VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

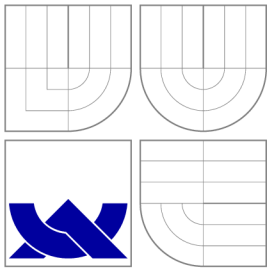
ALGORITMICKÉ OBCHODOVÁNÍ NA BURZE S VYUŽITÍM DAT Z TWITTERU

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

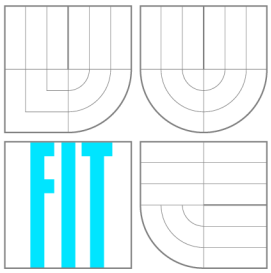
AUTOR PRÁCE
AUTHOR

Bc. JAKUB KŘÍŽ

BRNO 2015



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ALGORITMICKÉ OBCHODOVÁNÍ NA BURZE S VYUŽITÍM DAT Z TWITTERU

ALGORITHMIC TRADING USING TWITTER DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. JAKUB KŘÍŽ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. SZŐKE IGOR, Ph.D.

BRNO 2015

Abstrakt

Práce se zabývá tvorbou systému, který na základě analýzy historických burzovních dat a zpráv z Twitteru predikuje budoucí vývoj trhu. Tweety ze dvou různých sad jsou analyzovány pomocí náladových slovníků nebo přes rekurentní neuronovou síť. Z výsledků této analýzy a technické analýzy burzovních dat je pomocí vrstvené neuronové sítě prováděna predikce. Dle predikce poté systém vytvoří a otestuje obchodní strategii. V rámci práce je navržen a implementován celý systém, který pomocí dat z analýzy tweetů dosáhl zvýšení výnosu některých obchodních strategií o více než 25 %. Toto zlepšení však platí jen pro konkrétní data a časové období.

Abstract

This master's thesis describes creation of prediction system. This system predicts future market development based on stock exchange data and twitter messages analysis. Tweets from two different sources are analysed by mood dictionaries or via recurrent neural networks. This analysis results and technical analysis of stock exchange data results are used in multilayer neural network for prediction. A business strategy is created and tested based on results of this prediction. Design and implementation of prediction system is described in this thesis. This system achieved revenue increase more than 25 % of some business strategies by tweets analysis. However this improvement applies for certain data and timeframe.

Klíčová slova

burza, obchodování, Twitter, neuronová síť, technická a fundamentální analýza, predikce, automatický obchodní systém

Keywords

stock market, trading, Twitter, neural network, technical and fundamental analysis, prediction, automatic trading system

Citace

Jakub Kříž: Algoritmické obchodování na burze
s využitím dat z Twitteru, diplomová práce, Brno, FIT VUT v Brně, 2015

Algoritmické obchodování na burze s využitím dat z Twitteru

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing. Igora Szöke, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jakub Kříž
27. května 2015

Poděkování

Rád bych poděkoval svému vedoucímu diplomové práce za přínosné rady a návrhy při tvorbě této práce. Také děkuji panu Michalu Illichovi za poskytnutí datových sad. Nakonec bych rád poděkoval Kristiánovi za půjčení stroje, na kterém jsem analyzoval data a za připomínky ke grafické prezentací práce.

Obsah

Kapitola 1

Úvod

Obchodování na burze je často považováno za zajímavý způsob zhodnocení peněz. Proto je přitahováno velkým množstvím lidí, kteří si přejí rychlé zbohatnout. Každý se snaží získat nějakou výhodu, která by zlepšila jeho profit a minimalizovala rizika. Někdo při obchodování věří vlastnímu citu a zkušenostem, jiný zase používá moderní výpočetní techniku, analýzu trhu a matematické funkce. Všichni úspěšní obchodníci mají ale společnou jednu věc a tou je alespoň základní obchodní plán, podle kterého řídí své obchody.

V dnešní době je už podle autorů [7] většina obchodů prováděna pomocí automatických obchodních systémů, které řídí propracované obchodní plány. Tyto systémy se obvykle snaží na trhu hledat opakující se vzory, podle kterých predikují budoucí vývoj trhu a obchodují bez přítomností obchodníka.

Twitter je nejpoužívanější mikroblogovací služba na světě, proto je možné, že nějakým způsobem reflektuje dění na trhu. Právě tuto závislost bych chtěl v rámci diplomové práce najít.

Cílem této práce je vytvořit systém, který bude analyzovat zprávy z Twitteru a historická burzovní data, na základě kterých bude predikovat budoucí vývoj trhu. Systém by měl hledat závislosti mezi cenou akcií a tweety, kterou využije k vylepšení obchodní strategie.

Pro tvorbu takového systému je nejdříve nezbytné seznámit se z teorií z oblasti burzy, Twitteru a neuronových sítí. Také je potřeba obstarat vhodná data, na kterých se bude systém učit a testovat. Poté se musí specifikovat způsob analýzy tweetů a způsob predikce, kterou bude provádět neuronová síť. Nakonec by se měla otestovat funkčnost výsledného systému.

Kapitola 2 představuje burzovní trhy, popisuje vznik burzy a její rozdělení. Dále je vysvětlen způsob obchodování, princip kontraktů a význam obchodních systémů. Kapitola končí přehledem vlastností trhu, které se používají v jeho analýze. Další kapitola 3 se zabývá Twitterem. Uvádí jeho použití, základní znaky a způsob analýzy. Kapitola 4 obsahuje výklad teorie k neuronovým sítím. Nastiňuje princip fungování umělého neuronu, spojování do sítě a její učení. Kapitola 5 popisuje datové sady, se kterými bude systém pracovat, a způsob jejich předzpracování. Jednu sadu burzovních dat a dvě různé sady zpráv z Twitteru. V kapitole 6 jsou navrženy základní prvky systému a jejich spolupráce. Jsou zde popsány dva způsoby analýzy Twitteru a tvorba neuronové sítě. Další kapitola 7 potom přibližuje způsob implementace jednotlivých částí výsledného systému a použité knihovny. Poslední kapitola 8 je zaměřena na experimenty, ve kterých se celý systém otestuje. Cílem experimentů je najít nejvhodnější parametry pro chod systému a ohodnotit obchodní strategie. V závěru jsou zhodnoceny výsledky práce a navrhnuty směry dalšího vývoje.

Kapitola 2

Burzovní trhy

Trh obecně můžeme definovat jako místo, kde se setkává nabídka a poptávka po různých produktech. Finanční trh zabezpečuje pohyb peněz a kapitálu mezi ekonomickými subjekty. Hlavní funkcí trhu je přesun volných finančních prostředků od subjektů s jejich přebytkem k subjektům s jejich nedostatkem. [13]

Burzovní trh neboli burza je tedy organizovaný trh. Zjednodušeně řečeno je to místo, kde se může obchodovat.

V této kapitole bude nejprve popsán způsob vzniku burzy. Poté bude detailněji přiblížena samotná burza a s čím se na ní obchoduje. Dále budou nastíněn způsob jakým lze na burze obchodovat a různé techniky obchodování. V další části budou rozebrána automatické obchodní systémy (AOS). V závěru kapitoly budou představeny základní vlastnosti burzy a shrnuty různé metody burzovní analýzy.

2.1 Vznik burzy

První zmínky o burze se obecně datují do 14. století, kdy začali zejména Italští obchodníci obchodovat se směnkami a zlatými/stříbrnými mincemi. V 16. století byla v Antverpách založena organizace, která je považována za první burzu s cennými papíry na světě. V 17. století pak začali postupně vznikat i další burzy v Evropě a Americe, později v 19. století pak po celém světě. [13]

Naproti tomu komoditní burzy, kde se obchoduje s futures kontrakty (viz 2.3), začali vznikat v Japonsku přibližně v 17. století. V té době potřebovali japonští pěstitelé rýže zajistit financování sezónního pěstování rýže. K zajištění produkce museli získat potřebný kapitál pro pokrytí nákladů na sadbu, mzdu zaměstnanců a sklizeň. Problém byl v tom, že neměly záruku úspěšné sklizně, a proto jim žádná banka nechtěla půjčit.

Pokud však farmář dostatečný kapitál neměl, mohl doposud nevypěstovanou rýži „předprodat“. Farmář tedy předpokládal, že dokáže vypěstovat jisté množství rýže v požadované kvalitě. Odběratel zase věděl, že bude toto množství rýže v budoucnu (po sklizni) potřebovat.

Farmář na sebe vzal jisté riziko a zavázal se odběrateli k budoucímu dodání stanoveného množství rýže za předem domluvenou cenu - podepsali spolu kontrakt s budoucím datem dodání. Domluvená cena obvykle pokryla předem plánované náklady a zahrnovala i patřičný zisk pro farmáře.

Tato smlouva byla výhodná pro obě strany. Farmář měl jistého kupce, ještě předtím než zasel první semínko. Obchodník pak dopředu věděl, jaké budou jeho přesné náklady.

Netrvalo dlouho a začalo se obchodovat se samotnými futures kontrakty, protože se to ukázalo jako skvělí způsob rychlého zbohatnutí. Kontrakty začali nakupovat lidé, kteří rýži nepotřebovali a neměli s ní vůbec nic společného. Dělali to v naději, že bude špatná úroda a oni budou moci své levně nakoupené kontrakty draze prodat s co nejvyšším ziskem. Ovšem i toto obchodování mělo své rizika. Pokud byla úroda bohatá, cena rýže klesla a tito spekulanti museli prodat své kontrakty se ztrátou.[6]

Z tohoto způsobu obchodování se zrodila komoditní burza a dnes se futures kontrakty na mnoho typů zboží běžně obchodují na burzách po celém světě.

2.2 Burza

Burza je instituce, která má na starosti organizaci trhu s investičními nástroji. Účastníkům obchodů burza poskytuje nejen potřebné informace, ale především záruku bezpečnosti investic a plnění závazků. Burzovní obchod je přesně vymezen platnými zákony v daném státě a na jejich dodržování dohlíží kontrolní orgány burzy. Burza striktně reguluje burzovní trh.[9]

Uzavírat obchody jsou zpravidla oprávnění jen členové burzy, kteří musí plnit přísné podmínky. Ostatní obchodníci musí uzavírat obchody pouze skrze jejich prostřednictví. Obchod na burze většinou probíhá v přesně určenou denní dobu.

Princip obchodování na burze je jednoduchý. Chcete levně nakupovat a draze prodávat. O ceně právě rozhoduje stav nabídky a poptávky. Když nabídka převažuje poptávku, cena klesá a naopak, pokud poptávka převyšuje nabídku, potom cena roste. Takto stanovená cena se nazývá kurz. Informace o aktuálním kurzu je veřejně přístupná stejně jako další informace, například objem obchodů.

Největšími světovými burzami jsou New York Stock Exchange, NASDAQ Stock Market a London Stock Exchange. V České republice je největší burzou Burza cenných papírů Praha.

