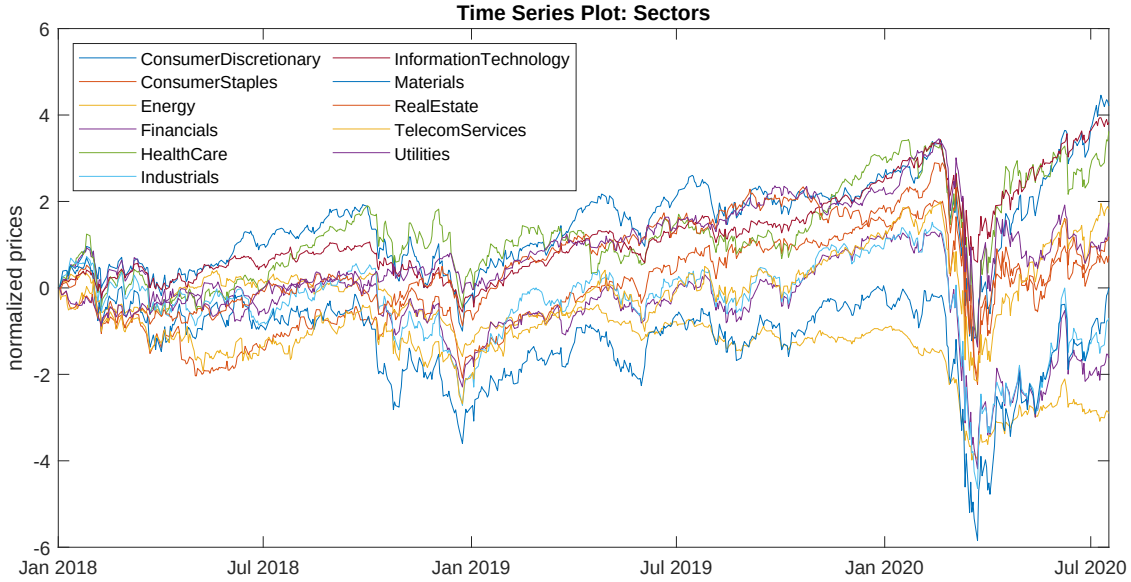
Datum	S&P 500	VIX	Zpravodajské weby
3/16/20	-11.98 %	35.76 %	Brought to the brink by coronavirus, airlines seek emergency aid. Hyundai Motor's China plant sales in February fall 97 % from a year ago. EU regulators ask companies to delay merger filings over coronavirus disruption
3/12/20	-9.51 %	33.66 %	Wall Street dazed and confused after worst day since 1987. Plunging Wall Street stocks end record bull run. Wall Street empties out as New York City declares state of emergency.
3/9/20	-7.60 %	26.12 %	Wall Street clobbered as crude plummets, virus crisis deepens. Exclusive: Trading in VIX options froze after open – CBOE. Investor fears rise over recession, bear market as coronavirus spreads in U.S.
3/24/20	9.38 %	0.13 %	Fed's stimulus eases global market fears, gets cash flowing. Dow soars over 11 % in strongest one-day performance since 1933. Dollar pares losses as investors wait on stimulus bill.
3/13/20	9.29 %	-26.62 %	Drumbeat of bad coronavirus news starts to hit U.S. auto dealers. Bill Gates steps down from Microsoft board. Late Wall Street rally leads global stocks higher; oil also jumps
4/6/20	7.03 %	-3.39 %	NYSE in talks with SEC to ease listing rules during coronavirus volatility. Fed says it will provide financing against new U.S. 'payroll protection' loans. Wall Street soars on hopes of slowing coronavirus deaths.

Tab. 4.2: Ukázka titulků v den s abnormálními výnosy
Zdroj: vlastní zpracování

Tabulka 4.2 shromažďují data, kdy přední americký indikátor ekonomiky zaznamenaly nejvyšší a nejnižší výnosy. K výnosům akciovému indexu jsou také dohledány změny indexu VIX. Může být vhodné prozkoumat titulky zpráv kolem dat uvedených v předchozích tabulkách, aby se zjistilo, zda tyto abnormální výnosy pocházejí z pozitivních nebo negativních výroků. Tabulka dále ukazuje, že záporné výnosy jsou v absolutních číslech vyšší než ty pozitivní a rozsah výnosů je značně rozkolísán, a to z obou stran.

S&P 500 je široce založený index, který odráží dynamiku amerického akciového trhu. Existují také sektorové indexy, které ukazují stav různých tržních odvětví a průmyslových odvětví. V tuto chvíli se indexy počítají pro 11 sektorů, 24 průmyslových skupin, 67 průmyslových odvětví a 156 pododvětví. Vážení sektoru S&P 500 jsou pravidelně přezkoumávána. Dále je pozornost zaměřena na sektory. Stejným způsobem však lze postupovat i pro detailnější členění. Na obrázku 4.4 je zachycen vývoj sektorů za sledované období. Pro vizualizaci do jednoho grafu byla využita normalizace cen, neboť každý sektorový index je obchodován za jinou cenovou úroveň. I zde na grafu je patrný strmý pokles v době propuknutí a nekontrolovaného

šíření pandemie na území USA. Po prudkém zhroucení trhů nastalo postupné oživení. Největší pokles v této době zaznamenal sektor materiálů, energetiky a finančnictví. Utlumilo se například stavebnictví a tím pádem i sektor materiálu včetně stagnace o finanční služby. Naopak nejvýraznější oživení po pandemické krizi nastalo u sektoru spotřebních služeb, informačních technologií a zdravotní péče. Jedná se zcela o logický vzestup těchto sektorů, neboť v této nestabilní situaci se omezil pohyb osob a při izolaci davu hrály informační technologie důležitou roli při včasném sdílení informací a šíření prevence a kontroly digitální pandemie. Také docházelo k hojnějšímu využívání informačních technologií, lidé více objednávali z internetu spotřební zboží a také se vkládali naděje do zdravotnictví a vývoji vakcín proti koronaviru. Celkově tyto odvětví nejenže vydržely nepříznivý dopad v den incidentu, ale také prokázaly silnou schopnost vyrovnat se – ceny akcií do určité míry vzrostly napříč těmito odvětvími.



Obr. 4.4: Vývoj sektorů za sledované období
Zdroj:vlastní zpracování

Z tabulky 4.3, kde jsou uvedeny deskriptivní statistiky jednotlivých sektorů, lze vyčíst, že průměrný výnos u všech sektorů vykazuje nulovou hodnotu za sledované období. Nepatrně záporné hodnoty výnosů u některých sektorů jsou zaznamenány prostřednictvím mediánu, který vykazuje vyšší vypovídající schopnost než prostý průměr. Nicméně oba ukazatele lze považovat za shodné, a tedy lze tvrdit, že v časové řadě se dlouhodobě nevyskytují extrémní hodnoty. Nejvyššího výnosu dosáhl sektor energetiky (25.1%), sektor nemovitostí (19.8%), finančních služeb (16.3%) a informačních technologií (16.2%). Ostatní sektory dokázali maximálně zhodnotit finanční prostředky o 10-16%. Nejvýraznější ztráta je zaznamenána také u sektoru

energetiky (14%), finančnictví (11.7%) a služeb (11.6%).

	SE	**Min**	**Medián**	**Max**	**Skewness**	**Kurtosis**	**J-B test**
Spotřební sektor	0.02	-0.08	0.00	0.14	1.53	0.00	4106.9*
Spotřební zboží	0.01	-0.08	0.00	0.10	0.67	14.88	3812.2 *
Energetika	0.03	-0.14	0.00	0.25	2.35	24.59	13016.3*
Finance	0.02	-0.12	0.00	0.16	1.27	15.50	4341.1*
Zdravotní péče	0.01	-0.07	0.00	0.11	0.77	10.72	1650.3*
Průmysl	0.02	-0.11	0.00	0.13	1.05	12.49	2517.5*
IT	0.00	-0.11	0.00	0.16	1.05	13.58	3104.8*
Materiál	0.02	-0.10	0.00	0.13	1.08	12.09	2327.7*
Reality	0.02	-0.08	0.00	0.20	2.41	29.26	19000.2*
Telekomunikace	0.02	-0.08	0.00	0.12	1.00	9.77	1327.4*
Veřejné služby	0.02	-0.12	0.00	0.13	0.68	17.87	5948.2*

*označuje p-hodnotu nižší než 0.01

Tab. 4.3: Deskriptivní statistiky sektorů
Zdroj:vlastní zpracování

Míra asymetrie měřená prostřednictvím ukazatele skewness vykazuje u všech sektorů hodnotu blízkou nule, což značí symetrické rozdělení výnosů, které jsou mírně vychýleny směrem doleva. Velmi často je hodnota kurtosis porovnávána s normálním rozdělením, která je definovaná hodnotou 3. V tomto případě je u všech sektorů hodnoty vyšší jako 3, což značí, že vícero dat je soustředěno na okraji. Jarque-Bera test tento předpoklad potvrdit, vykázaná p-hodnota vede k zamítnutí nulové hypotézy o normalitě dat. Stejně jako indikátor ekonomiky zastoupen akciovým indexem, tak také jednotlivé sektory nemají normální rozdělení pravděpodobnosti výnosů.

Další zajímavým statistickým údajem o sektorech je korelační matice, která zahrnuje sestavení srovnávací tabulky k odhalení jakékoli zajímavé závislosti a provázanosti sektorů. Korelační matice je široce používána při finanční a vícerozměrné analýze při řešení velkého počtu proměnných a určitě si zaslouží zmínku. Korelační tabulka není nic jiného než užitečné způsoby přenosu korelačních koeficientů mezi proměnnými. V tomto ohledu jsou vypočítány korelační koeficienty pro výnosy. Bylo provedeno statistické testování korelačních koeficientů, přičemž všechny údaje vykazovali p-hodnotu nižší než 0.01, což vede k zamítnutí nulové hypotézy, která předpokládá, že sektory jsou vzájemně nekorelované. Z tabulky 4.4 je patrná vysoká závislost jednotlivých sektorů, což deklarují vysoké hodnoty korelačních koeficientů. Hodnota korelačního koeficientu blízká hodnotě 1 udává dokonalou pozitivní závislost, která se vyskytuje u všech sektorů. Nejnižší provázanost s ostatními sektory vykazuje sektor služeb s hodnotou korelačního koeficientu blízkou 0.6. Ostatní sektory vykazují velmi silnou závislost mezi sebou navzájem měřeno korelačním koeficientem nad hodnotu 0.7. Je tedy zřejmé, že jednotlivé šoky, ať již negativní či pozitivní, se

mezi sektory budou navzájem šířit se stejným směrem a obdobnou intenzitou.

	SS	SZ	E	F	ZP	P	IT	M	RE	TS	VS
Spotřební sektor	1										
Spotřební zboží	0.71	1									
Energetika	0.73	0.57	1								
Finance	0.83	0.73	0.82	1							
Zdravotní péče	0.82	0.81	0.67	0.79	1						
Průmysl	0.86	0.74	0.83	0.92	0.82	1					
IT	0.92	0.73	0.70	0.82	0.85	0.84	1				
Materiál	0.84	0.74	0.8	0.89	0.81	0.92	0.83	1			
Reality	0.71	0.79	0.63	0.76	0.75	0.76	0.7	0.75	1		
Telekomunikace	0.83	0.73	0.65	0.75	0.78	0.75	0.82	0.74	0.66	1	
Veřejné služby	0.57	0.8	0.51	0.64	0.69	0.64	0.59	0.65	0.84	0.57	1

Tab. 4.4: Korelační matice výnosů jednotlivých sektorů
Zdroj:vlastní zpracování

Nicméně je potřeba upozornit na skutečnost, že tato vzájemná závislost je posouzena pouze za sledované období, v delším časovém horizontu může být korelace odlišná. Zejména korelace aktiv se při korekcích, panikách, silném medvědím trhu atd. se korelace dramaticky zvyšuje, takže nechrání portfolio ve chvílích, kdy je to nejvíce potřeba.

4.1.2 Textová data

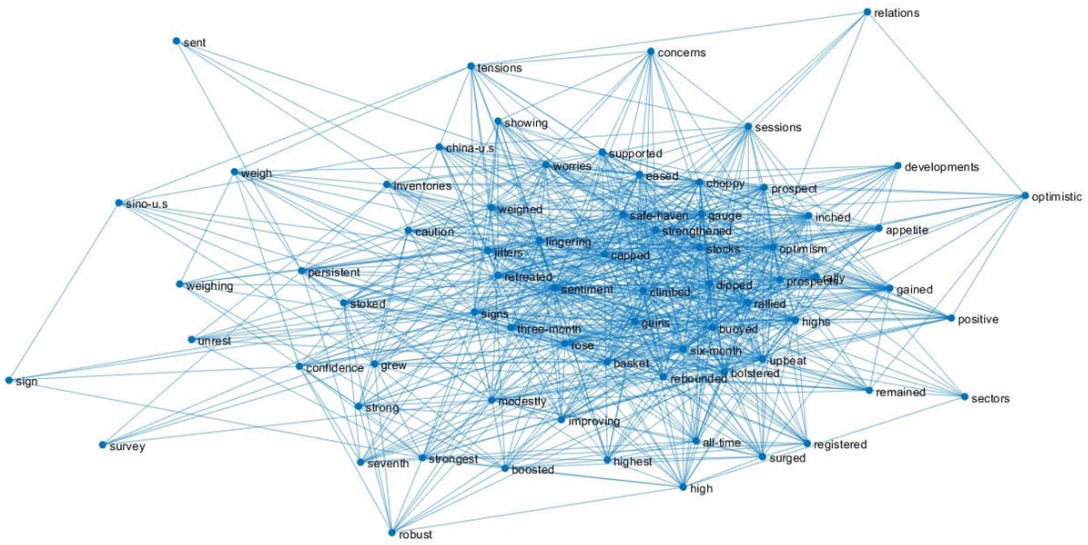
Texty mohou obsahovat především faktické (např. finanční výsledky) nebo názorové (např. pozitivní nebo negativní vztah k firmě či produktu) informace, případně obojí. V druhém případě tak můžeme zkoumat, zda pohyb ceny akcie souvisí s tím, jaké emoce, nálady a názory mají lidé, obchodující na akciových trzích, resp. zda a jak se jimi nechají ovlivnit. Mezi těmito dvěma typy dokumentů nebude explicitně rozlišováno, i když hrubé rozlišení poskytnou samotné zdroje textů – dá se očekávat, že novinové články budou obsahovat faktické, zatímco statusy na sociálních sítích spíše názorové informace.

Vzhledem ke skutečnosti, že mnohé z přijatelných a dříve analyzovaných zpravodajských serverů a sociálních sítí, které se jeví jako vhodné pro analýzu akciových trhů, mají omezený počet stažení za den i přes vývojářské rozhraní API. Zejména kvůli nedávným změnám zásad Twitteru by neměly být veřejně dostupné datové soubory obsahující tweety, což významně komplikuje tuto analýzu. Zásady Twitteru aktuálně relativně přísně omezují počet tweetů a období posledních několika dní, kdy jsou tweety k dispozici, bohužel je toto období příliš krátké na to, aby mohlo být použité pro účel disertační práce. Z toho důvodu jsou textová data pro účely

disertační práce využita z databáze Kaggle shromážděná od jiných výzkumníků, kteří jsou řádně citováni. Zpracování textové analýzy a následné analýzy sentimentu je provedeno prostřednictvím Text Analytics Toolbox™, které integruje speciální funkce k tomuto účelu.

Zpravodajské servery

Jako první logický zdroj dat se nabízejí zpravodajské servery (webové portály), zaměřující se na akciové trhy. Tyto servery publikují různé novinové články, pojednávající o jednotlivých firmách i o akciovém trhu celkově. Sentiment článků o finančních novinkách odráží a řídí výkonnost akciového trhu. Proto při použití zpracování přirozeného jazyka (NLP) prostřednictvím těchto titulků lze pozorovat jak pozitivita/negativita skóre každý den koreluje se zisky/ztrátami akciového trhu.



Obr. 4.5: Propojení s klíčovým slovem „sentiment“
Zdroj:vlastní zpracování

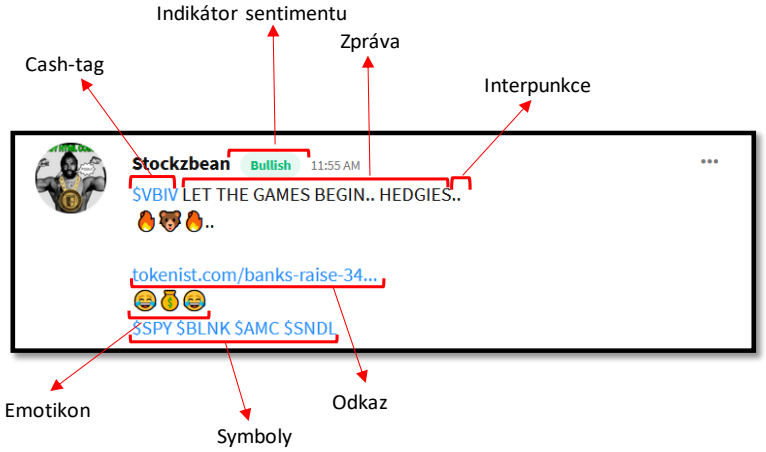
Titulky v získaných datových sadách, které byly zveřejněny z oficiálních webů CNBC, Guardian a Reuters, odrážejí přehled americké ekonomiky a akciového trhu každý den za poslední rok až 2 roky. Datové sady byly získány z databáze Kaggle od Pham (2020). Data získaná z agentury Reuters obsahuje celkem 32 770 článků od konce března 2018 do 19. července 2020. Data získaná z CNBC obsahují přibližně 3 070 zveřejněných článků od konce prosince 2017 do 19. července 2020. Tato datová sada obsahuje přibližně 3 070 zveřejněných článků. Data získaná z Guardian Business obsahuje 21 500 shromážděných článků za sledované období od konce prosince 2017 do 19. července 2020. V tabulce 4.5 jsou zachyceny ukázky online finanční zpráv.

Server	Datum	Nadpis
Reuters	4/29/20	Facebook sees 'signs of stability' in ad spending after coronavirus drop
Guardian	4/29/20	Some John Lewis stores may not reopen after lockdown, admits boss
CNBC	4/29/20	Jim Cramer reacts to AMD, Starbucks, Boeing, GE and Alphabet earnings reports
CNBC	4/29/20	Nucor CEO sticks by dividend, predicts steel price bottom in second quarter
Guardian	4/29/20	Business live Stock markets hit seven-week highs as Fed leaves interest rates
Reuters	4/29/20	U.S. economy faces hard slog back from pandemic, Fed chief says

Tab. 4.5: Ukázka zpravodajské datové sady
Zdroj:vlastní zpracování

StockTwits

Stocktwits je největší sociální síť pro investory a obchodníky s více než pěti miliony členů komunity a miliony měsíčních návštěvníků. Jako rozhodující hlas „sociálního financování" je Stocktwits nejlepším způsobem, jak zjistit, co se právě děje na trzích. Otevřená a upřímná povaha zpráv uživatelů nabízí okamžitý pohled na jejich názory, postoje a nelibosti. Z těchto StockTwits příspěvků lze pomocí specifických nástrojů a technik extrahovat sentiment investorů, ať už na individuální bázi, nebo agregovaný napříč skupinou uživatelů.



Obr. 4.6: Ukázka příspěvku na StockTwits
Zdroj: StockTwits (2021)

K extrakci požadovaných StockTwits příspěvků vztahujících se k vybraným společnostem lze využít funkci, která byla nedávno přidána do služby StockTwits a která se nazývá „cash-tags". Cash-tag je značka používaná pro tweety spojené s trhem. Skládá se z předpony $ a nabídky akcií. V případě StockTwits, uživatelé vytvářejí svá vlastní slova, zkratky, slangová slova, vkládají adresy URL a specifickou terminologii (Singh a Kumari, 2016). Ukázka příspěvku jednoho uživatele na StockTwits

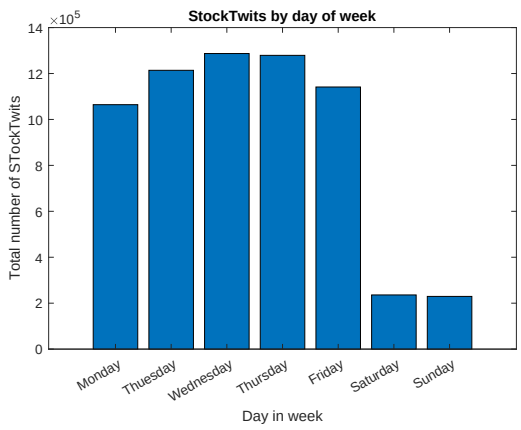
je uvedena na obrázku 4.6.

Vzhledem k povaze StockTwits v reálném čase nabízí ideální zdroj veřejných dat. Získaná textová datová sada obsahuje údaje o akciových titulech, které byly vyselektovány na základě cash-tagu, jedná se například o akciové společnosti jako Apple, Adobe, Tesla, Visa atp. Každý soubor obsahuje symbol akcií, zprávu, datum a čas a ID uživatele pro příslušné zprávy. Rozsáhlá datová sada je volně k dispozici od autorů Jaggi a kol. (2021). Datová sada StockTwits má téměř 6,4 milionu záznamů rozložených do deseti let. Pro disertační práci je tento soubor dat omezen tak, aby zahrnoval pouze příspěvky za sledované období. Ukázkové příspěvky zahrnuté v datové sadě jsou uvedeny níže v tabulce 4.6.

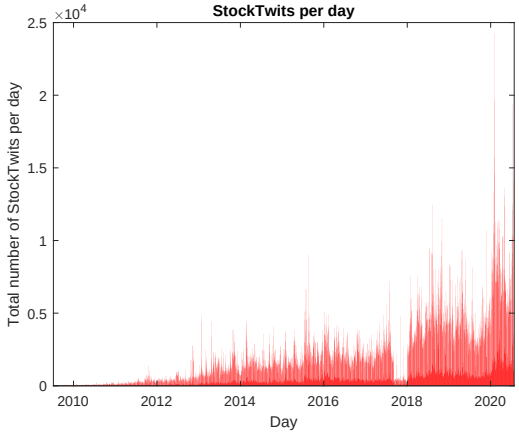
Datum a čas	**Symbol a zpráva**	**ID uživatele**
4/29/20 1:06	$TSLA like place bets plz	1925568
4/29/20 1:10	$SPY damn day vix options pretty cheap	1104480
4/29/20 8:46	$GOOGL extends fact checking us combat coronavirus fake inf.	2762379
4/29/20 13:57	$GOOG wall street jumps hopes potential coronavirus drug via	2257034
4/29/20 14:26	$AAPL another pump dump like yesterday	1595916
4/29/20 16:21	$SPY vix slowly creeping would nice market reacted accordingly	2091724
4/29/20 19:37	$TSLA new article tesla reports q1 earnings bell	1555408
4/29/20 21:05	$TSLA wow congrats	2899881

Tab. 4.6: Ukázka StockTwits datové sady
Zdroj: vlastní zpracování

Obrázek 4.7 představuje rozdělení frekvencí zpráv zveřejňovaných podle dne v týdnu, kdy jsou zprávy zveřejňovány StockTwits. Obchodní dny jsou vyneseny jako modré pruhy a neobchodní dny jsou vyneseny jako červené pruhy. Je patrné, že nejméně příspěvků je na sociálně platformě zveřejňováno právě v neobchodní dny. Pokles je znatelný a v procentním vyjádření představuje o 80% příspěvků méně o víkendu ve srovnání v pracovní dny. V průměru počet příspěvků za obchodní dny převyšuje počet 2 000 za den. Obrázek 4.8 vizualizuje rozdělení počtu StockTwits za celé sledované období. Je zřejmý dlouhodobě rostoucí trend v počtu zveřejňování myšlenek, názorů, postojů ohledně vývoje akciových trhů. Největší počet příspěvků byl zveřejněn 4. února 2020, kdy se za jeden den objevilo neuvěřitelných 24 285 příspěvků. Hned následující den, tj. 5. února 2020 byl zveřejněn druhý největší počet 19 676 příspěvků. Obecně rok 2020 lze považovat za převratný z hlediska počtu sdílení názorů uživatelů platformy. Lze to považovat za logické, neboť s probíhající pandemickou situací, rozšiřující se počítačovou gramotností a uzavření ekonomik, mělo mnoho uživatelů tendenci se podílet o své názory ohledně budoucího vývoje, který byl dosti nejasný. Navíc v témže roce probíhala intenzivní politická kampaň ohledně prezidentských voleb v USA, což mohla také zvýšit intenzitu diskuze.

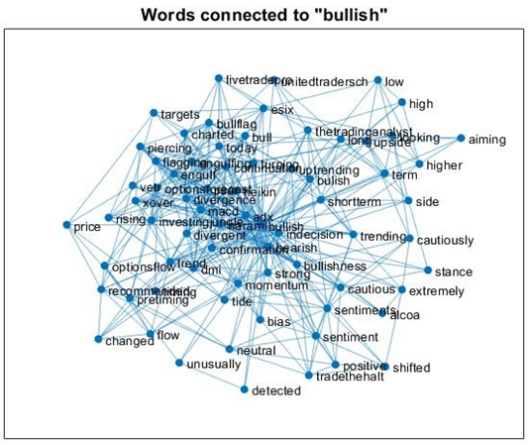


Obr. 4.7: Ukázka průměrného počtu příspěvků na den
Zdroj: vlastní zpracování

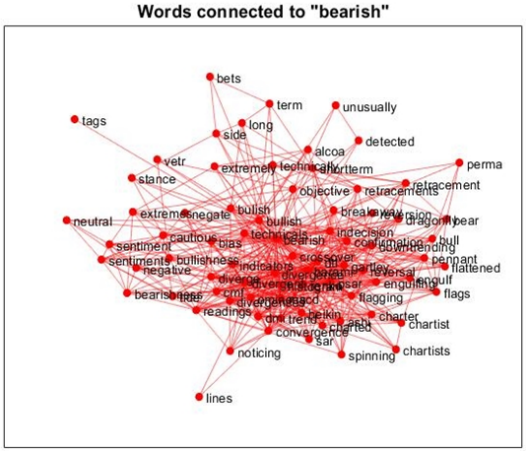


Obr. 4.8: Počet StockTwits za sledované období
Zdroj: vlastní zpracování

Obrázky 4.9, resp. 4.10 vizualizují zjednodušený graf představující propojení mezi klíčovým slovem „Bullish“, resp. „Bearish“, což označuje býčí, resp. medvědí trh. Rovněž jak je uvedeno v detailním popisu příspěvku zveřejňovaném na Stocktwits, jsou tato označení využívána k detekci indikátoru sentimentu, která využívají jednotliví uživatelé této platformy k označení vývoje akciového trhu. Uzly v grafu označují jednotlivá slova, která jsou propojena prostřednictvím hran a váženy kosinovou vzdáleností se slovem klíčovým. K vizualizaci je nastaveno hodnota 50, což znamená, že je vyobrazeno 50 nejbližších slovních spojení s klíčovým slovem.



Obr. 4.9: Vizualizace propojení s klíčovým slovem „bullish“
Zdroj: vlastní zpracování



Obr. 4.10: Vizualizace propojení s klíčovým slovem „bearish“
Zdroj: vlastní zpracování

Z hlediska kosinové vzdálenosti jsou jako nejbližší indikovaná slova propojena s „Bullish“ slova označují „sentiment“, „positive“, „bullishness“, „strong“, „uptren-

ding“, „rising“ apod. Tedy slova indikující silný a rostoucí trh s očekáváním pozitivního výhledu na slibné zisky a profity z realizovaných investic. Naopak silně propojená slova s „Bearish“, tedy medvědím trhem jsou slova jako „sentiment, „negative“, „bearishness“, „downtrending“, „indecision“, „retracements“, „cautious“, „negate“ aj., což jednoznačně označuje klesají tendenci trhu a negativní náladu šířící se trhem, navíc je upozorňováno na nerozhodnost a opatrnost investorů při investování v této fázi trhu.

4.2 Textová analýza

Následuje zpracování výše popsaných textových dat z nestrukturované podoby do podoby strukturované, tak, aby mohla být následně provedena analýza sentimentu, potažmo, aby mohlo být z těchto textových zpráv extrahováno skóre sentimentu investorů.

4.2.1 Předzpracování textových zdrojů

Fáze předzpracování dat je důležitou fází v procesu dolování a zpracování textových dat. Způsob předzpracování textových zdrojů je detailně popsán v podkapitole 1.2.3. Tento krok má za cíl odstranění irelevantních termínů vyskytujících se ve zdrojových dokumentech. Vhodným výběrem technik předzpracování lze zlepšit přesnost klasifikace. K tomuto účelu je zvolena funkce preprocessText, která provede předzpracování textových data tak, aby mohla být použita pro následnou analýzu sentimentu.

Tato funkce v sobě integruje následující dílčí kroky předzpracování:

1. Převedení textových dat na malá písmena pomocí funkce `lower`.
2. Tokenizace text pomocí `tokenizedDocument`.
3. Odstranění interpunkce pomocí `erasePunctuation` a stop slov (slova, která neobsahují žádné relevantní informace například „a“, „of“ a „the“) pomocí `removeStopWords`.
4. Odstranění slova se 2 nebo méně znaky a s 15 a více znaky pomocí funkce `removeShortWords`, resp.`removeLongWords`.
5. Využití part-of-speech tags (POS) pomocí `addPartOfSpeechDetails`. Tato funkce přiřadí každému slovu (tokenu) ve větě POS tag obsahující informaci o slovním druhu, popřípadě další důležité informace.
6. Lemmatizace slova pomocí `normalizeWords`.

Předpřipravený textový obsah je poté transformován do vektorů numerických funkcí pomocí přístupu TF-IDF, který dává charakteristickým výrazům větší váhu.

BoW	Počet	Bigram	Počet	Trigram	Počet
(buy)	495374	(look, like)	68268	(buy, buy, buy)	7733
(stock)	447323	(next, week)	63920	(medium, light, skin)	6676
(market)	441673	(long, term)	35119	(hand, light, skin)	6457
(vix)	410888	(short, term)	32406	(aapl, new, article)	6079
(go)	392278	(tsla, short)	31986	(aapl, look, like)	5627
(short)	391911	(skin, tone)	31855	(fold, hand, light)	5255
(get)	386777	(stock, market)	27174	(backhand, index, point)	5139
(today)	357844	(roll, floor)	25922	(loudly, cry, face)	3908
(day)	356629	(tsla, short)	25447	(goog, new, article)	3621
(good)	314326	(tsla, buy)	22817	(face, smile, eye)	3573
(sell)	312575	(tsla, look)	20563	(mouth, face, money)	3545
(like)	308819	(tsla, bear)	20335	(intraday, support, resistance)	3430
(look)	284198	(floor, laugh)	19013	(support, resistance, spx)	3387
(call)	282310	(elon, musk)	17847	(beam, face, smile)	3200
(week)	259995	(new, article)	17158	(red, heart, selector)	3182
(high)	250213	(tsla, get)	16976	(heavy, check, mark)	3081
(time)	247668	(appl, buy)	15828	(white, heavy, check)	3006
(long)	221714	(money, bag)	14847	(net, change, performer)	2360
(make)	212514	(light, skin)	14556	(percent, net, change)	2360
(new)	175944	(tsla, sell)	14367	(bottom, percent, net)	2359

Tab. 4.7: Počet BoW, Bigramů a Trigramů v korpusu
Zdroj:vlastní zpracování

Pro jemnější analýzu jsou extrahovány funkce BoW a N-gramů z korpusu. Konkrétně každá objevující se sekvence slov o délce „1“, „2“ a „3“ je extrahována z tweetů a vytváří slovník slov a frází. Pro vytvoření modelu Bag-of-Words je využita funkce `bagOfWords`, zatímco N-gramy jsou vygenerování funkcí `bagOfNgrams` a nastavením parametru `NGramLengths` na hodnotu 2 a 3. V tabulce 4.7 je uvedeno top 20 nejčastěji se vyskytujících slov, bigramů a trigramů a jejich konkrétní počet výskytů v celém korpusu. Výpis slov lze jednoduše vygenerovat funkcí `topkwords` a `topkngrams`. Některé jednoduché statistické analýzy ukazují, že korpus vykazuje řadu termínů, které jsou příliš časté a naznačují jisté aspekty související s aktivitami na trhu jako je například „buy“, či umocnění výzvy k nákupní investiční strategii u trigramů „buy, buy, buy“. Zatímco BoW naznačuje jednotlivé aktivity na trhu bez zjevné a jednoznačné souvislosti, N-gramy dávají do kontextu například konkrétní společnost „tesla, short“, „aapl, new, article“ či významnou osobnost „elon musk“. Stejnou analýzu lze provést i pro ostatní společnosti potažmo celý akciový trh. Lze

například vzít v úvahu různé aspekty společnosti Tesla, jako jsou aspekty k nákupu či prodeji, identifikovat sentiment a využít ho k určení nejpozitivnějších a negativních tweetů s ohledem na konkrétní aspekty společnosti a následnému generování nákupních a prodejních signálů.

Vzhledem k rozsáhlé datové textové sadě, výpočetní náročnosti a časovému omezení na vypracování je využito výhradně BoW.

4.3 Analýza sentimentu

Za předpokladu, že jsou textová data transformována z nestrukturované podoby do podoby strukturované, je možné přistoupit k samotné analýze sentimentu, která je popsána v podkapitole 1.3. K tomuto účelu je zvolen hybridní přístup kombinující metody strojového učení a slovníkový přístup založený na negativních a pozitivních slovech.

4.3.1 Klasifikační techniky

Následující část práce aplikuje různé metody strojového učení s učitelem za účelem natrénování klasifikátorů pro analýzu sentimentu pomocí anotovaného seznamu pozitivních a negativních sentimentálních slov a předcvičeného vkládání slov tzv. word embedding. Vložení předem předcvičeních slov hraje v tomto pracovním postupu několik rolí: převádí slova na číselné vektory a tvoří základ pro klasifikátor. Následuje využití klasifikátoru k předpovědi sentimentu jiných slov pomocí jejich vektorové reprezentace a pomocí těchto klasifikací vypočítat sentiment textu. Trénink a používání klasifikátoru sentimentu má čtyři kroky:

1. Načtení předcvičeného vkládání slov.
2. Načtení lexikonů se seznamem pozitivních a negativních slov.
3. Trénování a testování klasifikátor sentimentu pomocí vektorů slov obsahující lexikon pozitivních a negativních slov.
4. Vypočítání průměrného skóre sentimentu slov v textu.

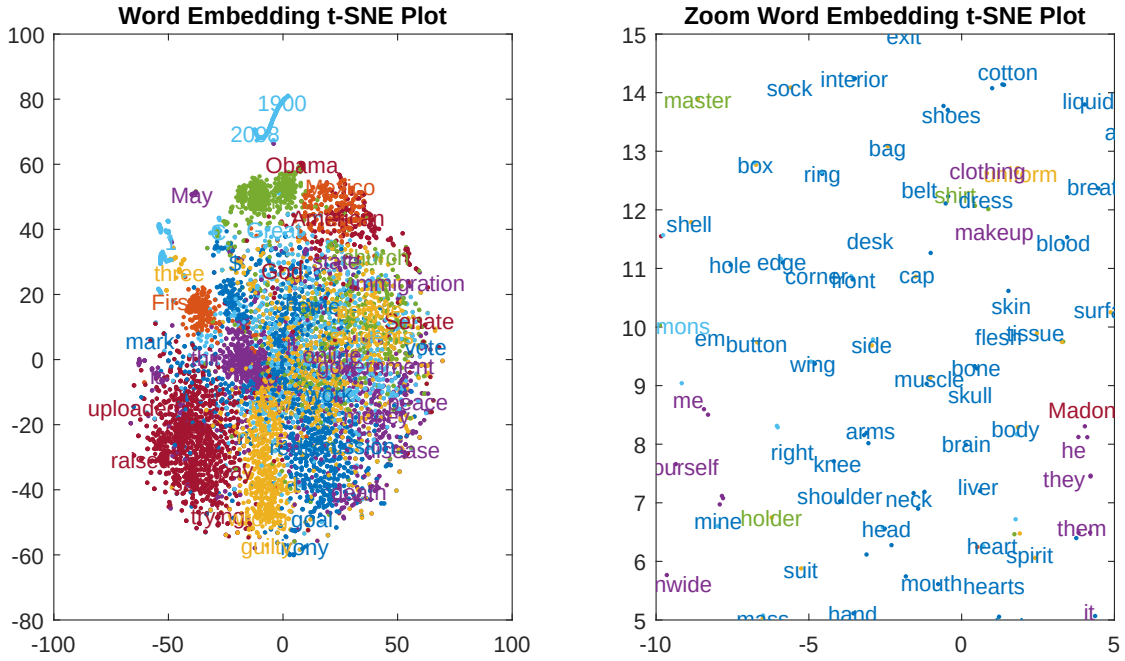
Předcvičené vkládání slov

Vkládání slov mapuje slova ve slovníku na číselné vektory. Tato vložení mohou zachytit sémantické podrobnosti slov, takže podobná slova mají podobné vektory. Rovněž modelují vztahy mezi slovy pomocí vektorové aritmetiky. K načtení předcvičeného vkládání slov je využita funkce `fastTextWordEmbedding`. Tato funkce vyžaduje model Text Analytics Toolbox ™ pro balíček podpory FastText English

16 miliard tokenů pro vkládání slov. FastText je knihovna s otevřeným zdrojovým kódem, která poskytuje efektivní a škálovatelné knihovny pro textovou analýzu.

Následující obrázek 4.11 ukazuje, jak vizualizovat vkládání slov pomocí 2-D vestavěné funkce `t-SNE`. Tato funkce obsahuje algoritmus pro redukci rozměrů, který je vhodný pro vizualizaci vysokodimenzionálních dat. Vkládání slov mapuje slova ve slovníku na skutečné vektory. Vektory se pokoušejí zachytit sémantiku slov, takže podobná slova mají podobné vektory. K převodu slov na vektory slouží funkce `word2vec`.

Pro ukázku je zvoleno prvních 20 000 slov a je vytvořena matice slovních vektorů o délce 300. Následně je provedeno shlukování prostřednictvím funkce `kmean` na 15 shluků, přičemž je využito euklidovské vzdálenosti při stanovení jednotlivých shluků. Kvůli čitelnosti textových dat se ve výchozím nastavení nezobrazuje všechna vstupní slova a místo toho zobrazuje značky bez slovního popisu.



Obr. 4.11: Vizualizace vkládání slov včetně přiblížení
Zdroj: vlastní zpracování

Lexikony

Pro analýzu sentimentu je zvoleno několik druhů lexikonů, neboť volba vhodného lexikonu je nezbytná pro správnou analýzu sentimentu. Jak vyplývá z kritického přezkumu, téměř všechny relevantní studie využívají jeden maximálně dva lexikony pro extrakci sentimentu z textových dat s následnou aplikací na akciovém trhu. Pro disertační práci je zvoleno celkem šest lexikonů. Z hlediska obecných slovníků je vybrán psychologický slovníky Harvard General Inquirer IV-4 (HGIV-4) vyvinut Stone

a kol. (1966) dostupný na `http://www.wjh.harvard.edu/~inquirer/`. Názorový slovníky (Opinion Lexicon) od Hu a Liu (2004) získán z `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon`. VADER lexikon sentimentu od Hutto a Gilbert (2014), který je přímo integrován ve využívaném softwaru MATLAB a AFINN od Nielsen (2011). Vzhledem k následné aplikaci sentimentu se zamřením na finanční oblast jsou zvoleny také zástupy finančních slovníků jako je Loughran-McDonald lexikon od stejnojmenných autorů Loughran a McDonald (2011) dostupný na `https://sraf.nd.edu/textual-analysis/resources/`. A na závěr je zvolen FinanceSentiment Lexikon, který je také plně integrován v MATLABu. Předpokládá se, že poslední dva jmenované lexikony by měli mít vyšší přesnost při aplikaci na textová data z finančních portálů a příspěvků na StockTwits.

Vybrané lexikony jsou načteny pomocí funkce `readLexicon`. Výstupními daty jsou tabulka s proměnnými obsahující slova a štítek obsahující kategorické označení polarity sentimentu jako pozitivní nebo negativní. Pro následné trénování klasifikátor sentimentu, je nezbytné převést slova z lexikonů na vektory slov pomocí předcvičeného vkládání slov. Dále je lexikon rozdělen náhodně na datovou sadu určenou pro testování a pro trénování. Dle literatury například Nti a kol. (2020) či Batra a Daudpota (2018) se doporučuje využít poměru 80% pro trénování a 20% pro testování klasifikátoru strojového učení. K tomuto úkonu je využita funkce `cvpartition`. Prostřednictvím této funkce lze definovat náhodný oddíl na datové sadě. Tento oddíl je použit k definování tréninkových a testovacích sad pro ověření statistického modelu pomocí křížové validace. Po rozdělení na dvě datové sady jsou slova převedeny na vektory slov pomocí funkce `word2vec`.

Klasifikátory sentimentu

Pro trénování a testování je zvoleno několik technik strojového učení. Konkrétně je ke klasifikaci zvolena metoda podpůrných vektorů, rozhodovací stromy, naivní Bayes, k-nejbližších sousedů, neuronové sítě a generalizovaný aditivní model. Tyto metody byly zvoleny na základě kritického přezkumu literatury. Níže aplikované metody strojového učení s učitelem jsou detailně popsány v podkapitole 1.3.3 a je využit Statistics and Machine Learning Toolbox™. Každá ze zvolených metod klasifikace je aplikována na všechny zvolené lexikony pozitivních a negativních slov. Následně jsou všechny klasifikátory komparovány prostřednictvím evaluačních metrik popsaných v podkapitole 1.4.3.

Support Vectore Machine

Nejprve je ke klasifikaci sentimentu zvolena metoda podpůrných vektorů (SVM), který klasifikuje slovní vektory do pozitivních a negativních kategorií. K natrénování

klasifikátoru SVM je využita funkce `fitcsvm`. Po natrénování je model testován a jsou predikovány štíty označující pozitivní a negativní slova v testovací sadě.

Z matice záměn jsou následně spočítány evaluační metriky, které slouží ke komparaci vybraných metod klasifikace. Z tabulky 4.8 je patrné, že nejpřesnější klasifikace sentimentu prostřednictvím SVM je docíleno pro slovník Opinion Lexikon s přesností 96.86%, dále AFINN, resp. Harvard IV-4 Lexikon s přesností 96.55%, resp. 95.18%. Velmi příznivých výsledků ovšem docílili také slovníky FinanceSentiment Lexikon a VADER. Nejhůře byl klasifikován sentiment u speciálního slovníku pro finanční oblast Loughran-McDonald. Důležitým evaluačním ukazatelem je také ztráta klasifikace, což je měřítko kvality zobecnění. Jeho interpretace závisí na ztrátové funkci a vážicím schématu, nicméně obecně lepší klasifikátory přinášejí menší hodnoty ztrát. Nezvykle vysokou hodnot ztráty vykazuje Loughran-McDonald lexikon. Ostatní slovníky vykazují přijatelnou ztrátu okolo 3-8%. Obecně lze poznamenat, že strojové učení SVM vykazuje velice přesné výsledky klasifikace a je schopen se naučit klasifikovat štíty z poskytnutých slovníků pozitivních a negativních slov.

	Precision	**Recall**	**F1**	**Accuracy**	**Loss**
Opinion Lexikon	0.9552	0.9435	0.9493	0.9686	0.0312
Loughran-McDonald Lexikon	0.2206	0.2376	0.2288	0.2395	0.7599
Harvard IV-4 Lexikon	0.9496	0.9531	0.9514	0.9518	0.0482
FinanceSentiment Lexikon	0.9846	0.9410	0.9623	0.9313	0.0696
VADER	0.9114	0.9061	0.9087	0.9220	0.0782
AFINN	0.9372	0.9728	0.9547	0.9655	0.0325

Tab. 4.8: Evaluace klasifikace SVM
Zdroj: vlastní zpracování

Rozhodovací stromy

Rozhodovací stromy umožňují také klasifikaci polarity sentimentu. Funkce `fitctree` vrací přizpůsobený binární klasifikační rozhodovací strom na základě vstupních proměnných obsažených ve slovnících. Výstupy testovací sady klasifikace sentimentu prostřednictvím rozhodovacích stromů jsou graficky znázorněny prostřednictvím matice záměn.

Rozhodovací stromy poskytují nižší přesnost, resp. vyšší chybovost klasifikace než SVM. Konkrétní numerické údaje evaluačních metrik lze vyčíst z tabulky 4.9. Nejvyšší přesnost dosáhl speciální finanční slovník FinanceSentiment Lexikon s hodnotu accuracy 83.52%, zatímco na druhé straně druhý analyzovaný finanční slovník Loughran-McDonald Lexikon vykazuje přesnost klasifikace 20.19%. Ostatní obecné slovníky jsou schopny rozpoznat polaritu na 73-80%. Zatímco velikost ztráty se pohybuje v rozmezí 16-26% u zkoumaných slovníků.

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexikon	0.6517	0.6616	0.6566	0.7899	0.2076
Loughran-McDonald Lexikon	0.2017	0.2092	0.2053	0.2019	0.7981
Harvard IV-4 Lexikon	0.7410	0.7437	0.7423	0.7446	0.2553
FinanceSentiment Lexikon	0.9105	0.9049	0.9077	0.8352	0.1662
VADER	0.6956	0.6837	0.6896	0.7332	0.2676
AFINN	0.6859	0.7661	0.7238	0.7972	0.1951

Tab. 4.9: Evaluace klasifikace rozhodovacích stromů
Zdroj: vlastní zpracování

Naivní Bayes

Pro natrénování modelu naivního Bayese je využita funkce `fitcnb`. Po vyškolení modelu na trénovací sadě je model testován. Obdobně jako u předchozích dvou klasifikátorů, je i pro naivního Bayese vytvořena tabulka evaluačních metrik viz tabulka 4.10.

	Precision	Recall	F1	Accuracy	Loss
Opinion Lexikon	0.8731	0.8864	0.8797	0.9264	0.0726
Loughran-McDonald Lexikon	0.4475	0.5059	0.4749	0.4995	0.5046
Harvard IV-4 Lexikon	0.8885	0.9114	0.8998	0.9018	0.0981
FinanceSentiment Lexikon	0.8673	0.9590	0.9109	0.8489	0.1514
VADER	0.8709	0.8294	0.8496	0.8686	0.1313
AFINN	0.8848	0.9037	0.8942	0.9189	0.0788

Tab. 4.10: Evaluace klasifikace naivního Bayese
Zdroj: vlastní zpracování

Precision lze interpretovat jako počet pozitivních slov, které byly opravdu klasifikovány jako pozitivní. V tomto případě se ukazatel dosahuje více než 85%. Recall udává, kolik slov bylo podchyceno klasifikátorem správně. Metrika F1 kombinuje oba výše popsané ukazatele, čím je dosaženo hodnoty blíže k 1, tím je dosaženo lepší klasifikace polarity. Přesnost klasifikace je velice obdobná jako v případě SVM a pohybuje se okolo 90%. V případě Loughran-McDonald je klasifikace správná pouze z 50%. To znamená, že v polovině případů nebyl naivní Bayese schopen správně klasifikovat polaritu slov obsažených v tomto slovníků.

K-nejbližších sousedů

Další metodou klasifikování, která se také výzkumných studií využívá, je metoda k-nejbližších sousedů. Nicméně lze na úvod konstatovat, že tato metoda není příliš rozšířená na rozdíl od předcházejících, které jsou zvoleny ke klasifikaci v této

disertační práci. Pro provedení klasice slouží funkce `fitcknn`, která slouží ke kategorizacím dat s více třídami, popřípadě `fitcecoc`, který je vhodnější pro binární klasifikaci.

Z matice záměn jsou dále vypočítány ukazatelé, které jsou zaneseny do tabulky 4.11. Z tabulky lze vyčíst, že klasifikátor k-nejbližších sousedů vykazuje lepší výkonnost než rozhodovací stromy, na druhou stranu tento klasifikátor není schopen překonat SVM a naivního Byese. Avšak je nezbytné konstatovat, že za těmito metodami příliš nezaostává, ba naopak poskytuje přijatelný výkon z hlediska vysoké přesnosti a nízké chybovosti.

	Precision	**Recall**	**F1**	**Accuracy**	**Loss**
Opinion Lexikon	0.8507	0.8529	0.8518	0.9087	0.0902
Loughran-McDonald Lexikon	0.2647	0.2525	0.2585	0.2234	0.7778
Harvard IV-4 Lexikon	0.8849	0.8723	0.8786	0.8786	0.1215
FinanceSentiment Lexikon	0.9383	0.9354	0.9368	0.8874	0.1136
VADER	0.8478	0.8255	0.8365	0.8588	0.1414
AFINN	0.8743	0.9126	0.893	0.9189	0.078

Tab. 4.11: Evaluace klasifikace k-NN
Zdroj: vlastní zpracování

Neuronové sítě

Dále je využita funkce `fitcnet` k trénování dopředné, plně propojené neuronové sítě pro klasifikaci. První plně připojená vrstva neuronové sítě má připojení ze síťového vstupu a každá následující vrstva má připojení z předchozí vrstvy. Každá plně připojená vrstva vynásobí vstup váhovou maticí a poté přidá zkreslený vektor. Po každé plně spojené vrstvě následuje aktivační funkce. Konečná plně spojená vrstva a následná aktivační funkce softmax produkují výstup sítě, jmenovitě klasifikační skóre a předpovězené označení polarity sentimentu.

	Precision	**Recall**	**F1**	**Accuracy**	**Loss**
Opinion Lexikon	0.9129	0.9107	0.9118	0.9456	0.0538
Loughran-McDonald Lexikon	0.3151	0.3247	0.3198	0.3147	0.6853
Harvard IV-4 Lexikon	0.9137	0.9104	0.912	0.9125	0.0875
FinanceSentiment Lexikon	0.9506	0.9506	0.9506	0.9121	0.0886
VADER	0.8690	0.8558	0.8623	0.8818	0.1185
AFINN	0.9005	0.9297	0.9149	0.9351	0.0625

Tab. 4.12: Evaluace klasifikace neuronových sítí
Zdroj: vlastní zpracování

Výstupy klasifikátoru založeného na neuronových sítí jsou v tabulce 4.12 srovnávající výkon prostřednictvím evaluačních metrik. Z vygenerovaných výstupů vy-

plývá, že neuronové sítě jsou svým výkonem porovnatelné s výkonem SVM a naivního Bayese. Přesnost klasifikace se pohybuje okolo 90% s velice nízkou chybovostí okolo 8%.

Generalizovaný aditivní model

Následně je trénován generalizovaný aditivní model pro binární klasifikaci sentimentu. Zásadní pro správnou klasifikaci je nalezení optimálních parametrů s ohledem na křížové ověření. To vyžaduje určení maximálního počtu rozhodovacích rozdělení (nebo větvových uzlů) pro každý strom prediktoru. Nalezení počtu stromů, což je v podstatě ekvivalentní počtu iterací podporujících gradient pro lineární výrazy pro prediktory. Následně je model trénován na základě optimálních parametrů. K tomu je využita funkce `fitcgam`.

	Precision	**Recall**	**F1**	**Accuracy**	**Loss**
Opinion Lexikon	0.8806	0.9100	0.8951	0.9363	0.0626
Loughran-McDonald Lexikon	0.2311	0.2355	0.2333	0.2234	0.7768
Harvard IV-4 Lexikon	0.9101	0.8908	0.9004	0.9000	0.1001
FinanceSentiment Lexikon	0.9753	0.9349	0.9547	0.9176	0.0835
VADER	0.8671	0.8654	0.8662	0.8859	0.1145
AFINN	0.9005	0.9101	0.9053	0.9270	0.0712

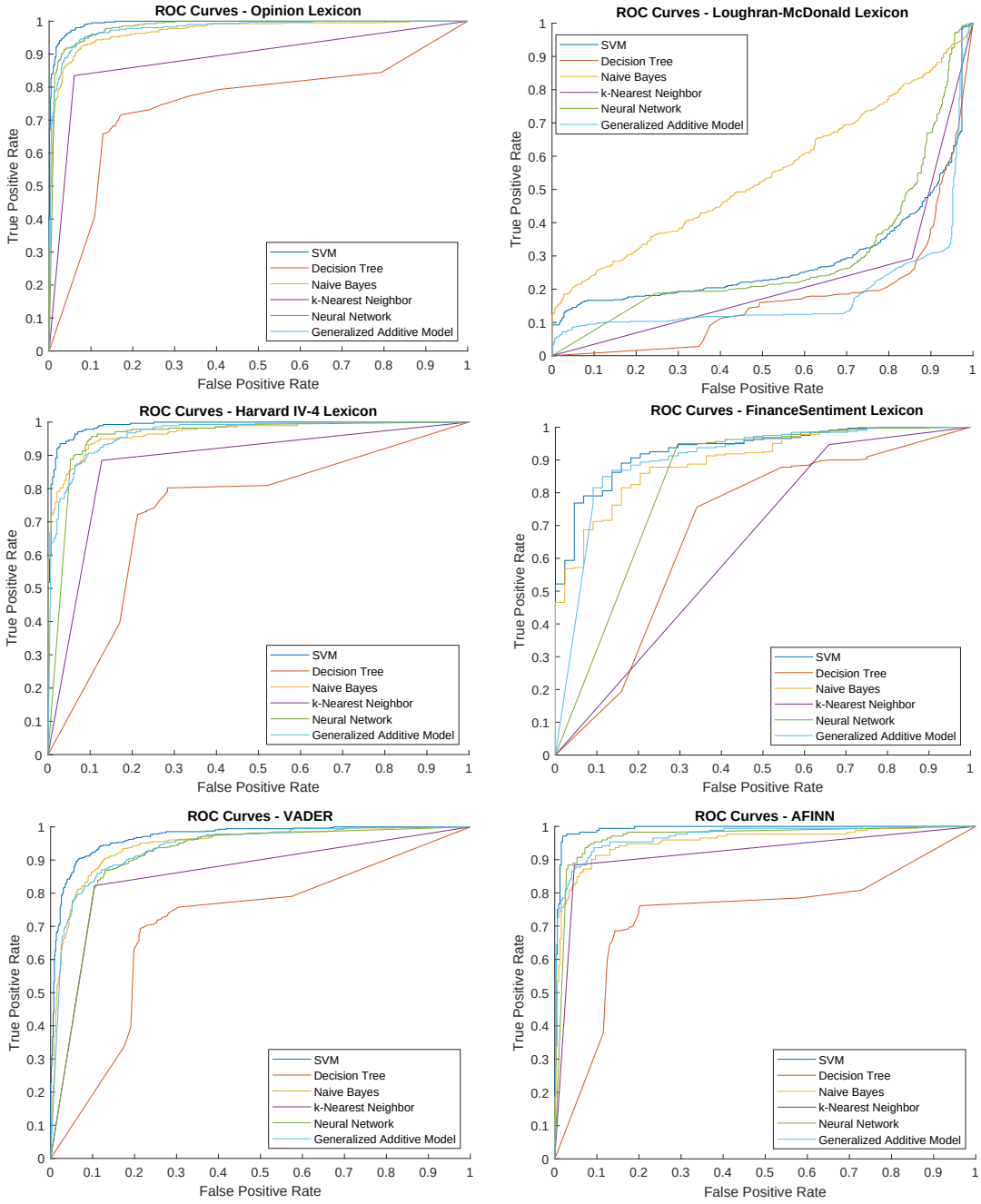
Tab. 4.13: Evaluace klasifikace gen, aditivního modelu
Zdroj: vlastní zpracování

Z tabulky 4.13 lze vyčíst evaluační metriky generalizovaného aditivního modelu. Nejlepších výsledků dosáhl z hlediska přesnosti klasifikace Opinion Lexikon a AFINN s hodnotou ukazatele accuracy 93.63%, resp. 92.7%. Nejhůře byl modelem klasifikován opět finanční slovník Loughran-McDonald. Obecně tato metoda vykazuje velmi uspokojivé výsledky a překová rozhodovací stromy i k-nejbližších sousedů. Tato metoda není příliš využívána v dřívějších studiích, nicméně na základě rozboru lze konstatovat, že je schopen konkurovat prosazovaným metodám jako je SVM či neuronové sítě.

Komparace klasifikátorů

Následně jsou všechny aplikované klasifikátory společně komparovány prostřednictvím ROC křivky na obrázku 4.12. ROC křivka *(angl. Receiver Operating Charakteristic Curve)* slouží k hodnocení a také grafickému znázornění výkonu klasifikátorů při klasifikaci do dvou tříd, tedy při hodnocení slov obsažených v analyzovaných slovnících na pozitivní a negativní. V ideálním případě by ROC křivka s perfektní diskriminační schopností kopírovala levý horní rok ROC prostoru. Již na základě

předchozích výpočtů je jako nejslabší klasifikátory jevily rozhodovací stromy a k-nejbližších sousedů, což dále potvrzeno i ROC křivkou, která se u těchto klasifikátorů značně odchyluje od levého horního rohu ROC prostoru u všech analyzovaných lexikonů. Lze tedy konstatovat, že rozhodovací stromy a k-nejbližších sousedů má nižší diskriminační schopnost než ostatní analyzované klasifikátory.



Obr. 4.12: ROC křivky jednotlivých klasifikátorů
Zdroj: vlastní zpracování

Plocha pod křivkou ROC je označována jako AUC, což lze také chápat jako měřítko kvality testu. AUC je nejběžnější kvantitativní index popisující ROC křivku.

Výstupy tohoto ukazatele jsou v numerické podobě zaneseny v tabulce 4.14.

	SVM	**Rozh. stromy**	**naivní Bayes**	**k-NN**	**Neur. sítě**	**Gen. aditivní model**
Opinion Lexikon	0.9938	0.7463	0.9711	0.8874	0.9819	0.9790
Loughran-McDonald Lexikon	0.2843	0.1669	0.5439	0.2187	0.2888	0.1805
Harvard IV-4 Lexikon	0.9899	0.7350	0.9689	0.8786	0.9560	0.9667
FinanceSentiment Lexikon	0.9316	0.6902	0.8979	0.6439	0.8312	0.9031
VADER	0.9698	0.7046	0.9448	0.8590	0.9080	0.9394
AFINN	0.9923	0.7355	0.9547	0.9185	0.9674	0.9707

Tab. 4.14: AUC všech klasifikátorů
Zdroj:vlastní zpracování

Přesnost testu klasifikace hodnocení jako výborná by měla, přesáhnou hodnotu 90%, což je případ například SVM, naivního Bayese, neuronových sítí a generalizovaného aditivního modelu. Dobrých výsledků přesnosti testu v rozmezí 70% až 80% dosáhly rozhodovací stromy a k-nejbližších sousedů. Tyto výsledky jsou patrné u všech zvolených lexikonů vyjma již několikrát zmiňovaného lexikonu od Loughran-McDonald, který je vytvořen speciálně pro finanční doménu. Obecně lze konstatovat, že výsledky naznačují nadřazenost naivní Bayes vůči všem ostatním klasifikátorům. Konkrétně dosahuje průměrně o 25% vyšší přesnosti než rozhodovací stromy, v průměru o 15% k-nejbližších sousedů a o 6% neuronové sítě a generalizovaný aditivní model. Téměř totožných výsledků dosáhl tento model s SVM, kde se kvalita binární klasifikace lišila v průměru pouze o 2% ve prospěch naivního Bayese. Ovšem v případě, že není uvažován nejhůře natrénovaný slovník od Loughran-McDonald. Je nejvyšší přesnost klasifikace přisuzována metodě SVM, která je v průměru o 2% až 3% vyšší než u naivního Bayese a generalizovaného aditivního modelu. Lze konstatovat, že z hlediska binární klasifikace zřetelně dominují a povzbudivých výsledků dosahují SVM, naivní Bayes i generalizovaný aditivní model a jedná se o vhodné modely pro klasifikaci sentimentu.

Vypočítání skóre sentimentu

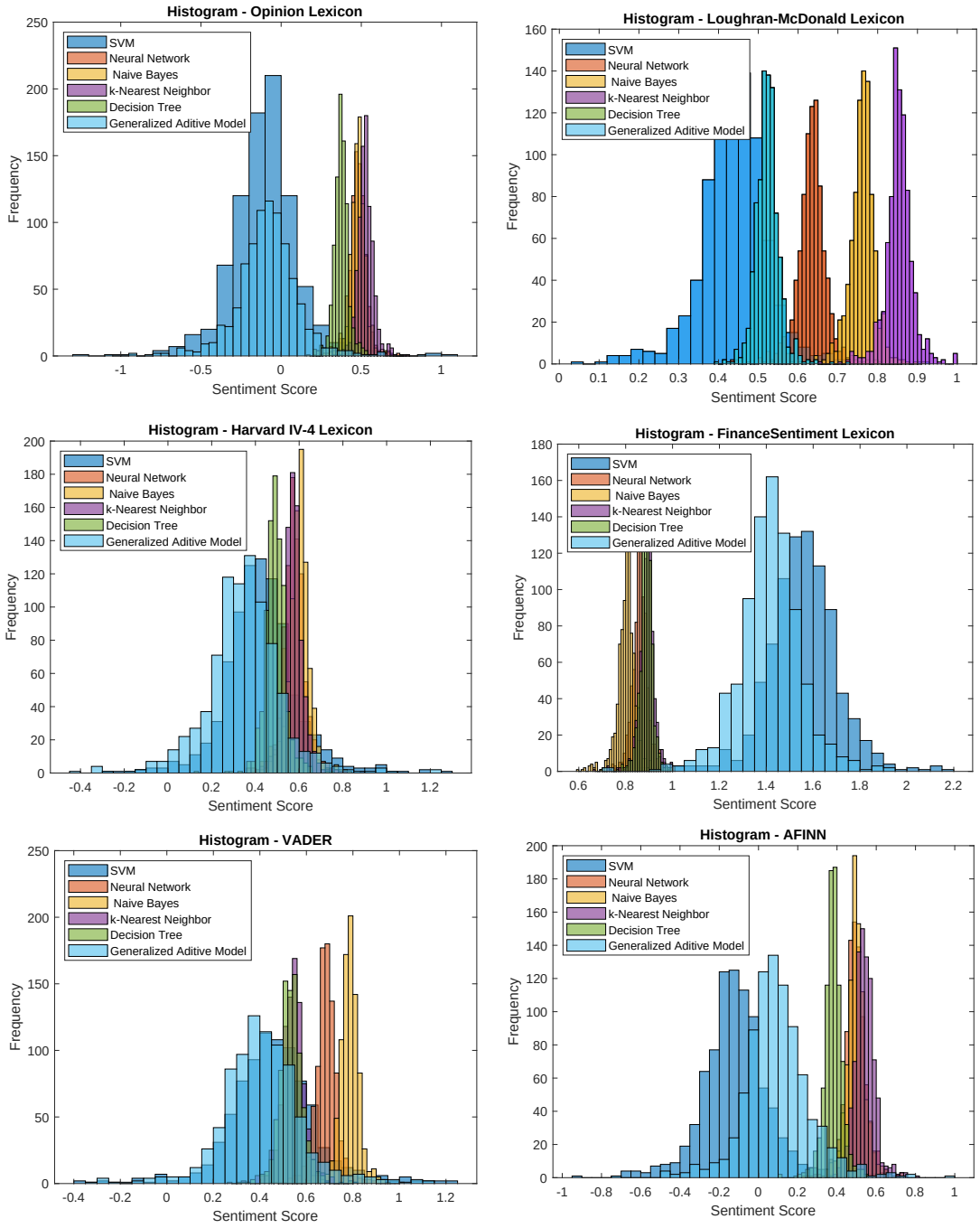
Po natrénování a testování klasifikátorů je pro zvolenou textovou sadu predikováno skóre sentimentu každého slova v textu a následně je stanoveno průměrné skóre sentimentu za každý sledovaný den. K tomu je nezbytné využít předzpracovanou textovou sadu. Tento proces zabere několik hodin v závislosti na rozsahu textu. Poté jsou z textu odebrána slova, která se neobjevují v knihovně předcvičeného vkládání slov prostřednictvím funkce `emb`. Pro vizuální posouzení správnosti klasifikace sentimentu na novém textu, je klasifikován sentiment u slov, která se vyskytují v textu,

ale nikoliv v tréninkových datech. K tomuto účelu lze využít funkci `wordcloud`, která vyobrazí slovní oblaka, což poslouží k manuálnímu posouzení, zda se klasifikátor chová dle očekávání. V případě správnosti klasifikace na novém textu je posléze vypočítáno průměrné skóre sentimentu.

U každé věty textového dokumentu jsou slova převedena na slovní vektory, na kterých je předpovídáno skóre sentimentu prostřednictvím funkce `predict`. Je nezbytné poznamenat, že predikování skóre sentimentu u rozsáhlých datových sad trván několik hodin. Následně je transformováno skóre pomocí funkce transformace a poté je vypočítáno střední skóre sentimentu. Tento proces je proveden na zvoleném textovém dokumentu pro všechny dříve analyzované slovníky pozitivních a negativních slov a také pro všechny dříve analyzované klasifikátory. Pro snadné vizuální posouzení získaného sentimentu, je zvolen histogram, ve kterém je zaznamenána distribuce skóre sentimentu. Na obrázku 4.13 lze vidět jednotlivé histogramy vygenerované prostřednictvím funkce`histogram`, přičemž každý jednotlivý obrázek představuje slovník a jednotlivé histogramy uvnitř obrázku udávají distribuci sentimentu získanou prostřednictvím šesti klasifikátorů, které jsou barevně rozlišeny.

Z obrázku jsou patrná značné rozdíly v rozložení distribuci skóre sentimentu nejen u jednotlivých slovníků, ale také u jednotlivých klasifikátorů. Na první pohled je zřejmá téměř shodná klasifikace a ohodnocení sentimentu prostřednictvím SVM a generalizovaného aditivního modelu, při použití všech slovníků. Oba klasifikátory jsou znázorněny modrou barvou a je patrné překrývání se hustoty distribuce sentimentu. U slovníku Harvard IV-4 je navíc patrné velice obdobné odhodnocení textového dokumentu i prostřednictvím neuronových sítí, naivního Bayese, k-nejbližších sousedů i rozhodovacích stromů, o čemž svědčí shodná distribuce skóre sentimentu, neboť jednotlivé funkce hustoty se téměř překrývají. Ostatní slovníky vykazují značné vychýlení distribuce skóre sentimentu. Toto vychýlení je dominantní u Opinion Lexikonu, AFINN a finančních slovníků Loughran-McDonald a FinanceSentiment.

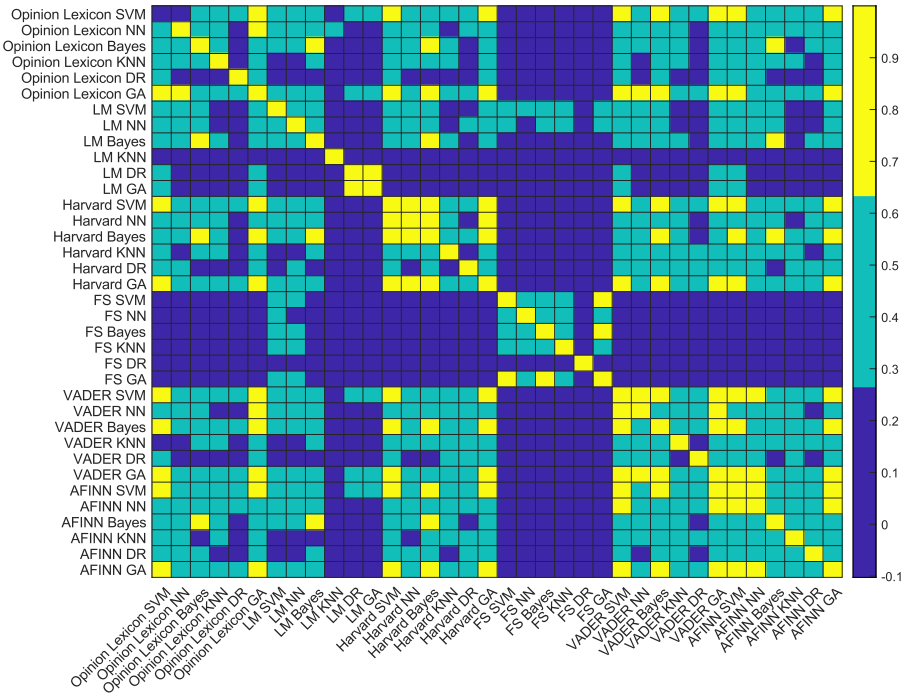
Lze konstatovat, že u většiny slovníků pozitivních a negativních slov, které jsou klasifikovány prostřednictvím neuronových sítí, naivního Bayese, k-nejbližších sousedů i rozhodovacích stromů dochází k nadhodnocování skóre sentimentu oproti klasifikátorům SVM a generalizovaného aditivního modelu. Výjimku tvoří pouze slovník FinanceSentiment, u kterého naopak došlo k podhodnocení skóre sentimentu oproti SVM a generalizovaného aditivního modelu. Z toho důvodu je nezbytné upravit rozdělení na pozitivní a negativní sentimentu a nespoléhat se na obecné pravidlo, že skóre větší než 0 svědčí o pozitivním sentimentu, zatímco skóre menší než 0 poukazuje na negativní sentiment. Z distribuce rozdělení skóre sentimentu je tato skutečnost patrná a je nezbytné s tímto poznatkem dále pracovat, aby mohl být sentiment správně uplatněn pro následnou predikci akciových trhů. To znamená, že popisky



Obr. 4.13: Distribuce skóre sentimentu
Zdroj: vlastní zpracování

tříd je nezbytné modifikovat a nastavit hranice pro jednotlivé třídy ručně. Lze například rozhodnout, že u slovníku VADER a klasifikátoru SVM skóre sentimentu pod 0.4 bude považováno za negativní, skóre sentimentu nad 0.4 bude považováno za pozitivní. Obdobně by se postupovalo u dalších slovníků. Cílem je nastavit hranice, které budou relevantní u každého slovníku a každého klasifikátoru, aby se dosáhlo nejvyšší přesnosti klasifikace.

Při kritickém přezkumu literatury vyplynulo, že výzkumníci se spoléhají povětšinou výhradně na jeden slovník, popřípadě využívají ke klasifikování polarity sentimentu výhradně jednu metodu či techniku. Zejména tato část výzkumu práce s textovým dokumentem a ohodnocení sentimentu sehrává klíčovou roli. Nesprávně zvolený slovník, který detailně nepostihuje zkoumanou problematiku textu, může poskytnout nesprávné výstupy se kterými je následně pracováno a vyvozovány nepravdivé závěry.



Obr. 4.14: Korelační matice sentimentu
Zdroj: vlastní zpracování

Dále je zkonstruována korelační matice sentimentu na obrázku 4.14. V této matici jsou souhrnně zaznamenána Pearsonova korelace mezi denním rozlišením časových řad sentimentu pro každý slovník a klasifikátor sentimentu. Velmi slabou korelaci (0 až 0.2) vykazuje slovník FinanceSentiment a Loughran-McDonald v porovnání se všemi ostatními slovníky bez ohledu na zvolnou techniku klasifikátoru. Většina slovníku vykazuje střední korelaci v rozmezí 0.3 až 0.6. Velmi silná korelace je patrná u slovníků Harvard IV-4, které jsou klasifikovány prostřednictvím SVM, neuronových sítí a naivního Bayese. Obdobně silnou závislost vykazuje také VADER a AFINN. Celkově obrázek ukazuje, že slovníky obsahují různorodé pozitivní a negativní slova a tím pádem měří či odráží něco jiného.

Kromě použití výše uvedených ukazatelů přesnosti k vyhodnocení účinnosti jednotlivých slovníků a modelů je nutné prověřit, zda jsou rozdíly ve výkonnosti statisticky významné. V této studii je, dle doporučení Bifeta a kol. (2015), využit

Wilcoxonův test.

H_0: Finanční a obecný slovník generují stejně přesné skóre sentimentu
H_1: Finanční slovník generuje přesnější skóre sentimentu než obecný slovník

Je zvolena hladina významnosti 0,05. Poté je rozhodnutí, zda nulová hypotéza je zamítnuta, či nikoli, založeno na výsledné p-hodnotě. Standardně platí, pokud je p-hodnota větší než 0,05, nulová hypotéza není zamítnuta. V opačném případě, pokud je p-hodnota menší než 0,05, bude nulová hypotéza zamítnuta na úrovni spolehlivosti 95%.

Výsledky Wilcoxonova testu na seznamu testovacích dat pro každou kombinaci finančních a obecných slovníků jsou shrnuty v tabulce 4.15. Z tabulky je patrné, že u tréninkových dat jsou p-hodnoty velké a vyšší než prahová hodnota 0,05. Nelze tedy zamítnout nulovou hypotézu ve prospěch alternativní, která tvrdí, že speciální slovníky pro finanční oblast generují přesnější skóre sentimentu investorů z textových dat. Není tedy signifikantně prokázána lepší schopnost finančních slovníků k označení textových dat na pozitivní a negativní slova. Celkově lze říci, že zde předložená data poskytují důkaz, že finanční slovníky negenerují přesnější hodnoty než obecné slovníky. Na základě výsledků Wilcoxova testu lze tvrdit, že FinanceSentiment lexikon je přesnější než slovník Loughran-McDonald.

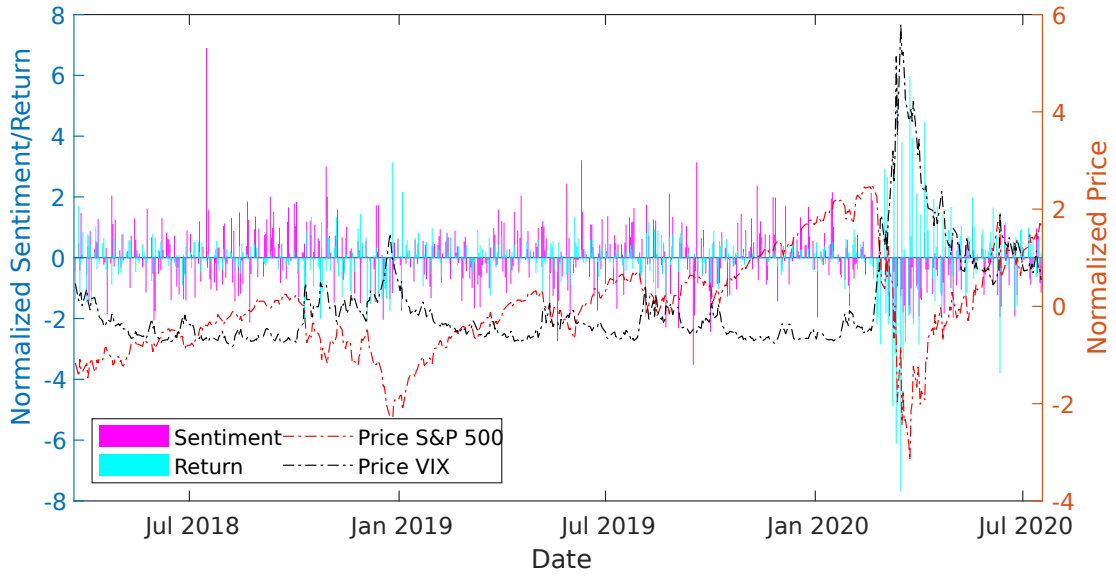
Typ slovníku	**Hodnota alfy**	**P-hodnota**	**Nulová hypotéza**
Finance Sentiment→Opinion Lexicon	0.0500	0.6582	Nezamítnuta
Finance Sentiment→AFINN	0.0500	0.7101	Nezamítnuta
Finance Sentiment→Loughran-McDonald	0.0500	0.0002	Zamítnuta
Finance Sentiment→VADER	0.0500	0.3394	Nezamítnuta
Finance Sentiment→Harvard	0.0500	0.7133	Nezamítnuta
Loughran-McDonald→Opinion Lexicon	0.0500	1.0000	Nezamítnuta
Loughran-McDonald→AFINN	0.0500	1.0000	Nezamítnuta
Loughran-McDonald→Finance Sentiment	0.0500	0.9998	Nezamítnuta
Loughran-McDonald→VADER	0.0500	0.9992	Nezamítnuta
Loughran-McDonald→Harvard	0.0500	1.0000	Nezamítnuta

Tab. 4.15: Wilcoxonův test jednotlivých slovníků
Zdroj:vlastní zpracování

Při následné analýze je, pro snížení výpočetní a časové náročnosti, využit pouze jeden slovník, resp. jedno skóre sentimentu. Je zvolen nejpřesnější slovník, a sice Opinion lexikon, jenž při klasifikaci prostřednictvím SVM dosáhl téměř 100% přesnosti. Nicméně veškeré následující operace lze provést i s ostatními slovníky, což je, vzhledem k časovému omezení této praci, nesplnitelné, avšak je tím alespoň poskytnut prostor pro následný výzkum v této oblasti.

4.4 Analýza vlivu sentimentu na akciové trhy

Po kalkulaci sentimentu investorů následuje analýza jeho vlivu na akciový index S&P500, který je přijat jako proxy amerického akciového trhu. Na obrázku 4.15 je do jednoho vykreslen sestrojený indikátor sentimentu na základě vytěžených online zpráv, výnos akciového indexu a dále vývoj samotné ceny akciového indexu a indexu VIX. Levá osa vykazuje normalizovaný výnos a sentimentu, pravá osa zobrazuje normalizovanou cenu akciového indexu a indexu VIX.



Obr. 4.15: Vliv sentimentu na akciové trhy
Zdroj: vlastní zpracování

Kromě toho se index sentimentu investorů mění velmi rychle a má různé trendy v různých časových obdobích. Zejména v pandemické situaci je patrný výrazně a po delší dobu patrný záporný sentiment, což je také projevilo strmým pádem amerického akciového trhu, neboť na trhu převládala vysoce pesimistická a nejistá nálada, která vyvolala vlnu prodejů.

Jak je dobře známo, obchodní strategie, ať již se sentimentem nebo bez sentimentu investorů přináší zisk, pouze pokud by mohla poskytnout určitou předvídatelnost budoucích změn cen akcií, vzhledem k velké variabilitě dat na akciovém trhu. Pro následnou predikci je tudíž potřeba posoudit existenci kointegrace, kauzality a lead or lag efektu sentimentu, aby mohl být následně využit k predikci akciového trhu.

4.4.1 Účinek kauzality a kointegrace mezi sentimentem a akciovým trhem

Nejprve je proveden Engle-Grangerův test kointegrace s finanční časovou řadou a sentimentem. Jinými slovy, pokud mají dvě časové řady tendenci udržovat mezi sebou konstantní rozdíl v dlouhém období, to znamená, že jsou tyto řady kointegrované. Lze říct, že mezi těmito indexy existuje dlouhodobý rovnovážný vztah. Tento test je proveden za pomoci funkce `gcitest`, která je integrována v Econometrics Toolbox™. Je testována nulová hypotéza proti alternativní hypotéze na hladině významnosti 5%:

H_0: Mezi sentimentem investorů a výnosem akciového trhu neexistuje kointegrace
H_1: Mezi sentimentem investorů a výnosem akciového trhu existuje kointegrace

Jak je patrné z tabulky 4.16, je zamítnuta nulová hypotéz o neexistenci kointegračních vztahů mezi akciovým indexem a sentimentem investorů. Výstupem testu kointegrace je p-hodnota, která nabývá hodnot v intervalu [0, 1]. Ve statistickém testování hypotéz je p-hodnota měřítkem toho, kolik důkazů existuje proti nulové hypotéze. Pokud je p-hodnota nižší než zvolená hladina významnosti 5%, nulová hypotéza je zamítnuta a výsledek je statisticky významný. Na druhou stranu velká p-hodnota představuje slabý důkaz proti nulové hypotéze; nulovou hypotézu tedy nelze zamítnout. V tomto případě vykazuje p-hodnota údaj hluboko pod zvolenou hladinou významnosti 5%. Rovněž t-statistika vykazuje vyšší hodnotu než kritická hodnota zanesená v tabulce. Znamená to, že existují kointegrační vztahy mezi akciovým trhem a sentimentem investorů. Nicméně to samo o sobě neznamená, že sentiment je příčinou pohybu akciových trhů.

	t-statistika	**Kritická hodnota**	**F-statistika**	**p-hodnota**
SP 500	-32.65	-3.35	19.16	1.40E-05
VIX	-26.92	-3.35	13.08	3.20E-04

Tab. 4.16: Testování Engle-Garngerovy kointegrace
Zdroj:vlastní zpracování

Ačkoliv sentiment investorů a výnosy akciového trhu společně kointegrují, je otázkou, zda je jedna časová řada užitečná při předpovídání druhé. Následuje tak diskuze o příčinném vztahu mezi výnosem akcií a sentimentem investora. Je aplikován statistický test ke zjištění, zda sentimenty vyjádřené v textových zprávách obsahují prediktivní informace o budoucích hodnotách vývoje akciového trhu. Za tímto účelem je proveden Grangerův test kauzality, což je statistický test hypotéz pro zjištění, zda je jedna časová řada účinná pro předpovídání jiné časové řady (Granger, 1969). Pokud se o časové řadě X říká, že je Grangerově příčinná časové

řadě Y, pak informace v minulých hodnotách X pomáhají předpovídat hodnoty Y lépe než pouze informace v minulých hodnotách Y. Proto budou mít zpožděné hodnoty X statisticky významnou korelaci s Y.