2.2.1 Typy burz

Burzy se podle obchodovaného aktiva dělí především na:

Komoditní

Obchoduje se zde převážně s komoditami, jako jsou například ropa, kukuřice, cenné kovy ve formě futures kontraktů. Komoditní obchodování je jedno z nejlikvidnějších, protože s komoditami obchoduje obrovské množství spekulantů. Je tedy v podstatě kdykoliv možné prodat nebo koupit jakoukoli komoditu.

Nerozlišuje se zboží různých dodavatelů, protože jeho kvalita je garantována.

Kromě futures kontraktů se obchodují také komoditní deriváty jako jsou opce. Opce se nezavazují k budoucímu obchodu, ale dávají k němu přednostní právo.

Akciové

Burzy s cennými papíry - akciemi. Akcie potvrzuje, že její majitel je akcionář, tj. že vložil určitý majetkový podíl do akciové společnosti. Což mu dává právo podílet se na zisku společnosti formou dividendy a účastnit se na řízení společnosti.

Obchodují se zde také akciové deriváty, jejichž hodnota se odvozuje s hodnoty skupiny akcí a odráží tak stav určité oblasti trhu.

Forex

K burzám bývá často nesprávně zařazován i tento mezinárodní obchodní systém pro obchod s měnovými páry. Řídí se stejnými pravidly nabídky a poptávky jako burza, ale vlastní burzu nemá a není tedy nijak centrálně regulován.

Je to systém vzájemně propojených bank, korporací, vládních institucí a brokerských společností. Umožňuje obchodníkům vydělávat na pohybech vzájemných kurzů mezi měnami. Jeho střední kurzy se považují za oficiální světové kurzy. Jednoznačně nejlikvidnější a v počtu obchodů největší trh světa.

2.3 Futures kontrakt

Futures kontrakt je jakási smlouva a budoucím prodeji zboží. Přesněji, podle standardní definice [6], lze futures kontrakt jednoduše charakterizovat jako dohodu dvou stran o nákupu/prodeji standardizovaného množství podkladového aktiva v předem specifikované kvalitě za danou cenu k danému budoucímu datu.

Akcie představuje majetkové právo, na rozdíl od toho futures kontrakt představuje povinnost dodat nebo odebrat podkladové aktivum. Ve skutečnosti k fyzickému předání obvykle nedojde, protože se kontrakt dále prodá.

Futures kontrakty je možné obchodovat výhradně na burzách. Všechny futures kontrakty jsou tzv. standardizovány, což znamená, že se podkladová aktiva obchodují ve standardizovaných jednotkách a v jasně definované kvalitě. Jeden kontrakt je minimální jednotka jakou lze obchodovat, jeho cena záleží na konkrétním podkladovém aktivu.

Pákový efekt

Zatímco na koupi či prodej akcie potřebujeme vždy 100 % její ceny, u futures kontraktů tomu je jinak. Stačí na burze složit pouze zálohu tzv. *margin*, díky které je možné kontrolovat až dvacetinásobně hodnotnější podkladové aktivum.

Tento efekt dělá komodity tolik atraktivní. Obchodník může za nepatrnou zálohu nakoupit obrovské množství komodit, na kterých pak i při minimální změně kurzu může vydělat nebo prodělat několikanásobně více než při klasickém investování. [16]

2.4 Způsob obchodování

Jak už jsem zmínil, na burze mohou obchodovat jen její členové. Proto musí mít většina obchodníků svého brokera, který zprostředkovává obchod. Obchodník mu tak může telefonicky nebo dnes už většinou on-line zadávat obchodní příkazy. Broker je pak skrze svého člena předává na burzu, kde uskuteční obchod. Za tuto službu si účtuje poplatek, kterému se říká komise. U intradenního obchodování se pohybuje do 10$.

Obchodování probíhá vstupováním do obchodních pozic, ve kterém kontrolujete kontrakty daného aktiva.

Dlouhá pozice - předpokládáte **růst** ceny a **nakupujete**. Chcete koupit levně, prodat dráž.

Krátká pozice - předpokládáte **pokles** ceny a **prodáváte**. Chcete prodat draze (aniž by jste ho vlastnili), koupit levněji.

Vydělávat tedy můžete ať jde cena nahoru nebo dolů.

Dalšími důležitými příkazy, které může obchodník zadávat jsou:

Stop-loss

Předem definovaná krajní hranice, při které se automaticky vystoupí ze ztrátové pozice. Stop-loss bývá označován [6] jako základní ochrna proti finančnímu krachu, kdy inkasujeme malou ztrátu dříve, než se rozroste do ztráty obří. Nikdy by se nemělo obchodovat bez předem definovaného stop-lossu.

Stop-loss by se měl umisťovat především s ohledem na obchodovanou formaci (jako důkaz nefunkční formace) a jeho výše by se měla odvíjet od celkového kapitálu (1-5 %).

Profit-target

Obchodní příkaz, který funguje na stejné principu jako stop-loss. Rozdílem je, že se jako hranice pro výstup z pozice udává minimální dosažený zisk. Používá se jen někdy a jeho výše se určuje podle předpokládaného maximálního zisku.

Oba tyto příkazy se mohou v průběhu obchodování průběžně posouvat.

2.4.1 Typy obchodníků

Každý způsob obchodování má své vlastní obchodní strategie. V literatuře [6] se obchodníci dělí podle časového horizontu v jakém obchodují:

Dlouhodobý poziční

Na pomezí mezi obchodníkem a investorem, který nevydělává samotným obchodováním. Drží své třeba i několik týdnů až měsíců a věnuje tak obchodování minimum času. Často obchoduje akcie a jeho zisk plyne i z dividend.

Poziční

Poziční obchodník drží své pozice otevřené déle než den, obvykle několik dní, a podstupuje tak větší riziko náhlé změny kurzu. Sleduje primárně denní a týdenní grafy. Tento obchodník zažívá méně stresu, protože má dostatek času promyslet každý obchod. Příležitostí k obchodování má maximálně několik týdně.

Intradenní

Obchoduje pouze během dne a své pozice drží pouze několik minut, maximálně pár hodin, nikdy však přes noc. Pracuje převážně s několika minutovými grafy a spekuluje tak na intradenní výkyvy trhů, které mohou být velmi ziskové. Intradenní obchodník se vystavuje menšímu riziku, protože neustále sleduje vývoj trhu a může při nepříznivé situaci okamžitě zareagovat a vystoupit z pozice. To znamená, že musí obchodování věnovat hodně času. Intradenní tradeři vstupují do trhu a vystupují z trhu i několikrát denně a často mají jen několik desítek vteřin na rozhodnutí.

Skalper

Nejnáročnější způsob obchodování. Drží své pozice jen několik sekund a díky obrovskému kapitálu profitují i na drobných pohybech kurzu. Časově velice náročné, protože často denně uskuteční desítky obchodů.

2.4.2 Obchodní systém

Pro každého úspěšného obchodníka je nejdůležitější plánování obchodů. Obchody sa plánují na základě obchodního systému, což je souhrn pravidel, kterými se obchodník řídí . Tyto pravidla definují jaké akce na trhu provádět v určitých situacích trhu a stavech obchodníka.

Tento systém může být navržen na základě analýzy trhu, matematických a ekonomických souvislostí nebo zkušeností obchodníka. Měl dávat jistou statistickou výhodu v trhu, která povede k dlouhodobým ziskům. Potvrzení použitých teorií a funkčnost celého systému se důkladně testuje na historických datech.

Důležitým kritériem pro měření úspěšnosti obchodního systému není vysoká pravděpodobnost předpovědi vývoje trhu, ale tzv. RRR (risk reward ratio) neboli poměr mezi riskem a očekávaným ziskem. Obchodníci se ho snaží minimalizovat, tak aby brali obchody jen s vyšším potenciálem zisku než velikostí risku.

Nepátrejte po vysoké pravděpodobnosti, ale raději po vysokém RRR. [16]

2.4.3 Automatické obchodní systémy

Existují různé přístupy k obchodování. Podle [6] se dělí podle účasti obchodníka na:

Diskreční obchodování

Při tomto přístupu k obchodování se klade velký důraz na subjektivní úsudek obchodníka. Obchodník může mít detailní obchodní plán, kterého se drží, ale konečné rozhodnutí vždy závisí jen na samotném obchodníkovi.

Úspěšnost tohoto způsobu obchodování je dána převážně zkušenostmi obchodníka a jeho citem pro trh.

Mechanické obchodování

Mechanické obchodování je založeno na přesných a jednoznačných pravidlech o řízení obchodu. Vypracovaný obchodní plán popisuje všechny situace, které mohou na trhu nastat a udává přesný postup jak se v takových situacích zachovat. Je tedy přesně dané kdy, jak a kolik obchodovat.

Výhodou mechanického obchodování je, že ho lze zautomatizovat, čímž vznikne automatický obchodní systém (AOS). Tvorba takového systému, který je schopen dlouhodobě generovat zisk, je obtížný úkol. Přínosem AOS je snadné zpětné testování na historických datech, možnost obchodovat na rychlejším timeframu než zvládne člověk a hlavně nezávislost na obchodníkovi. Obchodník tedy nemusí být při obchodování přítomen.

Nejlepší obchodníci tyto způsoby kombinují.[7] Většinu času se řídí přesnými pravidly svého obchodního plánu a nechávají si jen malý prostor pro vlastní rozhodování, čímž dosahují větší efektivity než samotné AOS, protože jsou schopny odhalit nezvyklé situace trhu.

Cílem této práce je vytvořit část AOS, která bude mít za úkol analyzovat data a predikovat vývoj trhu.

2.5 Burzovní grafy

Grafy se na burze používají pro vizualizaci pohybu cen a obchodní aktivity. Existuje mnoho typů grafů a způsobů jak se na data z trhu dívat. Grafy jsou základním nástrojem technické

analýzy a odlišné přístupy využívají různé typy grafů a dat.

Obrázek 2.1: Ukázka svíčkového grafu. [Zdroj: [16]]

Na likvidních trzích, kdy se obchoduje v podstatě neustále, se cena mění každých několik vteřin. Na grafech je proto důležité jejich časové měřítko - timeframe. Obvykle je na svislé ose grafu škála hodnot a na vodorovné čas. Čas je v těchto grafech rozdělen podle timeframe do intervalů a každý z nich agreguje data za dané časové období. Použitý timeframe má u obchodníků vliv především na frekvenci obchodů a používání stop-lossu. Protože rychlejší timeframe nabídne více příležitostí k obchodu, ale zato menší průměrný zisk na obchod.

Minimální hodnota o kterou se může trh pohnout se označuje jako *tick*. Je to nejmenší povolená jednotka o kterou se může změnit cena kontraktu . Hodnota ticku je individuálně definovaná pro každé aktivum. Různé burzy mohou mít různé velikosti kontraktů a také různé hodnoty nebo velikosti tiků pro stejné aktiva.