Tento test je proveden za pomoci funkce `gctest`, která je integrována opět v Econometrics Toolbox™. Tabulka 4.17 testuje Grangerovu kauzalitu směrem od sentimentu k výnosům, resp. směrem od výnosů k sentimentu investorů. V panelu A je testována nulová hypotézu proti alternativní hypotéze na hladině významnosti 5%, přičemž je uvažováno zpoždění 1-5 dnů:

H_0: Sentiment investorů nezpůsobuje Grangerovu kauzalitu výnosů
H_1: Sentiment investorů způsobuje Grangerovu kauzalitu výnosů

Tabulka 4.17 udává hodnotu F-statistiky. K přijetí či zamítnutí výše uvedené hypotézy je nezbytně nutné určit také kritickou hodnotu z F-distribuce, ta činí pro panel A a B hodnotu 3.86. Dále platí, pokud je hodnota F-testu vyšší než kritická hodnota, nulovou hypotéza je zamítnuta, jinými slovy sentiment vykazuje Grangerovu příčinu na výnos.

	Panel A: S → V					**Panel B: V → S**				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
S&P 500	3.48	4.31	3.87	1.45	1.45	0.02	0.40	0.40	0.40	0.40
VIX	15.26	5.53	5.53	5.53	5.53	0.42	1.29	1.29	1.29	1.29

Tab. 4.17: Testování Grangerovy kauzality
Zdroj:vlastní zpracování

Uvedený výsledek naznačuje, že výkonnost akciových trhů je ovlivněna sentimentem investorů. U akciového indexu je tato skutečnost patrná až do zpoždění 3 dny, zatímco u indexu VIX způsobuje sentiment kauzální vztah po všechna analyzovaná časová zpoždění. Je tak zřejmé, že akciové trhy vstřebávají názory a postoje investorů v relativně krátkém časovém horizontu, zatímco index strachu je ovlivňován po delší časové období než samotný akciový index představující barometr americké ekonomiky. Tento výsledek ospravedlňuje předpoklad, že sentiment investorů bude s největší pravděpodobností hybnou silou nadměrných výnosů akcií. Toto zjištění je podobné Ni a kol. (2015). To znamená, že sentiment investorů má významný vliv na výnosy akciového trhu. Tento výsledek je v souladu se závěry Bakera a Wurglera (2006).

V panelu B je testována nulová hypotézu proti alternativní hypotéze na hladině významnosti 5%, přičemž je opět uvažováno zpoždění 1-5 dnů:

H_0: Výnos nezpůsobuje Grangerovu kauzalitu sentimentu investorů
H_1: Výnos způsobuje Grangerovu kauzalitu sentimentu investorů

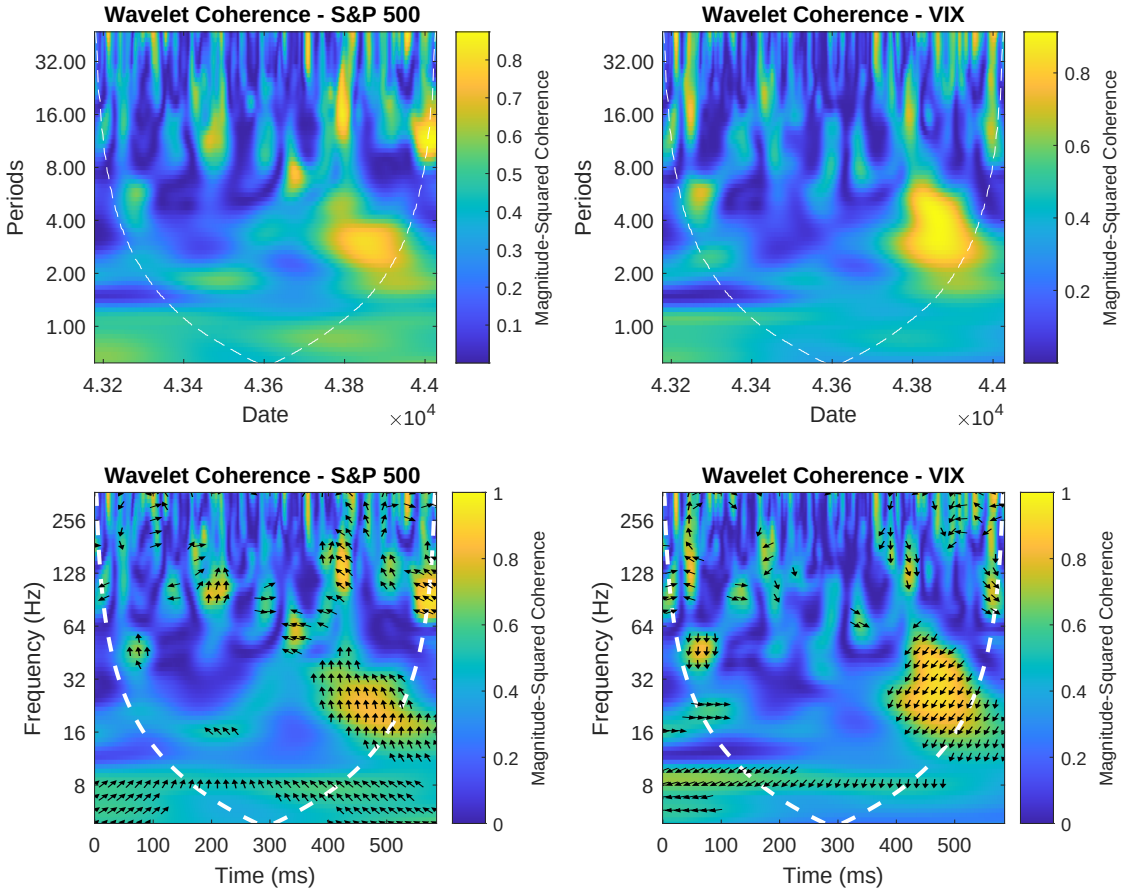
Dopady výnosů akciových trhů na sentiment investorů jsou méně jasné. Neexistuje žádný kauzální vztah od výnosu směrem k sentimentu. Výsledky odhalují významnou jednosměrnou kauzalitu plynoucí ze změny sentimentu investorů k výnosům akciového trhu. Nulová hypotéza žádné nelineární kauzality probíhající od výnosů k sentimentu není většinově odmítnuta. Výsledky testů příčinné kauzality tedy naznačují, že vliv krátkodobého sentimentu investora na výnosy akciového trhu jsou komplexní. Lze tedy konstatovat, že investiční obchodní aktivity jsou spojeny s náladou investorů. Získání hlubších znalostí o povaze kauzality může poskytnout cenný pohled na účastníky trhu. Pokud například příčinná souvislost naznačuje, že došlo ke zvýšení předvídatelnosti s ohledem na průměrný výnos. Pak mohou tyto informace přispět k realizaci nadměrných výnosů, zatímco zvýšená předvídatelnost průměrné směrodatné odchylky by následně mohla zvýšit schopnost odhadnout riziko akciového trhu.

4.4.2 Sentiment investora a akciový trh: lead or lag efekt

V této části disertační práce je k posouzení dynamických vztahů mezi sentimentem investorů a výnosem amerického akciového trhu aplikována vlnková analýza k identifikaci vzájemných pohybů mezi finanční časovou řadou. Kromě toho vlnková analýza vysvětluje, jak časové řady souvisejí na různých frekvencích a jak tyto interakce postupují v čase a napříč časovými měřítky, jak popisuje Janková (2020). Zejména je využita vlnková koherence a metoda fázového rozdílu, ke zkoumání vzájemné pohyby a vztahů mezi prodlevou sentimentem investorů a akciovým indexem nebo indexem VIX v časové a frekvenční oblasti současně. Vlnková koherence tak umožňuje rozlišovat mezi krátkodobou a dlouhodobou dynamikou společného pohybu. Na obrázku 4.16 je znázorněna odhadovaná vlnková koherence pro sentiment investorů a akciový trh na levé straně a sentiment a index strachu VIX na pravé straně. Pro demonstrativní účely je zvolen sentiment vygenerovaný ze slovníku Opinion lexikon klasifikován SVM.

K vyhodnocení vlnkové koherence je využita funkce `wcoherence`, která je dostupná ve speciální knihovně Wavelet Toolbox™. Horizontální osa představuje časové intervaly, zatímco vertikální osa představuje měřítko frekvence. Hodnota koeficientu koherence je mezi 0 a 1. Oblast se žlutou (modrou) barvou vykazuje silný (slabý) souběžný pohyb při nízkých a vysokých frekvencích. Bílá plná čára izoluje statisticky významnou oblast na 5% hladině významnosti. Vzhledem k hraničnímu efektu je oblast mimo ohraničený kužel méně informativní, a proto je pozornost zaměřena primárně na oblast uvnitř bílého ohraničení. Šipky ve všech vlnkových koherenčních mapách ukazují jak korelační, tak i relační průběh mezi dvěma časovými řadami. Šipky směřující doprava → znamenají, že proměnné jsou ve fázi (cyklický

účinek na sebe navzájem). ↗ = sentimenty investorů vedou či určují pohyb trhu; ↘ = nálady investorů zaostávají za trhem; šipky směřující doleva ← znamenají, že proměnné jsou mimo fázi (anticyklický účinek). ↖ = sentiment investorů zaostávají; ↙ sentimenty investorů vedou trh, jak popisuje Dajčman (2013) a Ye a kol. (2020).



Obr. 4.16: Vlnková koherence
Zdroj: vlastní zpracování

Lze vypozorovat, že v kontextu sentimentu a trhu šipky směřují nahoru ve střední frekvenci, zatímco v nízkých a vysokých frekvencí je patrný směr nahoru a mírně doleva. To označuje pozitivní, tedy cyklický vztah mezi sentimentem investorů a akciovým trhem. Mírně směřování doleva označuje stav, že sentimentu určuje pohyb akciového trhu. Jiný směr šipek je patrný pro kontext sentimentu investorů a indexu strachu. Zde je zřetelný směr šipek dolů s mírným sklonem na levou stranu. To znamená, že sentiment investorů a index VIX vykazuje negativní vztah a jsou ve vzájemném anticyklickém účinku. Opět mírný levostranný sklon určuje, že sentiment vede index VIX. Se zvyšující se nejistotou na trhu se mezi investory rozšíří negativní postoj k tržním podmínkám, což se projeví zvýšením indexu VIX a poklesu na trzích. Sentiment investorů úzce souvisí s výkonem akciového trhu.

Dále lze identifikovat vztah mezi prodlevou sentimentu investora a akciového trhu, potažmo indexu VIX pomocí analýzy fázových rozdílů. Zejména šipky ve významné a žlutě podbarvené oblasti naznačuje, že sentiment investorů ovlivňuje akciový trh ve střednědobém horizontu (od 16 do 32 měsíců) na počátku roku 2020. Jinými slovy sentiment je velmi silný v době medvědího trhu a působí na trh nikoliv pouze v krátkém, ale i střednědobém horizontu. Obdobné závěry lze vyvodit i u indexu VIX. Od 32 do 64 měsíců však „studená“ zóna s relativně malým koeficient soudržnosti nenaznačuje žádný významný společný pohyb. Kromě toho před rokem 2020 lze vidět, že šipky pouze ve spodní části, tedy ve vysokofrekvenční oblasti (méně než 8 měsíců), což naznačuje, že sentiment investorů může být v krátkodobém horizontu nižší v době býčího trhu.

Během pandemické situaci, která se na trzích projevila prudkým pádem, vlnková koherence jasně zaznamenává velký ostrov žluté barvy lokalizovaný na nízkých a vysokých frekvencích, což znamená silný vztah mezi sentimentem investorů a výnosy akciovým trhem, který je reprezentován akciovým indexem S&P 500. Tento výsledek lze vysvětlit skutečností, že během krizové situaci jsou investoři pesimističtí a jejich nadměrný pesimismus a nejistota se projevuje výprodejem držených aktiv a následným negativním vlivem na trhu.

Prostřednictvím vlnkové koherence je prokázáno, že sentiment extrahovaných z textových dokumentů, které jsou zveřejňovány na internetu, vykazuje tzv. leading efekt neboli sentiment investorů vede a určuje směr pohybu trhu. Důležité bylo prozkoumat také kontext s indexem strachu VIX, neboť se lze domnívat, že tento index již v sobě integruje sentiment investorů ještě před zveřejněním zásadních informací a následnou ji odpovídající reakcí investorů na trzích. V tom případě by bylo nesmyslné prostřednictvím sentimentu předvídat vývoj akciového trhu, neboť by k tomu byl relevantnější index VIX. Signifikantně je tak prokázáno, že sentiment ovlivňuje akciového trhu a projevuje se také do indexu strachu VIX, nikoliv naopak.

4.5 Tvorba expertního modelu

Akciový trh je dynamický systém charakteristický vysoce nelineárním, dynamickým a komplikovaným chování, což znesnadňují investorům rychlé rozhodnutí o správných investicích. Je nezbytné vyvinout inteligentní expertní systém, který investorům poslouží jako podpora pro investiční rozhodování na akciových trzích, čím se snížení rizika při rozhodování investorů a napomůže jim maximalizovat zisky. Tato práce je zaměřena na fuzzy logiku, konkrétní na tvorbu modelu type-2 fuzzy logiky, která je schopna zahrnout nejistotu, nelinearitu a šum, které se ve finančních časových řadách vyskytují. Na základě sekundárního výzkumu vyplynulo, že pro predikci vývoje akciového trhu nebyl prozatím tento přístup využit zejména v kontextu integrace

vytvořené sentimentu investorů extrahovaném z textových zpráv. Lze se domnívat, že type-2 fuzzy logika je více intuitivnější a schopna lépe modelovat lidský úsudek se zahrnutím dodatečné nejistoty, která může vyplývat, ať již z nejednoznačných a vágních textových dat potažmo z nich extrahovaného sentimentu, tak z nejistého vývoje akciového trhu, jak uvádí Janková a Dostál (2021a). K vytvoření fuzzy modelu je využit Fuzzy Logic Toolbox™.

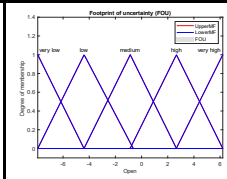
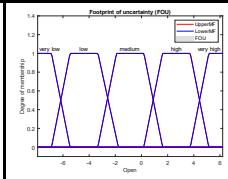
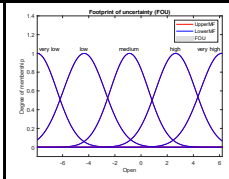
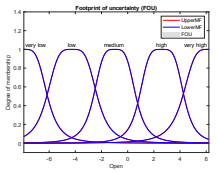
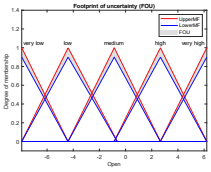
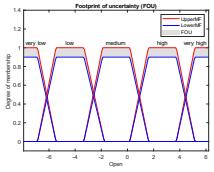
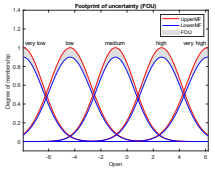
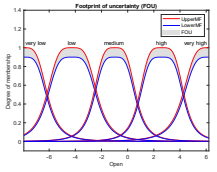
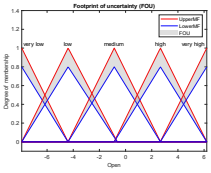
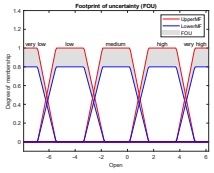
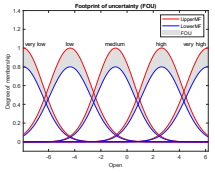
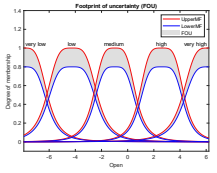
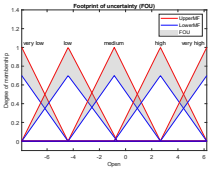
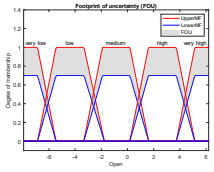
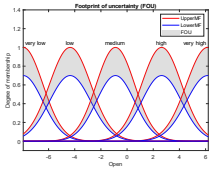
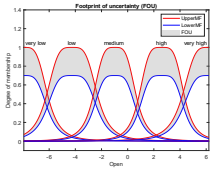
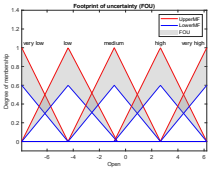
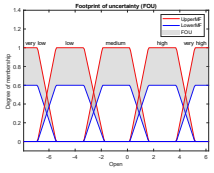
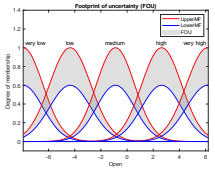
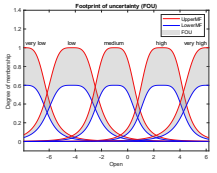
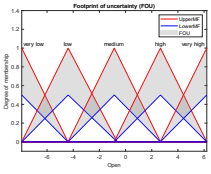
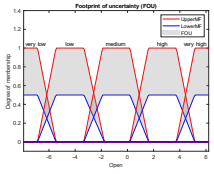
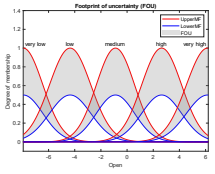
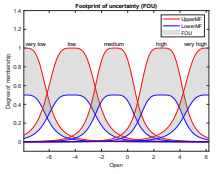
4.5.1 Návrh expertního modelu

Je tvořen type-2 fuzzy model typu Sugeno. K tomu je využita funkce `sugfistype2`. Popřípadě je alternativně možné vytvořit type-1 fuzzy model a prostřednictvím funkce `convertToType2` konvertovat na type-2. Fuzzy model typu Sugeno je zvolen z několika zásadních důvodů. Jednak má větší flexibilitu v návrhu systému oproti druhému hojně rozšířenému typu Mamdani. Navíc typ Sugeno disponuje větší výpočetní efektivitou a je možné ho integrovat s neuronovými sítěmi prostřednictvím ANFIS modelu.

Při tvorbě expertního modelu je kladen výrazný důraz na funkce členství, které jsou ve většině výzkumných prací zanedbávány a nedostatečně prozkoumány, což může způsobit značné problémy s funkčností modelu. Vyjádření funkcí členství závisí jak na subjektu (jak hluboká je zkušenost výzkumníka), tak na kontextu (kde je řešení problému). Z toho důvodu jsou postupně využity funkce členství s rozdílnou mírou nejistoty, která se postupně zvyšuje. Rovněž není využit pouze jeden typ funkce členství, jak je obvyklé v drtivé většině studií, ale postupně jsou vytvořeny funkce členství typu trojúhelníkového, lichoběžníkového, Gaussovského a zvonkového. Neupínání se výhradně na jednu funkci členství je záměrné, neboť kromě znalostní báze jsou funkce členství zásadní pro správnou funkčnost modelu, resp. pro poskytování přesných výstupů. Nesprávné popsání vstupních dat prostřednictvím funkcí členství může poskytovat nesprávné výstupy, resp. investiční signály.

Funkce členství jsou vytvořeny prostřednictvím funkce `fismftype2`. Při kompilaci fuzzy množin se používají parametry faktoru změny měřítka dolní funkce členství zadaného jako kladný skalár menší nebo rovný 1. Pomocí `LowerScale` se definuje maximální hodnotu dolní funkce členství. Dále je nastavena zpožděná hodnota prostřednictvím `LowerLag` pro dolní funkci členství. Toto zpoždění definuje bod, ve kterém se hodnota dolní funkce členství začíná zvyšovat od nuly na základě hodnoty horní funkce členství. V rámci tohoto parametru lze zadat zpožděné hodnoty mezi 0 a 1 včetně. Například hodnota zpoždění 0.1 znamená, že dolní funkce členství se stane pozitivní, když má funkce horního členství hodnotu členství 0.1. V závislosti na hodnotě `LowerLag` může být skutečná maximální hodnota dolní funkce členství menší než `LowerScale`. Jestliže je hodnota zpoždění nula, dolní funkce členství se

začne zvyšovat ve stejném bodě jako horní funkce členství. Postupně je vytvořeno celkem šest modelů, které se liší mírou nejistoty obsažené ve fuzzy funkcích členství popisujících vstupní proměnné.

Model	Trojúhelníkové funkce členství	Lichoběžníkové funkce členství	Gaussovské funkce členství	Zvonkové funkce členství
Klasický fuzzy model				
Type-2 fuzzy model s 10 % nejistotou				
Type-2 fuzzy model s 20 % nejistotou				
Type-2 fuzzy model s 30 % nejistotou				
Type-2 fuzzy model s 40 % nejistotou				
Type-2 fuzzy model s 50 % nejistotou				

Obr. 4.17: Funkce členství fuzzy modelů
Zdroj: vlastní zpracování

Jednotlivé typy funkce členství jsou volání prostřednictvím příkazů. Trojúhelníková funkce členství je jednou z nejzákladnější funkce a volána příkazem `trimf`. Lichoběžníkového funkce je sestavena pomocí `trapmf`. Gaussova funkce členství je vytvořena prostřednictvím příkazu `gaussmf`. Zvonková funkce je vyvolána příkazem `gbellmf`. Na obrázku 4.17 jsou vykresleny jednotlivé funkce členství prostřednictvím příkazu `plotmf`, který se liší typem a nejistotou pro první vstup fuzzy modelu.

Analogicky jsou vytvořeny funkce členství pro další vstupy do fuzzy modely. Je patrné, jak se postupně s přidáváním nejistoty rozšiřuje vzdálenost mezi dolní a horní funkcí členství. Každá vstupní proměnná je reprezentována funkcí příslušnosti, která se skládá z fuzzy množin obsahující celkem pět lingvistických fuzzy hodnot či atributů: VL – velmi nízká, L – nízká, M – střední, H – vysoká a VH – velmi vysoká.

V návaznosti na rozličné funkce členství je vytvořeno několik reprezentativních fuzzy modelů. Celkem je tedy vytvořeno a testováno pět type-2 fuzzy modelů pro každý typ členské funkce, které se liší právě úrovní a mírou obsažené nejistoty. K vytvořením type-2 fuzzy modelům jsou dále také vytvořeny jejich jednodušší protějšky a to type-1 fuzzy modely.

- Model 1: Klasický fuzzy model, resp. IT2FLS s 0% nejistotou
- Model 2: IT2FLS s 10% nejistotou
- Model 3: IT2FLS s 20% nejistotou
- Model 4: IT2FLS s 30% nejistotou
- Model 5: IT2FLS s 40% nejistotou
- Model 6: IT2FLS s 50% nejistotou

Báze znalostí představuje pravidla ve tvaru „jestliže – potom“ vyjadřuje odborné znalosti o zkoumaném vztahu. Tato pravidla představují znalostní základnu, která popisuje chování systému. Antecedent zahrnuje všechny skutečné kombinace jazykových hodnot vstupních proměnných. Výsledkem je vyhodnocení všech kombinací, to znamená přiřazení lingvistických hodnot výstupním proměnným. Při tvorbě expertního modelu v této práci je ke stanovení báze znalostí využita umělé inteligence, konkrétně jsou pravidla generována prostřednictvím ANFIS, jehož podstata je popsána v podkapitole 1.4.2. K tomuto kroku je přistoupeno účelově, neboť stanovení pravidel prostřednictvím autorky práce by nemusela dostatečně pokrýt řešenou problematiku, navíc nejsou k dispozici experti, kteří by znalostní bázi stanovili, v neposlední řadě hraje také roli hledisko časové. Navíc vytvořením znalostní báze prostřednictvím umělých neuronových sítí umožní expertní model jednoduše revidovat a podrobit dalšímu zkoumání jinými výzkumníky. Celkem je vytvořeno 625 pravidel, na základě, kterých jsou vytvořené modely vyhodnoceny.

Další modely s vyšší neurčitostí nejsou považovány za relevantní pro tuto studii, neboť nadměrné zahrnutí nejistoty podstatně sníží použitelnost modelu. Vytvořený model je vyhodnocen prostřednictvím funkce `evalfis`. Jedná se o defuzzifikační fázi modelu na základě, kterého je vygenerováno závěrečné hodnocení, která slouží jako podpora pro rozhodování pro investory, zda investovat či nikoliv.

4.5.2 Aplikace expertního modelu

Je demonstrováno potenciální praktické uplatnění vytvořeného expertního modelu. Většina výzkumů zabývajících se předpovědí akciových trhů založených na online datech má převážně teoretickou povahu a při zvažování návratnosti investic nebere v úvahu reálná omezení.

I když některé z těchto faktorů jsou zde zohledněny, není navržena úplná investiční strategii, která by mohla být aplikována na akciovém trhu tak, jak je. Obchody nejsou prováděny za reálných tržních podmínek. Expertní model je založen na předpokladech a zjednodušeních, která jsou co nejpraktičtější. Na reálných obchodních platformách a v reálných obchodních prostředích však existuje mnoho faktorů, které mohou ovlivnit model obchodování, který teoreticky funguje perfektně.

Pro podrobnější prozkoumání prediktivní sílu sentimentu investorů, je vytvořeno několik modelů fuzzy logiky, aby se prozkoumaly různé problémy, například zda mohou indikátory denního sentimentu investora předpovídat budoucí výnosy a zda denní investor indikátory sentimentu mohou zlepšit výkonnost předpovědi ohledně budoucích výnosů. S ohledem na výzkum Li a kol. (2020) je pozornost zaměřena na předpověď budoucích výnosů nikoliv na předpověď uzavírací ceny nebo otevírací ceny, neboť jejich využití neposkytlo v jejich práci uspokojivé výsledky. Predikčními objekty této práce jsou směry výnosů; tj. směrem nahoru, když je $R_t > 0$ a je doporučováno koupit nebo směrem dolů, když $R_t < 0$ a je doporučováno prodat. V této studii je testováno skóre sentimentu investorů, který v předcházející části poskytl nejpřesnější výsledky klasifikace. Dále je zkoumán, zda model fuzzy model s různou úrovní neurčitosti obsaženou ve funkcích členství může zlepšit predikční výkon akciového trhu se zahrnutím skóre sentimentu investora. Výsledky tedy mohou změnit závěr o tom, zda má sentiment investorů určitou moc předpovídat budoucí pohyb akciového trhu. Predikční výkon každého trénovaného modelu na základě tréninkového vzorku se hodnotí na základě přesnosti předpovědi směru změn cen trhu (nahoru nebo dolů) v testovaném vzorku z hlediska accuracy, precission, recall a F – skóre.

V základním fuzzy modelu vstupní časová řada obsahuje historická data transakcí tržního indexu, jako jsou uzavírací ceny, otevírací ceny a objem obchodování. Tento základní vstup je zapsán jako $X_(t)$, přičemž je využita historická finanční časová řada uzavíracích cen, otevíracích cen a objemu obchodování akciového indexu za den. Je zahrnut indikátory sentimentu investorů vygenerovaný z online, který byl zpracován v podkapitole 4.3.

$$X_{t-1} = (close_{t-n,\cdots,1}, open_{t-n,\cdots,1}, volume_{t-n,\cdots,1}, sentiment_{t-n,\cdots,1}) \tag{4.2}$$

Pro snížení výpočetní náročnosti a zrychlení výpočtů je spuštěn Parallel Server. Pro plynulou predikci jsou zkontrolována chybějící finanční data a vyplněná lineární kombinací dat předcházejících. Opět je nezbytné rozdělit datovou sadu na trénovací a testovací. Trénovací období je 22.3.2018 až 31.3.2020. Pro překonání problém s učením modelů, jsou data normalizována. Sestupy gradientu proto mohou konvergovat rychleji a s vyšší přesností, neboť normalizovaná data vyloučí zkreslení modelu směrem k prvku s větší číselnou hodnotou.

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7230	0.2770	0.7741	0.7232	0.7478
Type-2 fuzzy model s 10 % nejistotou	0.721	0.2790	0.7741	0.7207	0.7464
Type-2 fuzzy model s 20 % nejistotou	0.7191	0.2809	0.7926	0.7110	0.7496
Type-2 fuzzy model s 30 % nejistotou	0.721	0.2790	0.8185	0.7038	0.7568
Type-2 fuzzy model s 40 % nejistotou	0.7092	0.2908	0.8407	0.6837	0.7542
Type-2 fuzzy model s 50 % nejistotou	0.6739	0.3261	0.8741	0.6413	0.7398
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7092	0.2908	0.7222	0.7276	0.7249
Type-2 fuzzy model s 10 % nejistotou	0.7112	0.2888	0.7259	0.7286	0.7273
Type-2 fuzzy model s 20 % nejistotou	0.7092	0.2908	0.7296	0.7243	0.7269
Type-2 fuzzy model s 30 % nejistotou	0.7112	0.2888	0.7407	0.722	0.7313
Type-2 fuzzy model s 40 % nejistotou	0.721	0.2790	0.7778	0.7192	0.7473
Type-2 fuzzy model s 50 % nejistotou	0.7112	0.2888	0.8148	0.6940	0.7496
Gaussovské funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7269	0.2731	0.7667	0.7314	0.7486
Type-2 fuzzy model s 10 % nejistotou	0.7289	0.2711	0.7667	0.7340	0.7500
Type-2 fuzzy model s 20 % nejistotou	0.7073	0.2927	0.7111	0.7300	0.7205
Type-2 fuzzy model s 30 % nejistotou	0.6739	0.3261	0.6444	0.7131	0.677
Type-2 fuzzy model s 40 % nejistotou	0.6306	0.3694	0.5519	0.6898	0.6132
Type-2 fuzzy model s 50 % nejistotou	0.6012	0.3988	0.5185	0.6573	0.5797
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7230	0.2770	0.7889	0.7172	0.7513
Type-2 fuzzy model s 10 % nejistotou	0.7250	0.2750	0.8222	0.707	0.7603
Type-2 fuzzy model s 20 % nejistotou	0.6110	0.3890	0.9185	0.5849	0.7147
Type-2 fuzzy model s 30 % nejistotou	0.5639	0.4361	0.9074	0.5543	0.6882
Type-2 fuzzy model s 40 % nejistotou	0.5403	0.4597	0.8852	0.5407	0.6713
Type-2 fuzzy model s 50 % nejistotou	0.5305	0.4695	0.8852	0.5347	0.6667

Tab. 4.18: Trénování fuzzy modelů bez sentimentu
Zdroj: vlastní zpracování

Následuje hodnocení vytvořených expertních modelů pro predikci směru nominálních výnosů. Odpovídající výsledky jsou podrobně uvedeny v tabulkách 4.18 až 4.21. První tabulka odráží výkonnost predikce bez zahrnutí sentimentu. Průměrná

přesnost všech analyzovaných metod bez sentimentu dosahuje hodnoty 71.8%. Z hlediska jednotlivých tvarů fuzzy funkcí členství využitích pro predikci, vykazuje fuzzy model s trojúhelníkovou funkcí členství přesnost z hlediska accuracy průměrnou hodnotu 71.12%, model s trapezoidní funkcí členství průměrnou přesnost 71.21%, model s Gaussovou funkcí členství průměrnou přesnost 67.8% a nakonec model se zvonovou funkcí průměrnou přesnost 61.6%. Lze si povšimnout, že u fuzzy modelů bez zahrnutí sentimentu vykazují nevyšší accuracy modely type-1 fuzzy modely, tedy model neobsahující dodatečnou nejistotu a type-2 fuzzy model se zahrnutím dodatečné 10% nejistoty. U ostatních modelů s rostoucím stupněm neurčitosti dochází také k poklesu přesnosti predikce. Jednoznačně je tento trend patrný u zvonkových funkcí členství. Absolutně nejvyšší přesnost vykazuje type-2 fuzzy model s 10% a Gaussovskými funkcemi členství s hodnotou 72.89%. Paradoxně u trapezoidních funkcí vykazuje nejlepší přesnost model s vysokou neurčitostí, konkrétně s nepřesností až 40%. Nicméně takový model lze považovat za nevhodný, neboť takový model již postrádá vypovídajícím schopnost.

Nicméně accuracy je definována jako jednoduše počet správně kategorizovaných případů dělený celkovým počtem případů. Může být užitečná, ale nebere v úvahu jemnosti nerovnováhy tříd. Obecně accuracy má tu výhodu, že je velmi snadno interpretovatelná, ale nevýhodou je, že není robustní, když jsou data nerovnoměrně rozložena nebo pokud jsou s konkrétním typem chyby spojeny vyšší náklady. Z toho důvodu jsou v tabulkách také vypočítání ukazatelé Precision, Recall a především F-skóre. Skóre F1 je ve skutečně vyšší (zhruba o 2-4%) než accuracy, což říká, že jednotlivé třídy vykazují dobrou citlivost a jsou rovnoměrně rozloženy, neboť není patrná výrazná asymetrie u Precision a Recall.

V tabulce 4.19 jsou uvedeny metriky přesnosti predikce vývoje akciového trhu za využití zpožděného sentimentu investorů o jeden den. Průměrná accuracy všech analyzovaných modelů s denním zpožděním sentimentu investorů dosahuje hodnotu 71.6%, kdežto průměrná hodnota F-skóre je 75.9%. Z hlediska jednotlivých typů fuzzy funkcí členství využitích k predikci, vykazují shodnou nejvyšší průměrnou přesnost type-2 fuzzy modely s 10% nejistotou a Gaussovými i zvonkovými funkcemi členství 80.35%. Za nimi následuje klasický type-1 fuzzy model s hodnotou accuracy 79.96%. Zejména u Gaussových a zvonkových funkcí členství je zjevný strmý propad přesnosti s narůstající nejistotou, jenž zkresluje průměrnou přesnost predikce u těchto typů funkcí členství. Z trojúhelníkových a trapezoidních funkcí tento pokles přesnosti není tak markantní. U trapezoidních funkcí je patrná vyvážená přesnost napříč úrovní nejistoty obsažené mezi jednotlivými horními a dolními funkcemi členství. V případě zpožděného sentimentu je již patrná lehká asymetrie distribuce mezi falešně pozitivním a falešně negativním označením k nákupu či prodeji. Je proto vhodné se zaměřit na F1, neboť toto skóre bere v úvahu falešně pozitivní i falešně

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7917	0.2083	0.8148	0.7971	0.8059
Type-2 fuzzy model s 10 % nejistotou	0.7937	0.2063	0.8148	0.8000	0.8073
Type-2 fuzzy model s 20 % nejistotou	0.7898	0.2102	0.8259	0.7880	0.8065
Type-2 fuzzy model s 30 % nejistotou	0.7466	0.2534	0.7741	0.7545	0.7642
Type-2 fuzzy model s 40 % nejistotou	0.7151	0.2849	0.7333	0.7306	0.7320
Type-2 fuzzy model s 50 % nejistotou	0.6857	0.3143	0.7222	0.6964	0.7091
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7623	0.2377	0.8111	0.7578	0.7835
Type-2 fuzzy model s 10 % nejistotou	0.7662	0.2338	0.8111	0.7631	0.7864
Type-2 fuzzy model s 20 % nejistotou	0.7662	0.2338	0.8111	0.7631	0.7864
Type-2 fuzzy model s 30 % nejistotou	0.7662	0.2338	0.7963	0.7706	0.7832
Type-2 fuzzy model s 40 % nejistotou	0.7760	0.2240	0.7852	0.7910	0.7881
Type-2 fuzzy model s 50 % nejistotou	0.7603	0.2397	0.7222	0.8058	0.7617
Gaussovské funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7996	0.2004	0.8481	0.7897	0.8179
Type-2 fuzzy model s 10 % nejistotou	0.8035	0.1965	0.8556	0.7911	0.8221
Type-2 fuzzy model s 20 % nejistotou	0.7505	0.2495	0.8407	0.7299	0.7814
Type-2 fuzzy model s 30 % nejistotou	0.6798	0.3202	0.8963	0.6419	0.7481
Type-2 fuzzy model s 40 % nejistotou	0.6444	0.3556	0.8481	0.6206	0.7167
Type-2 fuzzy model s 50 % nejistotou	0.6130	0.3870	0.7741	0.6058	0.6797
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7996	0.2004	0.837	0.7958	0.8159
Type-2 fuzzy model s 10 % nejistotou	0.8035	0.1965	0.8852	0.7760	0.8270
Type-2 fuzzy model s 20 % nejistotou	0.5815	0.4185	0.9296	0.564	0.7021
Type-2 fuzzy model s 30 % nejistotou	0.5403	0.4597	0.8852	0.5407	0.6713
Type-2 fuzzy model s 40 % nejistotou	0.5187	0.4813	0.8667	0.5282	0.6564
Type-2 fuzzy model s 50 % nejistotou	0.5265	0.4735	0.9000	0.5317	0.6685

Tab. 4.19: Trénování fuzzy modelů se sentimentem t-1
Zdroj: vlastní zpracování

negativní výsledky. Intuitivně to není tak snadné pochopit jako accuracy, ale F1 je obvykle užitečnější, zvláště pokud existuje ve vzorku nerovnoměrné rozdělení tříd, jak již bylo naznačeno výše. Z hlediska F1 se průměrná přesnost zlepšila o například u zvonkových funkcí členství až o 10%, což je zapříčiněné zejména značnou nevyváženou distribucí jednotných tříd matice záměn u vyšších stupňů neurčitosti. U ostatních tříd metod je pozorováno zlepšení o 2-5%.

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6817	0.3183	0.7704	0.6753	0.7197
Type-2 fuzzy model s 10 % nejistotou	0.6817	0.3183	0.8111	0.6636	0.7300
Type-2 fuzzy model s 20 % nejistotou	0.6483	0.3517	0.8519	0.6233	0.7199
Type-2 fuzzy model s 30 % nejistotou	0.6424	0.3576	0.8778	0.614	0.7226
Type-2 fuzzy model s 40 % nejistotou	0.6228	0.3772	0.8481	0.6026	0.7046
Type-2 fuzzy model s 50 % nejistotou	0.5972	0.4028	0.8222	0.5858	0.6841
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6306	0.3694	0.6926	0.6404	0.6655
Type-2 fuzzy model s 10 % nejistotou	0.6306	0.3694	0.6926	0.6404	0.6655
Type-2 fuzzy model s 20 % nejistotou	0.6287	0.3713	0.7037	0.6355	0.6678
Type-2 fuzzy model s 30 % nejistotou	0.6267	0.3733	0.7037	0.6333	0.6667
Type-2 fuzzy model s 40 % nejistotou	0.6228	0.3772	0.7111	0.6275	0.6667
Type-2 fuzzy model s 50 % nejistotou	0.6287	0.3713	0.7148	0.6328	0.6713
Gaussovské funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7033	0.2967	0.7667	0.7017	0.7327
Type-2 fuzzy model s 10 % nejistotou	0.7014	0.2986	0.7519	0.7049	0.7276
Type-2 fuzzy model s 20 % nejistotou	0.6346	0.3654	0.5037	0.7234	0.5939
Type-2 fuzzy model s 30 % nejistotou	0.5796	0.4204	0.4741	0.6400	0.5447
Type-2 fuzzy model s 40 % nejistotou	0.5737	0.4263	0.5815	0.6015	0.5913
Type-2 fuzzy model s 50 % nejistotou	0.5599	0.4401	0.6889	0.5706	0.6242
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7033	0.2967	0.7815	0.6964	0.7365
Type-2 fuzzy model s 10 % nejistotou	0.6916	0.3084	0.7815	0.6828	0.7288
Type-2 fuzzy model s 20 % nejistotou	0.6267	0.3733	0.9296	0.5948	0.7254
Type-2 fuzzy model s 30 % nejistotou	0.5756	0.4244	0.9407	0.5595	0.7017
Type-2 fuzzy model s 40 % nejistotou	0.5422	0.4578	0.8926	0.5416	0.6741
Type-2 fuzzy model s 50 % nejistotou	0.5363	0.4637	0.8333	0.5409	0.6560

Tab. 4.20: Trénování fuzzy modelů se sentimentem t-2
Zdroj: vlastní zpracování

V tabulce 4.20, resp. 4.21 je uvedena přesnost predikce s využitím sentimentu zpožděného o dva, resp. o tři dny. V případě zahrnutí vyššího zpoždění v sentimentu je pozorována nižší přesnost accuracy kvůli závažné nerovnováze třídy. V případě dvoudenního zpoždění sentimentu je přesnost predikce z hlediska accuracy pouze 62.8%, v případě F1 68%. U třídenního zpoždění je accuracy 60.4% a F1 60.8%. To značení, že s rostoucím zpoždění sentimentu investorů dochází také ke zvyšujícímu se asymetrickému rozdělení mezi jednotlivými třídami matice záměn. Kupříkladu type-2 fuzzy model se zvonkovými funkcemi členství a 50% nejistotou zachycuje recall 63.16%, zatímco precision 4.4% neboť je podhodnocena jedna z kategorií matice záměn. Obdobně lze dané výstupy shledat i u dalších podkategorií zvonkových funkcí členství. Obecně lze konstatovat, že zvyšujícím se zpoždění sentimentu se snižuje

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6444	0.3556	0.7630	0.6378	0.6948
Type-2 fuzzy model s 10 % nejistotou	0.6444	0.3556	0.7704	0.6361	0.6968
Type-2 fuzzy model s 20 % nejistotou	0.6346	0.3654	0.7370	0.6338	0.6815
Type-2 fuzzy model s 30 % nejistotou	0.5894	0.4106	0.6333	0.6085	0.6207
Type-2 fuzzy model s 40 % nejistotou	0.5678	0.4322	0.5333	0.605	0.5669
Type-2 fuzzy model s 50 % nejistotou	0.5501	0.4499	0.4741	0.5953	0.5278
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6090	0.3910	0.8481	0.5917	0.6971
Type-2 fuzzy model s 10 % nejistotou	0.6090	0.3910	0.8444	0.5922	0.6962
Type-2 fuzzy model s 20 % nejistotou	0.6071	0.3929	0.8296	0.5926	0.6914
Type-2 fuzzy model s 30 % nejistotou	0.6071	0.3929	0.8222	0.5936	0.6894
Type-2 fuzzy model s 40 % nejistotou	0.6090	0.3910	0.8037	0.5978	0.6856
Type-2 fuzzy model s 50 % nejistotou	0.6012	0.3988	0.7926	0.5928	0.6783
Gaussovské funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7033	0.2967	0.8333	0.6798	0.7488
Type-2 fuzzy model s 10 % nejistotou	0.7053	0.2947	0.8444	0.6786	0.7525
Type-2 fuzzy model s 20 % nejistotou	0.6464	0.3536	0.7222	0.6500	0.6842
Type-2 fuzzy model s 30 % nejistotou	0.6169	0.3831	0.7074	0.6221	0.662
Type-2 fuzzy model s 40 % nejistotou	0.5422	0.4578	0.5556	0.5703	0.5629
Type-2 fuzzy model s 50 % nejistotou	0.4951	0.5049	0.3593	0.5359	0.4302
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6916	0.3084	0.8259	0.6697	0.7396
Type-2 fuzzy model s 10 % nejistotou	0.6994	0.3006	0.8185	0.6800	0.7429
Type-2 fuzzy model s 20 % nejistotou	0.6208	0.3792	0.6667	0.636	0.6510
Type-2 fuzzy model s 30 % nejistotou	0.5324	0.4676	0.3593	0.5988	0.4491
Type-2 fuzzy model s 40 % nejistotou	0.5010	0.4990	0.0889	0.7500	0.1589
Type-2 fuzzy model s 50 % nejistotou	0.4794	0.5206	0.0444	0.6316	0.0830

Tab. 4.21: Trénování fuzzy modelů se sentimentem t-3
Zdroj: vlastní zpracování

přesnost predikovat pohybovat trh směrem nahoru nebo dolů. Jinými slovy rostoucí zpoždění neposkytuje tak povzbudivé výsledky, jako v případě nižšího zpoždění.

Je zjištěno, že přidání indikátorů sentimentu investora vytvořených algoritmem binární klasifikace může sloužit jako prediktor budoucího vývoje akciového trhu. Vytvořením expertního modelu fuzzy logiky, lze docílit s více než 80% přesností predikce vývoje akciového trhu při aplikace type-2 fuzzy logiky se zahrnutím dodatečné 10% nejistoty. Povzbudivé výsledky poskytují jak Gaussovské, tak zvonkové funkce členství. Obdobné výsledky ovšem s nižší přesnosti vykazuje klasický type-1 fuzzy logika. Při aplikace type-2 fuzzy logiky je ovšem nezbytné upozornit, že s rostoucí mírou nejistoty se snižuje i přesnost predikce, z toho důvodu je nezbytné mít na zřeteli úroveň nejistoty zahrnuté ve funkcích členství.

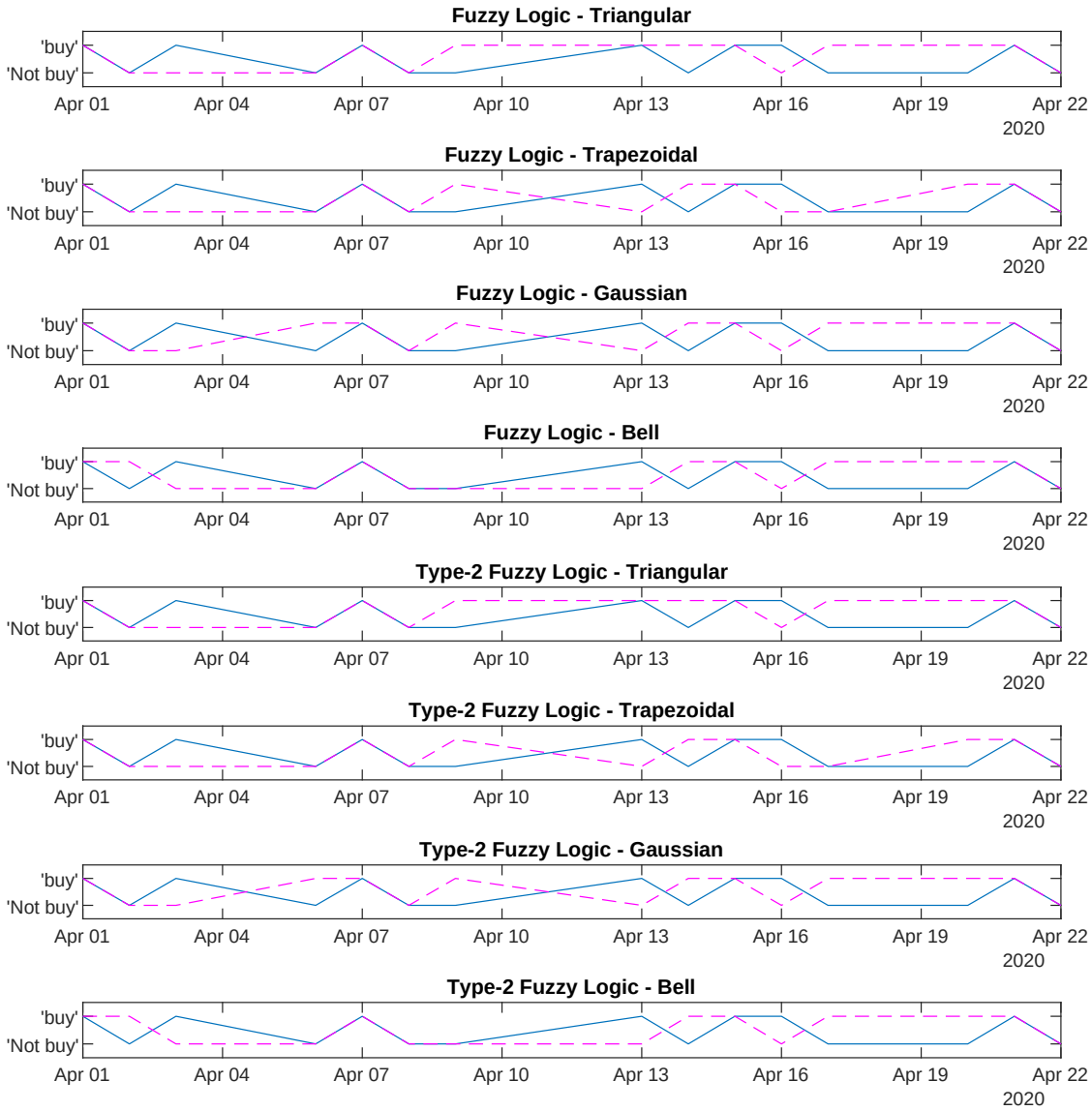
Dále je prokázáno, že se zvyšujícím se zpožděním zahrnutým do modelu prudce klesá přesnost predikce až o 20%. Na základě výsledků prognózy budoucího vývoje akciového trhu, je přesnost modelů se zahrnutím zpožděného sentimentu o více než jeden den nepřijatelná; přesnost většiny modelů je mírně nad 60% a u některých modelů dokonce pod 50%. Výsledky ukazují, že sentiment extrahovaný textem má určitou mírnou prediktivní sílu pro budoucí pohyb akciového trhu ovšem s ohledem na velikost zpoždění.

4.5.3 Investiční strategie

Rozšiřují se důkazy o tom, že ceny akcií přehnaně reagují nebo nedostatečně reagují na zveřejňované informace, což naznačuje, že pravděpodobně bude existovat zisková obchodní strategie, která investuje na základě jejich minulých výnosů a sentimentu. Koncept této výzkumné práce je postaven na studii Tetlocka a kol. (2008), kteří zjistili, že obchodní strategie založená na negativních slovech ve zprávách o konkrétních firmách může vydělat abnormální výnosy. Pro důkladnější testování schopnost vydělat zisky na základě konkrétních názorů, myšlenek a mínění investorů ve zprávách StockTwits, je navržena obchodní strategie založena na sentimentu investorů, o nichž se předpokládá, že má potenciální vliv na rozhodnutí o prodeji, nákupu nebo držení na akcií. Na rozdíl od studie Tetlock a kol. (2008), kteří pomocí jednoduchého kvantitativního měřítka jazyka předpovídali účetní zisky a výnosy akcií pouze na základě negativních slov, tato studie zvažuje kolektivní použití tweetového jazyka, přičemž při predikci StockTwits jsou zohledněna pozitivní i negativní slova obchodní strategie. Jednoduše je nezbytné zhodnotit ziskovost extrahovaného sentimentu z online zpráv a určit, zda je pro investory v zásadě efektivní a poskytuje vodítko při správném, přesném a výnosném rozhodnutí týkajícím se akciového indexu.

Tvorba investiční strategie

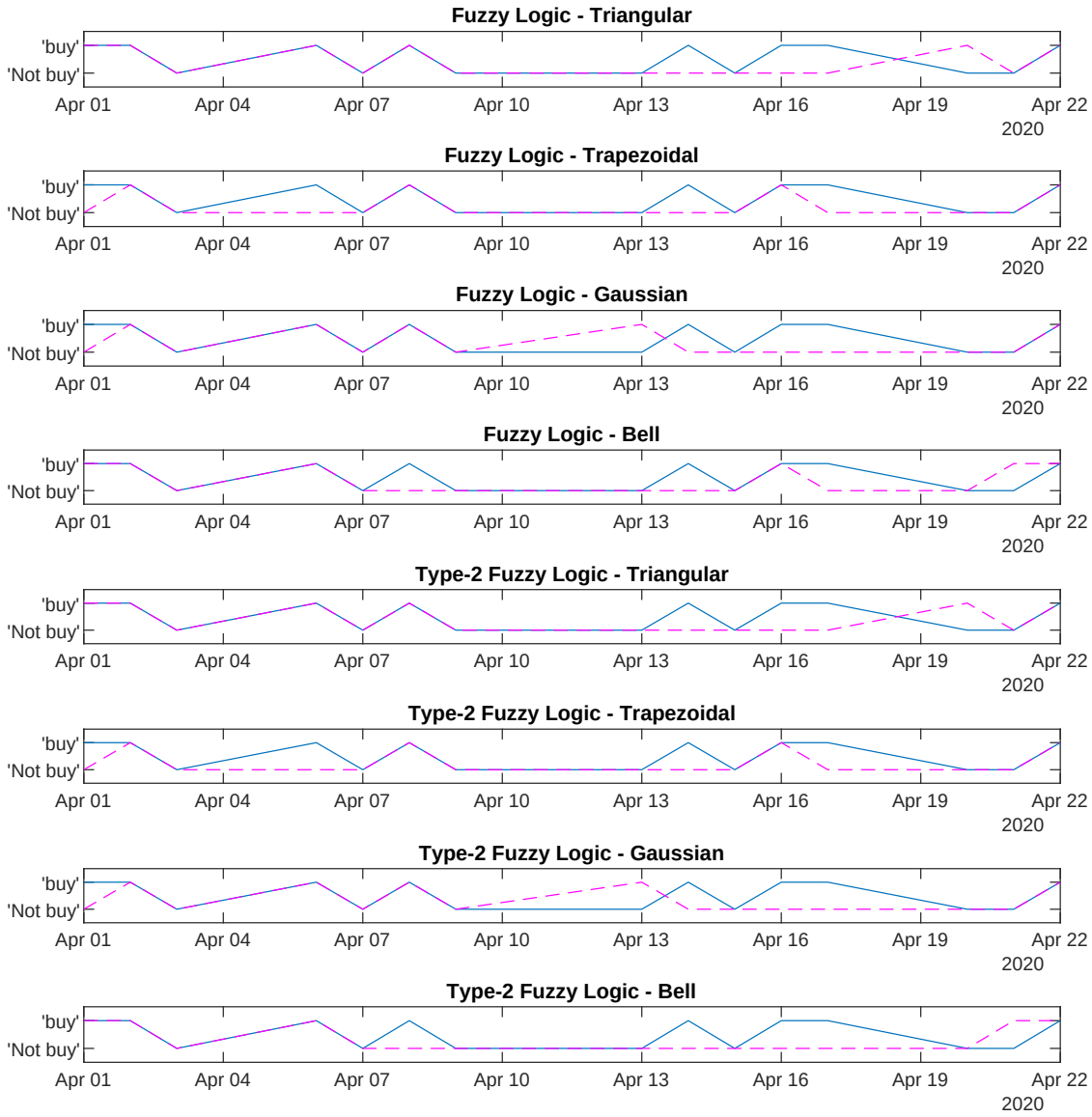
Po natrénování zvolených fuzzy modelů jsou tak vybráni zástupci jednotlivých tříd, kteří dosáhli nejvyšší přesnosti. Je zvolena strategie „zisk> ztráta“, to znamená následující. Pokud je skutečné doporučení koupit a výstup predikce modelu rovněž signalizuje příkaz k nákupu, je realizován zisk. Nicméně v případě, že je skutečné doporučení nekupovat či neinvestovat do akciového indexu a výstup z predikce indikuje nákupní signál, je realizována ztráta. Na závěr je provedeno zpětné testování, které je zautomatizováno, aby se aktualizovalo nové modelování o nová aktualizovaná data každého dne k předpovědi rozhodnutí následujícího obchodního dne. Dále je pracováno výhradně se zpožděním sentimentu o jeden den a jsou zvoleny nejpřesnější zástupci fuzzy modelů každého typu fuzzy funkce členství.



Obr. 4.18: Strategie fuzzy modelů bez sentimentu
Zdroj: vlastní zpracování

Výstupy zvolených strategií za zvolený vzorek testovacích dní jsou vyobrazeny na obrázku 4.18 a 4.19, přičemž první obrázek zachycuje predikci vybízející k prodejní či nákupní strategii bez zahrnutí sentimentu, zatímco druhý vyobrazuje investiční strategii s využitím zpožděného sentimentu. Ostatní strategie se zpožděním sentimentu, tj. o dva a tři dny jsou vloženy v příloze práce. Modrá plná čára vyobrazuje původní stav označující koupi či prodej, přerušován čára zobrazuje strategii predikovanou konkrétním modelem.

Lze vypozorovat, že nejvíce nákupních signálů generuje model type-1 fuzzy logiky s trapezoidními funkcemi členství, přičemž za testované období tento model vygeneroval v 36 případech příkaz k nákupu či investování do akciového indexu, type-2



Obr. 4.19: Strategie fuzzy modelů se sentimentem
Zdroj: vlastní zpracování

fuzzy model se stejným typem funkcí s počtem nákupních signálů 35. Jedná se tedy o nejaktivnější investiční strategii Na druhé straně za konzervativnější či pasivnější strategii lze považovat model fuzzy logiky nižšího i vyššího stupně sa zvonkovými funkcemi členství, popřípadě s Gaussovými funkcemi, kde lze za testované období vypozorovat 28, resp. 31 příkazů k nákupní strategii. V závěru lze shrnout, že v případě expertních modelů bez zahrnutí sentimentu je dominancí investiční strategie signál k prodeji.

Obdobných výsledků v prezentované testovací sadě, dosáhla strategie integrující sentiment investorů. Avšak při zahrnutí sentimentu dochází ke změně u nejaktivnější strategie na fuzzy model i type-2 fuzzy model s trojúhelníkovými funkcemi členství.

Nákupní signál je generován v celkem 38 případech, což dává převahu nákupních signálů v testovací sadě nad prodejními signály. Naopak opětovně nejpasivnější investiční strategie je dosaženo při využití type-2 fuzzy modelu se zvonkovými funkcemi členství. V tomto modelu je stanoven příkaz k nákupu pouze ve 22 dnech. V případě Gaussových a trapezoidních funkcí členství je dosaženo podobného počtu příkazů k nákupu či prodeji, přičemž stěžejní převahu má doporučení neinvestovat.

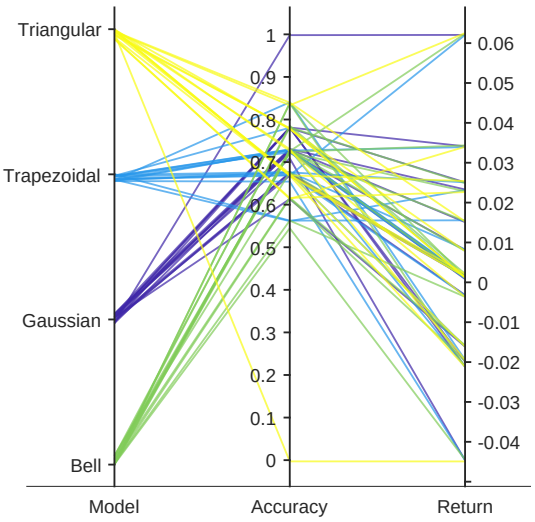
Zajímavostí je, že u vyššího zpoždění sentimentu se podstatně zvyšuje počet signálů k nákupu či investování. To znamená, že se prosazuje aktivnější investiční strategie a jak je patrné z přílohy práce, některé modely za testované období vysílají ve více než 60% testovaných dnů nákupní signál. Jinými slovy, se zahrnutím vyššího zpoždění sentimentu je patrná aktivnější investiční strategii a namísto neinvestování do akciového indexu je doporučeno nakupovat akcie v indexu, a nikoliv držby hotovostního aktiva.

Dále je vykreslen paralelní graf nebo graf paralelních souřadnic sumarizující modely, přesnost a dosažený výnos ze zvolené investiční strategie. K tomuto účelu je zvolena funkce `parallelplot`. Paralelní graf umožňuje studovat vlastnosti predikce pro několik kvantitativních proměnných. Jeho síla spočívá v tom, že proměnné mohou být dokonce zcela odlišné: různé rozsahy, a dokonce různé jednotky. Tento typ grafu nahrazuje konvenční trojrozměrné ortogonální soustavu rovnoběžných os, přičemž každý trojrozměrný bod je reprezentován křivkou v paralelních souřadnicích, a je velmi užitečný pro zobrazení vysokodimenzionálních dat.

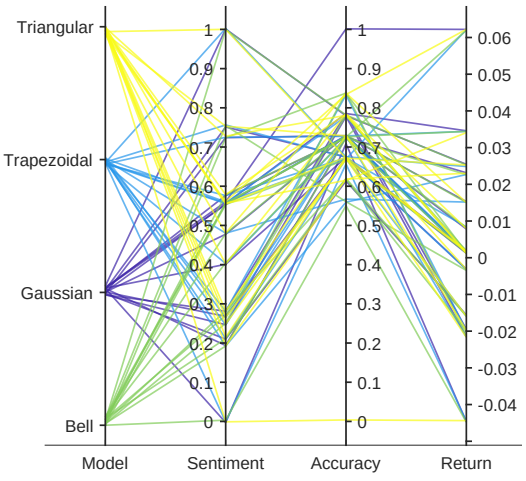
Obrázek 4.20 vykresluje paralelní graf bez sentimentu, zatímco obrázek 4.21 se zpožděným sentimentem o jeden den. V případě obrázku bez sentimentu jsou k dispozici výhradně tři svislé osy představující zvolený model či techniku predikci, resp., pro přehledné zobrazení, pouze jejich zástupce, dále přesnost zvolené techniky predikci v zastoupení ukazatele accuracy (který je pro lepší zobrazení upraven na hodnotu 0 až 1, označující nejnižší až nejvyšší přesnost) a následný výnos vygenerovaný aplikací dané investiční strategie. Je patrné, že například fuzzy model s Gaussovou funkcí členství měla přesnost predikce v jeden testovací den nejvyšší a vygenerovala nejvyšší denní výnos 6%. Nicméně i ostatní funkce členství, integrované ve fuzzy modelu, dokázaly vygenerovat stejný výnos při nižší přesnost. Jak již bylo zmíněno výše, vhodnější by bylo zanést do grafu ukazatel F1, neboť je schopen obsáhnout asymetrické rozdělení do tříd nákup/prodej. Na druhé straně trapezoidní, Gaussovi i zvonkové funkce členství s relativně vysokou přesností vykázaly při predikci konkrétního dne velice nízkou přesnost, což odráží skutečnost, že při zvolené těchto technik v daný den byl dosažen nejnižší výnos, resp. ztráta ve výši 4%.

V případě obrázků se zpožděným sentimentem přibyla jedna svislá osa označující právě vytvořený sentiment z předcházející části práce. Rovněž stejně jako přesnost je i sentiment normalizován do rozmezí od nuly do jedné. Lze si povšimnout, že

při vzedmutí vlny pesimismu, což označuje hodnota blízké nule, se při špatném natrénování modelu dosahuje ztráta, ale také při vyšší přesnosti predikci modelu v rozmezí 60% až 8%, je dosaženo z realizované investiční strategie ztráty, popřípadě denního výnosu lehce nad 1%. Při optimističtější náladě investor a obdobné přesnosti predikce je schopna zvolená strategie vygenerovat výnos okolo 2-4% za den.



Obr. 4.20: Paralelní graf bez sentimentu
Zdroj: vlastní zpracování



Obr. 4.21: Paralelní graf se sentimentem
Zdroj: vlastní zpracování

Zpětné testování investiční strategie

Posoudit výkonnost strategie založené na skóre sentimentu vytvořené v předchozí části musí být hodnocena na základě srovnávacích strategií obchodování. Je nezbytné připomenout, že účelem tohoto výzkumu je zjistit, zda by obchodní strategie založené na sentimentu vytěženém z fóra StockTwits a online finančních zpráv integrovaných do expertního modelu mohla odhalit ziskové příležitosti. Dále je nezbytné konstatovat, že toto srovnání investičních strategií a experimentální simulace je prozatímní a není vyčerpávající, protože je experimentováno pouze s výběrem základních obchodních strategií, ale poskytuje indikaci skutečné hodnoty navrhované metodiky. Experimentování a zpětné testování investiční strategie vytvořeného skóresentimentu investorů je komparováno s benchmarkem prostřednictvím Financial Toolbox™. Výsledky jsou poté analyzovány a je porovnán výkon každé strategie.

Investiční strategie je sestavena prostřednictvím objektu `backtestStrategy` definovaných ve funkcích skriptu. Každá strategie využívá funkci rebalance k rozhodování o obchodování na základě příslušných signálů. Signály vyžadují dostatečné koncové údaje pro výpočet obchodních signálů. Všechna koncová data jsou zachycena v předpočítaných obchodních signálech. Skutečné strategie tedy potřebují pouze

2denní hloubku pohledu k tomu, aby se obchodní rozhodnutí mohla vyhodnotit, když signály překročí prahové hodnoty obchodování. Jak je u tohoto typu analýzy obvyklé, transakční náklady jsou obvykle ignorovány (Zhang a Skiena, 2010). Nelze však vyloučit dopad těchto transakčních nákladů na zisk při implementaci v reálném světě. Tato studie proto sleduje Hu a kol. (2015) zvážením transakčních nákladů k vyhodnocení výkonu investiční strategie. Všechny strategie platí 25 bazických bodů transakčních nákladů na nákup a prodej. Bezriziková sazba získaná z neinvestované hotovosti je 1% ročně. Pro investiční strategie je zvolena funkce `crossoverRebalanceFunctiony`. To z toho důvodu, že oba obchodují na svých příslušných signálech stejným způsobem (kupují, když signál jde z 0 na 1, prodávají, když signál jde z 1 na 0.

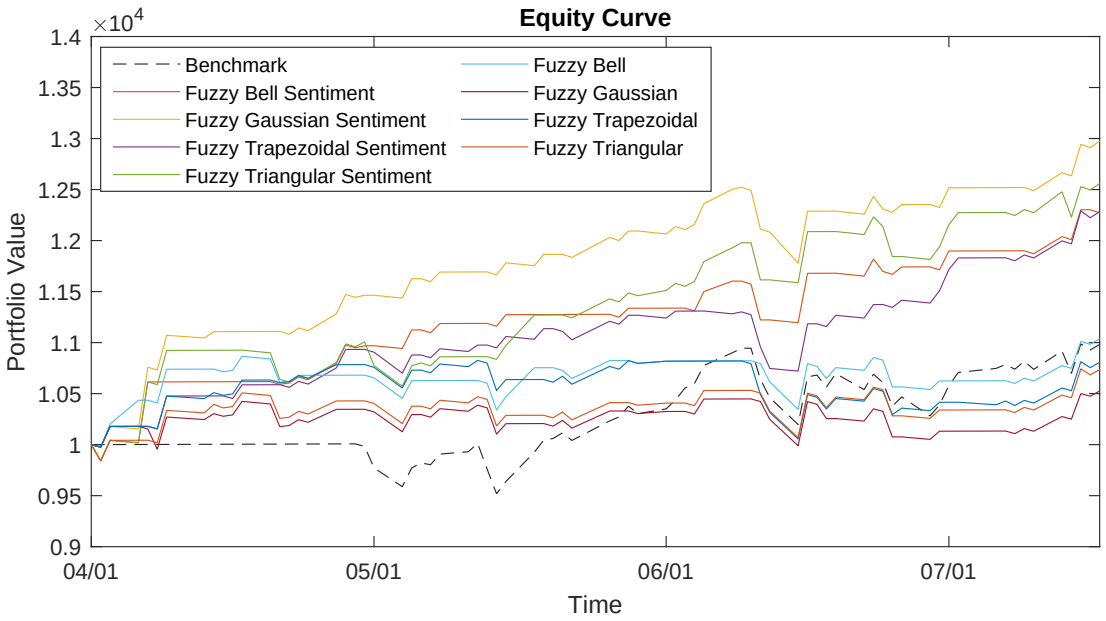
Nastavení zpětného testování investiční strategie

Jako benchmark se používá jednoduchá strategii akciového indexu k určení, zda obchodní signály poskytují cenné informace o budoucích výnosech akciového indexu. Veškeré obchodné signály jsou agregovány a synchronizovány do jednoho výsledného obchodního signálu. Pomocí `backtestEngine` je vytvořen modul zpětného testování a poté pomocí `runBacktest` je spuštěno zpětné testování investiční strategie.

Simulace zpětného testování

K prozkoumání výsledku zpětného testování je zvolena křivka ekvity k vizualizaci výkonu investičních strategií. Tato křivka charakterizuje vývoj vloženého kapitálu v čase. Ukázka ekvitní křivky obchodních signálů generované prostřednictvím expertního fuzzy modelu s různým typem funkcí členství je vykreslena na obrázku 4.22. Je nezbytné upozornit, že jsou zvoleny výhradně fuzzy modely, které dosáhly nejvyšší přesnosti. K tomuto kroku je přistoupeno, aby byla jednoduše pozorovatelná výkonnost fuzzy modelů. Přičemž je uvažováno o počáteční investici 10 000 USD, která je zhodnocována v průběhu sledovaného testovaného období 1.4.2020 do 17.7.2020. Křivka tak zobrazuje kumulované zhodnocení vložených finančních prostředků.

Pokud jde o pouze zhodnocení kapitálu, je třeba poznamenat, že výkonnost sentimentální strategie již od první čtvrtiny testovaného období začíná fungovat poměrně dobře. Prudký nárůst této strategie probíhá v poslední čtvrtině časového horizontu, kde je patrný výrazný rozdíl ve zhodnocení vloženého kapitálu investiční strategie založené na sentimentu oproti investiční strategii bez sentimentu investorů. Lze si povšimnout, že strategie založená na sentimentu je schopna se lépe popasovat s nenadálými propady na trhu, resp. lépe využít následné ziskové příležitosti po propadu trhu. Ve skutečnosti investiční strategie bez sentimentu funguje lépe než jednoduchá srovnávací strategie založená na benchmarku, nicméně od konce května 2020



Obr. 4.22: Simulace vývoj vloženého kapitálu
Zdroj: vlastní zpracování

tato strategie zaostává za benchmarkem. Naprostou dominanci, z hlediska zhodnocení portfolia v podobě akciového indexu, vykazuje fuzzy model s Gaussovými funkcemi členství. Jedná se o model, který vykázal více než 80% přesnost predikce. Avšak vysoká přesnost predikce vývoje trhu ještě neznamená identifikaci ziskových investičních příležitostí. To lze deklarovat u fuzzy modelu se zvonkovými funkcemi členství, které vykazovali nejvyšší přesnost predikce na testovací sadě, nicméně je z obrázku patrné, že i přes vysokou přesnost, nebyl tento model schopen identifikovat nejvíce ziskové příležitosti. Nicméně je třeba dále tuto dominanci sentimentu analyzovat, aby mohlo být vyřčeno, že sentiment dokázal obsáhnout pesimistický či optimistický pohled investorů, který sdělovali na webu v tomto období a správnou reakcí vytěžila tato strategie maximum.

V tabulce 4.22, resp. 4.23 jsou uvedeny ukazatele výkonnosti komparovaných investičních obchodních signálů bez sentimentu, resp. se zahrnutým sentimentem investorů. Již dříve byl analyzován počet nákupních a prodejních signálů, je patrné, že reakce na tyto signály je třeba realizovat, což sebou přináší dodatečné transakční náklady. Co se týká ukazatelů výkonnosti, je v tabulce patrný celkový výnos, kterého jednotlivé investiční strategie dosáhly. Jedná se o neanualizovaný celkový výnos strategie, včetně poplatků, po celou dobu zpětného testování.

Výnos benchmarku, tedy srovnávací základny je za testované období 9.82%. Při pohledu na celkový výnos fuzzy modelů bez integrovaného sentimentu dokázal výnos benchmarku překonat pouze model se zvonkovými funkcemi členství s hodnotou 10.38%. U ostatních modelů bez sentimentu se celkový výnos pohybuje v rozmezí

	Benchmark	Fuzzy Triangular	Fuzzy Trapezoidal	Fuzzy Gaussian	Fuzzy Bell
Celkový výnos	0.0982	0.0737	0.0807	0.0526	0.1038
Průměrný výnos	0.0014	0.0010	0.0011	0.0008	0.0014
Volatilita	0.0131	0.0114	0.0113	0.0113	0.0116
Průměrný obrat	0.0068	0.2568	0.2770	0.2905	0.2230
Maximální obrat	0.5000	0.5000	0.5000	0.5000	0.5000
Maximální ztráta	0.0684	0.0439	0.0712	0.0439	0.0485
Průměrné nák. náklady	0.3381	6.6750	7.4973	7.6450	6.1416
Průměrné prod. náklady	0.0000	6.6911	7.1768	7.3249	5.8177

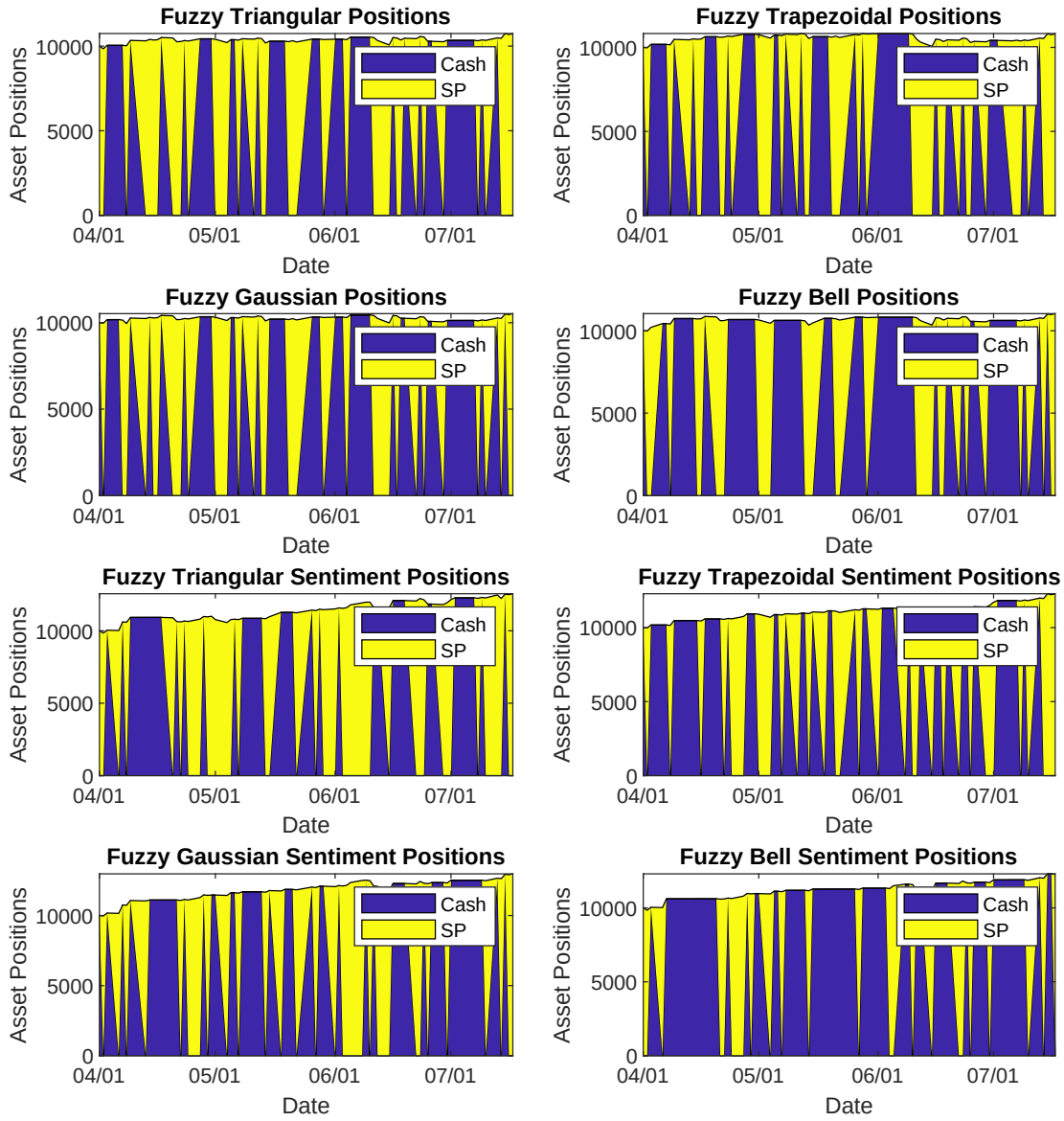
Tab. 4.22: Výkonnost investiční strategie bez sentimentu
Zdroj:vlastní zpracování

	Benchmark	Fuzzy Triangular	Fuzzy Trapezoidal	Fuzzy Gaussian	Fuzzy Bell
Celkový výnos	0.0982	0.2559	0.2284	0.2975	0.2272
Průměrný výnos	0.0014	0.0032	0.0028	0.0036	0.0028
Volatilita	0.0131	0.0139	0.0104	0.0122	0.0113
Průměrný obrat	0.0068	0.2432	0.3041	0.3041	0.2297
Maximální obrat	0.5000	0.5000	0.5000	0.5000	0.5000
Maximální ztráta	0.0684	0.0395	0.0521	0.0595	0.0351
Průměrné nák. náklady	0.3381	6.8911	8.6002	9.1268	6.4998
Průměrné prod. náklady	0.0000	6.9078	8.2823	8.8102	6.5152

Tab. 4.23: Výkonnost investiční strategie se sentimentem
Zdroj:vlastní zpracování

5% až 8%. Naproti tomu strategie založená na sentimentu investorů extrahovaném z textových zpráv se celkový výnos pohybuje v rozmezí 22% až 30%. Nejvyššího výnosu dosáhla investiční strategie založení na fuzzy modelu s Gaussovskými funkcemi členství. Naopak fuzzy model se zvonovými a trapezoidními funkcemi členství dosáhly shodného celkového výnosu 22%. Vyšší výkon sentimentální strategie by však mohl pouze počátkem vyššího rizika strategie. Z toho důvodu je také analyzována volatilita, což v podstatě reprezentuje neanualizovaná standardní odchylku výnosů strategie. Kolísavost výnosu je v podstatě u všech modelů naprosto totožná okolo 1%. Z toho důvodu nelze říci, že by strategie založená na sentimentu byla více rizikovou investiční strategií. Maximální ztrátu zaznamenala investiční strategie bez sentimentu u fuzzy modelu s trapezoidními funkcemi členství (7.12%) a benchmark (6.84%) zatímco nejnižší maximální ztrátu vykázala strategie založená na sentimentu u fuzzy modelu se zvonkovými funkcemi členství (3.51%). Důležité jsou také průměrné nákupní a prodejní náklady. Vzhledem ke skutečnosti, že strategie založená na sentimentu vykazovala velice intenzivní nákupní a prodejní politiku, je logická

také vysoká hodnota těchto ukazatelů.



Obr. 4.23: Graf denních pozic dle investičních signálů
Zdroj: vlastní zpracování

Změnu alokace strategií do akciového indexu a do hotovosti v průběhu času, lze vizualizovat pomocí plošného grafu denních pozic aktiv. Využita je k tomuto účelu pomocná funkce `assetAreaPlot` pro každou z analyzovaných strategií. Strategie založená na benchmarku držela první testovaný měsíc výhradně hotovost, až poté alokoval finanční prostředky do akciového indexu, který držel po zbytek testovaného období. Naproti tomu, jak je vyobrazeno na obrázku 4.23, strategie založená na sentimentu i bez sentimentu velmi střídavě alokuje prostředky do akciového indexu a do hotovosti dle generovaných signálů z fuzzy modelů. Zejména u fuzzy modelu se zvonkovými funkcemi členství, který generuje signály na základě indikátoru sentimentu,

drží hotovost příliš dlouho, což může mít následek na celkový výnos, který je nižší oproti ostatním modelů se sentimentem, i přesto, že právě zvonkové funkce vykázaly nejvyšší přesnost predikce. Hotovostní aktivum vydělává bezrizikovou sazbu, která je defaultně nastavena na 1%. Je patrné, že tuto strategie bude spíše preferovat vysoce aktivní investor.

Strategie sentimentu ve sledovaném období překonává srovnávací strategii benchmarku i investiční strategii bez integraci vytvořeného sentimentu investorů, který je extrahován z online finančních zpráv a příspěvků na sociální síti. Skutečnost, že strategie založená na sentimentu investorů také překonává strategii benchmarku, která nepoužívá informace o sentimentu, je počáteční indikací, že sentiment investora nese relevantní informace pro ziskovou investiční strategii. Jelikož strategie založená na sentimentu investorů je aktivní strategií, vykazuje vyšší obrat než spíše pasivní srovnávací strategie.

4.6 Komparace modelů predikceu

V předcházející části disertační práce byl vytvořen expertní systém založený na fuzzy logice prvního a druhého typu s různými funkcemi členství. Je nezbytné model navržený v této práci komparovat s více užívanými modely, které byly v různých formách aplikovány ve výzkumných pracích. Konkrétně je pozornost zaměřena na komparaci fuzzy logiky s modely rozhodovacích stromů, SVM, k-nejbližších sousedů, neuronových sítí a naivního Bayese, přičemž tyto metody jsou detailně rozčleněny.