Většina burz a brokerů poskytuje cenová data ve formátu OHLC. Tato zkratka vyjadřuje čtveřici hodnot, jaké jsou agregovány z jednoho časového intervalu:

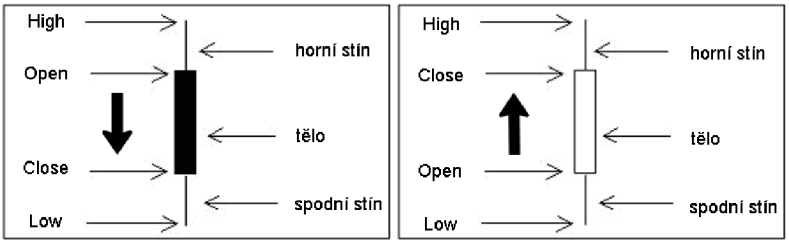
- Open - otevírací cena (cena na začátku intervalu)
- High - nejvyšší cena
- Low - nejnižší cena
- Close - uzavírací cena

Pro výpočet technických indikátorů (viz 2.7) se často používá jen uzavírací hodnota.

Tyto data se zobrazují pomocí sloupcového nebo **svíčkového** grafu (obr. 7.2), což jsou nejpoužívanější cenové grafy.

Jeden sloupec/svíčka (obr. 2.2) v nich reprezentuje právě jeden časový interval a v něm všechny čtyři příslušné hodnoty. Rozlišují se dva typy svíček podle toho, jestli cena v daném časovém úseku stoupla nebo klesla. Pokud je close cena vyšší než open cena, tělo svíčky je bílé nebo zelené, což značí nárůst ceny Pokud je to naopak, je tělo svíčky černé nebo červené.

Jiným používaným cenový graf je **tickový graf**. Jako tick data se označuje úplné, neagregované stanovení všech cen jednoho aktiva v průběhu času. Tick reprezentuje jeden provedený obchod, který se může skládat ze spárování více kontraktů. Tickové grafy jsou tedy složeny z jednotlivých provedených obchodů. Graf se tak vykresluje v závislosti na aktivitě trhu.



Obrázek 2.2: Svíčka - část svíčkového grafu.

Podobným často používaným grafem je **volume graf**, který se vytváří na základě počtu zobchodovaných kontraktů a ne obchodů. Lépe ukazuje aktivitu trhu vzhledem k množství obchodovaných kontaktů.

2.6 Analýza trhu

Úkolem analýzy trhu je předpověď budoucího vývoje trhu, na základě které můžeme generovat signály pro vstup a výstup z trhu. Existuje technická, fundamentální a psychologická analýza trhu. Psychologická se zabývá především psychologií davu a analyzuje chování lidí na určité situace na trhu.

Chtěl bych v rámci této práce zkombinovat technickou analýzu a fundamentální analýzu založenou na zprávách z Twitteru, které budou dále blíže popsány.

Technická analýza

Technická analýza je založena na studiu grafů. Vychází z předpokladu, že vnitřní hodnota aktiva je těžko vypočitatelná, a že se ve vývoji kurzu objevují opakující se obrazce. Techničtí obchodníci hledají v grafech různé formace, pracují s indikátory, používají matematické vzorce a analyzují cenový vývoj trhu. Nemusí znát ani podrobnosti o obchodovaném aktivu, vše co jim stačí je aktuální graf cenového vývoje. Někteří analyzují trh pouhým pohledem na graf, zatímco jiní využívají moderní statistické metody zkoumající vztahy mezi mnoha vlastnostmi trhu.

Dle [16] jsou u obchodníků oblíbené technické formace pozorovatelné na grafu pouhým okem (*patterns*). V průběhu historie obchodování bylo dokázáno, že určitě formace na grafu se objevují znovu a znovu, a že na základě takovýchto formací a ukazatelů je možné relativně spolehlivě předvídat, jakým směrem trh v příštích dnech půjde. Díky tomu také znají svá pravděpodobná rizika ještě před samotným vstupem do obchodu.

Fundamentální analýza

Fundamentální analýza ze zabývá velmi detailním zkoumáním základních a podstatných ekonomických, politických, sociálních, demografických faktorů a událostí, které určují vývoj burzovních kurzů.[12]

Zkoumá i samotné podkladové aktivum a předpokládá, že každé aktivum má nějakou vnitřní hodnotu, kolem které se pohybuje aktuální kurz. Vnitřní hodnota pomáhá zjistit zda je aktivum momentálně podhodnoceno nebo nadhodnoceno.

Jedná se tedy o komplexní přístup, který se snaží zahrnou všechny faktory z nejrůznějších odvětví, které mohou cenu nějak ovlivňovat.

2.7 Vlastnosti trhu

V rámci technické analýzy se vyžívá zejména trendů, formací na grafu, faktorů (vlastností) trhu a technických indikátorů.

Na základě těchto prvků, které pomáhají při rozhodování, se vytváří vhodná obchodní strategie. Nejpoužívanější jsou:

Trend

Trend je dlouhodobější neměnný směr trhu. O trendu se často říká, že je to nejlepší přítel obchodníka. Většina obchodních systémů je postavena především na včasném zachycení budoucího trendu. Dle trendu se trh dělí na býčí a medvědí. Jako trh býka se označuje stoupající trh. Naopak jako medvědí se označuje trh, který ma tendenci klesat. Většinu času však trh netrenduje a jde tzv. do strany. [11]

Likvidita

Velmi důležitý faktor trhu, který lze chápat jako momentální schopnost trhu plnit závazky neboli uskutečnit obchod. Dostatečně likvidní trh znamená, že obchodník může pozici otevřít i uzavřít podstatě v libovolné situaci. Je-li trh málo likvidní, dochází k výrazně horšímu plnění příkazů, což vede k horší ceně.

Volatilita

Další důležitý faktor, který udává jak živý a rychlý daný trh je. Určuje se na základě rychlosti a velikosti změny ceny. Na volatilním trhu se dá rychleji vydělat, ale stejně tak rychle i prodělat.

Volume

Technický indikátor, který udává objem uzavřených obchodů za daný časový úsek. Udává, jak hodně se v daný okamžik obchodovalo a tím odráží likviditu trhu.

Open interest

Tento indikátor vyjadřující množství otevřených kontraktů v daném trhu v dané časové periodě. Tedy kolik je v daném momentě drženo dvojic dlouhé a k ní příslušné krátké pozice. Stejně jako volume říká, jak vysoká je likvidita daného trhu. Čím vyšší open interest, tím je trh likvidnější a pro nás tudíž zajímavější.

Pokud open interest i volume současně stoupají, znamená to potvrzení trendu. Pokud open interest i volume současně klesají, znamená to, že je trend docela možná na svém konci a každou chvíli tak může dojít k jeho otočení. [7]

Grafické vzory

Obchodníci často analyzují trh pouhým pohledem na graf. Hledají v něm jisté grafické formace, které by měly napovědět další vývoj trhu. Pomáhají si třeba trendovými čárami, které kopírují směr trendu, a hledají v nich určité trojúhelníkové formace.

Dále jsou často používány techniky založené na úrovních spupport a resistance (S/R), protože dle [16] jsou na trhu vídány neustále a fungují vždy dobře.

Vycházejí z vlastnosti lidské psychiky upínat se k určitým bodům trhu. Support je v cenovém grafu jakási hranice nebo dno, pod které už obchodníci nechtějí prodávat. Resistance je oblast opačná, tedy strop, nad kterým už obchodníci nechtějí nakupovat.

Z grafu jde poznat tak, že se cena zasekla na určité hodnotě, od které se odráží. Po nějaké době bývají tyto hranice zpravidla proraženy, což obvykle značí nástup nového trendu. Čím déle se cena do této úrovně drží, tím bývá daná hranice silnější.

Dle [16] mají S/R úrovně tendenci se v grafu opakovat, čímž nám poskytují mapu oblasti historické nabídky a poptávky, která se v budoucnu může opakovat.

Díky nim může mít obchodník lepší představu o maximálně dosažitelném zisku svého obchodu a umísťuje do těchto úrovní profit-target.

2.7.1 Indikátory

Indikátor neboli ukazatel je obvykle matematická funkce, které zvýrazní některou vlastnost časové řady.

Nejznámější a nejpoužívanější jsou tyto indikátory:

Klouzavý průměr

Klouzavý průměr - MA (moving average) je jeden ze základních indikátorů technické analýzy. Podle [16] je jeho funkčnost časem ověřená a ve své obchodní strategii ho používá spousta obchodníků.

Představuje aritmetický průměr n posledních zavíracích cen, kde n je perioda klouzavého průměru, která udává počet historických hodnot z nichž se počítá. Používá se k vyhlazení průběhu ceny v čase, čímž redukuje šum.

Nejjednodušší pravidlo pro jeho obchodování je, že pokud cena protne klouzavý průměr směrem dolů, je toto považováno za signál k prodeji. Pokud cena protne klouzavý průměr směrem nahoru, je toto považováno za signál k nákupu.

Obvykle se ale používá kombinace více klouzavých průměrů z různou periodou, které zachycují různá historické období a navzájem se tak doplňují.

Někdy se takto označuje jednoduchý klouzavý průměr - SMA (simple moving average), který se počítá podle vzorce **??**:

$$SMA_t(n) = \frac{P_t + P_{t-1} + \cdots + P_{t-n-1}}{n} \tag{2.1}$$

kde n je počet historických hodnot, z nichž se klouzavý průměr počítá a P_t je uzavírací cena v čase t.

Kromě něj se také hojně používá exponenciální klouzavý průměr - EMA.

Jeho výpočet je podobný jednoduchému klouzavému průměru s tím rozdílem, že starší hodnoty mají ve výpočtu menší váhu. Nejvyšší váhu má tedy nejnovější hodnota a směrem ke starším hodnotám klesá jejich váha exponenciálně. Jeho výpočet je dán rekurzivním vzorcem 2.1

$$EMA_t(n) = EMA_{t-1} + \frac{2}{n+1}(P_t - EMA_{t-1}) \tag{2.2}$$

kde n je počet historických hodnot, ze kterých se počítá, P_t je uzavírací cena v čase t a EMA_{t-1} je hodnota v předchozím časovém intervalu .

RSI

Indikátor relativní síly trhu (Relative Strength Index). Slouží především jako doplněk obchodní strategie, který měří sílu a rychlost pohybu cen.

Jeho cílem je identifikovat překoupené a přeprodané trhy. Indikátor vrací hodnotu v rozmezí 0 až 100. Pokud je hodnota nad 80 je trh považován za překoupený a pokud je menší než 20 je přeprodaný. Výstup z těchto oblastí je považován za signál k provedení obchodu.

Vychází s počtu historických cen, které jsou nad a pod cenou aktuální. Stejně jako MA má parametr, který nastavuje jeho periodu, tedy počet historických hodnot, ze kterých se počítá. Pro jeho výpočet se používá vzorců 2.2 a 2.3.

$$RSI(n) = 100 - \frac{100}{1 + RS(n)} \tag{2.3}$$

$$RS(n) = \frac{\text{průměrná velikost pohybu nahoru v posledních } n \text{ hodnotách}}{\text{průměrná velikost pohybu dolů v posledních } n \text{ hodnotách}} \tag{2.4}$$

CCI

CCI (Commodity Channel Index) je univerzální indikátor trhu, kteří měří odchylku aktuální ceny od své průměrné hodnoty. Může sloužit k identifikaci, že je trh překoupený/přeprodaný nebo k detekci nového trendu.