Odpovídající výsledky jsou podrobně uvedeny v tabulkách 4.24-4.27. Tabulka 4.24 odráží výkonnost predikce fuzzy modelů bez zahrnutí sentimentu. Oproti přesnosti trénování fuzzy modelů, dochází k poklesu jak ukazatele accuracy, tak F-skóre.

Průměrná přesnost všech analyzovaných fuzzy modelů bez sentimentu dosahuje hodnotu accuracy 64% a hodnotu F-skóre 68%. Z hlediska jednotlivých tvarů funkcí členství, vykazuje nejvyšší přesnost trojúhelníková a trapezoidní s hodnotou okolo 70%. Absolutní nejvyšší hodnotu F-skóre vykázal type-2 fuzzy model s trojúhelníkovou funkcí členství a integrací 40% nejistoty. Lze poznamenat, že v případě predikce bez sentimentu investorů, vykazuje type-2 fuzzy logika s dodatečnou nejistotou zahrnutou ve funkcí členství lepší výsledky, resp. vyšší přesnost než klasická type-1 fuzzy logika.

Zda jsou uvedené výsledky uspokojivé, je nezbytné fuzzy modely komparovat s jinými expertními modely, které se běžně ve výzkumných a vědeckých studiích aplikují. Konkrétně ke komparaci investiční strategie bez sentimentu investorů, je analyzována přesnost modelů v tabulce 4.25. Průměrná hodnota accuracy zkoumaných modelů je pouze 48%, resp. F-skóre 27%. Jedná se tedy o signifikantní rozdíl ve výkonnosti predikce vývoje akciového trhu bez zahnrutí sentimentu investorů.

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6757	0.3243	0.6923	0.6923	0.6923
Type-2 fuzzy model s 10 % nejistotou	0.6757	0.3243	0.6923	0.6923	0.6923
Type-2 fuzzy model s 20 % nejistotou	0.6757	0.3243	0.6923	0.6923	0.6923
Type-2 fuzzy model s 30 % nejistotou	0.6622	0.3378	0.7179	0.6667	0.6914
Type-2 fuzzy model s 40 % nejistotou	0.7092	0.2908	0.8407	0.6837	0.7542
Type-2 fuzzy model s 50 % nejistotou	0.6486	0.3514	0.7692	0.6383	0.6977
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6757	0.3243	0.6667	0.7027	0.6842
Type-2 fuzzy model s 10 % nejistotou	0.6892	0.3108	0.6923	0.7105	0.7013
Type-2 fuzzy model s 20 % nejistotou	0.7027	0.2973	0.7179	0.7179	0.7179
Type-2 fuzzy model s 30 % nejistotou	0.6892	0.3108	0.7436	0.6905	0.7160
Type-2 fuzzy model s 40 % nejistotou	0.6757	0.3243	0.7179	0.6829	0.7000
Type-2 fuzzy model s 50 % nejistotou	0.6486	0.3514	0.7179	0.6512	0.6829
Gaussovské funkce	Accuracy	Loss	Precision	Recall	F1
Klasický fuzzy model	**0.6351**	**0.3649**	**0.6923**	**0.6429**	**0.6667**
Type-2 fuzzy model s 10 % nejistotou	0.6351	0.3649	0.6923	0.6429	0.6667
Type-2 fuzzy model s 20 % nejistotou	0.6216	0.3784	0.6667	0.6341	0.6500
Type-2 fuzzy model s 30 % nejistotou	0.6081	0.3919	0.5897	0.6389	0.6133
Type-2 fuzzy model s 40 % nejistotou	0.6622	0.3378	0.6667	0.6842	0.6753
Type-2 fuzzy model s 50 % nejistotou	0.6081	0.3919	0.6667	0.6190	0.6420
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6216	0.3784	0.7179	0.6222	0.6667
Type-2 fuzzy model s 10 % nejistotou	0.6351	0.3649	0.7436	0.6304	0.6824
Type-2 fuzzy model s 20 % nejistotou	0.6216	0.3784	0.8462	0.6000	0.7021
Type-2 fuzzy model s 30 % nejistotou	0.5135	0.4865	0.7692	0.5263	0.625
Type-2 fuzzy model s 40 % nejistotou	0.5405	0.4595	0.8205	0.5424	0.6531
Type-2 fuzzy model s 50 % nejistotou	0.5270	0.4730	0.8205	0.5333	0.6465

Tab. 4.24: Testování fuzzy modelů bez sentimentu
Zdroj: vlastní zpracování

Navíc je patrné také asymetrické rozdělení do jednotlivých tříd matice záměn, jak je patrné z ukazatelů precision a recall. Přesnost nad 50% vykazuje SVM, ostatní modely predikují vývoj akciového trhu s přesností pod 50%.

V tabulce 4.26 jsou uvedeny metriky přesnosti predikce vývoje akciového trhu prostřednictvím fuzzy logiky za využití zpožděného sentimentu investorů o jeden den. Průměrná přesnost všech analyzovaných fuzzy modelů s denním zpožděním sentimentu investorů dosahuje hodnoty accuracy 65% a F-skóre 69%. Nejvyšší průměrnou přesnost vykazuje fuzzy model s trapezoidními funkcemi členství 72.29% a Gaussovskými funkcemi členství 71.60%. Absolutně nejvyšší přesnost prokázal type-1 i type-2 fuzzy model s 10% nejistotou a Gaussovskými funkcemi členství. F-skóre u těchto modelů je shodné 78.57%, což se blíží přesnosti trénování, které

	Accuracy	Loss	Precision	Recall	F1
Fine Tree	0.4595	0.5405	0.3636	0.3871	0.3750
Medium Tree	0.4730	0.5270	0.3939	0.4063	0.4000
Course Tree	0.4865	0.5135	0.1515	0.3333	0.2083
Linear Discriminant	0.5000	0.5000	0.2121	0.3889	0.2745
Linear SVM	0.4800	0.5200	0.0294	0.1429	0.0488
Quadratic SVM	0.6216	0.3784	0.3636	0.6316	0.4615
Cubic SVM	0.5135	0.4865	0.3333	0.4400	0.3793
Fine Gaussian SVM	0.5000	0.5000	0.1212	0.3333	0.1778
Medium Gaussian SVM	0.5600	0.4400	0.0294	1.0000	0.0571
Course Gaussian SVM	0.5600	0.4400	0.0294	1.0000	0.0571
Fine k-NN	0.4730	0.5270	0.3636	0.4000	0.3810
Medium k-NN	0.4324	0.5676	0.1515	0.2632	0.1923
Course k-NN	0.5600	0.4400	0.0294	1.0000	0.0571
Cosine k-NN	0.4459	0.5541	0.2121	0.3182	0.2545
Cubic k-NN	0.4189	0.5811	0.1212	0.2222	0.1569
Weighted k-NN	0.4324	0.5676	0.1818	0.2857	0.2222
Ensemble Boosted Tree	0.4595	0.5405	0.3939	0.3939	0.3939
Ensemble Subspace Discriminant	0.4730	0.5270	0.1212	0.2857	0.1702
Ensemble Subspace k-NN	0.4730	0.5270	0.3636	0.4000	0.381
Ensemble RUS Boosted Trees	0.4189	0.5811	0.3333	0.3438	0.3385
Narrow Neural Networks	0.4189	0.5811	0.2424	0.3077	0.2712
Medium Neural Networks	0.4459	0.5541	0.4242	0.3889	0.4058
Wide Neural Networks	0.4054	0.5946	0.3939	0.3514	0.3714
Bilayered Neural Networks	0.5405	0.4595	0.3636	0.4800	0.4138
Trilayered Neural Networks	0.4865	0.5135	0.4848	0.4324	0.4571
Gaussian Naive Bayes	0.4189	0.5811	0.1212	0.2222	0.1569
Kernel Naive Bayes	0.5270	0.4730	0.1818	0.4286	0.2553

Tab. 4.25: Testování modelů predikce bez sentimentu
Zdroj: vlastní zpracování

přesáhla u těchto modelů 80%. Opětovně lze vypozorovat, že type-2 fuzzy logika dosahuje stejných, ne-li lepších výsledků, jinými slovy obsažení jisté úrovně nejistoty ve funkcích členství poskytuje přesnější výsledky. Nicméně stejně jako bylo poznamenáno u tréninku, je nezbytné vhodně nastavit výši nejistoty, neboť je patrné, že s rostoucí nejistotou se od určité úrovně začíná přesnost predikce snižovat.

Z hlediska jednotlivých expertních modelů využitých v tabulce 4.27, je průměrná hodnota přesnosti srovnatelná s fuzzy logikou. Konkrétně průměrná hodnota accuracy všech modelů je 66.6% a F-skóre 69%. Co se týče jednotlivých tříd metod, tak nejvyšší přesnost vykazuje kernel naivní Bayes, přičemž průměrné F-skóre je 72%. Průměrné přesnosti predikce 70% dosáhl také SVM a k-NN. Ostatní modely vykazují průměrnou přesnost pod touto hodnotou. Nejnižší přesnost vykazují rozho-

dovací stromy a neuronové sítě. Absolutně nejvyšší hodnota F-skóre, které pokrývá i asymetrické rozdělení, vykazuje lineární SVM s hodnotou 77.78% a dvouvrstvá neuronová síť s hodnotou 76.92%. Lze pozorovat, že v každé třídě metod se objevuje alespoň jedna metoda, jejíž přesnost přesahuje 70%. Dále jsou patrné značné rozdíly v přesnosti predikce například u neuronových sítí, kde je patrný rozdíl téměř 15%, což je již zásadní rozdíl ve výkonnosti a může způsobit značné rozdíly ve výnosnosti investiční strategie.

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6892	0.3108	0.6410	0.7353	0.6849
Type-2 fuzzy model s 10 % nejistotou	0.6622	0.3243	0.6410	0.6944	0.6667
Type-2 fuzzy model s 20 % nejistotou	0.6757	0.2973	0.6667	0.7027	0.6842
Type-2 fuzzy model s 30 % nejistotou	0.7027	0.3649	0.6667	0.7429	0.7027
Type-2 fuzzy model s 40 % nejistotou	0.6351	0.3649	0.6154	0.6667	0.6400
Type-2 fuzzy model s 50 % nejistotou	0.5405	0.4595	0.5385	0.5676	0.5526
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7027	0.2973	0.7692	0.6977	0.7317
Type-2 fuzzy model s 10 % nejistotou	0.7027	0.2973	0.7692	0.6977	0.7317
Type-2 fuzzy model s 20 % nejistotou	0.7027	0.2973	0.7436	0.7073	0.7250
Type-2 fuzzy model s 30 % nejistotou	0.7027	0.2973	0.7179	0.7179	0.7179
Type-2 fuzzy model s 40 % nejistotou	0.6892	0.3108	0.6923	0.7105	0.7013
Type-2 fuzzy model s 50 % nejistotou	0.7297	0.2703	0.6923	0.7714	0.7297
Gaussovské funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.7568	0.2432	0.8462	0.7333	0.7857
Type-2 fuzzy model s 10 % nejistotou	0.7568	0.2432	0.8462	0.7333	0.7857
Type-2 fuzzy model s 20 % nejistotou	0.7027	0.2973	0.8462	0.6735	0.7500
Type-2 fuzzy model s 30 % nejistotou	0.6081	0.3919	0.8462	0.5893	0.6947
Type-2 fuzzy model s 40 % nejistotou	0.5811	0.4189	0.7692	0.5769	0.6593
Type-2 fuzzy model s 50 % nejistotou	0.5541	0.4459	0.6923	0.5625	0.6207
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.6892	0.3108	0.8205	0.6667	0.7356
Type-2 fuzzy model s 10 % nejistotou	0.6757	0.3243	0.8974	0.6364	0.7447
Type-2 fuzzy model s 20 % nejistotou	0.5811	0.4189	0.8462	0.5690	0.6804
Type-2 fuzzy model s 30 % nejistotou	0.5405	0.4595	0.8205	0.5424	0.6531
Type-2 fuzzy model s 40 % nejistotou	0.5000	0.5000	0.7949	0.5167	0.6263
Type-2 fuzzy model s 50 % nejistotou	0.5270	0.4730	0.8205	0.5333	0.6465

Tab. 4.26: Testování fuzzy modelů se sentimentem t-1
Zdroj: vlastní zpracování

V příloze disertační práce jsou uvedeny také další tabulky zaznamenávající přesnost predikce s vyšším zpožděním, jak pro modely fuzzy logiky, tak také pro ostatní expertní modely. V případě obsažení vyššího zpoždění v sentimentu je pozorována nižší přesnost accuracy kvůli závažné nerovnováze třídy.

V případě dvoudenního zpoždění sentimentu je přesnost predikce u fuzzy modelů z hlediska accuracy pouze 48%, v případě F1 52%. U dalších, běžně využívaných, expertních modelů je přesnost z hlediska accuracy pouze 51%, v případě F1 57%. To značení, že s rostoucím zpoždění sentimentu investorů dochází také ke zvyšujícímu se asymetrickému rozdělení mezi jednotlivými třídami matice záměn. Kupříkladu medium gasussian SVM zachycuje precision 100%, zatímco recall 53%, neboť je podhodnocena kategorie falešně pozitivních případů. Obdobně lze dané výstupy shledat i u dalších podkategorií SVM. Obecně lze konstatovat, že zvyšujícím se zpoždění sentimentu se snižuje přesnost predikovat pohybovat trh směrem nahoru nebo dolů. Jinými slovy rostoucí zpoždění neposkytuje tak povzbudivé výsledky, jako v případě nižšího zpoždění.

	Accuracy	**Loss**	**Precision**	**Recall**	**F1**
Fine Tree	0.6216	0.3784	0.6154	0.6486	0.6316
Medium Tree	0.5541	0.4459	0.5897	0.5750	0.5823
Course Tree	0.7162	0.2838	0.6667	0.7647	0.7123
Linear Discriminant	0.7027	0.2973	0.6667	0.7429	0.7027
Linear SVM	0.7838	0.2162	0.7179	0.8485	0.7778
Quadratic SVM	0.6622	0.3378	0.641	0.6944	0.6667
Cubic SVM	0.6216	0.3784	0.6154	0.6486	0.6316
Fine Gaussian SVM	0.7162	0.2838	0.6923	0.7500	0.7200
Medium Gaussian SVM	0.7027	0.2973	0.7179	0.7179	0.7179
Course Gaussian SVM	0.5270	0.4730	1.0000	0.5270	0.6903
Fine k-NN	0.6892	0.3108	0.7436	0.6905	0.7160
Medium k-NN	0.6892	0.3108	0.6667	0.7222	0.6933
Course k-NN	0.527	0.473	1.0000	0.527	0.6903
Cosine k-NN	0.7297	0.2703	0.7692	0.7317	0.7500
Cubic k-NN	0.7162	0.2838	0.6667	0.7647	0.7123
Weighted k-NN	0.7297	0.2703	0.7179	0.7568	0.7368
Ensemble Boosted Tree	0.5405	0.4595	1.0000	0.5342	0.6964
Ensemble Subspace Discriminant	0.7162	0.2838	0.7179	0.7368	0.7273
Ensemble Subspace k-NN	0.6757	0.3243	0.6923	0.6923	0.6923
Ensemble RUS Boosted Trees	0.6486	0.3514	0.641	0.6757	0.6579
Narrow Neural Networks	0.6351	0.3649	0.5897	0.6765	0.6301
Medium Neural Networks	0.6892	0.3108	0.7179	0.7000	0.7089
Wide Neural Networks	0.5811	0.4189	0.5897	0.6053	0.5974
Bilayered Neural Networks	0.7568	0.2432	0.7692	0.7692	0.7692
Trilayered Neural Networks	0.6081	0.3919	0.6154	0.6316	0.6234
Gaussian Naive Bayes	0.6892	0.3108	0.6923	0.7105	0.7013
Kernel Naive Bayes	0.7568	0.2432	0.6667	0.8387	0.7429

Tab. 4.27: Testování modelů predikce se sentimentem t-1
Zdroj: vlastní zpracování

Je zjištěno, že přidání indikátorů sentimentu investora vytvořených algoritmem binární klasifikace může vylepšit pouze modely s nízkou přesností předpovědi, jako jsou modely lineárního diskriminantu a Naïve Bayes, se zlepšením přesnosti o 0.2 až 0.4%. Přidání indikátoru sentimentu však nemůže zlepšit modely s vynikající přesností předpovědi, které jsou založeny na vstupních datech denních cen a objemu obchodování, jako jsou modely SVM, k-nejbližších sousedů, rozhodovací stromy a neuronové sítě. Navíc je prokázáno, že se zvyšujícím se zpožděním zahrnutým do modelu prudce klesá přesnost predikce až o 20%. Na základě výsledků prognózy výnosů akciového indexu, je přesnost modelů se zahrnutím zpožděného sentimentu o více než jeden den nepřijatelná; přesnost většiny modelů je mírně nad 50% a u některých modelů dokonce pod 50%. Výsledky ukazují, že sentiment extrahovaný textem má určitou mírnou prediktivní sílu pro budoucí výnosy akciového trhu ovšem s ohledem na velikost zpoždění. Dále bylo prokázáno, že modely fuzzy logiky jsou schopny vhodným nastavením funkcí členství a nejistoty v nich obsažených vylepšit predikci a jsou schopny konkurovat klasickým expertním modelům predikce, které jsou standardně využívané ve výzkumných studiích. Zde je dále nezbytné upozornit, že uspokojivých výsledků dosáhl zejména model type-2 fuzzy logiky s integrací 10% nejistoty, které je schopen konkurovat klasické type-1 fuzzy logice, ba co více, dokáže jeho přesnost převýšit.

5 Diskuse disertační práce

V této kapitole je celá studie diskutována v souvislosti se stanovenými hypotézy a kontextem dřívějších relevantních výzkumných studií. Rovněž jsou diskutována omezení a problémy spojené s implementací výzkumného procesu ve zkoumané problematice. Dále jsou uvedeny možná doporučení pro budoucí směřování studia.

5.1 Diskuze

Na základě syntézy informací získaných ze sekundárního a empirického výzkumu lze do určité míry konstatovat závěry ohledně obecného charakteru vlivu sentimentu investorů extrahovaného z textových zpráv na akciové trhy včetně jeho implementace do expertního modelu. Nicméně spolehlivost a míra zobecnění těchto závěru je limitována nejenom zúženým na americký akciový trh, ale i skutečností, že expertní model je založen na zjednodušených předpokladech. Rovněž je nezbytné zodpovědět položené výzkumné otázky a hypotézy stanovené na počátku disertační práce, které přispívají k jedinečnosti provedeného výzkumu a k rozvoji a prohloubení současného stavu poznaní v předmětné problematice.

5.1.1 Souhrn výsledků provedeného sekundárního výzkumu

V rámci sekundárního výzkumu bylo nejprve představeno teoretické pozadí disertační práce, které poskytuje základní a nezbytný vhled do řešené problematiky. Finanční pozadí výzkumu je zaměřeno na akciový trh, který je předmětem zájmu disertační práce. Akciový trh je nepostradatelnou součástí finančního systému, který podporuje ekonomický růst přitahováním zdrojů financování z odvětví a subjektů s přebytkovým kapitálem k odvětvím a subjektům s deficitním kapitálem. Jako vhodné měřítko cenové výkonnosti akciových portfolií jsou označovány akciové indexy, které jsou v této práci brány jako proxy celého akciového trhu. Pozornost mnoha výzkumníků a finančních analytiků se upírá zejména na akciové indexy s cílem co nejpřesnější predikce jejich budoucího vývoje s vidinou vyšších zisků a eliminaci rizika spojeného s investováním. Avšak dle hypotézy efektivního trhu (EMH), ceny cenných papírů vždy plně odrážejí všechny dostupné informace na efektivních kapitálových trzích. Ve skutečnosti je však obtížné vysvětlit několik iracionálních jevů, jako je nákup na vzestupu a prodej na sestupném trendu, pomocí efektivní teorie trhu. Behaviorální finance studují finanční trh na základě individuálního chování a nálad a naznačují, že iracionální odchylky rozhodování mohou být systematické a interakce mezi investory povede k celkové odchylce akciového trhu v budoucnosti.

To vede k myšlence předpovědět budoucí vývoj akciového trhu prostřednictvím sentimentu investorů, který integruje nálady, postoje, názory investorů ohledně budoucího vývoje.

Akciové trhy byly opakovaně studovány, aby získaly užitečné vzory a předpovídaly jejich pohyby. Těžba textových dokumentů a časových řad současně, například předpovídání pohybů cen akcií na základě obsahu nových informací, je objevujícím se tématem v oblasti dolování dat a komunita pro těžbu textů. Prognóza vývoje cen akcií se těší velké oblibě pouze na základě technické a fundamentální analýzy dat. Číselná data časových řad však obsahují pouze událost a nikoli příčinu, proč k ní došlo. Textová data, jako jsou novinové články, mají bohatší informace, a proto využívání textových informací zejména vedle číselných dat časové řady zvyšuje kvalitu vstupu a očekávají se lepší předpovědi z tohoto druhu vstupu, nikoli pouze z číselných údajů. Informace o zprávě společnosti nebo nejnovější zprávy mohou dramaticky ovlivnit cenu akcií. Po dlouhou dobu byly oficiální zprávy publikované investičními analytiky nebo odborníky důležité a dostatečné zdroje, pro rozhodování o koupi či prodeji akciových titulů do portfolia. S nárůstem počtu IT zařízení a růstem používání rozličných online platforem se však při rozhodování o investicích do akcií dynamicky sdílejí informace nebo individuální úmysl prostřednictvím online komunit a sociálních sítí. Sociální data jsou široce používána k odhalení spokojenosti zákazníků s produkty nebo službami nebo jejich stížností, ačkoli použitelnost sociálních dat je omezena na správu zákazníků, řízení kvality nebo vývoj nových produktů. Zejména v posledních letech prudce roste zájem o zkoumání sociálních sítí a jejich použití k analýze trendů v investování do akcií, ačkoli většina investorů může při rozhodování o nákupu nebo prodeji akcií využít různé informace zveřejněné na webových stránkách nebo online komunitách. Vzhledem k tomu, že stále více osobních názorů je k dispozici online, ukazují různé studie, že tyto druhy analýz lze automatizovat a mohou přinést užitečné výsledky.

Těžba textu a následná analýza sentimentu, je počítačová analýza textových dat, která nabízí skvělou příležitost k prozkoumání vlivu na akciové trhy, potažmo jako vhodný prostředek pro jejich predikci. To je motivem k provedení systematický přehled o stavu a vývoji aplikací pro dolování textu a analýze sentimentu na akciových trzích a také jako systematický přehled pro ty, kteří uvažují o použití technik dolování textu ve svém vlastním výzkumu. Celkově lze říci, že tento přehled a výzkumné priority, které jsou stanoveny, mohou usnadnit smysluplné využití dolování textu a analýzu sentimentu, jak začátečníkům, tak zkušeným vědcům a přispět tak k metodickému bohatství v oblasti výzkumu na akciových trzích i mimo něj. Krom toho je přestaven také expertní systém fuzzy logiky, který je následně aplikován.

K prozkoumání současného stavu poznání byla provedena kritická rešerše literatury čerpající zdroje z primárních vědeckých databází. Bylo získáno velké množství

vědeckých článků a příspěvků se vztahem k předmětné problematice. Ačkoliv oblast sentimentální analýzy prostřednictvím textových dat na akciovém trhu se v posledních letech dostává do popředí, vyplývají jistá bílá místa neboli vědecká mezera v současně publikované literatuře, které je potřeba zaplnit, popřípadě současný hlavní proud rozšířit a posílit. Vysoké procento stávajících diskutovaných studií se spoléhala na extrakci textových informace z renomovaných serverů či sociální sítě Twitter či StockTweet či z internetových diskusních fór odděleně. Ve velmi nízkém zastoupení jsou k dispozici studie kombinující více textových zdrojů. Avšak dle provedeného sekundárního výzkumu dále vyplývá, že je vhodné pracovat s vícero textovými zdroji. Z toho důvodu je v praktické části pracováno s rozličnými textovými dokumenty s cílem zvýšit přesnost predikce akciového trhu. Navíc je patrný rozličný a nejednoznačný vliv sentimentu na akciové trhy. Jedni autoři tvrdí, že sentiment je prediktorem vývoje trhu, zatímco jiný poukazují na skutečnost, že trh reaguje dříve, než jsou zveřejněny články a příspěvky, na které by mohli investoři reagovat. Zásadní mezera ve výzkumu se objevila v aplikaci expertních systémů v předmětné problematice. Zcela opomíjena je fuzzy logika, která dosahuje na akciových trzích pozoruhodných výsledků. Tato disertační práce se pokusila zaplnit vědecké mezery ve výzkumu řešením těchto limitů a omezení, které byly identifikovány na základě kritického přezkumu literatury a řešeny prostřednictvím výzkumných otázek a hypotéz.

5.1.2 Souhrn výsledků provedeného empirického výzkumu

V empirické části výzkumu je nejprve pro analýzu sentimentu zvolena textová data pocházející z online finančních zpráv a příspěvků zveřejněných na finanční sociální síti StockTwits. Tato textová data včetně tržních dat, reprezentující přední americký akciový index S&P 500, jsou popsány prostřednictvím explorační analýzy. Následně je nezbytné textová data z nestrukturované podoby převést prostřednictvím kroků předzpracování na strukturovaná data, která lze zpracovat prostřednictvím metod strojového učení. Z kritického přezkumu vyplynula jistá bílá místa, která jsou v této části práce řešena. Jedním z problémů byla volba jediného lexikonu pozitivních a negativních slov ke stanovení sentimentu investorů. V této disertační práce je zvoleno lexikonů celkem šest. Jedná se o obecnější lexikony i lexikony speciálně vytvořené pro finanční doménu. Autorka práce se domnívá, stejně jako dříve citovaní autoři, že volba slovníků je jedním z aspektů, který má zásadní vliv na správně stanovené skóre sentimentu investorů. Druhým aspektem přímo souvisejícím s konkrétní hodnotou skóre sentimentu je volba klasifikátoru. V této práci je rozlišováno pouze mezi pozitivními a negativní slovy a prostřednictvím binárních klasifikátorů jsou stanoveny konkrétní číselné hodnoty ohodnocující průměrnou hodnotu sentimentu každé věty.

Konkrétně ke klasifikaci sentimentu jsou zvoleny metody SVM, k-NN, rozhodovací stromy, neuronové sítě a generalizovaný aditivní model. Je třeba konstatovat, že narozdíl od výzkumu Sieringa (2012) vyšší přesnosti klasifikace je patrná u obecných lexikonu namísto seznamů slov souvisejících s financemi jako je vysoce opěvovaný slovník Loughran-McDonald, který v případě binární klasifikace neobstál. V této práci byly porovnány různé algoritmy učení. Bylo zjištěno, že třída SVM fungovala nejlépe při řešení binární klasifikace sentimentu investorů.

Rozporuplnost současných výstupů se týká také vlivu sentimentu investorů na akciové trhy. To je zásadní problém, neboť v případě, že sentiment nemá vliv, popřípadě zaostává za vývoje akciového trhu, není možné jej integrovat do expertního modelu s cílem predikce. V takovém případě by to bylo kontraproduktivní. Prostřednictvím Grangerovy kauzality a vlnkové koherence je potvrzeno, že StockTwits a online finanční zprávy obsahují cenné informace a předcházejí obchodním aktivitám na akciových trzích. Al Nasseri a kol. (2015) také zjistili, že příspěvky investorů zveřejňované na sociální platformě určují následné pohyby na akciovém trhu. Což je v rozporu s výzkumem Hwang a Kim (2019), kteří tvrdí, že ceny akcií mají větší dopad na sentiment článků. Jinými slovy, ceny akcií reagovaly na sociální problémy dříve než články. Zjištěné výstupy této práce jsou v souladu se závěry Siering (2012).

Dále je vytvořen model expertního systému. Konkrétně je v disertační práci navržen a aplikován model fuzzy logiky. Kromě klasické type-1 fuzzy logiky je v posledních době patrný značný rozmach type-2 fuzzy logiky v různých vědních oborech. Fuzzy logika sama o sobě disponuje vynikajícími vlastnosti zejména při aplikaci na vysoce volatilní prostřední. Je navrženo několik modelů fuzzy logiky, přičemž největší pozornost je zaměřena úroveň nejistoty zahrnuté ve funkcích členství type-2 fuzzy logiky. Krom toho výkonnost fuzzy modelu je možné ovlivnit také typ fuzzy funkce členství. Je patrné, že v doposud publikovaných a zde řešených výzkumných prací tento model nebyl řešen. Z toho důvodu autorka práce považuje za vhodné prozkoumat nejen neurčitosti, ale také jednotlivé typy funkcí členství s ohledem na jeho výkonnost. K řešení je využit fuzzy model typu Sugeno, který je schopen integrovat adaptivní neuro-fuzzy inferenční systém. V rámci tohoto systému je stanovena expertní báze pravidel prostřednictvím neuronových sítí. Následně je implementována investiční strategie s cílem predikce vývoje akciového trhu prostřednictvím sentimentu investorů.

Xie a Jiang (2017) identifikovali zpoždění efektu zpráv na akciovém trhu na méně než dva dny. V této studii bylo implementováno zpoždění až tři dny pro účely predikce. Nicméně přesnost modelů predikce s delším, než jednodenním zpoždění se zásadně zhoršila. Z toho důvodu byla aplikována obchodní strategie výhradně na sentiment s denním zpožděním, neboť vyšší zpoždění neposkytuje relevantní výstupy. Lze konstatovat, že většina investorů reaguje na zveřejněné zprávy s maximálně jed-

nodenním zpožděním. Je patrné, že investoři si zveřejněné zprávy ověřují, popřípadě provádějí další potřebné analýzy, aby neprovedli unáhlený obchod.

Výzkum navazuje na Gross-Klussmann a kol. (2019), kteří tvrdí že signály odborného sentimentu mohou přinést vyšší výnosy očištěné o riziko než klasické signály založené na ceně. V této souvislosti naše výsledky ukazují, že sentiment generuje vyšší kumulovaný kapitál než benchmark či model bez sentimentu investorů. Zjištění tohoto výzkumného článku mohou přinést slibný pohled na potenciální poskytnutí mechanismu podpory investic pro analytiky, investory a jejich kolegy. V praxi by to mohlo být použito ke stanovení přesného času, kdy mají být akcie drženy, přidávány (kupovány) nebo odebírány (prodávány) z portfolia, čímž se pro investora získá maximální návratnost investice. To by mohlo ušetřit čas a úsilí a povede to k informovanějšímu investičnímu rozhodnutí na akciovém trhu.

5.1.3 Zodpovězení výzkumných otázek

V průběhu interpretace jednotlivých výsledků analýz a jejich syntézy byly již provedeny odpovědi potřebné k vyřešení výzkumného problému v různé formě výzkumných otázek, tak jak byly definované na začátku disertační práce. Je nutno zdůraznit, že první dvě otázky mají obecnější charakter a jejich odpovědi vychází z výsledků sekundárního výzkumu.

VO1: Jsou zjištěné poznatky vlivu sentimentu na akciové trhy jednoznačné a konzistentní?

Sekundární výzkum prostřednictvím literární rešerše provedený v kapitole 2 poukázal na různé rozporuplnost vlivu sentimentu na akciových trzích. Z literární rešerše je zřejmé že integrace sentimentu ze sociálních médií by mohla pomoci zlepšit predikci akciového trhu. Shi a kol. (2018) zjistili, že sentiment investorů má krátkodobé pozitivní účinky a střednědobé reverzní účinky na akciový trh. Avšak Hwang a Kim (2019) zjistili, že ceny akcií mají větší dopad na sentiment článků. Jinými slovy, ceny akcií reagovaly na sociální problémy dříve než články. Kromě toho nemusí být snadné předpovídat ceny akcií pomocí zpravodajských článků. Predikce ceny akcií je ještě obtížnější, pokud jsou zahrnuty minulé a budoucí články. Nguyen a kol. (2015) dále uvádějí, že jednoduchým předpokladem účinnosti funkce sentimentu je, že analýza sentimentu nemusí poskytnout žádné další informace, pokud lze pohyb akcií dobře předpovědět pouze historickou cenou. V takových případech mohou k předpovědi stačit pouze historické ceny a integrace sentimentu nemusí přesnost příliš zlepšit. Z toho důvodu lze i na základě názoru Gross-Klussmann a kol. (2019), tvrdit, že dosud neexistuje shoda ohledně implementace sentimentálních signálů do investičních strategií a jeho zřejmého a jednoznačného vlivu na pohyb akciového trhu. Z výše

uvedeného lze tvrdit, že doposud není ve vědecké komunitě shoda ohledně vlivu sentimentu na akciové trhy.

VO2: Jaká existují bílá místa v současném světovém výzkumu v oblasti predikce vývoje akciového trhu s ohledem na sentiment investorů?

Kritický přezkum literatury také odhaluje několik výzev v současném stavu poznání a identifikuje tzv. bílá místa či mezery ve světovém výzkumu v oblasti predikce vývoje akciového trhu s ohledem na sentiment investorů. Mezi ně patří věrohodnost finančních zpráv nebo, jinak řečeno, kvalita informací spojená se slovními projevy, jak uvádí Feuerriegel a Gordon (2018). I když se finanční zprávy jeví jako významná hnací síla oceňování akcií, stále existuje velká část nevysvětlitelných odchylek. Tuto složku se šumem lze snížit lepší technikou předpovídání, i když zbytková složka se šumem může být dokonce nepředvídatelná, zvláště když jsou signály nejasné nebo hlučné. Z matematického hlediska je výzkumné prostředí velmi náročné kvůli matici vysokodimenzionálního prediktoru, která může snadno vést k nadměrnému vybavení.

Co se týče standardních modelů užívaných při predikci vývoje akciového trhu, tak techniky navržené v literatuře poskytly přijatelné výsledky a zajímavé informace o sentimentální analýze a vztahu mezi akciovým trhem a sentimentální analýzou. Průzkum o těžbě textu pro predikci akciových trhů dospěl k závěru, že SVM a Naïve Bayes jsou výzkumníky silně upřednostňováni, zatímco NN a fuzzy logika jsou v této fázi v oblasti prediktivní těžby textů na akciových trzích výrazně nedostatečně prozkoumány, přestože NN ukázaly slibný potenciál pro textovou klasifikaci a analýzu sentimentu, kdežto fuzzy logika prozatím nebyla v tomto kontextu využita vůbec, i přes slibnou schopnost pracovat s vágními údaji, nejistým a chaotickým prostředí akciového trhu.

VO3:Jaký vliv na skóre sentimentu má lexikon pozitivních a negativních slov?

Z provedené analýzy sentimentu jsou patrné značné rozdíly v rozložení distribuci skóre sentimentu nejen u jednotlivých slovníků. Na první pohled je zřejmá téměř shodná klasifikace a ohodnocení sentimentu prostřednictvím SVM a generalizovaného aditivního modelu, při použití všech slovníků. U slovníku Harvard IV-4 je navíc patrné velice obdobné odhodnocení textového dokumentu i prostřednictvím neuronových sítí, naivního Bayese, k-nejbližších sousedů i rozhodovacích stromů, o čemž svědčí shodná distribuce skóre sentimentu. Ostatní slovníky vykazují značné vychýlení distribuce skóre sentimentu. Toto vychýlení je dominantní u Opinion Lexikonu, AFINN a finančních slovníků Loughran-McDonald a FinanceSentiment.

Z toho důvodu je nezbytné upravit rozdělení na pozitivní a negativní sentimentu a nespoléhat se na obecné pravidlo, že skóre větší než 0 svědčí o pozitivním sentimentu, zatímco skóre menší než 0 poukazuje na negativní sentiment. Z distribuce

rozdělení skóre sentimentu je tato skutečnost patrná a je nezbytné s tímto poznatkem dále pracovat, aby mohl být sentiment správně uplatněn pro následnou predikci akciových trhů. To znamená, že popisky tříd je nezbytné modifikovat a nastavit hranice pro jednotlivé třídy ručně. Cílem je nastavit hranice, které budou relevantní u každého slovníku a každého klasifikátoru, aby se dosáhlo nejvyšší přesnosti klasifikace.

Prostřednictvím Pearsonova koeficientu korelace byl spočten vliv mezi denním rozlišením časových řad sentimentu pro každý slovník a klasifikátor sentimentu. Velmi slabou korelaci (0 až 0.2) vykazuje slovník FinanceSentiment a Loughran-McDonald v porovnání se všemi ostatními slovníky bez ohledu na zvolnou techniku klasifikátoru. Většina slovníku vykazuje střední korelaci v rozmezí 0.3 až 0.6. Velmi silná korelace je patrná u slovníků Harvard IV-4, které jsou klasifikovány prostřednictvím SVM, neuronových sítí a naivního Bayese. Obdobně silnou závislost vykazuje také VADER a AFINN. Celkově lze říci, že slovníky obsahují různorodé pozitivní a negativní slova a tím pádem měří či odráží něco jiného a poskytují rozdílné skóre sentimentu.

H1: Speciální finanční slovník generuje přesnější skóre sentimentu jako obecný slovník.

Pro testování této hypotézy je zvolen Wilcoxonův neparametrický test. Z výsledků testu je patrné, že u tréninkových dat jsou p-hodnoty finančních slovníků velké a vyšší než prahová hodnota 0,05. Nelze tedy zamítnout nulovou hypotézu ve prospěch alternativní, která tvrdí, že specialně vytvořené slovníky pro finanční oblast generují přesnější skóre sentimentu investorů z textových dat. Není tedy signifikantně prokázána lepší schopnost finančních slovníků k označení textových dat na pozitivní a negativní slova. Celkově lze říci, že zde předložená data poskytují důkaz, že finanční slovníky negenerují přesnější hodnoty než obecné slovníky. Na základě výsledků Wilcoxova testu lze tvrdit, že FinanceSentiment lexikon je přesnější než slovník Loughran-McDonald.

VO4:Jak souvisí volba binární klasifikátor s přesností kalkulace skóre sentimentu s ohledem na zvolený lexikon?

Přesnost klasifikace hodnocení jako výborná by měla, přesáhnou hodnotu přesnosti 90%, což je případ například SVM, naivního Bayese, neuronových sítí a generalizovaného aditivního modelu. Dobrých výsledků přesnosti testu v rozmezí 70% až 80% dosáhly rozhodovací stromy a k-nejbližších sousedů. Tyto výsledky jsou patrné u všech zvolených lexikonů vyjma lexikonu od Loughran-McDonald, který je vytvořen speciálně pro finanční doménu. Obecně lze konstatovat, že výsledky naznačují nadřazenost naivní Bayes vůči všem ostatním klasifikátorům. Konkrétně dosahuje průměrně o 25% vyšší přesnosti než rozhodovací stromy, v průměru o 15%

k-nejbližších sousedů a o 6% neuronové sítě a generalizovaný aditivní model. Téměř totožných výsledků dosáhl tento model s SVM, kde se kvalita binární klasifikace lišila v průměru pouze o 2% ve prospěch naivního Bayese. Ovšem v případě, že není uvažován nejhůře natrénovaný slovník od Loughran-McDonald. Je nejvyšší přesnost klasifikace přisuzována metodě SVM, která je v průměru o 2% až 3% vyšší než u naivního Bayese a generalizovaného aditivního modelu. Lze konstatovat, že z hlediska binární klasifikace zřetelně dominují a povzbudivých výsledků dosahují SVM, naivní Bayes i generalizovaný aditivní model a jedná se o vhodné modely pro klasifikaci sentimentu.

VO5: Jaký vliv má skóre sentimentu na pohyb akciových trhů?

Jak je dobře známo, obchodní strategie, ať již se sentimentem nebo bez sentimentu investorů přináší zisk, pouze pokud by mohla poskytnout určitou předvídatelnost budoucích změn cen akcií, vzhledem k velké variabilitě dat na akciovém trhu. Pro následnou predikci je tudíž potřeba posoudit existenci kointegrace, kauzality a lead or lag efektu sentimentu, aby mohl být následně využit k predikci akciového trhu. K tomuto účelu jsou stanoveny dvě doprovodné výzkumné hypotézy, které byly otestovány.

H2: Mezi sentimentem investorů a výnosem akciového trhu existuje kointegrace.

Pro otestování této hypotézy je zvolen Engle-Grangerův test kointegrace. Na základě výstupů testu je na hladině významnosti 5% zamítnuta nulová hypotéz o neexistenci kointegračních vztahů mezi finanční časovou řadou a sentimentem investorů. Znamená to, že existují kointegrační vztahy mezi akciovými trhy a sentimentem investorů. Nicméně to samo o sobě neznamená, že sentiment je příčinou pohybu akciových trhů.

H3: Sentiment investorů způsobuje Grangerovu kauzalitu výnosů akciového trhu.

Na základě Grangerova testu kauzality je zamítnuta nulová hypotéza na hladině významnosti 5%, která udává, že sentimentu investorů nezpůsobuje Grangerova kauzalitu výnosů akciového trhu. Uvedený výsledek naznačuje, že výkonnost akciových trhů je ovlivněna sentimentem investorů. U akciového indexu je tato skutečnost patrná až do zpoždění 3 dny, zatímco u indexu VIX způsobuje sentiment kauzální vztah po všechna analyzovaná časová zpoždění. Je tak zřejmé, že akciové trhy vstřebávají názory a postoje investorů v relativně krátkém časovém horizontu, zatímco index strachu je ovlivňován po delší časové období než samotný akciový index představující barometr americké ekonomiky.

Dále je k zodpovězení VO5 ještě využita vlnková koherence. Prostřednictvím vlnkové koherence je prokázáno, že sentiment extrahovaných z textových dokumentů, které jsou zveřejňovány na internetu, vykazuje tzv. leading efekt neboli sentiment

investorů vede a určuje směr pohybu trhu.

VO6:Jaká míra neurčitosti ve fuzzy funkcích členství je vhodná pro predikci vývoje akciového trhu?

Lze vypozorovat, že type-2 fuzzy logika dosahuje stejných, ne-li lepších výsledků než type-1 fuzzy logika, jinými slovy obsažení jisté úrovně nejistoty ve funkcích členství poskytuje přesnější výsledky. Na základě trénování i testování type-1 a type-2 fuzzy modelu lze konstatovat, že nejvyšší přesnost predikce vývoje akciového trhu vykazuje model s integrací maximálně 10% nejistoty zahrnuté mezi dolní a horní funkcí členství. Nicméně je nezbytné upozornit, že s vyšší integrací nejistoty prudce klesá přesnost predikce. Z toho důvodu je na základě výstupů disertační práce doporučováno vhodně nastavit výši nejistoty, neboť zahrnutí příliš velké nejistoty snižuje přesnost predikce a může vést k chybným závěrům, potažmo k méně ziskové investiční strategii.

VO7: Který typ fuzzy funkce členství nejlépe odpovídá charakteru vstupních dat?

Povzbudivé výsledky poskytují jak Gaussovské, tak zvonkové funkce členství. Obdobný závěr poskytli i Janková a kol. (2021b), kteří tvrdí, že oba dva výše zmíněné typy funkcí patřičně reflektují popis vstupních dat pocházejících z akciového trhu a jsou vhodné zejména pro analýzu akciového trhu. Tyto funkce jsou identifikovány jako nejvhodnější z hlediska popisu charakteru vstupních dat, tak vykazované přesnosti predikce vývoj akciového trhu. Nicméně opětovně je nezbytné vhodné nastavení i úrovně nejistoty obsažené v těchto funkcích, neboť zejména u Gaussových a zvonkových funkcí členství je zjevný strmý propad přesnosti s narůstající nejistotou, který zkresluje průměrnou přesnost predikce u těchto typů funkcí členství. Z trojúhelníkových a trapezoidních funkcí tento pokles přesnosti není tak markantní. U trapezoidních funkcí je patrná vyvážená přesnost napříč úrovní nejistoty obsažené mezi jednotlivými horními a dolními funkcemi členství.

VO8: Jak ovlivňuje investiční strategii integrace sentimentu?

Strategie založená na sentimentu je schopna se lépe popasovat s nenadálými propady na trhu, resp. lépe využít následné ziskové příležitosti po propadu trhu. Ve skutečnosti investiční strategie bez sentimentu funguje lépe než jednoduchá srovnávací strategie založená na benchmarku, nicméně od konce května 2020 tato strategie zaostává za benchmarkem. Výnos benchmarku, tedy srovnávací základny je za testované období 9.82%. U ostatních modelů bez sentimentu se celkový výnos pohybuje v rozmezí 5% až 8%. Naproti tomu strategie založená na sentimentu investorů extrahovaném z textových zpráv se celkový výnos pohybuje v rozmezí 22% až 30%. Skutečnost, že strategie založená na sentimentu investorů také překonává strategii benchmarku, která nepoužívá informace o sentimentu, je počáteční indikací, že sen-

timent investora nese relevantní informace pro ziskovou investiční strategii. Jelikož strategie založená na sentimentu investorů je aktivní strategií, vykazuje vyšší obrat než spíše pasivní srovnávací strategie.

VO9: Který model poskytuje přesnější výsledky predikce vývoje akciového trhu?

Je zjištěno, že přidání indikátorů sentimentu investora vytvořených algoritmem binární klasifikace může vylepšit pouze modely s nízkou přesností předpovědi, jako jsou modely lineárního diskriminantu a Naïve Bayes, se zlepšením přesnosti o 0.2 až 0.4%. Přidání indikátoru sentimentu však nemůže zlepšit modely s vynikající přesností předpovědi, které jsou založeny na vstupních datech denních cen a objemu obchodování, jako jsou modely SVM, k-nejbližších sousedů, rozhodovací stromy a neuronové sítě. Navíc je prokázáno, že se zvyšujícím se zpožděním zahrnutým do modelu prudce klesá přesnost predikce až o 20%. Na základě výsledků prognózy výnosů akciového indexu, je přesnost modelů se zahrnutím zpožděného sentimentu o více než jeden den nepřijatelná; přesnost většiny modelů je mírně nad 50% a u některých modelů dokonce pod 50%. Výsledky ukazují, že sentiment extrahovaný textem má určitou mírnou prediktivní sílu pro budoucí výnosy akciového trhu ovšem s ohledem na velikost zpoždění. Dále bylo prokázáno, že modely fuzzy logiky jsou schopny vhodným nastavením funkcí členství a nejistoty v nich obsažených vylepšit predikci a jsou schopny konkurovat klasickým expertním modelům predikce, které jsou standardně využívané ve výzkumných studiích. Zde je dále nezbytné upozornit, že uspokojivých výsledků dosáhl zejména model type-2 fuzzy logiky s integrací 10% nejistoty, které je schopen konkurovat klasické type-1 fuzzy logice, ba co více, dokáže jeho přesnost převýšit.

5.2 Limity provedeného výzkumu

V následující části disertační práce jsou diskutovány hranice a limity, které byly zjištěny v průběhu vypracování předkládaného výzkumu.

Dostupnost textových dat

Jedním z hlavních problémů, je skutečnost, že textové datové soubory, které jsou aktuálně dostupné na trhu, jsou omezené nebo nejsou dostupné zdarma. Jedním z dalších problémů týkajících se tohoto procesu výzkumu je nedostatek vhodných a adekvátních databází obsahující již dříve shromážděná data. Zejména při vypracování této disertační práce se narazilo na omezení sociální sítě Twitter, která minulý rok začala prosazovat značné omezenou dostupnost příspěvků. V kombinaci s restriktivní politikou mnoho sociálních sítí a časovou nároční získání milionů příspěvků na sociálních sítí bylo nezbytné v této práci využít korpus od jiných autorů.

Technické vybavení

Nejdůležitějším omezením implementace výzkumného procesu je nedostatek vhodného softwaru pro dolování dat. Jelikož je výzkumná studie zcela založena na aplikaci technik dolování dat a textu, je nezbytné k implementaci různých kroků v procesu výzkumu využít mocné nástroje a software. Protože nebyl k dispozici žádný takový software pro dolování dat, je k napsání kódů pro algoritmy související s různými kroky v procesu výzkumu použít software MATLAB a jeho toolboxy. Rovněž se při zpracování narazilo na problém s hardwarovým vybavením při zpracování rozsáhlých textových datových souborů.

Časová a výpočetní náročnost

Samotné zpracování kódu a skriptů v softwaru pro mnoho dílčích úkolů, které v práci byly provedeny, je samo o sobě obtížný a časově náročný úkol. Při rozsáhlé datové sadě (v řádech milionů) dochází k neúměrnému zvyšování času potřebného k výpočtu. S větším souborem dat během několika měsíců by se mohl objevit optimální časový rámec agregace sentimentu. Pro obchodní model založený na tomto přístupu by měl optimální časový rámec velký význam pro maximalizaci zisku. Další časová náročnost je při získávání samotných textových dat.

Schopnost generalizace

Navíc, jak uvádí Swamy (2017), model analýzy sentimentu založený na názorech veřejnosti z jedné země nemusí fungovat ve všech případech (generalizace). Pochopení nebo analýza sentimentu lidí z globálního hlediska se proto může zdát složité, protože sentiment do značné míry ovlivňuje několik faktorů. Mezi nimi je zejména kulturní víra a praxe, rozdíly v politikách a regulačních rámcích mezi různými národy, rozdíly v úrovních propracovanosti akciových trhů, mimo jiné. Analýza sentimentu tedy může být jednodušší u zemí nebo regionů, které sdílejí kulturní spřízněnost nebo blízkost než z globální úrovně.

Výběr slovníků sentimentu

Jak bylo analyzováno v předkládané práci, zásadní roli v detekci sentimentu sehrává výběr slovníku. Slovník sentimentu je třeba vylepšit pro každý trh, potažmo odvětví pomocí informací z široké škály databází. Jak již bylo zmíněno výše, existuje dvojznačnost slov podle typu trhu či odvětví, protože každé prostředí má odlišné vlastnosti. Ačkoli testované slovník sentimentu v této práci byly jak obecné, tak přímo související s finanční doménou, měl by být dále tato oblast rozpracována a vylepšena, neboť je zásadní pro správnou identifikaci sentimentu.

Důvěryhodnost zveřejňovaných zpráv

Dalším zásadním limitujícím faktorem je důvěryhodnost informací, neboť někteří

investoři mohou zejména záměrně psát a poskytovat nesprávné informace za účelem manipulace s cenou akcií. To může způsobit nesprávné investiční aktivity a poškodit jednotlivé investory. Navíc je patrné, že některé klíčové události nejsou prezentovány veřejně nebo jsou skryté či zveřejněné s podstatným časovým odstupem, což ztěžuje detekci těchto skrytých událostí, což může způsobit vážnou ztrátu. Takové informace lze poté považovat za zastaralý zdroj pro predikci budoucího vývoje akciového trhu a prevenci velkých ztrát.

Vliv jiných faktorů

Ceny akcií jsou ovlivňovány mnoha faktory, včetně makroekonomie a různých novinek. Tento výzkum se však zaměřuje pouze na pocity uživatelů (s využitím jejich komentářů). Aby bylo dosaženo co nejlepšího možného predikčního modelu pro akcie, měli by být agregovány všechny informace o nich, včetně novinek a periodických zpráv společnosti, ale cílem práce je dosáhnout co nejlepší přesnosti pouze s využitím názorů a komentářů uživatelů v sociálních médiích.

Expertní model

Další omezení se týká zejména aplikovaného expertního modelu fuzzy logiky. Zásadním problémem této metody je skutečnost, že fuzzy logika nedisponuje žádnou pamětí. Kromě toho mohou být výsledky fuzzy modelu zkresleny díky volbě počtu členských funkcí a jejich atributů. Stanovení nebo „vyladění“ dobré funkce členství a fuzzy pravidel není vždy snadné. Dokonce i po rozsáhlém testování je obtížné říci, kolik funkcí příslušnosti je opravdu potřebných.

5.3 Námět pro budoucí výzkum

Tento výzkum identifikoval pozitivní vztah mezi sentimentem v sociálních médiích a jeho schopností předpovídat vývoj cen akciového trhu. Tato studie se zaměřila výhradně na amerických akciový trh v období 2018 až 2020. Výsledky ukazují, že minulé signály sentimentu investorů jsou konkurenceschopné ve srovnání se signály kombinujícími minulé výnosy z akciových indexů a lze je použít podobným způsobem. Z této studie vyplývají různé implikace a doporučení pro možné budoucí směřování výzkumu v této oblasti. Jak je patrné z množství literatury dostupné na sociálních médiích a vztahů sentimentu investorů v sociálních médiích a výkonem ceny akcií, jde stále o poměrně nový studijní obor. Výsledky předchozích studií i této studie otevírají dveře mnoha možnostem budoucího výzkumu.

Náměty výzkumu jsou následující:

- Tato studie by měla být opakována i v rozvíjejících se zemích, aby byly identifikovány potenciální rozdíly mezi rozvíjejícími se a rozvinutými zeměmi v kontextu predikce akciového trhu pomocí sentimentu v příspěvcích sociálních médiích a sledovat výkonnost expertního modelu.
- Při provádění analýzy sentimentu nebyly zohledněny demografické údaje uživatelů StockTwits. Budoucí výzkum by mohl kategorizovat tweety podle různých demografických kritérií a opakovat studii, aby se otestovalo, zda lze na základě různých demografických kategorií najít různé vztahy.
- Vhodné by bylo dále prozkoumat integraci expertních signálů generovaných ze sentimentu investorů s technickými signály, popřípadě se strukturovanými informacemi, jako jsou ekonomické ukazatele, aby se zjistilo, zda lze dosáhnout vynikajících výsledků. Tím je možné získat holističtější pochopení faktorů ovlivňujících změny v investičním procesu (teoretická perspektiva) a také zlepšili prediktivní potenciál stávajících technik (praktická perspektiva).
- Měla by být provedena podrobnější studie, ale tentokrát s hodinovými změnami sentimentu v sociálních médiích a změnami cen akcií. To akademikům umožní zjistit, jak rychle trh reaguje na sentiment sociálních médií. Investoři a manažeři mohli tyto poznatky využít k zavedení nezbytných zásahů, aby se negativní sentiment změnil v pozitivní sentiment.
- Budoucí studie by rovněž měly odstranit četné spamové tweety, než bylo možné provést přesnou analýzu. Rovněž dezinformace záměrně sdělované autory příspěvků včetně v nich obsažený sarkasmus a slangové výrazy mohou mít potenciálně velmi negativní dopad na přesnou kalkulaci sentimentu a následného určení správného vlivu na akciové trhy. Potenciální dopad stojí za prozkoumání.
- V této studii byla pozorována rozdílná přesnost jednotlivých slovníků. S tím souvisí naléhavá potřeba tvorby speciálního slovníku zaměřeného na výrazy vyskytující se v příspěvcích na konkrétní sociální platformě včetně jeho průběžného aktualizování o nově vznikající výrazy, jako je například „covid-19“ a jiné výrazy, které skrz pandemickou situaci se staly široce užívanými.

6 Přínosy disertační práce

Přínosy disertační práce spočívají v naplnění stanovených cílů a lze je formulovat jak v rovině teoretické, tak v rovině praktické. Vzhledem ke skutečnosti, že praxe nemže existovat bez teorie, je třeba zdůraznit vzájemnou provázanost obou přínosů. Na základ jejich průniku je pak možné formulovat nové poznatky, které lze využít v pedagogickém procesu. Přínos disertační práce tak lze spatřit ve třech dimenzích.

6.1 Přínosy pro vědu a výzkum

Tato studie poskytuje ucelený přehled a demonstraci v současnosti hojně rozšířených metod sloužících pro predikci vývoje akciového trhu a přispívá k podpoře rozhodování pomocí rozvíjejících se sociálních sítí pro výzkumné komunity. Zde vytvořené modely a přístup se mohou stát základem pro budoucí výzkum, kde vědci i odborníci z praxe mohou považovat za plodnou oblast výzkumu, aby věnovali pozornost rozmachu finančních blogů při porozumění významné roli sentimentu, zejména sentimentů v oblasti mikro blogů, v předpovídání cenového chování na akciových trzích. Tato studie speciálně přispívá ke dvěma různým skupinám výzkumné komunity: komunitě finančního výzkumu a komunitě pro dolování dat. Teoretické šetření prezentované v této práci přispívá k finanční literatuře při posilování vazeb s referenčními disciplínami při řešení probíhající debaty mezi hypotézou efektivního trhu (EMH). Poskytuje podporu teorii behaviorálního financování v existenci různých typů investorů na finančních trzích a jejich sentimentálního účinku jejich obchodního chování při ovlivňování cenových změn.

Mimo jiné lze přínosy disertační práce pro vědu a výzkum shrnout do následujících bodů:

- v práci je využita bibliometrická analýza identifikující vzájemné souvislosti vyselektovaných vědeckých a výzkumných prací v předmětné problematice, což může poskytnout inspiraci a vhled do vzestupu popularity problematiky predikce vývoje akciového trhu prostřednictvím sentimentální analýzy,
- je proveden detailní kritický přezkum mezinárodních vědeckých a výzkumných článků a příspěvků, včetně ucelených přehledů postupů a využitých přístupů na základě čeho jsou identifikována tzv. bílá míst a poskytuje tak náměť na budoucí směřování výzkumu,
- disertační práce analyzuje a hodnotí lexikony pozitivních a negativních slov včetně binárních klasifikátorů ke stanovení skóre sentimentu z online finančních zpráv a příspěvků na sociální sítích a poukazuje na rozdílnou výkon-

nost nejen binárních klasifikátorů, ale také samotných lexikonů a doporučuje věnovat této problematice větší pozornost, neboť nesprávně stanovené skóre sentimentu ovlivňuje následnou predikci vývoje trhu,

- vzhledem k nekonzistentním výsledkům je analyzován vliv sentimentu na akciové trhy, zde je potřeba zdůraznit zejména vlnkovou koherenci, která je při identifikaci vlivu indikátorů sentimentu (již dříve vytvořených indikátorů jako je VIX) již aplikována viz například Janková (2020b), nicméně disertační práce rozšiřuje tuto oblast zájmu o identifikaci vlivu vlastního sentimentu investorů, který byl zkonstruován z textových zpráv,
- stěžejní přínos pro vědeckou komunitu lze shledat v tvorbě expertního type-2 fuzzy modelu, které byl v této souvislosti nedostatečně prozkoumán a aplikován. V podstatě v analyzované literatuře tento model není vůbec zmíněn a je zcela opomíjen. Vyšší typ fuzzy logiky je detailně rozebrán z hlediska typu využitých funkcích členství a také úrovně nejistoty integrované do modelu,
- v neposlední řadě lze vědecký přínos spatřit v následné komparaci vytvořeného modelu type-1 a type-2 fuzzy logiky s celou řadu jiných expertních modelů, které jsou standardně, dle literatury, v této oblasti aplikovány. Je tak poskytnut rozsáhlý přehled výkonnosti různých modelů, který doposud nebyl vědeckou komunitou nabídnut.

6.2 Přínosy pro praxi

Práce si nekladla za cíl prezentovat zaručený recept na zcela úspěšnou predikcí vývoje akciového trhu s integrací sentimentu investorů, který je extrahován z textových online zpráv, protože akciový trh je charakteristický vysoce nestabilním vývojem a disponuje specifickými charakteristikami. Nicméně, předpoklad možné přenositelnosti získaných poznatků nabízí zejména široké investorské veřejnosti přínos minimálně charakteru podpory pro rozhodování ve formě expertního modelu s identifikací možných ziskových příležitostí, i když nezaručí postačující podmínky k definitivnímu úspěchu.

Získané poznatky mohou využít v zásadě dvě skupiny subjektů. První jsou investiční fondy, banky, centrální banky či výzkumná pracoviště, které dostanou do rukou nástroj ke sledování a vyhodnocování nálady na akciovém trhu a souvislosti této nálady s pohybem trhu. Druhou skupinou jsou individuální investoři či spekulanti, kteří se snaží načasovat nákup akcie na nejvhodnější dobu. Pro ty by bylo užitečné vědět, do jaké míry lze z publikovaných textů usoudit, jakým směrem a o kolik se v budoucnosti změní cena akcie. Je vytvořena metodika, která by mohla poskytnout

pokyny investorům a dalším finančním profesionálům pro konstrukci a vyvážení jejich investičních portfolií.

6.3 Přínosy pro pedagogickou oblast

Disertační práce může být využita jako ucelený zdroj informací pro výuku na Fakultě podnikatelské VUT v Brně pro prezenční i kombinovanou formu studia. Zejména pro studenty ekonomicko-manažerských oborů na bakalářském a magisterském stupni v předmětech zabývající se teorií finančních trhů či kvantitativními metodami v ekonomii. Ale také studentům manažersko-informatických oborů v předmětech jako jsou pokročilé metody v rozhodování a operační a systémová analýza. Výstupy jsou použitelné jak pro teoretické znalosti, které můžou poskytnout teoretický podklad, ale tak i prakticky využitelných informací, které umožní studentům lépe pochopit zkoumanou realitu. Nezanedbatelný přínos budou mít výstupy disertační práce při zpracovávání bakalářských a diplomových prací se zaměřením na analýzu a následně návrhů modelů na akciových trzích, potažmo i na jiných trzích finančního systému.

V neposlední řadě také pro širší odbornou veřejnost k seznámení se s trendy a metodami využívajících se v současnosti pro predikci vývoje akciového trhu. Tato práce může také nabídnout přehled analýzy sentimentu nováčkům a může poskytnout cenné informace pro zkušené vědce pro vzdělávací účely z oblasti aplikace fuzzy logiky. V širším měřítku lze práci použít jako podklad pro učebnici pro vysokoškoláky a absolventy počítačových věd, knihovníky nebo jako příručku pro odborníky pracující na příslušných aplikačních problémech při analýze a správě textových dat.

Závěr

Tématem disertační práce je „Expertní systém pro rozhodování na akciových trzích s využitím sentimentu investorů“ a pojednává o možnosti využití expertního systému fuzzy logiky pro predikci vývoje akciových trhů. Hlavním cílem disertační práce bylo navrhnout a aplikovat modelu expertního systému sloužící pro podporu investičního rozhodování na akciových trzích s využitím sentimentu investorů extrahovaného z nestrukturovaných textových zpráv. V práci je nejprve představeno teoretické pozadí výzkumu zaobírající se teorií finančních trhů včetně upozornění na neplatnost hypotézy efektivních trhů a prosazujícího se proudu behaviorálních financí. Behaviorální finance tvrdí, že investoři se řídí názorem ostatních a mají zjevnější proces šíření sentimentu. To zvyšuje nepopiratelnou důležitou prozkoumat proces tvorby sentimentu, který by pomohl pochopit, jak tento důležitý sentiment generuje a ovlivňuje obchodní strategie investorů. Jsou představeny základní principy a postup z oblasti těžby textu a extrakce sentimentu z textových dat včetně expertních systémů. V rámci sekundárního výzkumu je také proveden kritický přezkum literatury, který tvoří neodmyslitelnou součást každého výzkumu. Cílem této části bylo nalezení vědeckých mezer, tedy oblastí výzkumu, které jsou doposud neřešeny nebo řešeny nedostatečně. Pro nalezení těchto mezer je provedena bibiliometrická a obsahová analýza detailně popisují relevantní mezinárodní výzkumné studie. Mnoho studií a výzkumných prací naznačuje, že analýzu sentimentu veřejné nálady lze použít k předpovědi pohybu jednotlivých cen akcií, nicméně výsledky jsou stále rozporuplné. Na základě jejich identifikace jsou stanoveny výzkumné otázky a hypotézy přispívající k jedinečnosti tohoto výzkumu.

V předkládané disertační práci je pozornost zaměřena výhradně na akciový trh, ale lze jej použít na jiné trhy, které jsou do značné míry ovlivněny lidskými chování. Vzhledem k tomu, že akciový trh je citlivý na chování investorů i na stav firem, slouží sociální data jako důležitý ukazatel chování investorů. Aby bylo možné vytvořit vhodný expertní model, byl implementován výzkumný proces, který se skládá z různých kroků, včetně sběru dat, předzpracování dat, klasifikace sentimentu, predikce a vyhodnocení predikční schopnosti modelů. Tato disertační práce k analýze sentimentu investorů využila kanály online finanční platformy StockTwits a online finanční zprávy. Ke klasifikaci sentimentu je zvolen hybridní přístup kombinující binární klasifikátory a lexikony pozitivních a negativních slov. Z těchto dat je následně stanoveno skóre sentiment investorů, který slouží jako vstup do expertního modelu. Konkrétně v této disertační práci je za expertní systém zvolen model fuzzy logiky, jenž dosahuje na akciových trzích značně přijatelných výsledků. Jedinečnost tohoto výzkumu tak spočíval v návrhu, aplikaci a verifikaci modelu type-1 a type-2 fuzzy logiky. Zejména vyšší typ fuzzy logiky nebyl v této oblasti prozatím vůbec

prozkoumán. Z toho důvodu je vytvořeno několik modelů lišících se úrovní nejistoty a typ využité funkce členství. Je stanovena investiční strategie integrující do modelu tržní data a vytvořené skóre sentimentu. Je prokázáno, že integrací sentimentu do modelu se značně zvyšuje přesnost predikce, resp. integrací sentimentu je investor schopen generovat ziskovější investiční strategii. Výkonnost predikce fuzzy modelů je dále srovnávána s výkonností několika srovnávacích modelů, včetně SVM, k-NN, naivního Bayes a dalších. Z experimentů bylo pozorováno, že modely fuzzy logiky jsou schopny vhodným nastavením funkcí členství a nejistoty v nich obsažených vylepšit predikci a jsou schopny konkurovat klasickým modelům predikce, které jsou standardně využívané ve výzkumných studiích.

Zároveň z výsledků této práce je možno poskytnout určité návrhy v oblasti aplikovatelnosti expertních systémů integrací vytvořeného sentimentu investorů na akciových trzích a přinést tak nový pohled na využití fuzzy logiky s přínosy nejen v teoretické, ale i praktické rovině včetně využitelnosti ve vzdělávací činnosti fakulty.

Literatura

[1] AB. RAHMAN, Asyraf Safwan, Shuzlina ABDUL-RAHMAN a Sofianita MUTALIB. Mining Textual Terms for Stock Market Prediction Analysis Using Financial News. In: Soft Computing in Data Science. Singapore: Springer Singapore, 2017, p. 293-305. ISBN 9789811072413.

[2] ABDUL-MAGEED, Muhammad, Mona DIAB a Sandra KÜBLER. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer speech&language. LONDON: Elsevier, 2014, 28(1), p. 20-37. ISSN 0885-2308.

[3] ALLAHYARI, Mehdi, Seyedamin POURIYEH, Mehdi ASSEFI, Saied SAFAEI, Elizabeth D TRIPPE, Juan B GUTIERREZ a Krys KOCHUT. 2017

[4] AL NASSERI, Alya Al, Allan TUCKER a Sergio DE CESARE. Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. Expert systems with applications. OXFORD: Elsevier, 2015, 42(23), 9192-9210. ISSN 0957-4174.

[5] AL NASSERI, Alya, Allan TUCKER a Sergio DE CESARE. Big Data Analysis of StockTwits to Predict Sentiments in the Stock Market. In: Discovery Science. Cham: Springer International Publishing, 2014, s. 13-24. ISBN 9783319118116.

[6] AL-RAMAHI, Mohammad, Omar EL-GAYAR, Jun LIU, a Yen-Ling CHANG. Predicting big movers based on online stock forum sentiment analysis. 2015 Americas Conference on Information Systems, AMCIS.

[7] ALOSTAD, Hana a Hasan DAVULCU. Directional prediction of stock prices using breaking news on Twitter. Web Intelligence. 2017, 15(1), 1-17. ISSN 2405-6456.

[8] ANANDARAJAN, Murugan, Chelsey HILL a Thomas NOLAN. Practical Text Analytics. 2. Cham: Springer International Publishing, 2019. ISBN 9783319956626.

[9] ANGELOU, M. Maya Angelou: Her quotes, poetry and prose. © 2011[cit. 2021-8-2]. Dostupné z: https://www.cbsnews.com/news/maya-angelou-quotes-poetry-and-prose/

[10] ANTWEILER, Werner a Murray Z FRANK. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of finance (New York). 350 Main Street , Malden , MA 02148 , USA and 9600 Garsington Road , Oxford OX4 2DQ , UK: Blackwell Publishing, 2004, 59(3), 1259-1294. ISSN 0022-1082.

[11] ANTONS, David, Eduard GRÜNWALD, Patrick CICHY a Torsten Oliver SALGE. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D management. Oxford: Wiley Subscription Services, 2020, 50(3), 329-351. ISSN 0033-6807.

[12] APPEL, Orestes, Francisco CHICLANA, Jenny CARTER a Hamido FUJITA. A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-based systems. AMSTERDAM: Elsevier B.V, 2016, 108, 110-124. ISSN 0950-7051.

[13] ARLT, Josef a Markéta ARLTOVÁ. Ekonomické časové řady: [vlastnosti, metody modelování, příklady a aplikace]. Praha: Grada, 2007, 285 s. ISBN 978-80-247-1319-9.

[14] BAHARUDIN, Baharum, Lam Hong LEE a Khairullah KHAN. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology. 2010, 1(1), 4-20. ISSN 1798-2340.

[15] BAKER, Malcolm a Jeffrey WURGLER. Investor Sentiment and the Cross-Section of Stock Returns. The Journal of finance (New York). Malden, USA: Blackwell Publishing, 2006, 61(4), 1645-1680. ISSN 0022-1082.

[16] BATRA, Rakhi a Sher Muhammad DAUDPOTA. Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In: 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). IEEE, 2018, 2018, s. 1-5. ISBN 978-1-5386-1370-2.

[17] BHATTACHARYYA, S., DUTTA, P. Fuzzy Logic. KHAN, Mohammad Ayoub a Abdul Quaiyum ANSARI, ed. Handbook of Research on Industrial Informatics and Manufacturing Intelligence. IGI Global, 2012, 33-71. Advances in Civil and Industrial Engineering.

[18] BIFET, Albert, Gianmarco DE FRANCISCI MORALES, Jesse READ, Geoff HOLMES a Bernhard PFAHRINGER. Efficient Online Evaluation of Big Data Stream Classifiers. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [online]. New York, NY, USA: ACM, 2015, 2015-08-10, s. 59-68. ISBN 9781450336642.

[19] BIRBECK, Ellie a Dave CLIFF. Using Stock Prices as Ground Truth in Sentiment Analysis to Generate Profitable Trading Signals. IDEAS Working Paper Series from RePEc. 2018.

[20] BOIY, Erik a Marie-Francine MOENS. A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval. 2009, 12(5), 526-558. ISSN 1386-4564.

[21] BOLLEN, Johan, Huina MAO a Xiaojun ZENG. Twitter mood predicts the stock market. Journal of computational science. AMSTERDAM: ELSEVIER, 2011, 2(1), 1-8. ISSN 1877-7503.

[22] BONTA, Venkateswarlu, JANARDHAN, N. K. N. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. Asian Journal of Computer Science and Technology, 2019, 8(S2), 1-6. ISSN 2249-0701

[23] BOUKTIF, Salah, Ali FIAZ a Mamoun AWAD. Augmented Textual Features-Based Stock Market Prediction. IEEE access. Piscataway: IEEE, 2020, 8, 40269-40282. ISSN 2169-3536.

[24] BOUKTIF, Salah, Ali FIAZ a Mamoun AWAD. Stock Market Movement Prediction using Disparate Text Features with Machine Learning. In: 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). IEEE, 2019, 2019, s. 1-6. ISBN 978-1-7281-0003-6.

[25] BOYACIOGLU, Melek Acar a Derya AVCI. An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange. Expert systems with applications. OXFORD: Elsevier, 2010, 37(12), 7908-7912. ISSN 0957-4174.

[26] BRAŞOVEANU L. O. Piaţa de capital, Ed ASE, 2011, Bucureşti.

[27] BU H., XIE Z., LI J.H., WU J.J. Investor sentiment extracted from internet stock message boards and its effect on chinese stock market. Journal of Management Science in China, 2018, 21 (4), 86-101.

[28] BUDÍKOVÁ, Marie, Maria KRÁLOVÁ a Bohumil MAROŠ. Průvodce základními statistickými metodami. Praha: Grada, 2010, 272 s. ISBN 978-80-247-3243-5.

[29] BUSTOS, O a A POMARES-QUIMBAYA. Stock market movement forecast: A Systematic review. Expert systems with applications. Elsevier, 2020, 156, 113464. ISSN 0957-4174.

[30] CASTILLO, Oscar, Patricia MELIN a Juan R CASTRO. Computational intelligence software for interval type-2 fuzzy logic. Computer applications in engineering education. 2013, 21(4), 737-747. ISSN 1061-3773.

[31] CASTILLO, Oscar, Patricia MELIN, Janusz KACPRZYK a Witold PEDRYCZ. Type-2 Fuzzy Logic: Theory and Applications. In: 2007 IEEE International Conference on Granular Computing (GRC 2007). IEEE, 2007, 2007, s. 145-145. ISBN 0-7695-3032-X.

[32] CNBC: Stock Markets, Business News, Financials, Earnings [online]. © 2021[cit. 2021-8-1]. Dostupné z: https://www.cnbc.com/world/?region=world

[33] COYNE, Scott, Praveen MADIRAJU a Joseph COELHO. Forecasting Stock Prices Using Social Media Analysis. In: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech). IEEE, 2017, 2017, s. 1031-1038. ISBN 978-1-5386-1956-8.

[34] ČERNOHORSKÝ, Jan. Finance: od teorie k realitě. Praha: Grada Publishing, 2020. Finance (Grada). ISBN 978-80-271-2215-8.

[35] DAJČMAN, Silvo. Interdependence Between Some Major European Stock Markets - A Wavelet Lead/Lag Analysis. Prague Economic Papers. 2013, 22(1), 28-49. ISSN 12100455.

[36] DARABI, R., CHENARY-E BUKAT, H., VALIKHANI, M. Dimensions and Approaches of Behavioral Finance Theory, Iranian Accounting and Auditing Studies, 2016, 5.

[37] DARŠKUVIENĖ, Valdonė. Financial Markets. Vytautas Magnus University, 2010.

[38] DAŘENA, František, Jonáš PETROVSKY, Jan ŽIŽKA a Jan PŘICHYSTAL. Machine Learning-Based Analysis of the Association Between Online Texts and Stock Price Movements. Inteligencia Artificial. 2018, 21(61), 95-110. ISSN 1988-3064.

[39] DAS, Subrata a Arup DAS. Fusion with sentiment scores for market research. In 2016 19th International Conference on Information Fusion (FUSION), 2016, pp. 1003-1010.

[40] DAVE, Kushal, Steve LAWRENCE a David M. PENNOCK. Mining the peanut gallery. In: Proceedings of the twelfth international conference on World Wide Web - WWW. New York, New York, USA: ACM Press, 2003, 2003, s. 519-. ISBN 1581136803.

[41] DENSCOMBE, Martyn. The Good Research Guide (3rd edn). Buckingham: Open University Press, 2007.

[42] DERAKHSHAN, Ali a Hamid BEIGY. Sentiment analysis on stock social media for stock price movement prediction. Engineering applications of artificial intelligence. OXFORD: Elsevier, 2019, 85, 569-578. ISSN 0952-1976.

[43] DOMENICONI, Giacomo, Gianluca MORO, Andrea PAGLIARANI a Roberto PASOLINI. Learning to Predict the Stock Market Dow Jones Index Detecting and Mining Relevant Tweets. In: Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. SCITEPRESS - Science and Technology Publications, 2017, 2017, s. 165-172. ISBN 978-989-758-271-4.

[44] DOSTÁL, Petr. Advanced decision making in business and public services. Brno: Akademické nakladatelství CERM, 2011, 167 s., grafy, tab. ISBN 978-80-7204-747-5

[45] EDMANS, Alex, Diego GARCÍA a Øyvind NORLI. Sports Sentiment and Stock Returns. The Journal of finance (New York). Malden, USA: Blackwell Publishing, 2007, 62(4), 1967-1998. ISSN 0022-1082.

[46] ELIACIK, Alpaslan Burak a Nadia ERDOGAN. Influential user weighted sentiment analysis on topic based microblogging community. Expert systems with applications. OXFORD: Elsevier, 2018, 92, 403-418. ISSN 0957-4174.

[47] ELIACIK, Alpaslan Burak a Nadia ERDOGAN. User-weighted sentiment analysis for financial community on Twitter. In: 2015 11th International Conference on Innovations in Information Technology (IIT). IEEE, 2015, 2015, s. 46-51. ISBN 978-1-4673-8509-1.

[48] ELTON, E DWIN. J., Martin J. GRUBER, Stephen J. BROWN a William N. GROETZMANN. Modern Portfolio Theory and Investment Analysis, chapter Financial markets, 2011, (pp. 11–27). John Wiley&Sons: New York, 5. edition, ISBN 978-1-118-46994-1

[49] FAMA, Eugene F. Efficient Capital Markets: II. The Journal of Finance. 1991, 46(5). ISSN 00221082.

[50] FAMA, Eugene F. EFFICIENT CAPITAL MARKETS: A REVIEW OF THEORY AND EMPIRICAL WORK. The Journal of finance (New York). Oxford, UK: Blackwell Publishing, 1970, 25(2), 383-417. ISSN 0022-1082.

[51] FAMA, Eugene F. The Behavior of Stock-Market Prices. The Journal of business (Chicago, Ill.). Chicago, Ill: University of Chicago Press, 1965, 38(1), 34-105. ISSN 0021-9398.

[52] FEUERRIEGEL, Stefan a Julius GORDON. Long-term stock index forecasting based on text mining of regulatory disclosures. Decision Support Systems. AMSTERDAM: Elsevier B.V, 2018, 112, 88-97. ISSN 0167-9236.

[53] Finance Yahoo!: Standard&Poor's 500. Finance Yahoo! [online]. © 2021 [cit. 2021-8-1]. Dostupné z: https://finance.yahoo.com/quote/%5EGSPC?p=%5EGSPC

[54] FRIESEN, Geoffrey a Paul A WELLER. Quantifying cognitive biases in analyst earnings forecasts. Journal of financial markets (Amsterdam, Netherlands). Elsevier B.V, 2006, 9(4), 333-365. ISSN 1386-4181.

[55] GABAIX, Xavier, David I. LAIBSON, Guillermo MOLOCHE a Stephen Ernest WEINBERG. The Allocation of Attention: Theory and Evidence. SSRN Electronic Journal. ISSN 1556-5068.

[56] GAVETTI, Giovanni, Daniel A LEVINTHAL a Jan W RIVKIN. Strategy making in novel and complex worlds: the power of analogy. Strategic management journal. Chichester, UK: John Wiley&Sons, 2005, 26(8), 691-712. ISSN 0143-2095.

[57] GHAURI, Pervez a Kings GRØNHAUG. Research Methods in Business Studies: A Practical Guide (3rd edn). Harlow: Financial Times Prentice Hall, 2005.

[58] GIARRATANO, Josehph a Gary RILEY. Expert Systems-Principles and Programming, 4th ed., Course Technology, 2004. ISBN 978-0534384470.

[59] GRANGER, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica. Menasha, Wis: The Econometric Society, 1969, 37(3), 424-438. ISSN 0012-9682.

[60] GROSS-KLUSSMANN, Axel, Stephan KÖNIG a Markus EBNER. Buzzwords build momentum: Global financial Twitter sentiment and the aggregate stock market. Expert systems with applications. OXFORD: Elsevier, 2019, 136, 171-186. ISSN 0957-4174.

[61] GROSSMAN, Sanford J a Joseph E STIGLITZ. On the Impossibility of Informationally Efficient Markets. The American economic review. Menasha, Wis: The American Economic Association, 1980, 70(3), 393-408. ISSN 0002-8282.

[62] GROTH, Sven S a Jan MUNTERMANN. An intraday market risk management approach based on textual analysis. Decision Support Systems. AMSTERDAM: Elsevier B.V, 2011, 50(4), 680-691. ISSN 0167-9236.

[63] Guardian: News, sport and opinion from the Guardian's US edition [online]. © 2021 [cit. 2021-8-1]. Dostupné z: https://www.theguardian.com/international

[64] GUTHRIE, D., a kol. A closer look at skip-gram modelling. In:Proceedings of the5th international Conference on Language Resources and Evaluation (LREC-2006), 1–4.

[65] HAJEK, Petr. Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. Neural computing&applications. London: Springer London, 2018, 29(7), 343-358. ISSN 0941-0643.

[66] HÁJEK, Petr a Jana BOHÁČOVÁ. Predicting Abnormal Bank Stock Returns Using Textual Analysis of Annual Reports – a Neural Network Approach. In: Engineering Applications of Neural Networks. Cham: Springer International Publishing, 2016, s. 67-78. ISBN 9783319441870. ISSN 1865-0929.