Počítá se (viz. vzorec 2.4 porovnáním aktuální ceny s cenou průměrnou, která je spočtena klouzavým průměrem z několika posledních hodnot podle parametru indikátoru.

$$CCI(n) = \frac{1}{0.015} \frac{TP_t - SMA^{TP}(n)}{MD^{TP}(n)} \tag{2.5}$$

kde TP je typická cena složena, což je průměr hodnot low, high a close, SMA^TP je jednoduchý klouzaví průměr spočítaný s typické ceny a MD^TP je střední odchylka typické ceny.

Obvykle se jeho hodnota pohybuje v rozmezí od -100 do 100, ale není nijak limitována. Kladné hodnoty říkají, že je aktuální cena výší než průměrná a naopak.

Kapitola 3

Twitter

Twitter [1] je mikroblogovací sociální sít. Umožňuje svým uživatelům posílat krátké příspěvky, kterým se říká *tweety*.

Nejprve budou v této kapitole popsány základní principy a funkce Twitteru. Dále bude přiblíženo kým a jak je využíván a nakonec budou zváženy možnosti analýzy tweetů.

Obrázek 3.1: Logo sociální sítě Twitter.

3.1 Základy

Twitter založil Jack Dorsey v roce 2006. Od té doby se z něj, především díky jednoduchosti použití, stala nejpoužívanější mikroblogovací služba. Aktivně ji každý měsíc využívá 284 miliónů uživatelů, kteří dohromady vygenerují pře 500 milionů tweetů denně.[17]

Obrázek 3.2: První tweet.

Základním principem Twitteru je tedy posílání zpráv, které mají maximálně 140 znaků. Do těchto textových zpráv je také možné vložit odkazy, obrázky a videa. Každý registrovaný

[1] Twitter https://twitter.com/

uživatel má svoji stránku, na které publikuje vlastní příspěvky. Kromě toho má každý uživatel také svůj časový přehled tzv. Timeline, na které se mu zobrazují odebírané příspěvky od jiných uživatelů. Každý si tedy může nastavit, které uživatel chce sledovat (following) a tím odebírat jejich příspěvky, aby si je mohl přečíst. Navíc funkce *retweet* umožňuje „přeposlat“ tweet jiného uživatele na svou vlastní zeď, takže ho uvidí i tvoji odběratelé.

3.2 Použití

Hlavní myšlenkou je krátce sdílet nejaktuálnější informace svého okolí a to v reálném čase. Twitter je používán různými skupinami uživatelů k odlišnými účelům. Běžní jednotlivý uživatelé přes něj většinou komunikují s přáteli tím, že informují o své aktuální činnost nebo sdílí své zajímavé postřehy a myšlenky. Neziskové organizace a univerzity ho používají především k sdílení novinek a zpráv z oblasti jejích působení. V komerční sféře ho firmy využívají k publikování aktuálních firemních informací nebo skrz něj propagují své zajímavé produkty. Někdy vytvoří samostatný Twitter účet jednomu konkrétnímu, obvykle významnému, produktu, kde sdílejí jen zprávy, které se ho týkají. Tweety jim tak slouží jako bezplatná reklama. Informační a sdělovací služby už zařazují publikování na Twittru do součásti své práce. Obvykle zveřejní odkaz na celý článek spolu s jeho názvem, stručným obsahem nebo zajímavostí, čímž nalákají na své stránky další čtenáře, které by dané téma mohlo zajímat. Nejpopulárnější jsou na Twitteru stránky známých osobností, jako jsou například herci, zpěváci nebo politici. Touto cestou vedou veřejný deník svých činností nebo komentují veřejné dění.

Mnoho uživatelů však nepoužívá Twitter ke sdílení aktualit, ale jen jako pasivní zdroj příjmu informací. Dle [17] se 40 % uživatelů přihlašuje ke svému účtu pouze proto, aby si přečetli novinky, ale sami nic nepublikují.

Twitter je tedy sociální síť, sloužící spíše pouze k informování než k interakci mezi uživateli. Na každý tweet muže kdokoli zareagovat a odpovědět na něj, což ale primárně neslouží k vedení rozsáhlých diskuzí, ale jen ke sdělení stručného názoru k danému tweetu.

3.3 Vazby

Tweety nespojuje jen autor a vazba odpovědi na původní zprávu, ale uživatelé mohou do svých zpráv vkládat tzv. *hashtag*. Hashtag je slovo nebo krátká fráze, která začíná znakem „#“. Uživatel jim může zařadit tweet do nějakého konkrétního kontextu nebo oblasti zájmu. Díky tomu lze odlišit témata a typy jednotlivých Tweetů a seskupovat je do kategorií. To vede k přehlednější orientaci a vyhledávání tweetu určitého zaměření nebo z dané oblasti.

Podobně jako lze hashtagem označovat místa nebo události, nabízí také Twitter možnost označit konkrétní uživatele, tím že se do zprávy před jeho uživatelské jméno vloží znak „@“. Všechny tyto vazby pa mohou být využity v rámci analýzy.

3.4 API

Twitter poskytuje také programové rozhraní (API), které umožňuje vytvářet programy a webové aplikace, které s ním komunikují. Díky tomu můžeme skrze vlastní aplikaci nejen tweety automaticky vytvářet a zveřejňovat, ale také v nich vyhledávat a číst. Každý získaný tweet obsahuje nejen samotný text zprávy a jméno autora, ale spoustu dalších metadat jako je například jazyk, časové pásmo, datum a čas vytvoření, externí odkazy, velikosti obrázků,

počet retweetů atd. Celá struktura tweetu je v příloze A. Využití tohoto API má ale své limity. Za 15 minut můžeme provést jen 180 dotazů na Twitter.[18]

Limity velmi znesnadňují vytvoření nástroje pro dolování dat z Twitteru, protože pro získání potřebných informací je často třeba mnoha dotazů. Především je při dolování třeba své dotazy rozložit v čase, tak aby se nedosáhlo limitu. Proto by získání většího množství dat trvalo velmi dlouho.

3.5 Analýza Twitteru

Twitter je považován za dobrý zdroj dat k dalšímu studiu, protože obsahuje obrovské množství zpráv a tím i informací. Protože je používán mnoha lidmi napříč celou populací, dokáže zachytit všeobecné veřejné mínění, z kterého lze dále odvodit budoucí globální vývoj. Zaměřením na konkrétní oblasti lze potom analyzovat jen určitou část populace, která ma něco společného.

Velkou překážkou při strojovém zpracování tweetů je pochopit správný kontext. Sebelepší nástroj nedokáže odhalit ironii nebo jazyková specifika. Další problém je také odlišit důležité tweety od těch nedůležitých, což je při tak ohromném počtu tweetů zásadní problém.

Jednou z prvních důležitých prací, která sa zabývá zavilostí mezi Twitterem a burzou je [1]. Autoři v ní podle GPOMS (Google-Profile of Mood States) přiřazují jednotlivá slova tweetu do jedné z šesti nálad. Na základě toho určují míru veřejné nálady. Pomocí samoorganizujících fuzzy neuronových sítí našly nejlepší prediktivní korelaci mezi hodnotami DJIA (Dow Jones Industrial Average) a náladou, kterou označili jako „klid“. Podle jejich výsledků se tato nálada odráží na vybraném trhu za 3-4 dny.

V jiné práci [3] se slova v tweetech rozdělila do dvou kategorií - pozitivní (štěstí, naděje) a negativní (strach, nervozita). Při analýze se poté uvažoval poměr pozitivních a negativních tweetů vůči celkovému počtu. Korelační analýza našla korelaci mezi tímto poměrem a některými burzovními indexy. Dalším důležitým zjištěním bylo, že počet followerů a retweetů nemají na analýzu prakticky žádný vliv.

Autoři práce [2] si například vytvořily vlastní slovník s 5 tisíci nejobvyklejšími slovy, kde každému přiřadili úroveň štěstí a smutku. Podle nich určily pravděpodobnost štěstí jednoho tweetu a celého dne. S tímto slovníkovým způsobem dosáhli 60 % úspěšnosti v přesnosti předpovědi směru vývoje trhu, s posunem tweetů o 3 dny.

Další práce [5] byla postavena na první práci [1]. Na rozdíl od ní ale rozděluje slova do čtyřech nálad (klid, štěstí, poplach a laskavost) podle své vlastní databáze nálad. Nejlépe se osvědčili nálady štěstí a klid, u kterých dosáhli úspěšnosti predikce směru vývoje trhu 75.56 %, při posunutí tweetů o 3-4 dny.

Poslední práce [15] testuje více způsobů analýzy Twitteru, pro účely pozičního i intradenního obchodování. V jednom způsobu rozděluje tweety na pozitivní a negativní, v dalším zase hledá slova a slovní spojení, jejichž četnost bude odpovídat pohybům trhu. V rámci pozičního obchodování přišel autor například na to, že cena nejčastěji klesala, když se na Twitteru nejvíce psalo slovo „larges“. Ale u intradenního obchodování nenašel vůbec žádnou korelaci ani s jedním způsobem analýzy.

Všechny tyto práce ukazují, že existuje závislost mezi Twitterem a burzovním trhem. Vetšinou se jedná o nějakou globální náladu jednoho celého dne, která se na trhu odrazí během 3-5 dnů.

Kapitola 4

Neuronové sítě

Neuronová síť je výpočetní model inspirovaný přírodou, který se snaží napodobit chování lidského mozku. Slouží k hledání/modelování složitých vztahů mezi vstupy a výstupy systému. Chová se jako univerzální aproximátor funkcí, který se na základě vstupů a příslušných výstupů dokáže natrénovat a potom se chová jako sledovaný systém.

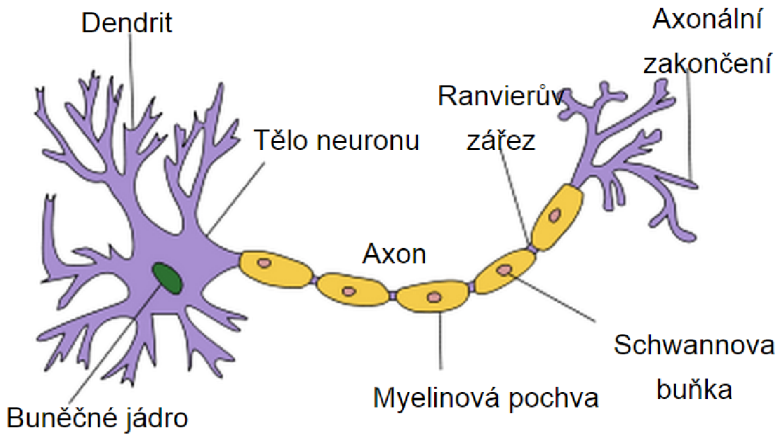
Neuronové sítě jsou jeden z nejúspěšnějších a nejrozšířenějších modelů pro strojové učení. Jsou používány zejména při rozpoznávání, predikci nebo v řízení autonomních systémů a dle [4] jsou vhodné i pro analýzu burzovních dat.