[67] HAKIM, Catherine. Research Design: Successful Designs for Social and Economic Research (2nd edn). London: Routledge 2000. ISBN 9780415223126

[68] HAO, Pei-yi, Chien-feng KUNG, Chun-yang CHANG a Jen-bing OU. Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane. Applied soft computing. Elsevier B.V, 2021. ISSN 1568-4946.

[69] HAN, Eui-Hong (Sam), George KARYPIS a Vipin KUMAR. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. In: Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, s. 53-65. ISBN 9783540419105.

[70] HATZIVASSILOGLOU, Vasileios a Kathleen R. MCKEOWN. Predicting the semantic orientation of adjectives. In: Proceedings of the 35th annual meeting on Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 1997, 1997, s. 174-181.

[71] HENDL, Jan. Kvalitativní výzkum: základní teorie, metody a aplikace. 2., aktualiz. vyd. Praha: Portál, 2008, 407 s. ISBN 978-80-7367-485-4.

[72] HIRSHLEIFER, David. Behavioral Finance. Annual Review of Financial Economics. 2015, 7(1), 133-159. ISSN 1941-1367.

[73] HU, Yong, Bin FENG, Xiangzhou ZHANG, E.W.T NGAI a Mei LIU. Stock trading rule discovery with an evolutionary trend following model. Expert systems with applications. OXFORD: Elsevier, 2015, 42(1), 212-222. ISSN 0957-4174

[74] HU, Minqing a Bing LIU. Mining and summarizing customer reviews. In: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04. New York, New York, USA: ACM Press, 2004, 2004, s. 168-. ISBN 1581138889.

[75] HUANG, Chenn-jung, Jia-jian LIAO, Dian-xiu YANG, Tun-yu CHANG a Yun-cheng LUO. Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert systems with applications. OXFORD: Elsevier, 2010, 37(9), 6409-6413. ISSN 0957-4174.

[76] HUŠEK, Roman. Ekonometrická analýza. Praha: Oeconomica, 2007, 367 s. ISBN 978-80-245-1300-3.

[77] HUTTO, C. J., GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media, 2014.

[78] HWANG, Eunjeong a Yong-hyuk KIM. Interdependency between the Stock Market and Financial News. 2019. ArXiv.org.

[79] CHANG, Pei-chann, Chin-yuan FAN a Jun-lin LIN. Trend discovery in financial time series data using a case based fuzzy decision tree. Expert systems with applications. OXFORD: Elsevier, 2011, 38(5), 6070-6080. ISSN 0957-4174.

[80] CHEN, Mu-yen a Ting-hsuan CHEN. Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena. Future generation computer systems. AMSTERDAM: Elsevier B.V, 2019, 96, 692-699. ISSN 0167-739X

[81] CHEN, Chun-Hao a Ping SHIH. A Stock Trend Prediction Approach based on Chinese News and Technical Indicator Using Genetic Algorithms. In: 2019 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2019, 2019, s. 1468-1472. ISBN 978-1-7281-2153-6.

[82] CHEN, Yang, Dazhi WANG a Wu NING. Forecasting by TSK general type-2 fuzzy logic systems optimized with genetic algorithms. Optimal control applications&methods [online]. HOBOKEN: WILEY, 2018, 39(1), 393-409. ISSN 0143-2087

[83] CHEN, Hsinchun a David ZIMBRA. AI and Opinion Mining. IEEE Intelligent Systems. 2010, 25(3), 74-80. ISSN 1541-1672.

[84] CHENLO, Jose M a David E LOSADA. An empirical study of sentence features for subjectivity and polarity classification. Information sciences. NEW YORK: Elsevier, 2014, 280, 275-288. ISSN 0020-0255.

[85] CHO, Heeryon, Songkuk KIM, Jongseo LEE a Jong-seok LEE. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. Knowledge-based systems. AMSTERDAM: Elsevier B.V, 2014, 71, 61-71. ISSN 0950-7051.

[86] CHEVRIE, François a François GUELY. Fuzzy logic- Cahier technique no 191, first issued, 1998.

[87] IGNATOW, Gabe a Rada MIHALCEA. Text Mining: A Guidebook for the Social Sciences . 2455 Teller Road, Thousand Oaks California 91320 : SAGE Publications, 2017. ISBN 9781483369341.

[88] IKONOMAKIS, Emmanouli, Sotiris M. KOTSIANTIS a TAMPAKAS, V. Text Classification Using Machine Learning Techniques, Wseas Transactions on Computers, 2005, 8(4), 966-974.

[89] JAGGI, Mukul, Priyanka MANDAL, Shreya NARANG, Usman NASEEM a Matloob KHUSHI. Text mining of stocktwits data for predicting stock prices. Applied System Innovation. Basel: MDPI, 2021, 4(1), 1-22. ISSN 2571-5577.

[90] JAMMALAMADAKA, Rao S., Jinwen QIU a Ning NING. Predicting a stock portfolio with the multivariate bayesian structural time series model: Do news or emotions matter? nternational Journal of Artificial Intelligence, 2019, 17(2), ISSN 0974-0635

[91] JANÍČEK, Přemysl a Jiří MAREK. Expertní inženýrství v systémovém pojetí. Praha: Grada, 2013, 592 s. ISBN 978-80-247-4127-7.

[92] JANKOVÁ, Zuzana a Petr DOSTÁL. Type-2 Fuzzy Expert System Approach for Decision-Making of Financial Assets and Investing under Different Uncertainty. Mathematical problems in engineering. Hindawi, 2021a, 2021, 1-16. ISSN 1024-123X.

[93] JANKOVA, Zuzana, Dipak Kumar JANA a Petr DOSTAL. Investment decision support based on interval type-2 fuzzy expert system. Inzinerinee ekonomika. 2021b, 32(2), 118-129. ISSN 1392-2785.

[94] JANKOVÁ, Z. Sentiment on the Stock Markets: Evidence from the Wavelet Coherence Analysis. Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration. 2020a, roč. 28, č. 3, s. 1-10. ISSN: 1804-8048.

[95] JANKOVÁ, Z. Literature Review of Fundamental and Technical Indicators Prediction of Financial Market Using Artificial Intelligence Technique. In Conference Proceedings DOKBAT 16th Annual International Bata Conference for Ph.D. Students and Young Researchers. 16. Zlin, Czech Republic: Tomas Bata University of Zlin, 2020b. s. 242-254. ISBN: 9788074549359.

[96] JANKOVÁ, Zuzana a Petr DOSTÁL. Expertní systém Type-2 fuzzy logika pro investiční analýzu. Scientific Papers of the University of Pardubice. Series D. Faculty of Economics and Administration. Pardibuce: University of Pardubice, Faculty of Economics and Administration, 2019, (47), 79. ISSN 1211-555X.

[97] JENSEN, Michael C. Some anomalous evidence regarding market efficiency. Journal of financial economics. Amsterdam: Elsevier B.V, 1978, 6(2), 95-101. ISSN 0304-405X

[98] JOSHI, N. S., ITKAT, S. A. A Survey on Feature Level Sentiment Analysis, In (IJCSIT) International Journal of Computer Science and Information Technologies, 2014, 5(4), 5422-5425.

[99] JUAREZ-OROZCO, Luis Eduardo, Octavio MARTINEZ-MANZANERA, Sergey V. NESTEROV, Sami KAJANDER a Juhani KNUUTI. The machine learning horizon in cardiac hybrid imaging. European Journal of Hybrid Imaging. 2018, 2(1). ISSN 2510-3636.

[100] JURA, P. Fuzzy logika v modelování a řízení dynamických systémů: současný stav, perspektivy a výuka: teze přednášky k profesorskému jmenovacímu řízení v oboru Technická kybernetika. 2005. Brno: VUTIUM.

[101] JURAFSKY, Dan a James H MARTIN. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. Upper Saddle River: Pearson Education, 2008, xxxi, 988 s. ISBN 978-0-13-187321-6.

[102] KAHNEMAN, Daniel. Thinking, fast and slow. New York: Farrar, Straus and Giroux, 2011, 499 s. ISBN 978-0-374-27563-1.

[103] KAPLANSKI, Guy a Haim LEVY. Sentiment and stock prices: The case of aviation disasters. Journal of financial economics. Elsevier B.V, 2010, 95(2), 174-201. ISSN 0304-405X.

[104] KARNIK, Nilesh N a Jerry M MENDEL. Centroid of a type-2 fuzzy set. Information sciences. NEW YORK: Elsevier, 2001, 132(1), 195-220. ISSN 0020-0255.

[105] KAUR, Amandeep a Vishal GUPTA. A Survey on Sentiment Analysis and Opinion Mining Techniques, Journal of Emerging Technologies in Web Intelligence, 2013, 5, 367-371.

[106] KAYACAN, Erdal, Andriy SARABAKHA, Simon COUPLAND, Robert JOHN a Mojtaba Ahmadieh KHANESAR. Type-2 fuzzy elliptic membership

functions for modeling uncertainty. Engineering applications of artificial intelligence. OXFORD: Elsevier, 2018, 70, 170-183. ISSN 0952-1976

[107] KEARNEY, Colm a Sha LIU. Textual sentiment in finance: A survey of methods and models. International review of financial analysis. NEW YORK: Elsevier, 2014, 33, 171-185. ISSN 1057-5219.

[108] KEDIA, Amand a Mayank RASU. Hands-On Python Natural Language Processing: Explore tools and techniques to analyze and process text with a view to building real-world NLP applications. Packt Publishing Ltd. 2020. ISBN 978-1838989590.

[109] KERVIN, John B. Methods for Business Research (2nd edn). New York: HarperCollins. 1999. ISBN 978-0060436360.

[110] KHANESAR, M. A, E KAYACAN, M TESHNEHLAB a O KAYNAK. Analysis of the Noise Reduction Property of Type-2 Fuzzy Logic Systems Using a Novel Type-2 Membership Function. IEEE transactions on systems, man and cybernetics. Part B, Cybernetics. PISCATAWAY: IEEE, 2011, 41(5), 1395-1406. ISSN 1083-4419.

[111] KHEDR, Ayman E, S.E SALAMA a Nagwa YASEEN. Predicting stock market behavior using data mining technique and news sentiment analysis. International journal of intelligent systems and applications. Hong Kong: Modern Education and Computer Science Press, 2017, 9(7), 22-30. ISSN 2074-904X

[112] KIM, Misuk, Eunjeong Lucy PARK a Sungzoon CHO. Stock price prediction through sentiment analysis of corporate disclosures using distributed representation. Intelligent data analysis. AMSTERDAM: IOS PRESS, 2018, 22(6), 1395-1413. ISSN 1088-467X.

[113] KIM, Soon-ho a Dongcheol KIM. Investor sentiment from internet message postings and the predictability of stock returns. Journal of economic behavior&organization. AMSTERDAM: Elsevier B.V, 2014, 107, 708-729. ISSN 0167-2681.

[114] KING, Robert G a Ross LEVINE. Finance and Growth: Schumpeter Might Be Right. The Quarterly journal of economics. CARY: MIT Press, 1993, 108(3), 717-737. ISSN 0033-5533.

[115] KLÍMEK, Petr a Roman KASAL. Počítačové zpracování dat v programu Statistica: studijní pomůcka pro distanční studium. Zlín: Univerzita Tomáše Bati ve Zlíně, 2007. ISBN 9788073185268.

[116] KOHAJDA, Michael a Milan BAKEŠ. Finanční systém, finanční trh nebo kapitálový trh jako definiční znak podoboru finančního práva? AUC IURIDIC. 2018, 2018(1), 27-34. ISSN 2336-6478.

[117] KOSKO, Bart Fuzzy engineering. Upper Saddle River, NJ: Prentice Hall, 1997, ISBN 978-0131249912.

[118] KOSTOLANY, André. Kostolanyho burzovní seminář pro kapitálové investory a spekulanty. [Havlíčkův Brod]: Mirage, 2000. ISBN 80-238-5969-2.

[119] KRAUS, Mathias a Stefan FEUERRIEGEL. Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems. AMSTERDAM: Elsevier B.V, 2017, 104, 38-48. ISSN 0167-9236.

[120] KUČERA, J., RADVAN, E., Filosofické aspekty metodologie vědy (Vybrané problémy), Brno: Vojenská akademie v Brně, 2000, s. 159.

[121] KUMAR, Akshi a Teeja Mary SEBASTIAN. Sentiment Analysis: A Perspective on its Past, Present and Future. International journal of intelligent systems and applications. Hong Kong: Modern Education and Computer Science Press, 2012, 4(10), 1-14. ISSN 2074-904X

[122] LEONDES, Cornelius. Fuzzy logic and expert systems applications. Academic Press, 1998. ISBN 9780080553191

[123] LI, Yelin, Hui BU, Jiahong LI a Junjie WU. The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. International journal of forecasting. Elsevier B.V, 2020, 36(4), 1541-1562. ISSN 0169-2070.

[124] LI, Xiao, Dehua SHEN, Mei XUE a Wei ZHANG. Daily happiness and stock returns: The case of Chinese company listed in the United States. Economic modelling. AMSTERDAM: Elsevier B.V, 2017, 64, 496-501. ISSN 0264-9993.

[125] LIEN MINH, Dang, Abolghasem SADEGHI-NIARAKI, Huynh Duc HUY, Kyungbok MIN a Hyeonjoon MOON. Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network. IEEE access. PISCATAWAY: IEEE, 2018, 6, 55392-55404. ISSN 2169-3536.

[126] LINDEROVÁ, Ivica, Petr SCHOLZ a Michal MUNDUCH. Úvod do metodiky výzkumu. Jihlava: Vysoká škola polytechnická Jihlava, 2016. ISBN 978-80-88064-23-7.

[127] LIŠKA, Václav. Doctorandus: průvodce budoucích Ph.D. Praha: Professional Publishing, c2004. ISBN 8086419606.

[128] LIU, Bing. Sentiment Analysis. Cambridge: Cambridge University Press, 2015. ISBN 9781139084789

[129] LIU Bing.Handbook of Natural Language Processing, Marcel Dekker, Inc. New York, NY, USA. 2009, ISBN 978-1-4200-8593-8.

[130] LIU, Bing. Web data mining: exploring hyperlinks, contents, and usage data. Berlin: Springer, 2007, xix, 532 s. ISBN 978-3-540-37881-5.

[131] LIU, Bing a Lei ZHANG. A Survey of Opinion Mining and Sentiment Analysis. AGGARWAL, Charu C. a ChengXiang ZHAI, ed. Mining Text Data. Boston, MA: Springer US, 2012, 2012-1-7, s. 415-463. ISBN 978-1-4614-3222-7

[132] LONG, Wen, Ye-ran TANG a Ying-jie TIAN. Investor sentiment identification based on the universum SVM. Neural computing&applications. London: Springer London, 2018, 30(2), 661-670. ISSN 0941-0643.

[133] LOUGHRAN, TIM a BILL MCDONALD. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance. 2011, 66(1), 35-65. ISSN 00221082.

[134] LUCAS, P. Principles of Expert Systems. Addison-Wesley, 1991.

[135] MÄNTYLÄ, Mika V, Daniel GRAZIOTIN a Miikka KUUTILA. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer science review. AMSTERDAM: ELSEVIER, 2018, 27, 16-32. ISSN 1574-0137.

[136] MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C., BYERS, A. H. Big Data: The Next Frontier for Innovation, Competition, and Productivity. Report. 2011, McKinsey Global Institute

[137] MAMDANI, Ebrahim H. Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis. IEEE transactions on computers. IEEE, 1977, C-26(12), 1182-1191. ISSN 0018-9340.

[138] MAMUN, Abdullahil, Mohammad HASMAT ALI, Nazamul HOQUE, Md Masrurul MOWLA a Shahanara BASHER. The Causality between Stock Market Development and Economic Growth: Econometric Evidence from Bangladesh. International Journal of Economics and Finance. 2018, 10(5). ISSN 1916-9728.

[139] MATHUR, Neha, Ivan GLESK a Arjan BUIS. Comparison of adaptive neuro-fuzzy inference system (ANFIS) and Gaussian processes for machine learning (GPML) algorithms for the prediction of skin temperature in lower limb prostheses. Medical engineering&physics. OXFORD: Elsevier, 2016, 38(10), 1083-1089. ISSN 1350-4533.

[140] MATSUBARA, Takashi, Ryo AKITA a Kuniaki UEHARA. Stock Price Prediction by Deep Neural Generative Model of News Articles. IEICE transactions on information and systems, 2018, E101.D(4), 901-908. ISSN 0916-8532.

[141] MAURO, Paolo. Stock returns and output growth in emerging and advanced economies. Journal of development economics. AMSTERDAM: Elsevier B.V, 2003, 71(1), 129-153. ISSN 0304-3878.

[142] MCCAIN, Katherine W. Mapping authors in intellectual space: A technical overview. Journal of the American Society for Information Science [online]. Washington, D.C: Wiley Subscription Services, Inc., A Wiley Company, 1990, 41(6), 433-443. ISSN 0002-8231.

[143] MCGURK, Zachary, Adam NOWAK a Joshua C HALL. Stock returns and investor sentiment: textual analysis and social media. Journal of economics

and finance. New York: Springer Nature B.V, 2020, 44(3), 458-485. ISSN 1055-0925.

[144] MEDASANI, Swarup, Jaeseok KIM a Raghu KRISHNAPURAM. An overview of membership function generation techniques for pattern recognition. International journal of approximate reasoning. NEW YORK: Elsevier, 1998, 19(3), 391-417. ISSN 0888-613X.

[145] MEESAD, Phayung a Jiajia LI. Stock trend prediction relying on text mining and sentiment analysis with tweets. In: 2014 4th World Congress on Information and Communication Technologies (WICT 2014). IEEE, 2014, 2014, s. 257-262. ISBN 978-1-4799-8115-1.

[146] MEHRA, Rajnish a Raaj SAH. Mood fluctuations, projection bias, and volatility of equity prices. Journal of economic dynamics&control. AMSTERDAM: Elsevier B.V, 2002, 26(5), 869-887. ISSN 0165-1889.

[147] MELIN, Patricia a Oscar CASTILLO. A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition. Applied soft computing. AMSTERDAM: Elsevier B.V, 2014, 21, 568-577. ISSN 1568-4946.

[148] MENDEL, J.M, R.I JOHN a F LIU. Interval Type-2 Fuzzy Logic Systems Made Simple. IEEE transactions on fuzzy systems. PISCATAWAY: IEEE, 2006, 14(6), 808-821. ISSN 1063-6706.

[149] MENDEL, Jerry M. Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions, Prentice-Hall, Upper-Saddle River, 2001. ISBN 978-3-319-51370-6

[150] MINER, Gary. ELDER, J., IV, FAST, A., HILL, T., NISBET, R., a DURSUN, D. Text Mining and Statistical Analysis for Non-structured Text Data Applications. Waltham, MA: Elsevier, 2012. ISBN 9780123869791.

[151] MITCHELL, Tom. Machine Learning. McGraw-Hill, 1997. ISBN 0070428077

[152] MOHAMED, Amr. Comparative Study of Four Supervised Machine Learning Techniques for Classification. International Journal of Applied Science and Technology. 2017. 7(2), 5-18.

[153] MOLNÁR, Zdeněk. Pokročilé metody vědecké práce. [Zeleneč]: Profess Consulting, 2012. Věda pro praxi (Profess Consulting). ISBN 978-80-7259-064-3.

[154] MORAES, Rodrigo, João Francisco VALIATI a Wilson P GAVIÃO NETO. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert systems with applications. OXFORD: Elsevier, 2013, 40(2), 621-633. ISSN 0957-4174.

[155] MORO, Gianluca, Roberto PASOLINI, Giacomo DOMENICONI, Andrea PAGLIARANI a Andrea ROLI. Prediction and Trading of Dow Jones from Twitter: A Boosting Text Mining Method with Relevant Tweets Identification. 976. Knowledge Discovery, Knowledge Engineering and Knowledge

Management. Cham: Springer International Publishing, 2019, s. 26-42. ISBN 9783030156398.

[156] NANN, Stefan, Jonas KRAUSS a Detlef SCHODER. Predictive Analytics On Public Data - The Case Of Stock Markets. ECIS2013 Completed Research, 2013. 102

[157] NASSIRTOUSSI KHADJEH, Arman, Saeed AGHABOZORGI, Teh YING WAH a David Chek Ling NGO. Text mining for market prediction: A systematic review. Expert systems with applications. OXFORD: Elsevier, 2014, 41(16), 7653-7670. ISSN 0957-4174.

[158] NIELSEN, Finn Årup. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: CEUR Workshop Proceedings. 2011, s. 93-98. ISSN 1613-0073.

[159] NGUYEN, Thien Hai, Kiyoaki SHIRAI a Julien VELCIN. Sentiment analysis on social media for stock movement prediction. Expert systems with applications. OXFORD: Elsevier, 2015, 42(24), 9603-9611. ISSN 0957-4174.

[160] NOVÁK, Vilém. Základy fuzzy modelování. Praha: BEN, 2000, 176 s. ISBN 80-7300-009-1.

[161] NTI, Isaac Kofi, Adebayo Felix ADEKOYA a Benjamin Asubam WEYORI. Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana. Applied Computer Systems (Online). Sciendo, 2020, 25(1), 33-42. ISSN 2255-8691.

[162] O'HARE, Neil, Michael DAVY, Adam BERMINGHAM, Paul FERGUSON, Páraic SHERIDAN, Cathal GURRIN a Alan F. SMEATON. Topic-dependent sentiment analysis of financial blogs. In: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09. New York, New York, USA: ACM Press, 2009, 2009, s. 9-. ISBN 9781605588056.

[163] OLECKÁ, Ivana a Kateřina IVANOVÁ. Metodologie vědecko-výzkumné činnosti. Olomouc: Moravská vysoká škola Olomouc, 2010. ISBN 978-80-87240-33-5.

[164] OLIVEIRA, Nuno, Paulo CORTEZ a Nelson AREAL. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert systems with applications. OXFORD: Elsevier, 2017, 73, 125-144. ISSN 0957-4174.

[165] OLIVEIRA, Nuno, Paulo CORTEZ a Nelson AREAL. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. In: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13. New York, New York, USA: ACM Press, 2013, 2013, s. 1- . ISBN 9781450318501.

[166] OPREAN, Camelia a Cristina TANASESCU. Effects of Behavioural Finance on Emerging Capital Markets. Procedia Economics and Finance. 2014, 15, 1710-1716. ISSN 22125671.

[167] OWEN, Louis a Finny OKTARIANI. SENN: Stock Ensemble-based Neural Network for Stock Market Prediction using Historical Stock Data and Sentiment Analysis. In: 2020 International Conference on Data Science and Its Applications (ICoDSA). IEEE, 2020, 2020, s. 1-7. ISBN 978-1-7281-8235-3.

[168] PAGOLU, Venkata Sasank, Kamal Nayan Reddy CHALLA, Ganapati PANDA a Babita MAJHI. Sentiment analysis of Twitter data for predicting stock market movements. International conference on signal processing, communication, power and embedded system, scopes 2016 - Proceedings, Institute of Electrical and Electronics Engineers Inc. 2017. 1345-1350.

[169] PARK, Cheol-ho a Scott H IRWIN. WHAT DO WE KNOW ABOUT THE PROFITABILITY OF TECHNICAL ANALYSIS? Journal of economic surveys. Oxford, UK: Blackwell Publishing, 2007, 21(4), 786-826. ISSN 0950-0804.

[170] Passino, Kevin a Stephen Yurkovich. Fuzzy control. Menlo Park, CA: Addison-Wesley, 1998, ISBN 0-201-18074-X.

[171] PHAM, L. Financial News Headlines Data. Kaggle. ©2020[cit. 2021-04-22]. Dostupné z: https://www.kaggle.com/notlucasp/financial-news-headlines

[172] PORSHNEV, Alexander, Ilya REDKIN a Nikolay KARPOV. Modelling Movement of Stock Market Indexes with Data from Emoticons of Twitter Users. In: Information Retrieval. Cham: Springer International Publishing, 2015, s. 297-306. ISBN 3319254847. ISSN 1865-0929.

[173] PRAVICA, D.a M. SPURR. Mathematical Modeling for the Scientific Method. USA: Jones&Bartlett Learning, 2011.

[174] PUNCH, Keith. Úspěšný návrh výzkumu. Vydání druhé. Praha: Portál, 2015, 230 stran; 20 cm. ISBN 978-80-262-0980-5.

[175] PURGSTALLER, R. Dynamic N-Gram Based FeatureSelection for Text Classification. Master's Thesis. 2018. Graz University of Technology.

[176] RAEI, R., a S. FALLAHPOUR. Behavioral Finance, Different Approaches in the Area of Finance, Financial Re-search, 2004, 77-106.

[177] RAVI, Kumar a Vadlamani RAVI. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-based systems. AMSTERDAM: Elsevier B.V, 2015, 89, 14-46. ISSN 0950-7051.

[178] REJNUŠ, Oldřich. Finanční trhy. 4., aktualiz. a rozš. vyd. Praha: Grada, 2014, 760 s. : il., grafy, tab. ISBN 978-80-247-3671-6.

[179] REN, Rui, Desheng Dash WU a Tianxiang LIU. Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. IEEE systems journal. PISCATAWAY: IEEE, 2019, 13(1), 760-770. ISSN 1932-8184.

[180] Reuters: News [online]. ©2021 [cit. 2021-8-1]. Dostupné z: https://www.reuters.com/

[181] REZAEI, N. a Z. ELMI, Z. Behavioral Finance Models and Behavioral Biases in Stock Price Forecasting. Advances in mathematical finance&applications, 2018, 3(4), 67-82.

[182] ROKACH, L a O MAIMON. Top-down induction of decision trees classifiers - a survey. IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews. PISCATAWAY: IEEE, 2005, 35(4), 476-487. ISSN 1094-6977.

[183] ROSS, Timothy J. Fuzzy logic with engineering applications. John Wiley, 2004.

[184] ROSS, Timothy J., BOOKER, J., a PARKINSON, J. Fuzzy logic and probability applications: Bridg-ing the gap, society for industrial and applied mathematics. Philadelphia, PA, 2003. ISBN 978-0898715255.

[185] RÜDIGER, Matthias, David ANTONS a Torsten Oliver SALGE. From Text to Data: On The Role and Effect of Text Pre-Processing in Text Mining Research. Academy of Management Proceedings. 2017, 2017(1). ISSN 0065-0668.

[186] RUSSELL MATTHEW A. Mining the Social Web. 2. O'Reilly, 2013. ISBN 978-1-449-36761-9.

[187] RUSSELL, James A. A circumplex model of affect. Journal of Personality and Social Psychology. 1980, 39(6), 1161-1178. ISSN 0022-3514.

[188] SAKHARE, N. N., IMAMBI, D. S., KAGAD, S., KAPADWANJWALA, T., MALEKAR, H., DALAL, M. Stock Market Prediction Using Sentiment Analysis. International Journal of Advanced Science and Technology, 2020, 29(4s), 1126 - 1133.

[189] SALLEH, Mohd Najib Mohd, Noureen TALPUR a Kashif HUSSAIN. Adaptive Neuro-Fuzzy Inference System: Overview, Strengths, Limitations, and Solutions. In: Data Mining and Big Data. Cham: Springer International Publishing, 2017, s. 527-535. ISBN 331961844X.

[190] SAUNDERS, Mark, Philip LEWIS a Adrian THORNHILL. Research methods for business students. Harlow u.a: Pearson, 2012. ISBN 1292016620.

[191] SAUNDERS, M. N. K., Philip LEWIS a Adrian THORNHILL. Research methods for business students. 4th ed. New York: Financial Times/Prentice Hall, 2007. ISBN 978-0-273-70148-4.

[192] SEDLÁKOVÁ, Renáta. Výzkum médií: nejužívanější metody a techniky. Praha: Grada, 2014, 539 s., grafy, tab. ISBN 978-80-247-3568-9.

[193] SEZER, Omer Berat a Ahmet Murat OZBAYOGLU. Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. Applied soft computing. AMSTERDAM: Elsevier B.V, 2018, 70, 525-538. ISSN 1568-4946.

[194] SHI, Yong, Ye-ran TANG, Ling-xiao CUI a Wen LONG. A text mining based study of investor sentiment and its influence on stock returns. Economic computation and economic cybernetics studies and research. BUCHAREST: ACAD ECONOMIC STUDIES, 2018, 52(1), 183-199. ISSN 0424-267X.

[195] SHI, Kansheng, Jie HE, Hai-tao LIU, Nai-tong ZHANG a Wen-tao SONG. Efficient text classification method based on improved term reduction and term weighting. The Journal of China Universities of Posts and Telecommunications. 2011, 18, 131-135. ISSN 10058885.

[196] SHU, Hui-chu. Investor mood and financial markets. Journal of economic behavior&organization. AMSTERDAM: Elsevier B.V, 2010, 76(2), 267-282. ISSN 0167-2681.

[197] SILGE, Julia a David ROBINSON. Text mining with R: A tidy approach. Sebastopol: O'Reilly Media, Inc. 2017. ISBN 978-1491981658.

[198] SIERING, M. "Boom"or "Ruin"–Does It Make a Difference? Using Text Mining and Sentiment Analysis to Support Intraday Investment Decisions. In: 2012 45th Hawaii International Conference on System Sciences. IEEE, 2012, s. 1050-1059. ISBN 9781457719257.

[199] SHILLER, Robert. Irrational Exuberance, Princeton University Press. USA: New Jersey, 2000.

[200] SIMOES, Carlos, Rui NEVES a Nuno HORTA. Using sentiment from Twitter optimized by Genetic Algorithms to predict the stock market. In: 2017 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2017, 2017, s. 1303-1310. ISBN 978-1-5090-4601-0.

[201] SINGH, Tajinder a Madhu KUMARI. Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science. 2016, 89, 549-554. ISSN 18770509.

[202] SMAILOVIC, Jasmina, Miha GRCAR, Nada LAVRAC a Martin ZNIDARSIC. Stream-based active learning for sentiment analysis in the financial domain. Information sciences. NEW YORK: Elsevier, 2014, 285(1), 181-203. ISSN 0020-0255.

[203] SMALL, Henry. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science. Washington, D.C: Wiley Subscription Services, Inc., A Wiley Company, 1973, 24(4), 265-269. ISSN 0002-8231

[204] STATMAN, Meir. Behavioral finance: Finance with normal people. Borsa Istanbul review. 2014, 14(2), 65-73. ISSN 2214-8450.

[205] STEWART, David W. a Michael A. KAMINS. Secondary research: information sources and methods. 2nd ed. Newbury Park: Sage Publications, 1993. ISBN 9780803950375.

[206] STONE, Philip J. a Earl B. HUNT. A computer approach to content analysis. In: Proceedings of the May 21-23, 1963, spring joint computer conference on - AFIPS '63 (Spring). New York, New York, USA: ACM Press, 1963, 1963, s. 241.

[207] SUCIU, Titus. FROM THE CLASSICAL FINANCE TO THE BEHAVIORAL FINANCE. Journal of Public Administration, Finance and Law. Technopress, 2015, 4(7), 80-88. ISSN 2285-2204.

[208] SUN, Yuan, Xuan LIU, Guangyue CHEN, Yunhong HAO a Zuopeng (Justin) ZHANG. How mood affects the stock market: Empirical evidence from microblogs. Information&management. Elsevier B.V, 2020, 57(5), 103181. ISSN 0378-7206.

[209] SWAMY, K.M. Sentiment Analysis with Tensorflow – Tensor Flow and Deep Learning Singapore, 2017.

[210] TASKIN, Ahmet a Tufan KUMBASAR. An Open Source Matlab/Simulink Toolbox for Interval Type-2 Fuzzy Logic Systems. In: 2015 IEEE Symposium Series on Computational Intelligence. IEEE, 2015, 2015, s. 1561-1568. ISBN 978-1-4799-7560-0.

[211] TEKIN, Bilgehan a Erol YENER. The causality between economic growth and stock market in developing and developed countries: Toda-Yamamoto approach. Theoretical and applied economics. General Association of Economists from Romania, 2019, XXVI(2), 79-90. ISSN 1841-8678.

[212] TETLOCK, PAUL C, Maytal SAAR-TSECHANSKY a Sofus MACSKASSY. More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of finance (New York). Malden, USA: Blackwell Publishing, 2008, 63(3), 1437-1467. ISSN 0022-1082.

[213] TETLOCK, PAUL C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of finance (New York). Malden, USA: Blackwell Publishing, 2007, 62(3), 1139-1168. ISSN 0022-1082.

[214] TRAWINSKI, Bogdan, Magdalena SMETEK, Zbigniew TELEC a Tadeusz LASOTA. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. International journal of applied mathematics and computer science. Versita, 2012, 22(4), 867-881. ISSN 1641-876X.

[215] TIREA, Monica a Viorel NEGRU. Classifying and quantifying certain phenomena effect. In: 2013 IEEE 11th International Symposium on Intelligent

Systems and Informatics (SISY). IEEE, 2013, 2013, s. 363-368. ISBN 978-1-4799-0305-4.

[216] TRIVERS R. Deceit and self-deception. In Man and Beast Revisited, ed. MH Robinson, L Tiger. Washington D.C.: Smithsonian Press, 1991.

[217] TURNEY, Peter D. Thumbs up or thumbs down?: semantic orientation applied tounsupervised classification of reviews. InProceedings of the 40th annual meeting onassociation for computational linguistics, 2002, pages 417–424. Association for Computational Linguistics.

[218] UROLAGIN, Siddhaling. Text Mining of Tweet for Sentiment Classification and Association with Stock Prices. In: 2017 International Conference on Computer and Applications (ICCA). IEEE, 2017, 2017, s. 384-388. ISBN 978-1-5386-2752-5.

[219] UYSAL, Alper Kursat a Serkan GUNAL. The impact of preprocessing on text classification. Information processing&management. OXFORD: Elsevier, 2014, 50(1), 104-112. ISSN 0306-4573

[220] VALITUTTI, A., STRAPPARAVA, C. a STOCK, O. Developing Affective LexicalResources.PsychNology Journal, 2004, 2(1), 61–83.

[221] VAN ECK, Nees Jan a Ludo WALTMAN. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. Dordrecht: Springer Netherlands, 2010, 84(2), 523-538. ISSN 0138-9130.

[222] VAN RIJSBERGEN, C. J. Information Retrieval. Butter-worths, 1975.

[223] WIEBE, Janyce, Theresa WILSON a Claire CARDIE. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation. Dordrecht: Springer, 2005, 39(2/3), 165-210. ISSN 1574-020X.

[224] WEISS, Sholom M, Nitin INDURKHYA a Tong ZHANG. Fundamentals of Predictive Text Mining. London: Springer London, Limited, 2015. ISBN 9781447167495.

[225] WEISS, Sholom M, Nitin INDURKHYA, Tong ZHANG a Fred J DAMERAU. Text Mining. New York, NY: Springer New York. 2005. ISBN 0387954333.

[226] WISNIEWSKI, Tomasz Piotr a Brendan LAMBE. The role of media in the credit crunch: The case of the banking sector. Journal of economic behavior&organization. AMSTERDAM: Elsevier B.V, 2013, 85(1), 163-175. ISSN 0167-2681.

[227] XIE, Yancong. Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method. Journal of Computers. 2017, 500-510. ISSN 1796203X

[228] XING, Frank Z, Erik CAMBRIA a Roy E WELSCH. Natural language based financial forecasting: a survey. Artificial Intelligence Review. Dordrecht: Springer Netherlands, 2018, 50(1), 49-73. ISSN 0269-2821.

[229] YE, Zhengke, Chunyan HU, Linjie HE, Guangda OUYANG a Fenghua WEN. The dynamic time-frequency relationship between international oil prices and investor sentiment in china: A wavelet coherence analysis. The Energy journal (Cambridge, Mass.). International Association for Energy Economics, 2020, 41(5), 251-270 . ISSN 0195-6574.

[230] YEKRANGI, Mehdi a Neda ABDOLVAND. Financial markets sentiment analysis: developing a specialized Lexicon. Journal of intelligent information systems. 2020. ISSN 0925-9902.

[231] YIN, Robert K.Case Study Research: Design and Method (3rd edn). London: Sage, 2003.

[232] ZADEH, Lofti. A. Fuzzy sets. Information and Control, 1965, 8, 338–353.

[233] ZARANDI FAZEL, M.H, B REZAEE, I.B TURKSEN a E NESHAT. A type-2 fuzzy rule-based expert system model for stock price analysis. Expert systems with applications. OXFORD: Elsevier, 2009, 36(1), 139-154. ISSN 0957-4174

[234] ZHAI, Cheng Xiang a Sean MASSUNG. Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. ACM, 2016.

[235] ZHAI, Yuzheng, Arthur HSU a Saman K HALGAMUGE. Combining news and technical indicators in daily stock price trends prediction. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2007, s. 1087-1096. ISBN 3540723943.

[236] ZHANG, Wei, Xiao LI, Dehua SHEN a Andrea TEGLIO. Daily happiness and stock returns: Some international evidence. Physica A. AMSTERDAM: Elsevier B.V, 2016, 460, 201-209. ISSN 0378-4371.

[237] ZHANG Wenbin a Steven SKIENA. Trading strategies to exploit blog and news sentiment. Fourth International Conference on Weblogs and Social Media, 2010, 186.

[238] ZHANG, G.Peter a Min QI. Neural network forecasting for seasonal and trend time series. European journal of operational research. AMSTERDAM: Elsevier B.V, 2005, 160(2), 501-514. ISSN 0377-2217

[239] ZHANG, Tong. Statistical behavior and consistency of classification methods based on convex risk minimization. The Annals of Statistics. 2004, 32(1). ISSN 0090-5364.

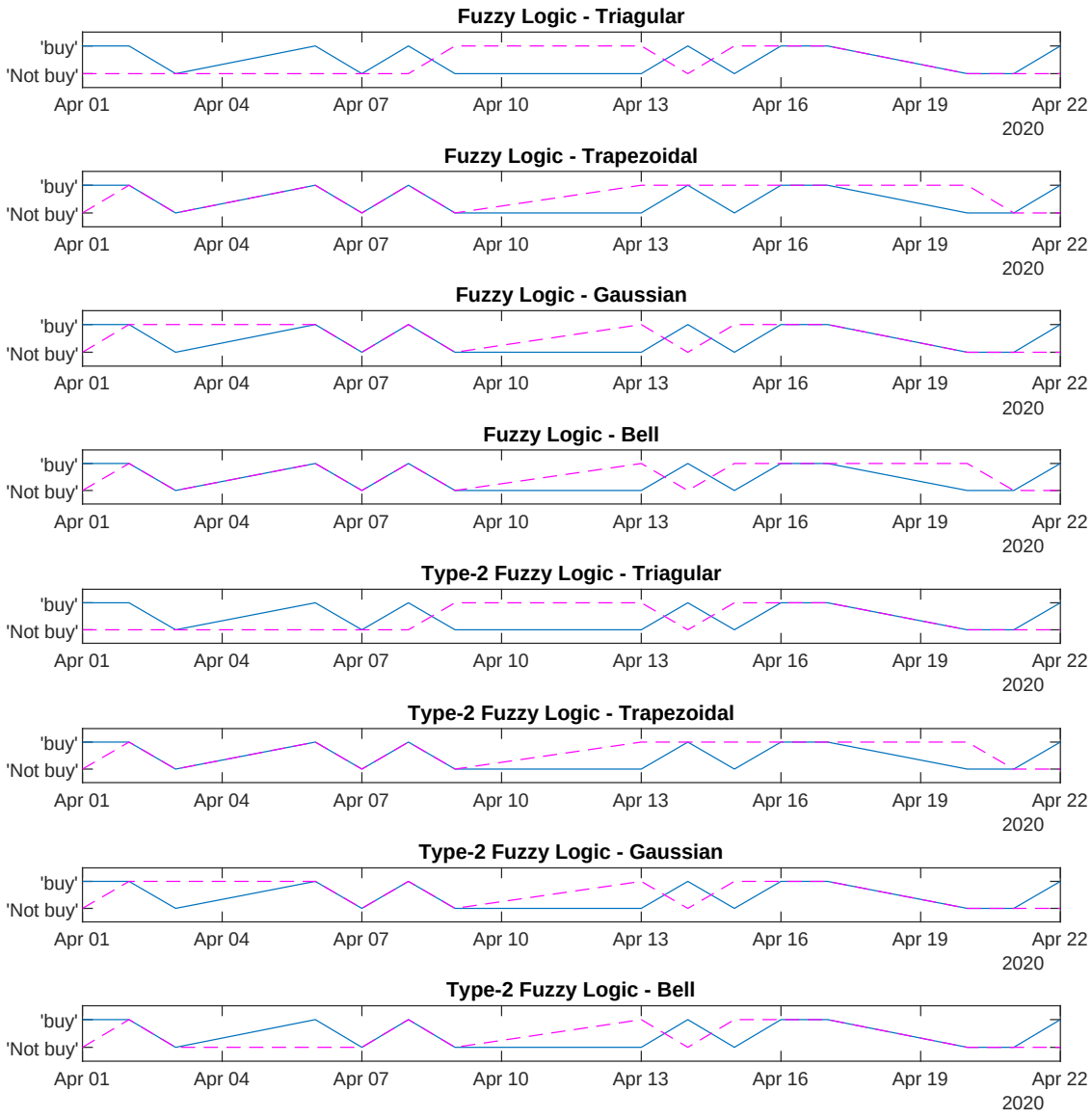
[240] ZHAO, Bo, Yongji HE, Chunfeng YUAN a Yihua HUANG. Stock market prediction exploiting microblog sentiment analysis. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016, 2016, s. 4482-4488. ISBN 978-1-5090-0620-5.