V této kapitole budou popsány základní prvky neuronové sítě, princip její funkce a způsob učení. Dále bude nastíněno jejich použití v oblasti burzy.

4.1 Biologický neuron

Neuron neboli nervová buňka je základním prvkem nervové tkáně, která zpracovává, přenáší a uchovává informace. Skládá se z těla, krátkých výběžků (dendritů) a dlouhého výběžku (axonu). Dendrity jsou vstupy neuronu, které přijímají elektrické vzruchy. Pokud součet jejich napětí přesáhne určitý práh, neuron se aktivuje a vyšle axonem elektrický impuls do ostatních neuronů, které jsou k němu přes své denrity připojeny .

Rozhraní mezi dendritem jednoho a axonem druhého neuronu se nazývá synapse. Každá synapse má svojí váhu, která určuje "důležitost"a tím i velikost předávaného signálu. Učení



Obrázek 4.1: Biologický neuron. [zdroj: http://cs.wikipedia.org/wiki/Neuron]

neuronu spočívá právě jen v nastavováním těchto synaptických vah.

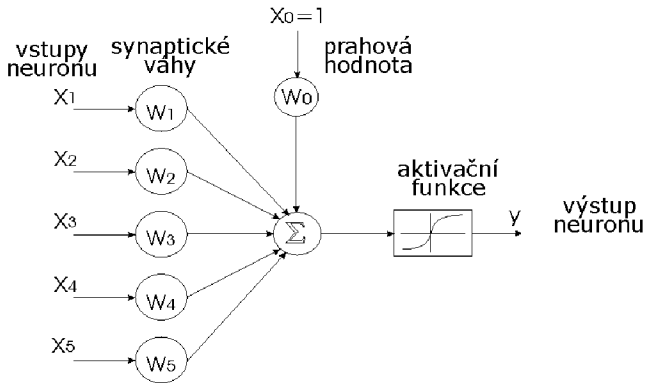
Lidský mozek má přibližně 10^{11} neuronů, každý neuron má obvykle 10^2 až 10^5 dendritů, což je celkem kolem 10^{14} synapsí [14].

4.2 Umělý neuron

Umělý neuron znázorněn na obrázku 4.2 je zjednodušený matematický model biologického neuronu, který napodobuje jeho chování. Obecně je jeho výstup y dán rovnicí 4.1, kde x je vstupní vektor, f je bázová funkce, g je aktivační funkce a u je vnitřní potencionál neuronu.

$$y = g(u) = g(f(x)) \tag{4.1}$$

Neuron funguje tak, že jsou na jeho vstupu vloženy hodnoty, které jsou modifikovány dle příslušných synaptických vah a sečteny bázovou funkcí. Tím dostaneme vnitřní potencionál neuronu, který je předán do aktivační funkce, která z něj spočítá výstupní hodnotu neuronu.



Obrázek 4.2: Model umělého neuronu

Bázová funkce

Bázová funkce f někdy označována jako sumarizační udává způsob, jakým jsou jednotlivé vstupy neuronu zkombinované. Výstupem bázové funkce je vnitřní potencionál neuronu u.

- **Lineární bázová funkce** - skalární součin vstupního vektoru x a váhového vektoru w .

$$u = f(x) = xw = \sum_{i=1}^{n} w_i x_i$$

- **Radiální bázová funkce** - vzdálenost vstupního vektoru x od váhového vektoru w.

$$u = f(x) = ||x - w|| = \sqrt{\sum_{i=1}^{n} w_i x_i}$$

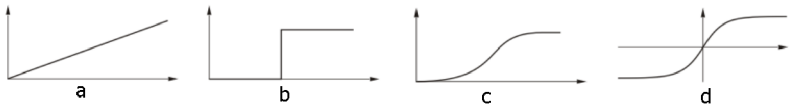
Aktivační funkce

Aktivační (přenosová) funkce g počítá výstup neuronu z jeho vnitřního potencionálu a převede jej do mezí, které jsou na výstupu neuronu očekávány.

V tabulce 4.1 jsou uvedeny nejčastěji se používají tyto aktivační funkce [10]:

Lineární funkce	$y = g(u) = au + b$
Skoková (prahová) funkce	$y = g(u) = \begin{cases} 0W & \text{pro } u < 0 \\ 1 & \text{pro } u > 0 \\ y_{old} & \text{pro } u = 0 \end{cases}$
Sigmoida	$y = g(u) = \frac{1}{1+e^{-\lambda u}}$
Hyperbolický tangenc	$y = g(u) = tanh(u)$

Tabulka 4.1: Aktivační funkce



Obrázek 4.3: Aktivační funkce neuronu: a) lineární, b) skoková, c) sigmoida, d) hyperbolický tangenc

4.3 Topologie sítě

Samostatný neuron s prahovou aktivační funkcí může sloužit jako nejjednoduší neuronová sít - Perceptron. Používá se jako primitivní klasifikátor, který je schopný řešit jen lineárně separovatelné problémy [10]. Perceptron totiž rozdělí vstupní prostor na dva podprostory a dokáže každý vstupní vektor zařadit do jednoho z nich.

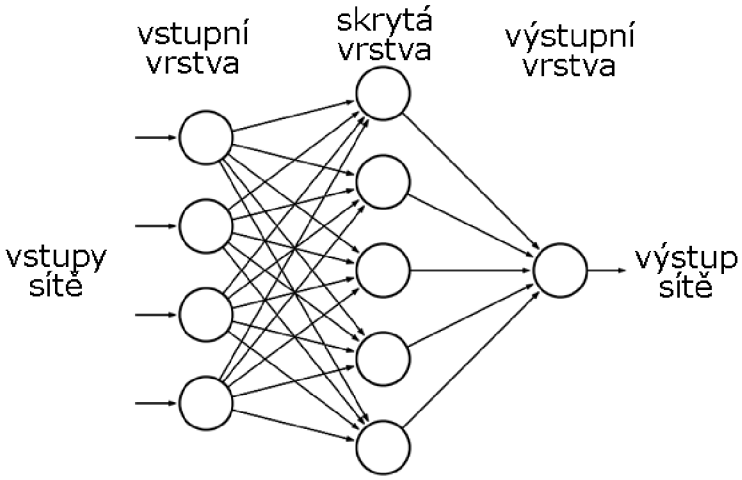
Pokud chceme řešit složitější problémy a nelineární funkce, potom je třeba použít více neuronů. Spojením několika neuronů vznikne neuronová síť, ve které si vhodně propojené neurony předávají signály.

Topologie neuronové sítě popisuje způsob jakým jsou neurony vzájemně propojeny a jejich počet. Způsob propojení jednotlivých neuronů hraje největší roli ve funkcionalitě a výkonu sítě. Složitější topologie s sebou nesou značnou míru problémů spojených s učením a tvorbou sítě.

Podle topologie lze rozlišit mnoho druhů modelů neuronových sítí, kde má každý své specifické využití. Nejznámější a nejpoužívanější topologie je vrstvená (perceptronová) síť, ve které je síť rozdělena na několik vrstev.

Obecně se skládá ze 3 typů vrstev - vstupní vrstvy, jedné nebo více skrytých vrstev a výstupní vrstvy (viz obrázek 4.4). Vrstvy jsou plně propojené, tj. všechny neurony vstupní vrstvy jsou propojeny se všemi neurony skryté vrstvy a všechny neurony skryté vrstvy jsou propojeny se všemi neurony výstupní vrstvy.

Vstupy sítě jsou přivedeny na pasivní neurony vstupní vrstvy, kteří je pouze rozdělí další vrstvě neuronů. Pracovní neurony ve skryté vrstvě vstupy transformují (využívají



Obrázek 4.4: Vícevrstvá dopředná síť

nelineární aktivační funkce, což umožňuje řešit nelineární funkce) a výsledek předají další vrstvě. Odezvou celé neuronové sítě na vstupy jsou výstupy neuronů ve výstupní vrstvě.

Podle směru toku dat se vrstvené sítě dělí na dopředné (acyklické) a rekurentní (cyklické). V dopředných sítích se signál šíří výhradně směrem od vstupních neuronů k výstupním. Neobsahuje žádné cykly a zpětné vazby, to znamená, že neexistuje propojení výstupu neuronu do vstupu neuronu ve stejné nebo předešlé vrstvě. Zato rekurentní síť musí obsahovat alespoň jednu zpětnou vazbu.

Nejpoužívanější a v rámci této práce použita je dopředná vrstevná síť (feedforward neural network) se zpětným šířením chyby, která je také označována dle učící metody backpropagation.

4.4 Učení neuronové sítě

Nejdůležitější vlastností a největší výhodou neuronových sítí je schopnost se učit. Není třeba přesně definovat algoritmus transformace vstupů na výstupy, ale místo toho se jej síť naučí sama.

Cílem učení neuronové sítě je nastavit všechny synaptické váhy a upravit topologii, tak aby se minimalizoval rozdíl mezi odezvou sítě na dané vstupy a výstupy požadovánými pro tyto vstupy. Fáze učení neuronové sítě se nazývá adaptivní. Po naučení neuronové sítě je síť ve fázi vybavování.

Učení se může probíhat dvěma způsoby, bez učitele a s učitele.

Při učení bez učitele známe vstupní hodnoty, ale očekávané/požadované výstupy neexistují. Síť hledá společné vlastnosti jednotlivých prvků, podle kterých je sama třídí do skupin. Na vstup sítě jsou postupně předkládány jednotlivé vstupní vektory a síť se nakonfiguruje tak, aby pro každý vstupní vektor existoval jednoznačně určený výstup.

Při učení s učitelem se využívá zpětné vazby. Síť se trénuje pomocí dvojic vstupní vektor a příslušný, požadovaný výstup.

Neuronové síti jsou překládány vstupní vektory. Na základě aktuálního nastavení sítě je zjištěna její aktuální odezva (výstup). Tu porovnáme s požadovaným výstupem a určíme chybu sítě. Poté spočítáme korekci (dle typu sítě), na základě které upravíme hodnoty vah,

tak abychom snížili velikost chyby. Toto se opakuje dokud síť nedosáhne námi stanovené minimální chyby, poté je síť adaptována Existuje tedy jakýsi vnější učitel, který porovnává požadované a skutečné výstupy sítě.

Backpropagation, neboli metoda zpětného šíření chyby, je základní a nejpoužívanější metoda pro učení neuronových sítí (kolem 80 % všech aplikací) [10]. Metoda je založena na principu učení s učitelem, kdy se při adaptaci vah vrstvených sítí využívá zpětného šíření chyby.

Samotný algoritmu učení je složen ze tří hlavních fází: dopředné šíření vstupního signálu, zpětné šíření chyby a aktualizace vah na vstupech neuronů. Tyto fáze se cyklicky opakují dokud není dosaženo požadované přesnosti sítě, maximálního počtu iterací nebo jiné podmínky, která ukončí proces učení. Ukončení trénování ve správnou chvíli je nejdůležitějším rozhodnutím v trénovacím procesu.[10]

Při dopředném šíření vstupního signálu klasicky předáme vstupní vektor neuronům vstupní vrstvy. Podle nastavení sítě se bude signál postupně šířit směrem k výstupní vrstvě, jejímž výstupem bude odezva sítě na vstup. Na základě rozdílu skutečného a požadovaného výstupu neuronové sítě je definována její chyba, pro kterou je vypočítán faktor. Zpětné šíření chyby spočívá v šíření vypočítaných faktorů do předcházejících vrstev. Protože je výstupní vrstva jediná, která může porovnat výstup s požadovanou hodnotou, musí šíření chyby začít od ní. Na rozdíl od předchozí fáze je signál šířen opačně, tedy z výstupní vrstvy směrem do vrstvy vstupní, proti „běžnému“ propojení neuronů. V poslední fázi je na základě faktorů v každém neuronu spočítána lokální chyba a podle ní upraveny vstupní váhy.

Množina dvojic vstupů a příslušných výstupů se obvykle rozdělí na trénovací a testovací množinu. Trénovací množinu používáme při učení sítě. Prvky této množiny vkládáme na vstup sítě a podle výstupů adaptujeme váhy. Testovací množinu používáme pro testování odezvy sítě. Pomocí ní se rozhodujeme, zda je síť naučená (má dostatečnou přesnost) nebo zda je třeba síť dál učit. Rozdělení na dvě množiny se provádí proto, aby byla neuronová síť schopná zobecňovat, tj. správně vyhodnocovat neznáme vstupy. Pokud bychom rozdělení neprovedli nebo učili síť příliš dlouho, tak může dojít k přeučení sítě. Potom by byla síť schopna správně reagovat pouze na známá data z natrénované množiny, ale špatně na data neznámá.

4.5 Analýza burzovních trhů

Podle [4] dosáhli neuronové sítě v posledních letech vysoké popularity v oblasti analýzy burzovních trhů. Díky své schopnosti generalizace a modelování nelineárních vztahů jsou považovány za ideální nástroj pro popis nelineárních systémů jakým jsou právě trhy.

Jejich činnost je založena na tom, že se určité situace na trhu často opakují. Neuronové sítě dokáži právě takové opakující se vzory chování detekovat a zapamatovat si je. Potom, když se trh dostane do podobného stavu, ve kterém byl už dříve, je naučená neuronová síť schopna s určitou pravděpodobností predikovat jeho budoucí vývoj. Vycházejí tedy z dříve získaných zkušeností a znalostí trhu.

Ve srovnání s moderními statistickými metodami a metodami regresní analýzy vycházejí neuronové sítě velice úspěšně. V některých testech je dokonce úspěšnost jejich předpovědi daleko vyšší. [4].

Neuronové sítě se používají buď samostatně nebo se ve snaze o dosažení lepších výsledků často kombinují s jinými metodami. Například jsou udávány neuronové sítě spolupracující s fuzzy logikou, modulární systémy složené z více sítí, kde každá využívá jiná data, rekurentní sítě, které si pamatují předchozí stavy trhu nebo neuronové sítě spolupracující

s expertními systémy.

Nejčastěji se však při obchodování podle [4] využívají právě vrstvené perceptronové sítě, učící se na principu backpropagation (viz 4.4).

Obvykle se jako vstupy neuronové sítě používají nejen historické hodnoty cen ale i odvozená data, jako jsou různé indikátory technické a fundamentální analýzy. Při tvorbě neuronové sítě určené k analýze trhu je právě nejobtížnější definovat vhodné vstupy. Tato činnost vyžaduje expertních znalosti trhu, četné analýzy a testování. Někdy je snaha tuto část zautomatizovat a zefektivnit, proto se k vybrání vhodných indikátorů a jejich optimalizaci používají například genetické algoritmy.

Kapitola 5

Datové sady

Důležitý základ každé analýzy jsou vhodná data, ze kterých se budou informace získávat a dále odvozovat. Ke své práci potřebuji datové sady, na kterých se bude vytvořený obchodní systém trénovat, a poté i testovat.

Systém bude tedy vycházet z technické analýzy, která vyžaduje cenová data z burzy, a z fundamentální analýzy Twitteru, pro kterou je potřeba velké množství tweetů. Také bude využito několik pomocných slovníků, které budou sloužit k analýze nálady a filtraci slov.

5.1 Burzovní data

Podařilo se mi získat historická cenová data akcií těchto velkých firem: Apple, Amazon, Google, Microsoft a Netflix. Konkrétně to jsou akcie *aapl, amzn, goog, msft, nflx*. Data jsou ve formátu OHLC + Volume daného trhu.

Datová sada pokrývá časové období téměř od začátku ledna do konce srpna 2014. Obsahuje kompletní informace z každého všedního dne (167 dní) z doby mezi 7:00 a 20:00 EST (čas burzy v New Yorku). Mezi 9:30 až 16:00, tedy v době kdy je burza otevřená, jsou na minutovém timeframu. Čím jsou od tohoto časového intervalu dále, tím jsou časové rozestupy mezi daty větší a nepravidelné. Informace mimo hlavní obchodní dobu napomáhají představit si zhruba vývoj trhu, a na základě toho rozjíždět některé technické indikátory s větší periodou.

Společnost	**Akcie**	**Celkem OHLC záznamů**
Apple	aapl	109 023
Amazon	amzn	82 379
Google	goog	80 940
Microsoft	msft	76 400
Netflix	nflx	82 915

Tabulka 5.1: Počet OHLC záznamů pro každou akcii.

Tyto akcie jsou dostatečně volatilní pro intraday obchodování. Někteří obchodníci [7] považují akcie velkých společností pro automatické obchodování nevhodná, protože je daná oblast trhu příliš nestálá a těžko predikovatelná. Obávají se také toho, že mohou být velice ovlivněna fundamentálními faktory, jako je třeba nálada, čehož bych chtěl právě využít.

Také jsou to velké společnosti, o kterých se obecně hodně mluví, a proto o nich bude mnoho zmínek na Twitteru.

5.2 Malá Twitter sada

První sada tweetů patří k burzovním datům. Každá akcie z burzovních dat (Apple, Amazon, Google, Microsoft a Netflix) má svoji vlastní sadu tweetů. Tweety do jednotlivých sad byly vybírány zvlášť podle nějakého klíčového slova souvisejícího s danou firmou. Například k akciím aapl společnosti Apple se vybíraly všechny tweety, které obsahovaly slova jako apple, aapl, iphone, ipad atd.

Sada obsahuje vybrané tweety od začátku ledna do konce října 2014 a to pouze z obchodních dnů. Celkem jsou v této datové sadě uloženy tweety z 218 dní. Jedná se o všední dny, ve kterých byla burza otevřena.

V každém dnu jsou uloženy pouze tweety v rozmezí osmi hodin od 13:00 do 21:00 UTC, což je 8:00 - 16:00 EST, tedy hodinu a půl před otevřením burzy až do jejího zavření. Tyto tweety jsou tedy ze stejného časového období jako burzovní data.

V této datové sadě je pouze čas publikace a text jednotlivých tweetů, bez dalších informací o autorovi či jiných metadat.

	Celkem tweetů	**Průměr za den**	**Průměr za hodinu**
Apple	100 754 550	462 176	57 772
Amazon	15 007 259	68 840	8 605
Google	67 257 736	308 521	38 565
Microsoft	6 563 210	30 106	3 763
Netflix	7 373 256	33 822	4 228

Tabulka 5.2: Počet tweetů v jednotlivých sadách.

5.3 Velká Twitter sada

Druhá sada tweetů je obsáhlejší než první. Získal jsem ji z veřejného archivu[1]. Jedná se o obrovskou sbírku všech tweetů, které pocházejí z různých zdrojů a jsou psány v různých jazycích. Autoři sady uvádí, že tweety získali z nějakého všeobecného twitter proudu (general twitter stream).

Data v této sadě jsou rozdělena po měsících a zabalena v archívech o průměrné velikosti 45 GB. V každém archivu je pro lepší orientaci a jednodušší použití vytvořena stromová struktura. Na nejvyšší úrovni jsou složky pro každý den, v nich jsou podsložky pro všechny hodiny daného dne. V složce hodin je pro každou minutu textový soubor, který konečně obsahuje tweety, které byly vytvořeny pouze v dané minutě. Každý tweet je uložen včetně všech metadat ve formátu JSON (viz příloha A), který lze také získat srze Twitter API (viz 3.4).

Archív s daty pro leden neexistuje, proto mám pouze osm archivů pro měsíce únor až září, které zachycují 234 dní - od 5. února do 30. září. Na rozdíl od malé sady nejsou tweety filtrovány podle klíčových slov ani nejsou vybrány jen z určitých dnů nebo časů. Datová sada tedy obsahuje kompletní záznam všech tweetů (z vybraných zdrojů) v každé minutě v průběhu celého dne a to v každém dni z vybraného období. Zachycuje tak kompletní informace o aktivitě a průběhu celého dne, tedy i z období, kdy se neobchodovalo.

V sadě připadne na jednu minutu průměrně 3 tisíce tweetů, což je 180 tisíc tweetů za hodinu. To je přes 4 milióny tweetů za den, které každý den průměrně vytvoří přes 3.5 miliónu unikátních denních uživatelů. Celkem je tedy v této sadě okolo 900 milionů tweetů.

[1] https://archive.org/details/twitterstream

5.4 Slovníky

Tweety, které budu analyzovat, jsou převážně v anglickém jazyce, proto potřebuji vhodné anglické slovníky. K analýze nálad (více v 6.2) použiji náladové slovníky, které jsou složeny ze samostatných slov.

Existují dva typy náladových slovníků. Jsou buď slovníky výčtové nebo slovníky hodnotové. Výčtový slovník obsahuje pouze seznam slov. Všechna slova v něm mají stejnou úroveň nálady, což znamená, že všechna slova jsou ve slovníku z hlediska nálady úplně stejná. Naproti tomu ve slovníku hodnotovém je ke každému slovu přiřazena hodnota, která udává míru nálady daného slova. Můžeme tedy vzájemně porovnat jednotlivá slova a určit, které představuje danou náladu více a které méně. Dalo by se říct, že výčtový slovník je hodnotový slovník, ve kterém mají všechna slova stejnou hodnota. Našel a připravil jsem si několik volně dostupných slovníků, které jsou dále popsány.

5.4.1 Hodnotové slovníky

Hodnoty nálady všech slovníků jsem lineárně transformoval do intervalu od mínus jedné do jedné. Tím všechny slovníky sjednotím a z transformované hodnoty bude hned zřejmá míra nálady, aniž bych věděl, co je to za konkrétní slovník a musel to neustále přepočítávat.