Seznam symbolů a zkratek

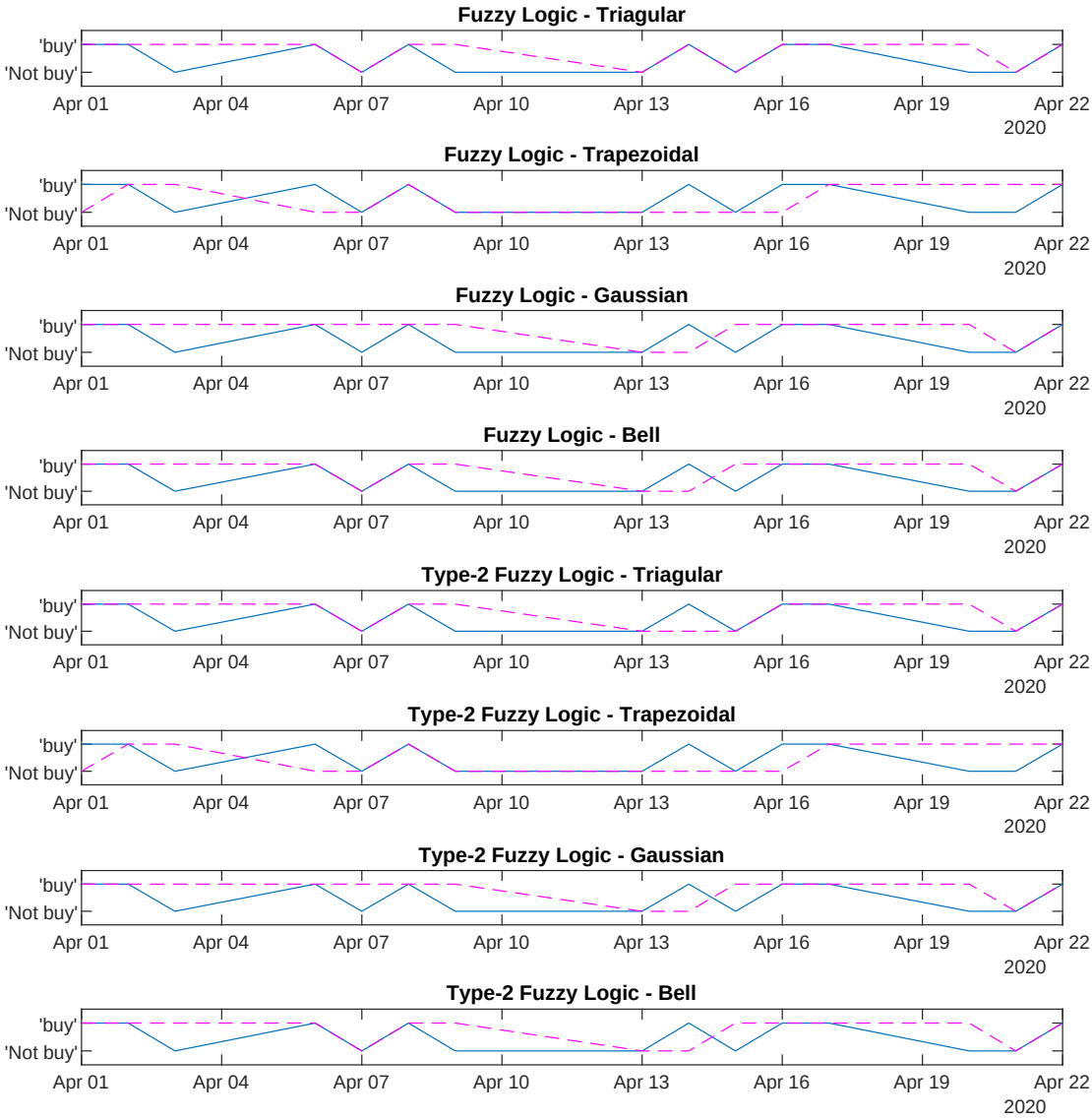
API	Application Programming Interface – Rozhraní pro programování aplikací
ANFIS	Adaptive Neuro-Fuzzy Inference System – Adaptivní neuro-fuzzy inferenční systém
BoW	Bag of words – Model pytle slov
DT	Decision Tree –Rozhodovací stromy
EMH	Efficient Market Hypothesis – Hypotéza efektivního trhu
ES	Expert System – Expertní systém
FL	Fuzzy logic – Fuzzy logika
FIS	Fuzzy inference system – Fuzzy inferenční systém
HTML	Hypertext Markup Language – Hypertextový značkovací jazyk
IDF	Inverse Document Frequency – Inverzní frekvence dokumentu
k-NN	k-Nearest Neighbors – k-nejbližších sousedů
LSTM	Long Short Term Memory – Neuronová síť s dlouhou krátk. pamětí
ML	Machine Learning – Strojové účení
NB	Naive Bayes – Naivní Bayes
NLP	Natural Language Processing – Zpracování přirozeného jazyka
NN	Neural Network – Neurnová síť
RNN	Recurrent Neural Network – Rekurentní neuronová síť
T1FLS	Type-1 Fuzzy Logic System – Systém type-1 fuzzy logiky
T2FLS	Type-2 Fuzzy Logic System – Systém type-2 fuzzy logiky
TF	Term Frequency – Frekvence termínu
S&P 500	Standard&Poor's 500 – Standard&Poor's 500
SVM	Support Vectore Machine – Metoda podpůrných vektorů

Seznam příloh

A Výstupy zvolené strategie



Obr. A.1: Strategie fuzzy modelů se sentimentem t-2
Zdroj: vlastní zpracování



Obr. A.2: Strategie fuzzy modelů se sentimentem t-3
Zdroj: vlastní zpracování

B Komparace expertních systémů

	Accuracy	Loss	Precision	Recall	F1
Fine Tree	0.6465	0.3535	0.6992	0.6548	0.6762
Medium Tree	0.6689	0.3311	0.7034	0.6803	0.6917
Course Tree	0.7136	0.2864	0.7966	0.7015	0.746
Linear Discriminant	0.7204	0.2796	0.7458	0.7303	0.7379
Linear SVM	0.7293	0.2707	0.7669	0.7328	0.7495
Quadratic SVM	0.7315	0.2685	0.7924	0.7248	0.7571
Cubic SVM	0.6085	0.3915	0.8941	0.5845	0.7069
Fine Gaussian SVM	0.7248	0.2752	0.7924	0.7165	0.7525
Medium Gaussian SVM	0.7204	0.2796	0.7712	0.7194	0.7444
Course Gaussian SVM	0.698	0.302	0.911	0.6535	0.7611
Fine k-NN	0.6577	0.3423	0.6822	0.6736	0.6779
Medium k-NN	0.698	0.302	0.8051	0.681	0.7379
Course k-NN	0.7136	0.2864	0.7627	0.7143	0.7377
Cosine k-NN	0.698	0.302	0.7712	0.692	0.7295
Cubic k-NN	0.6935	0.3065	0.8051	0.6762	0.735
Weighted k-NN	0.6734	0.3266	0.7288	0.6772	0.702
Ensemble Boosted Tree	0.7002	0.2998	0.7966	0.6861	0.7373
Ensemble Subspace Discriminant	0.7315	0.2685	0.7712	0.7339	0.7521
Ensemble Subspace k-NN	0.6309	0.3691	0.6568	0.6485	0.6526
Ensemble RUS Boosted Trees	0.7092	0.2908	0.7881	0.6992	0.741
Narrow Neural Networks	0.7114	0.2886	0.7881	0.7019	0.7425
Medium Neural Networks	0.6443	0.3557	0.6695	0.6611	0.6653
Wide Neural Networks	0.6667	0.3333	0.6695	0.69	0.6796
Bilayered Neural Networks	0.6667	0.3333	0.6737	0.6883	0.6809
Trilayered Neural Networks	0.6219	0.3781	0.6525	0.639	0.6457
Gaussian Naive Bayes	0.7002	0.2998	0.6949	0.7257	0.71
Kernel Naive Bayes	0.7069	0.2931	0.7712	0.7027	0.7354

Tab. B.1: Trénování modelů predikce bez sentimentu
Zdroj: vlastní zpracování

	Accuracy	**Loss**	**Precision**	**Recall**	**F1**
Fine Tree	0.6331	0.3669	0.6483	0.6538	0.6511
Medium Tree	0.6667	0.3333	0.7119	0.6747	0.6928
Course Tree	0.7136	0.2864	0.8305	0.6901	0.7538
Linear Discriminant	0.7226	0.2774	0.7542	0.7295	0.7417
Linear SVM	0.7293	0.2707	0.7712	0.7309	0.7505
Quadratic SVM	0.7159	0.2841	0.7881	0.7072	0.7455
Cubic SVM	0.7114	0.2886	0.7669	0.7098	0.7373
Fine Gaussian SVM	0.7092	0.2908	0.7712	0.7054	0.7368
Medium Gaussian SVM	0.7136	0.2864	0.7585	0.716	0.7366
Course Gaussian SVM	0.6935	0.3065	0.911	0.6495	0.7584
Fine k-NN	0.6465	0.3535	0.6695	0.6639	0.6667
Medium k-NN	0.6756	0.3244	0.7712	0.6667	0.7151
Course k-NN	0.7136	0.2864	0.7754	0.7093	0.7409
Cosine k-NN	0.6935	0.3065	0.7669	0.6882	0.7255
Cubic k-NN	0.6734	0.3266	0.7839	0.6607	0.7171
Weighted k-NN	0.6622	0.3378	0.7246	0.6654	0.6937
Ensemble Boosted Tree	0.698	0.302	0.7669	0.6935	0.7284
Ensemble Subspace Discriminant	0.7159	0.2841	0.75	0.7224	0.736
Ensemble Subspace k-NN	0.6376	0.3624	0.6949	0.6457	0.6694
Ensemble RUS Boosted Trees	0.6935	0.3065	0.7458	0.6957	0.7198
Narrow Neural Networks	0.6801	0.3199	0.7119	0.6914	0.7015
Medium Neural Networks	0.6264	0.3736	0.6441	0.6468	0.6454
Wide Neural Networks	0.6398	0.3602	0.6737	0.6543	0.6639
Bilayered Neural Networks	0.651	0.349	0.6695	0.6695	0.6695
Trilayered Neural Networks	0.6443	0.3557	0.6568	0.6652	0.661
Gaussian Naive Bayes	0.7047	0.2953	0.6992	0.7301	0.7143
Kernel Naive Bayes	0.7069	0.2931	0.7585	0.7075	0.7321

Tab. B.2: Trénování modelů predikce se sentimentem t-1
Zdroj: vlastní zpracování

	Accuracy	**Loss**	**Precision**	**Recall**	**F1**
Fine Tree	0.4787	0.5213	0.5042	0.5064	0.5053
Medium Tree	0.5034	0.4966	0.6737	0.523	0.5889
Course Tree	0.4765	0.5235	0.6229	0.5034	0.5568
Linear Discriminant	0.5056	0.4944	0.8644	0.5191	0.6486
Linear SVM	0.528	0.472	1	0.528	0.6911
Quadratic SVM	0.5145	0.4855	0.8898	0.5237	0.6593
Cubic SVM	0.519	0.481	0.6737	0.5354	0.5966
Fine Gaussian SVM	0.4922	0.5078	0.5932	0.5166	0.5523
Medium Gaussian SVM	0.5235	0.4765	0.9534	0.5269	0.6787
Course Gaussian SVM	0.528	0.472	1	0.528	0.6911
Fine k-NN	0.5235	0.4765	0.5254	0.5511	0.538
Medium k-NN	0.4855	0.5145	0.6102	0.5106	0.556
Course k-NN	0.5123	0.4877	0.7288	0.5276	0.6121
Cosine k-NN	0.5078	0.4922	0.6356	0.5282	0.5769
Cubic k-NN	0.481	0.519	0.6356	0.5068	0.5639
Weighted k-NN	0.5347	0.4653	0.5508	0.5603	0.5556
Ensemble Boosted Tree	0.4989	0.5011	0.6568	0.5201	0.5805
Ensemble Subspace Discriminant	0.5101	0.4899	0.911	0.5206	0.6626
Ensemble Subspace k-NN	0.4989	0.5011	0.5932	0.5224	0.5556
Ensemble RUS Boosted Trees	0.4855	0.5145	0.5297	0.5123	0.5208
Narrow Neural Networks	0.5414	0.4586	0.5975	0.5618	0.5791
Medium Neural Networks	0.5168	0.4832	0.5551	0.5413	0.5481
Wide Neural Networks	0.4832	0.5168	0.4915	0.511	0.5011
Bilayered Neural Networks	0.5436	0.4564	0.5678	0.5678	0.5678
Trilayered Neural Networks	0.5011	0.4989	0.5297	0.5274	0.5285
Gaussian Naive Bayes	0.5078	0.4922	0.4322	0.5426	0.4811
Kernel Naive Bayes	0.4765	0.5235	0.5805	0.5037	0.5394

Tab. B.3: Trénování modelů predikce se sentimentem t-2
Zdroj: vlastní zpracování

	Accuracy	**Loss**	**Precision**	**Recall**	**F1**
Fine Tree	0.4541	0.5459	0.4831	0.4831	0.4831
Medium Tree	0.5101	0.4899	0.7076	0.5268	0.604
Course Tree	0.4765	0.5235	0.7797	0.5027	0.6113
Linear Discriminant	0.4877	0.5123	0.8475	0.5089	0.6359
Linear SVM	0.528	0.472	1	0.528	0.6911
Quadratic SVM	0.5011	0.4989	0.9237	0.5154	0.6616
Cubic SVM	0.5101	0.4899	0.8305	0.5227	0.6416
Fine Gaussian SVM	0.5391	0.4609	0.7458	0.5466	0.6308
Medium Gaussian SVM	0.528	0.472	1	0.528	0.6911
Course Gaussian SVM	0.528	0.472	1	0.528	0.6911
Fine k-NN	0.4676	0.5324	0.5127	0.4959	0.5042
Medium k-NN	0.5056	0.4944	0.6822	0.5244	0.593
Course k-NN	0.4765	0.5235	0.7585	0.5028	0.6047
Cosine k-NN	0.519	0.481	0.6992	0.534	0.6055
Cubic k-NN	0.528	0.472	0.6949	0.5413	0.6085
Weighted k-NN	0.4497	0.5503	0.5169	0.4803	0.498
Ensemble Boosted Tree	0.4877	0.5123	0.6568	0.5116	0.5751
Ensemble Subspace Discriminant	0.5123	0.4877	0.9195	0.5216	0.6656
Ensemble Subspace k-NN	0.4787	0.5213	0.5593	0.5057	0.5312
Ensemble RUS Boosted Trees	0.4295	0.5705	0.4068	0.455	0.4295
Narrow Neural Networks	0.4631	0.5369	0.5	0.4917	0.4958
Medium Neural Networks	0.4989	0.5011	0.5297	0.5252	0.5274
Wide Neural Networks	0.4787	0.5213	0.4746	0.5068	0.4902
Bilayered Neural Networks	0.4966	0.5034	0.5212	0.5234	0.5223
Trilayered Neural Networks	0.4832	0.5168	0.5127	0.5105	0.5116
Gaussian Naive Bayes	0.5078	0.4922	0.75	0.5237	0.6167
Kernel Naive Bayes	0.434	0.566	0.589	0.4712	0.5235

Tab. B.4: Trénování modelů predikce se sentimentem t-3
Zdroj: vlastní zpracování

	Accuracy	**Loss**	**Precision**	**Recall**	**F1**
Fine Tree	0.589	0.411	0.641	0.6098	0.625
Medium Tree	0.589	0.411	0.5641	0.6286	0.5946
Course Tree	0.6849	0.3151	0.7436	0.6905	0.716
Linear Discriminant	0.4384	0.5616	0.5641	0.4783	0.5176
Linear SVM	0.4247	0.5753	0.7436	0.4754	0.58
Quadratic SVM	0.5205	0.4795	0.5641	0.55	0.557
Cubic SVM	0.5342	0.4658	0.5897	0.561	0.575
Fine Gaussian SVM	0.4932	0.5068	0.7436	0.5179	0.6105
Medium Gaussian SVM	0.5342	0.4658	1	0.5342	0.6964
Course Gaussian SVM	0.5342	0.4658	1	0.5342	0.6964
Fine k-NN	0.589	0.411	0.5897	0.6216	0.6053
Medium k-NN	0.3973	0.6027	0.4615	0.439	0.45
Course k-NN	0.5342	0.4658	1	0.5342	0.6964
Cosine k-NN	0.4658	0.5342	0.4872	0.5	0.4935
Cubic k-NN	0.4521	0.5479	0.4872	0.4872	0.4872
Weighted k-NN	0.5205	0.4795	0.6667	0.5417	0.5977
Ensemble Boosted Tree	0.6027	0.3973	0.8205	0.5926	0.6882
Ensemble Subspace Discriminant	0.4384	0.5616	0.6154	0.48	0.5393
Ensemble Subspace k-NN	0.6027	0.3973	0.6667	0.619	0.642
Ensemble RUS Boosted Trees	0.6575	0.3425	0.7436	0.6591	0.6988
Narrow Neural Networks	0.4521	0.5479	0.4872	0.4872	0.4872
Medium Neural Networks	0.3699	0.6301	0.3846	0.4054	0.3947
Wide Neural Networks	0.4658	0.5342	0.5128	0.5	0.5063
Bilayered Neural Networks	0.4795	0.5205	0.5385	0.5122	0.525
Trilayered Neural Networks	0.5479	0.4521	0.6667	0.5652	0.6118
Gaussian Naive Bayes	0.3836	0.6164	0.3333	0.4063	0.3662
Kernel Naive Bayes	0.4932	0.5068	0.641	0.5208	0.5747

Tab. B.5: Testování modelů predikce se sentimentem t-2
Zdroj: vlastní zpracování

	Accuracy	**Loss**	**Precision**	**Recall**	**F1**
Fine Tree	0.5278	0.4722	0.5263	0.5556	0.5405
Medium Tree	0.4861	0.5139	0.5789	0.5116	0.5432
Course Tree	0.5556	0.4444	0.6053	0.575	0.5897
Linear Discriminant	0.5556	0.4444	0.6053	0.575	0.5897
Linear SVM	0.6944	0.3056	0.7368	0.7	0.7179
Quadratic SVM	0.5556	0.4444	0.7368	0.56	0.6364
Cubic SVM	0.5694	0.4306	0.6053	0.5897	0.5974
Fine Gaussian SVM	0.5417	0.4583	0.6842	0.5532	0.6118
Medium Gaussian SVM	0.5556	0.4444	0.8421	0.5517	0.6667
Course Gaussian SVM	0.5278	0.4722	1	0.5278	0.6909
Fine k-NN	0.4167	0.5833	0.4211	0.4444	0.4324
Medium k-NN	0.5417	0.4583	0.6053	0.561	0.5823
Course k-NN	0.5278	0.4722	1	0.5278	0.6909
Cosine k-NN	0.625	0.375	0.7105	0.6279	0.6667
Cubic k-NN	0.611111	0.3889	0.6316	0.6316	0.6316
Weighted k-NN	0.5417	0.4583	0.6579	0.5556	0.6024
Ensemble Boosted Tree	0.5139	0.4861	0.6842	0.5306	0.5977
Ensemble Subspace Discriminant	0.5694	0.4306	0.6316	0.5854	0.6076
Ensemble Subspace k-NN	0.4028	0.5972	0.5526	0.4468	0.4941
Ensemble RUS Boosted Trees	0.5278	0.4722	0.4737	0.5625	0.5143
Narrow Neural Networks	0.5139	0.4861	0.5	0.5429	0.5205
Medium Neural Networks	0.5694	0.4306	0.6053	0.5897	0.5974
Wide Neural Networks	0.5556	0.4444	0.6053	0.575	0.5897
Bilayered Neural Networks	0.5417	0.4583	0.5263	0.5714	0.5479
Trilayered Neural Networks	0.4722	0.5278	0.5	0.5	0.5
Gaussian Naive Bayes	0.5417	0.4583	0.7105	0.551	0.6207
Kernel Naive Bayes	0.5	0.5	0.6579	0.5208	0.5814

Tab. B.6: Testování modelů predikce se sentimentem t-3
Zdroj: vlastní zpracování

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.527	0.473	0.5641	0.55	0.557
Type-2 fuzzy model s 10 % nejistotou	0.5135	0.4865	0.5641	0.5366	0.55
Type-2 fuzzy model s 20 % nejistotou	0.473	0.527	0.5385	0.5	0.5185
Type-2 fuzzy model s 30 % nejistotou	0.5	0.5	0.5897	0.5227	0.5542
Type-2 fuzzy model s 40 % nejistotou	0.5	0.5	0.5897	0.5227	0.5542
Type-2 fuzzy model s 50 % nejistotou	0.5	0.5	0.641	0.5208	0.5747
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.5541	0.4459	0.5128	0.5882	0.5479
Type-2 fuzzy model s 10 % nejistotou	0.5541	0.4459	0.5128	0.5882	0.5479
Type-2 fuzzy model s 20 % nejistotou	0.5541	0.4459	0.5128	0.5882	0.5479
Type-2 fuzzy model s 30 % nejistotou	0.5405	0.4595	0.5128	0.5714	0.5405
Type-2 fuzzy model s 40 % nejistotou	0.5135	0.4865	0.4615	0.5455	0.5
Type-2 fuzzy model s 50 % nejistotou	0.5135	0.4865	0.4615	0.5455	0.5
Gaussovské funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.4595	0.5405	0.4359	0.4857	0.4595
Type-2 fuzzy model s 10 % nejistotou	0.4595	0.5405	0.4359	0.4857	0.4595
Type-2 fuzzy model s 20 % nejistotou	0.4595	0.5405	0.4615	0.4865	0.4737
Type-2 fuzzy model s 30 % nejistotou	0.3919	0.6081	0.4615	0.4286	0.4444
Type-2 fuzzy model s 40 % nejistotou	0.4189	0.5811	0.4872	0.4524	0.4691
Type-2 fuzzy model s 50 % nejistotou	0.3919	0.6081	0.5128	0.4348	0.4706
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.5	0.5	0.4615	0.5294	0.4932
Type-2 fuzzy model s 10 % nejistotou	0.473	0.527	0.4872	0.5	0.4935
Type-2 fuzzy model s 20 % nejistotou	0.4459	0.5541	0.641	0.4808	0.5495
Type-2 fuzzy model s 30 % nejistotou	0.473	0.527	0.7179	0.5	0.5895
Type-2 fuzzy model s 40 % nejistotou	0.4595	0.5405	0.7179	0.4912	0.5833
Type-2 fuzzy model s 50 % nejistotou	0.4459	0.5541	0.6923	0.4821	0.5684

Tab. B.7: Testování fuzzy modelů se sentimentem t-2
Zdroj: vlastní zpracování

Trojúhelníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.4324	0.5676	0.4103	0.4571	0.4324
Type-2 fuzzy model s 10 % nejistotou	0.4189	0.5811	0.4103	0.4444	0.4267
Type-2 fuzzy model s 20 % nejistotou	0.4324	0.5676	0.3846	0.4545	0.4167
Type-2 fuzzy model s 30 % nejistotou	0.4054	0.5946	0.3846	0.4286	0.4054
Type-2 fuzzy model s 40 % nejistotou	0.4459	0.5541	0.4872	0.475	0.481
Type-2 fuzzy model s 50 % nejistotou	0.473	0.527	0.5128	0.5	0.5063
Lichoběžníkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.4595	0.5405	0.5897	0.4894	0.5349
Type-2 fuzzy model s 10 % nejistotou	0.473	0.527	0.6154	0.5	0.5517
Type-2 fuzzy model s 20 % nejistotou	0.473	0.527	0.6154	0.5	0.5517
Type-2 fuzzy model s 30 % nejistotou	0.473	0.527	0.6154	0.5	0.5517
Type-2 fuzzy model s 40 % nejistotou	0.4865	0.5135	0.6154	0.5106	0.5581
Type-2 fuzzy model s 50 % nejistotou	0.473	0.527	0.5897	0.5	0.5412
Gaussovské funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.4865	0.5135	0.3846	0.5172	0.4412
Type-2 fuzzy model s 10 % nejistotou	0.4865	0.5135	0.3846	0.5172	0.4412
Type-2 fuzzy model s 20 % nejistotou	0.527	0.473	0.4359	0.5667	0.4928
Type-2 fuzzy model s 30 % nejistotou	0.527	0.473	0.3846	0.5769	0.4615
Type-2 fuzzy model s 40 % nejistotou	0.5405	0.4595	0.4103	0.5926	0.4848
Type-2 fuzzy model s 50 % nejistotou	0.4459	0.5541	0.2051	0.4444	0.2807
Zvonkové funkce	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1**
Klasický fuzzy model	0.527	0.473	0.4359	0.5667	0.4928
Type-2 fuzzy model s 10 % nejistotou	0.527	0.473	0.4359	0.5667	0.4928
Type-2 fuzzy model s 20 % nejistotou	0.5676	0.4324	0.4103	0.64	0.5
Type-2 fuzzy model s 30 % nejistotou	0.5676	0.4324	0.2564	0.7692	0.3846
Type-2 fuzzy model s 40 % nejistotou	0.4865	0.5135	0.0769	0.6	0.1364
Type-2 fuzzy model s 50 % nejistotou	0.4595	0.5405	0.0256	0.3333	0.0476

Tab. B.8: Testování fuzzy modelů se sentimentem t-3
Zdroj:vlastní zpracování

C Odborný životopis

ZUZANA JANKOVÁ

Email: Zuzana.Jankova@vutbr.cz

VZDĚLÁNÍ

Vysoké učení technické v Brně, Fakulta podnikatelská

2017/dosud	Studijní program: Ekonomika a management (doktorský) Studijní obor: Řízení a ekonomika podniku
2016/2018	Studijní program: Ekonomika a management (magisterský) Studijní obor: Podnikové finance a obchod

Mendelova univerzita v Brně, Provozně ekonomická fakulta

2015/2017	Studijní program: Hospodářská politika a správa (magisterský) Studijní obor: Finance a investiční management
2012/2015	Studijní program: Hospodářská politika a správa (bakalářský) Studijní obor: Finance

PEDAGOGICKÉ PRAXE

ak. rok 2021/2022	– zimní semestr	Operační a systémová analýza Kvantitativní metody Statistika
ak. rok 2020/2021	– letní semestr	Pokročilé metody v rozhodování Pokročilé metody analýz a rozhodování
	– zimní semestr	Kvantitativní metody Statistika
ak. rok 2019/2020	– zimní semestr	Statistika
ak. rok 2018/2019	– letní semestr	Optimalizace a rozhodování
	– zimní semestr	Výpočetní metody

ZAHRANIČNÍ ZKUŠENOSTI

2/2020	Freemover, Ekonomická univerzita v Bratislavě, Slovensko
10/2019	zahraniční cesta, Tchaj-wan
7/2019	jazykový kurz angličtiny, Malta

PROJEKTY

2020/2021	Modelování a optimalizace podnikových procesů v podmínkách digitální transformace, **spoluřešitel**
2020	Využití umělé inteligence v podnikatelství IV., **navrhovatel**
2019	Využití umělé inteligence v podnikatelství III., **navrhovatel**
2018/2019	Informační a znalostní management v éře Průmyslu 4.0, **spoluřešitel**
2018	Kompetence lídrů vnímané jako žádoucí v éře nastupující digitální transformace, **spoluřešitel**

PRACOVNÍ ZKUŠENOSTI

	Vysoké učení technické v Brně, Fakulta podnikatelská
9/2020-dosud	Pozice: Lektor
	Náplň práce: Pedagogická činnost
	HUMUSOFT
1-6/2020	Pozice: Aplikační specialista
	Náplň práce: Ekonomické a finanční výpočty v software MATLAB
	Grant Capital
9/2017-1/2019	Pozice: Externí konzultant
	Náplň práce: Publikování článků v oblasti financí a investování
	Městský úřad Břeclav
7-8/2016	Pozice: Referent
	Náplň práce: Praxe na odboru sociálních věcí a školství

KURZY A ŠKOLENÍ

2/2020	**HUMUSOFT** Školení MATLAB II
1/2020	**Mathworks** Školení MATLAB Onramp, ML Onramp, Deep Learning Onramp
11/19	**HUMUSOFT** Školení MATLAB I
9/19	**HUMUSOFT, Technical Computing Camp** Druhé místo v soutěži o nejlepší studentský projekt
2017/2018	**Vysoké učení technické v Brně, ICV** Doplňující pedagogické vzdělání

D Přehled publikační činnosti

1. Původní vědecká práce v impaktovaném časopise zařazeném dle WoS
JANKOVÁ, Z.; DOSTÁL, P. Type-2 Fuzzy Expert System Approach for Decision-Making of Financial Assets and Investing under Different Uncertainty. Mathematical Problems in Engineering, 2021, roč. 2021, č. 1, s. 1-16. ISSN: 1563-5147.

JANKOVÁ, Z., JANA, D. K; DOSTÁL, P. Investment Decision Support Based on Interval Type-2 Fuzzy Expert System. Engineering Economics. 2021, roč. 32, č. 2, s. 118-129. ISSN 2029-5839.

2. Původní publikace ve Scopus
JANKOVÁ, Z. A Bibliometric Analysis of Artificial Intelligence Technique in Financial market. Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration. 2021, roč. 29, ISSN: 1804-8048 [In press].

JANKOVÁ, Z. Sentiment on the Stock Markets: Evidence from the Wavelet Coherence Analysis. Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration. 2020, roč. 28, č. 3, s. 1-10. ISSN: 1804-8048.

JANKOVÁ, Z.; DOSTÁL, P. Expertní systém type-2 fuzzy logika pro investiční analýzu. Scientific Papers of the University of Pardubice, Series D, 2019, roč. 47, č. 27, s. 79-90. ISSN: 1804-8048.

JANKOVÁ, Z.; DOSTÁL, P. Utilization of Artificial Intelligence for Sensitivity Analysis in the Stock Market. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 2019, roč. 67, č. 5, s. 1269-1283. ISSN: 1211-8516.

3. Příspěvek ve sborníku světového nebo evropského kongresu, sympózia, vědecké konference (příspěvky ve sbornících Web of Science nebo SCOPUS Conference Proceeding)
JANKOVÁ, Z.; DOSTÁL, P. Hybrid approach Wavelet seasonal autoregressive integrated moving averagemodel (WSARIMA) for modeling time series. In AIP Conference Proceedings. 2333. AIP Publishing, 2021. s. 090001-1 (090001-10 s.) ISBN: 978-0-7354-4077-7.

DOSTÁL, P.; JANKOVÁ, Z. COVID-19 Threats Evaluation of Countries via Fuzzy Interface Systems. In 35th IBIMA Conference proceedings. Seville, Spain: IBIMA, 2020. p. 2340-2351. ISBN: 978-0-9998551-4-0.

JANKOVÁ, Z. Comparison of Portfolios Using Markowitz and Downside Risk Theories on the Czech Stock Market. In Proceedings of the 7th International Conference: Innovation Management,Entrepreneurship and Sustainability. Prague, Czech Republic: Oeconomica, 2019. s. 291-303. ISBN: 978-80-245-2316-3.

JANKOVÁ, Z. Application of Artificial Neural Networks and Fuzzy Logic in Stock Trading. In Proceedings of the 33rd International Business Information Management Association Conference (IBIMA). Granada, Spain: IBIMA, 2019. s. 2610-2619. ISBN: 978-0-9998551-2-6.

JANKOVÁ, Z.; DOSTÁL, P. Analysis of Financial Market Using Soft Computing

Techniques. In Proceedings of the 16th International Scientific Conference. 1. Brno: Masaryk University, 2019. s. 201-209. ISBN: 978-80-210-9338-6.

JANKOVÁ, Z.; MAZÁNEK, L. Quantitative Research of Management Styles Applied By Leaders in Industrial Companies. In Proceedings of the 32nd International Business Information Management Association Conference (IBIMA). Seville, Italy: IBIMA, 2018. s. 5305-5317. ISBN: 978-0-9998551-1-9.

JANKOVÁ, Z. Black-Scholes Model Differential Equation and its Modifications for Valuation of Financial Derivatives. In Inovation Management and Education Excellence through Vision 2020. Milano, Italy: IBIMA, 2018. s. 801-811. ISBN: 978-0-9998551-0-2.

4. Odborné knihy a kapitoly v knihách

DOSTÁL, P.; JANKOVÁ, Z. Řízení kybernetických rizik. In Právní, kriminalistické a kybernetické aspekty kybernetické kriminality a bezpečnosti: Pocta Vladimíru Smejkalovi. Brno: Akademické nakladatelství CERM, 2021. s. 367-371. ISBN: 978-80-7623-065-1.

DOSTÁL, P.; JANKOVÁ, Z.; ŠEBESTOVÁ, M.; MACHŮ, E. Operační a systémová analýza: Pokročilé metody. Brno: Akademické nakladatelství CERM, 2020. 100 s. ISBN: 978-80-7623-030-9.

5. Publikace v odborném časopisu recenzovaném neimpaktovaném Jneimp, Jrec

JANKOVÁ, Z.; DOSKOČIL, R. Evaluation the Performance of Exchange Traded Funds (ETFs) Listed on the International Stock Markets. Ecoforum, 2021, roč. 10, č. 1, s. 1-8. ISSN: 2344-2174.

JANKOVÁ, Z.; DOSTÁL, P. Prediction of European Stock Indexes Using Neurofuzzy Technique. TRENDY EKONOMIKY A MANAGEMENTU, 2020, roč. 35, č. 1, s. 45-57. ISSN: 1802-8527.

6. Příspěvek ve sborníku národního nebo mezinárodního kongresu, sympózia, vědecké konference publikovaný

JANKOVÁ, Z. Effects of epidemic disease COVID-19 on sectors performance of stock markets. Workshop specifického výzkumu 2020. Brno, Česká republika: Vysoké učení technické v Brně, Fakulta podnikatelská, 2020. s. 94-101. ISBN: 978-80-214-5933-5.

JANKOVÁ, Z.; DOSKOČIL, R. Impacts of Federal Reserve System (FED) Economic Reports on US Financial Market Using Text Mining. Proceedings of the 14th International Scientific Conference INPROFORUMBusiness Cycles –more than Economic Phenomena. 14. České Budějovice, Czech republic: University of South Bohemia in České Budějovice, Faculty of Economics, 2020. s. 39-46. ISBN: 978-80-7394-824-5.

JANKOVÁ, Z. Literature Review of Fundamental and Technical Indicators Prediction of Financial Market Using Artificial Intelligence Technique. In Conference Proceedings DOKBAT 16th Annual International Bata Conference for Ph.D. Students and Young Researchers. 16. Zlin, Czech Republic: Tomas Bata University of Zlin, 2020. s. 242-254. ISBN: 9788074549359.

JANKOVÁ, Z. Artificial Intelligence Model as a Decision Support for Investing in Financial Products. Aktuárska veda v teórii a v praxi 2020. Brno, Česká republika: H. R. G. spol. s.r.o., 2020. s. 33-39. ISBN: 978-80-88320-36-4.

JANKOVÁ, Z. Identifikace optimálního portfolia na českém akciovém trhu. In MEZINÁRODNÍ MASARYKOVA KONFERENCE PRO DOKTORANDY A MLADÉ VĚDECKÉ PRACOVNÍKY. Hradec Králové: MAGNANIMITAS, 2019. s. 356-361. ISBN: 978-80-87952-27-6.

JANKOVÁ, Z.; DOSTÁL, P. Decision-Making Process Using Neuro-Fuzzy Model for Capital Market. In Perspectives of Business and Entrepreneurship Development in Digital Transformation of Corporate Business. Brno, Czech Republic: Faculty of Business and Management, Brno University of Technology, 2019. s. 175-182. ISBN: 978-80-214-5756-0.

JANKOVÁ, Z. Wavelet Analysis for Stock Market Forcasting. Interdisciplinární mezinárodní vědecká konference doktorandů a odborných asistentů QUAERE 2019. Hradec Králové, Czech Republic: Magnanimitas, 2019. s. 149-153. ISBN: 978-80-87952-30-6.

DOSTÁL, P.; JANKOVÁ, Z. Fuzzy Goal Programming Approach to Decision Making: Case of Moldova Stock Exchange. Teoria şi practica administrării publice. Moldova: Academia de Administrare Publica, 2019. s. 520-526. ISBN: 978-9975-3240-4-5.

JANKOVÁ, Z. Stock Price Prediction Problem Using Artificial Intelligence. In Knowledge on Economics and Management:Profit or Purpose. Olomouc, Czech Republic: Palacký University Olomouc, 2019. s. 56-62. ISBN: 978-80-244-5543-3.

JANKOVÁ, Z. Adaptive Neuro-Fuzzy Inference System (ANFIS) for Forceasting: The Case of the Czech Stock Market. In Conference Proceedings DOKBAT 15th Annual International Bata Conference for Ph.D. Students and Young Researchers. 15. Zlin, Czech Republic: Tomas Bata University of Zlin, 2019. s. 469-477. ISBN: 978-80-7454-893-2.

JANKOVÁ, Z. Wavelet Decomposition Analysis of Stock Market Movements. Brno, Czech Republic: Vysoké učení technické v Brně, 2019. s. 125-133.

JANKOVÁ, Z. Výkonnost burzovně obchodovaných fondů investujících do REIT. In Interdisciplinární mezinárodní vědecká konference doktorandů a odborných asistentů QUAERE 2018. Hradec Králové: MAGNANIMITAS, 2018. s. 396-403. ISBN: 978-80-87952-26-9.

JANKOVÁ, Z. Hedonická cena: vliv stavebních determinantů na prodejní cenu bytů v Brně. In Sborník příspěvků konference Junior Forensic Science Brno 2018. Brno: Vysoké učení technické v Brně, Ústav soudního inženýrství, Purkyňova 464/118, 612 00 Brno, 2018. s. 105-109. ISBN: 978-80-214-5621-1.

JANKOVÁ, Z. Využití fuzzy logiky při optimalizaci investičního portfolia. In Workshop specifického výzkumu 2018. Brno: Vysoké učení technické v Brně, Fakulta podnikatelská, 2018. s. 101-109. ISBN: 978-80-214-5705-8.

7. Publikace v odborném časopisu

JANKOVÁ, Z.; KADEROVÁ, A. Application of artificial intelligence model in financial markets. Ekonomika a informatika, 2020, roč. 18, č. 2, s. 67-74. ISSN: 1336-3514.

JANKOVÁ, Z. Drawbacks and Limitations of Black-Scholes Model for Options Pricing. Journal of Financial Studies & Research, 2018, roč. 2018, č. 1, s. 1-7. ISSN: 2166-000X.

MAZÁNEK, L.; SVOBODOVÁ, K.; SOUČKOVÁ, M.; JANKOVÁ, Z.; MEGOVÁ, S. Women in the Leadership Position and Their Relevant Competencies: Results of Exploratory Quantitaive Research Study. Journal of Strategic and International Studies, 2018, roč. Volume 8, č. 4, s. 48-54. ISSN: 2326-3636.

„Hluboko v člověku dřímou skryté síly. Takové síly, o nichž se mu nikdy nesnilo. Schopnosti, které by provedly revoluční změny v jeho životě, pokud by se mu podařilo je probudit ze spánku a přimět k akci.“

— Orison Swett Marden

Nenechávejte tyto síly už déle spát...