AFFIN[2]

Seznam slov, která byla manuálně ohodnocena náladou od mínus pěti (negativní) do pěti (pozitivní). Použil jsem největší verzi AFFIN-111, která obsahuje 2 477 různých slov.

LabMT[3]

Slovník byl vytvořen kombinací deseti tisíc (10 192) nejpoužívanějších slov v New York Times, Google Boock a Twitteru. Slova byla automaticky ohodnocena průměrnou hodnotou štěstí od nuly do desíti podle Mechanical Turk.

SentiWordNet[4]

Tento slovník je založen na WordNetu, což je velká lexikální databáze pro anglický jazyk, která slouží jako nástroj ke zpracování přirozeného jazyka. V této databázi se související slova seskupují do sad zvaných *synsets*, mezi kterými jsou vytvořeny další sémantické vztahy.

Náladový slovník tedy není jednoduchý seznam slov, ale jsou v něm navíc brány v potaz vazby mezi jednotlivými slovy. Proto tento slovník, na rozdíl od předchozích slovníků, nemám v textové podobě, ale využívám jeho implementace v knihovně pro zpracování přirozeného jazyka *NTLK*.

Výslednou náladou je potom dvojce čísel, ve které je zvlášť uvedena míra pozitivní a negativní nálady (nejsou navzájem závislé), obě v rozmezí od nuly do jedné.

[2] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
[3] http://neuro.imm.dtu.dk/wiki/LabMT
[4] http://sentiwordnet.isti.cnr.it/

5.4.2 Výčtové slovníky

Většina výčtových slovníků je složena ze dvou částí - seznamu pozitivních slov a seznamu negativních slov. Kromě toho existují také výčtové slovníky, které jsou zaměřeny na jednu konkrétní náladu, a mají tak pouze jeden seznam.

WordStat[5]

Klasický výčtový slovník, který má 5 536 pozitivních slov a 9 520 negativních. Je složen z několika starších náladových slovníků a rozšířen o synonyma a další související slova.

Opinion Lexicon[6]

Tento slovník nálad se používá převážně v akademické sféře, kde byl v roce 2004 vytvořen. Od té doby se průběžně doplňuje dalšími slovy podle výsledků navazujících akademických prací. Nyní obsahuje 2 006 pozitivních a 4 783 negativních slov.

Inquirer[7]

Obrovský slovník, který obsahuje skoro 12 tisíc slov a 180 různých kategorií, jako jsou nálady, oblasti použití nebo typy slov. Každá kategorie vlastně představuje samostatný výčtový slovník. Protože se jednotlivé kategorie navzájem nevylučují, může být jedno slovo ve více kategoriích. Ve slovníku je pro každé slovo uvedeno, do kterých všech kategorií patří.

Nepoužil jsem všechny kategorie, ale vybral jen několik, které jsou velké a významné. Největší a nejdůležitější jsou kategorie pozitivních a negativních slov. Kromě nich jsem také vybral několik kategorií obecných nálad, které obsahovaly nejvíce slov. Celkem je 6 kategorii, které jsou uvedeny v tabulce 5.3.

Název	Pozitivní	Negativní	Síla	Slabost	Aktivita	Pasivita
Originální označení	Positiv	Negativ	Strong	Week	Active	Passive
Počet slov	1637	2005	1475	647	1570	732

Tabulka 5.3: Vybrané kategorie ze slovníku Inquierer.

WordNet Affect[8]

Šest samostatných výčtových slovníků konkrétních nálad, které byly extrahovány z databáze WordNet. Jednotlivé slovníky a jejich velikosti jsou v tabulce 5.4

Nálada	Radost	Smutek	Nechuť/Odpor	Zloba	Překvapení	Strach
Originální označení	joy	sadness	disgust	anger	surprise	fear
Počet slov	400	202	53	255	71	147

Tabulka 5.4: Nálady ve slovníku WordNet Affect.

[5] http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/

[6] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[7] http://www.wjh.harvard.edu/~inquirer/homecat.htm

[8] http://wndomains.fbk.eu/publications/lrec2004.pdf

5.4.3 Stopslova

Mimo náladové slovníky využiji také slovník stopslov. Stopslova (StopWords) jsou často používaná slova, které nenesou žádnou významnou informaci. Obvykle mají pouze syntaktický význam. Většinou se jedná o předložky, spojky nebo zájmena. Tato slova jsou tedy pro analýzu neužitečná a jen ji zbytečně zatěžují, proto jsou z textu běžně vypouštěna.

Našel jsem si několik volně dostupných[9,10,11,12,13] seznamů anglických stopslov, které jsem zkombinoval. Výsledný slovník jsem spojil dohromady se seznamem stopslov, který je součástí knihovny pro zpracování jazyka *NTLK*. Celkem jsem získal unikátních 984 stopslov.

5.5 Předzpracování dat

Nedílnou součástí každé práce s daty je jejich předzpracování neboli příprava, která bývá často časově náročnější než samotná analýza. Obecně je cílem předzpracování opravit v datech problémy (chybějící a nesmyslné hodnoty), vybrat z nich pouze relevantní údaje, které budou potřeba pro další práci, a reprezentovat je v podobě, která bude vhodná pro další použití.

Čím jsou datové sady větší, tím je jejich příprava důležitější, protože vede ke zmenšení velikosti a tím i ke zrychlení samotného zpracování. Jelikož budu zpracovávat datové sady o desítkách GB, je jejich předzpracování nezbytné, jinak by se s nimi nedalo v rozumném čase pracovat.

Kromě toho můžeme v rámci předzpracování udělat základní analýzu, díky čemuž si uděláme o datech lepší představu. Můžeme například zjistit, jaké nesou informace, co obsahují za položky, jak jsou rozděleny, jakých nabývají hodnot, z jakého jsou období nebo jestli se jejich struktura v čase nějak nemění. Tyto údaje můžeme pak dále využít tak, aby se s daty lépe pracovalo. Dále je popsána příprava burzovních dat a předzpracování datových sad tweetů v několika fázích.

5.5.1 Předzpracování burzovních dat

Protože jsou burzovní data celkem malá (několik MB) a obsahují pouze položky, které budou potřeba, byla jejich příprava jednoduchá. Spočívala jen v doplnění chybějících hodnot. Je to třeba udělat, protože algoritmy pro výpočet indikátorů potřebují všechny hodnoty z počítaného intervalu, a kdyby jim nějaká chyběla, tak by nefungovaly správně.

Z doby, kdy byla burza otevřena, jsou v datové sadě téměř všechny údaje. Chybí maximálně 2 hodnoty za den, které jsem doplnil průměrem skutečných hodnot, mezi kterými tato mezera ležela.

Mimo burzovní dobu jsou mezi daty různé časové rozestupy. Každý rozestup jsem celý vyplnil stejnými průměrnými hodnotami ze skutečných krajních hodnot, mezi kterými byl rozestup. Cena se v této době obvykle příliš neměnila a držela na podobné úrovní několik minut, takže to nebude mít téměř vliv na rozjezd některých indikátorů.

[9] http://www.textfixer.com/resources/common-english-words.txt
[10] http://xpo6.com/list-of-english-stop-words/
[11] https://code.google.com/p/stop-words/
[12] http://www.ranks.nl/stopwords
[13] http://www.lextek.com/manuals/onix/stopwords1.html

5.5.2 Extrakce velké Twitter sady

Příprava velké sady byla složitější, protože obsahuje mnohem více dat a nepotřebných údajů. Navíc jsou všechny tweety z jednoho měsíce zabaleny do archívu, se kterým se špatně pracuje.

Proto jsem se rozhodl data z archivu extrahovat, ale zachovat jeho vnitřní strukturu složek pro dny a hodiny tak, aby se v datech dalo lépe orientovat. Takže jsem v archívu postupně prošel všechny složky a zpracoval textové soubory, ve kterých byly uloženy tweety z jedné minuty. Vybral jsem z tweetů jen potřebné informace a uložil je do stejně pojmenovaného souboru uvnitř stejné struktury složek, ale jen mimo archív.

Z každého tweetu (obsahoval všechna data a metadata v JSON viz příloha A) jsem vybral jen údaje o čase vytvoření, id autora a text tweetu, které jsem pak uložil v prosté textové podobě, oddělené pouze oddělovačem. Informace o autorech tweetů jsem ukládal zvlášť. Pro každý den jsem vytvořil seznam unikátních autorů, do kterého jsem uložil autorovo id, jméno, jazyk a počet followerů.

Protože v sadě byly tweety psané různými jazyky, některé byly dokonce psané arabsky nebo japonsky, narazil jsem na problém s kódováním různých znakových sad. Kvůli tomu jsem musel veškeré získané údaje převést do unicode a uložit v escape[14] formátu tak, aby byla jejich hodnota všude zobrazitelná.

Navíc proti dříve uvedenému počtu tweetů obsahovala tato datová sada dalších asi 5 % řídících tweetů, které jsem přeskočil a dál nezpracovával. Tyto tweety bez obsahu mají speciální význam a to obvykle ten, že informují o smazání nějakého tweetu.

5.5.3 Filtrace velké Twitter sady

Velkou Twitter sadu nebudu používat celou, protože je příliš obsáhlá (průměrně 50 tweetů během každé vteřiny) a analýza všech jejich tweetů by trvala neúnosně dlouho. Musím z ní tedy vybrat jen vhodnou část.

U této sady tweetů mi půjde spíše o analýzu globální nálady, než o zaměření se pouze na určitou oblast tak, jak je to v malé Twitter sadě, která je cílena na konkrétní společnosti.

Rozhodl jsem se vybrat jen tweety uživatelů, kteří mají alespoň určitý počet followerů. Jsou tak teoreticky schopni zachytit globální veřejné mínění lépe než tweety od uživatelů s méně followery. Protože více followerů znamená, že si tweet pravděpodobně přečte víc lidí, které může dál ovlivnit, a více followerů také většinou znamená kvalitnější obsah tweetů, takže autor nepíše žádné nesmyly, ale spíše hodnotně reaguje na aktuální světové dění Nebudu tak muset zbytečně analyzovat tweety uživatelů s pár followery, kteří jsou v rámci globální dění zcela nevýznamní.

Jako přijatelná mi přišla hranice tisíce followerů. Protože ve tweetech, které si může přečíst více než tisíc lidí, by se už mohl odrážet náznak globální nálady. Uživatelé s více než tisícem followerů vytvořili zhruba pětinu všech tweetů v sadě, což lze považovat za dostatečně reprezentativní vzorek dat.

Z velké Twitter sady jsem tedy vyfiltroval a pro další práci použil pouze tweety od uživatelů, kteří měli více než tisíc followerů.

5.5.4 Rozdělení a filtrace slov

V této fázi jsem měl z obou sad připraveny všechny tweety, které budu analyzovat. Z malé sady jsou to všechny tweety a z velké jen ty, které prošli filtrací. Další předzpracování už

[14] Ne-ASCII znaky jsou reprezentovány textovou podobou jejich hexadecimální hodnoty s prefixem $\backslash u$.

bude pro obě sady stejné.

Každý tweet je teď jen řetězec znaků. Aby jsme ho mohli analyzovat a zjistit podle slovníků náladu, je třeba z něj odstranit neužitečné znaky a rozdělit jej na jednotlivá slova.

Nejprve jsem odstranil všechny ne-ASCII znaky a zbytek převedl na malá písmena, protože jsou slova ve všech slovnících složena jen z malých ASCII znaků. Poté jsem pomocí regulárních výrazů odstranil všechny odkazy, celá uživatelská jména začínající „@“ a počáteční znak „#“ z hashtagů. Odkazy a uživatelská jména vyřadím, protože jsou obvykle unikátní a nemají tak žádnou vypovídající hodnotu. Zatímco hashtagy ponechám, protože jsou obvykle nejdůležitější slova v tweetu. Poté nahradím opakování tří a více stejných znaků za dva stejné znaky, protože se v žádném slově tři stejné znaky po sobě nevyskytují, ale někteří lidé několikrát znak zopakují, když chtějí slovo zdůraznit nebo projevit emoce.

Následuje rozdělení tweetu na jednotlivá slova. K tomu jsem použil velice užitečnou funkci `word_tokenize` z knihovny pro zpracování jazyka *NTLK*. Tato funkce je přímo určena k rozdělení anglického textu na slova a dokáže správně zpracovat i nejrůzněji uspořádaný text, který není korektní. Při rozdělování bere v úvahu nejen bílé znaky, ale i interpunkční znaménka a neznámé znaky, které pak označí jako samostatná slova.

Teď je každý tweet seznam slov. Dále jsem všechna slova filtroval tak, aby zůstala jen plnohodnotná slova, protože se ve tweetech objevují různé zkomoleniny, vymyšlená slova, nesmyslné posloupnosti znaků nebo zkratky, které jsou při analýze k ničemu. Nejprve jsem ze slov odstranil všechny znaky kromě písmen, číslic, pomlčky, podtržítka a apostrofu, tedy znaků ze kterých jsou slova normálně složena. Pak jsem odřízl případný apostrof, pomlčku nebo podtržítko ze začátku a konce slova, protože takové slovo neexistuje. Dál jsem vybral jen slova, která měla alespoň tři znaky a nebyla označená jako číslo.

Posledním krokem předzpracování dat z Twitteru bylo odstranění stopslov. Tweet byl už rozložen na jednotlivá plnohodnotná slova, takže jednoduše stačilo odstranit z něj ta slova, která byla ve slovníku stopslov.

5.5.5 Projekce přes slovníky

Základní myšlenkou projekce přes slovníky je snížit počet slov ve tweetech, ale přesto co nejvíce zachovat informaci o náladě. To lze provést tak, že v tweetech nechám jen ta slova, která se vyskytují v nějakém náladovém slovníku. Zmenší se tak počet slov, ale zachovám informaci o náladě, protože když mám slovo, které není v žádném slovníku, bude vždy ohodnoceno nulou. Proto když takové slovo úplně vynechám nemělo by to mít na analýzu nálady výrazný vliv.

Zkombinováním všech slov ze všech náladových slovníků jsem vytvořil kompletní slovník, který obsahoval 25 671 unikátních slov, u kterých je možné určit nějakou náladu. Projekci jsem provedl pro obě Twitter sady až potom, co jsem zcela dokončil jejich předzpracování. Každý tweet byl tedy seznam slov, z kterého jsem odstranil slova, která nebyla v kompletním slovníku.

Ve velké Twitter sadě jsem průměrně odstranil 60 % slov a v malé Twitter sadě to bylo průměrně pouze 35 % slov. Takže ve velké sadě zůstalo 40 % a v malé dokonce 65 % původních slov, což ukazuje, že je příprava datových sad kvalitní a dokáže tweet správně rozdělit a vyfiltrovat vhodná slova.

V tomto případě se nejednalo o klasické předzpracování dat, ale o vytvoření dalších paralelních datových sad, protože mám v plánu dál používat sady původní, a také ty které vznikly projekcí přes slovníky.

5.6 Převod času

Ve všech datových sadách, které jsem získal, byl čas uveden v jiném formátu nebo časové zóně. V akciových datech je čas rozdělen do dvou sloupců na den a hodinu s minutami, které jsou v časovém pásmu ET (UTC-5). V malé datové sadě tweetů je čas v číselné podobě jako den, hodina a minuta dohromady v časovém pásmu UTC Ve velké datové sadě tweetů to je z časem o něco složitější. U každého tweetu je uveden přesný časový údaj o jeho vytvoření v dlouhé textové formě, která obsahuje i příslušnou časovou zónu. Ale přímo v datové sadě jsou tweety rozděleny do složek podle dní, ale v jiné časové zóně posunuté o 6 hodin. Proto jsem musel při zpracování jednoho dne načíst i složku předchozího dne, ale časy brát jen podle údajů ve tweetech.

Aby se v další práci lépe z daty pracovalo, bylo potřeba převést všechny časové hodnoty na jeden formát a do jedné časové zóny. V mojí práci by stačilo rozlišovat čas jen na úrovni minut, protože s menším timeframem nebudu nikde pracovat. Já jsem se ale rozhodl, že budu všechny časové hodnoty uchovávat v sekundách ve formátu Unix Timestamp. Zvolil jsem si ho, protože je to univerzální způsob uchování času, pro který existuje mnoho funkcí pro snadnou manipulaci a převod na ostatní časové formáty. Navíc se u něj předejde častým pochybnostem o časové zóně, protože je implicitně v UTC.

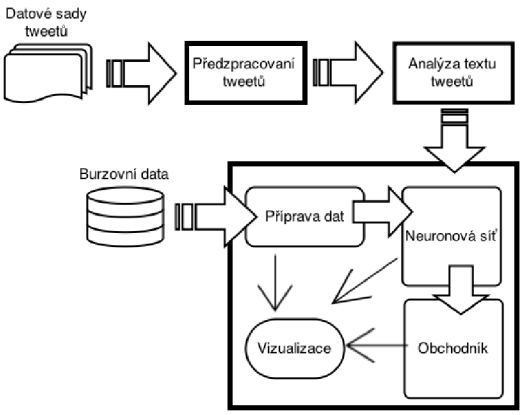
Kapitola 6

Návrh systému

Cílem této práce je vytvořit systém, který bude schopen analyzovat historická burzovní data a zprávy z Twitteru. Z výstupů analýzy by měl systém predikovat budoucí vývoj trhu. Na základě predikce pak systém vytvoří jednoduchou obchodní strategii, kterou otestuje.

Analýza burzovních dat bude spočívat ve výpočtu technických indikátorů a v hledání podobných opakujících se situací trhu (vzorů), které budou reprezentovány několika předchozími hodnotami ceny akcie. Analýza zpráv z Twitteru bude provedena dvěma různými metodami na obou sadách tweetů. V první metodě se bude podle několika náladových slovníků určovat průměrná nálada slov v tweetech. Druhá, trochu nestandardní metoda bude charakterizovat text tweetů pomocí rekurentní neuronové sítě.

Systém bude výsledky těchto analýz předávat do neuronové sítě, která tak bude nepřímo hledat vzájemné podobnosti a jejich souvislosti s budoucím pohybem ceny.



Obrázek 6.1: Schéma návrhu propojení jednotlivých částí systému.

Systém bude složen ze tří samostatných částí,jejichž spolupráce je znázorněna na obrázku 6.1. První dvě menší části systému budou mít za úkol pouze zpracovat a analyzovat zprávy z Twitteru.

První část bude obstarávat jen předzpracování dat z Twitteru tak, jak to bylo popsáno v kapitole 5.5. Tato část bude pracovat zcela samostatně a výsledek uloží do souborů, které poté bude zpracovávat další část systému. Druhá část systému bude vycházet z těchto předzpracovaných tweetů, které bude analyzovat. V této části budou probíhat obě metody analýzy a její výstupy budou vektory nálad a charakteristiky tweetů, které se opět uloží do souborů.

Třetí část bude největší a nejdůležitější. Bude složena s několika prvků, které budou společně obstarávat chod výsledného systému. Tyto prvky budou provádět zejména zpracování burzovních dat a technickou analýzu, predikci, testování obchodní strategie a vizualizaci. Jádrem této části bude neuronová síť, která by měla po naučení predikovat budoucí cenu.

Analýza tweetů tedy bude na rozdíl od analýzy burzovních dat umístěna do samostatné části, protože zpracování tweetů je mnohem časově náročnější než technická analýza. Tak bude analýza zpráv z Twitteru provedena pouze jednou a uloží se. Dále se pak budou využívat předem vypočítané vektory. Doba analýzy burzovních dat je naproti tomu vcelku zanedbatelná a může se provést i opakovaně. Kdyby se dopředu zpracovala i burzovní data, a spočítaly se všechny potřebné indikátory pro různé periody na všech použitých timeframech, tak by uložení těchto dat zabralo několikanásobně víc místa než datová sada samotná. Jejich načtení by pak nebylo o moc výhodnější než samotný výpočet, protože se jedná jen o několik jednoduchých matematických operací.

Dále jsou v této kapitole popsány podstatné části celého systému.

6.1 Zpracování burzovní dat

Jediným předzpracováním burzovních dat je doplnění chybějících hodnot viz 5.5.1. Dostaneme tak pro jeden den kolem 700 OHLCV hodnot zhruba mezi sedmou hodinou ranní a osmou večerní, které jsou na minutovém timeframu. Dál se bude pracovat s celým tímto dnem, tedy se všemi hodnotami, ze kterých se ale nakonec přímo použijí jen některé. Při predikci se budou používat jen hodnoty z doby, kdy byla burza otevřena. Proto se nakonec z každého dne vždy vystřihnou pouze hodnoty ze stejného časového úseku, které se předají do neuronové sítě. Uložené hodnoty mimo obchodní dobu budou sloužit pouze k rozjíždění indikátorů, ale přímo se nepoužijí.

Data jsou určena k obchodovaní na minutovém timeframu, ale systém bude analyzovat vliv tweetů na obchodování na několika různých timeframech. Systém bude určen hlavně pro intraday obchodování, kde bude obchodovat na nejpoužívanějších intraday timeframech 1, 2, 5 a 10 minut. Větší timeframe by už ztrácel smysl, protože bychom v rámci obchodního dne dostali jen několik hodnot.

Data bude nutné agregovat do větších timeframů. To se provede tak, že se vezmou všechny hodnoty (svíčky) menšího timeframu z intervalu o velikosti většího timeframu. Z nich se vezme maximální high a minimální low, jako nové high a low. Jako nová open hodnota bude použita open prvního a jako nová close bude close posledního. Tím získáme z daného intervalu novou svíčku na větším timeframu. Takto přepočítaná data se budou moci uložit pro každou akcii a timeframe zvlášť do samostatného souboru tak, aby se příště nemusely počítat, ale stačilo je jen načíst.

V rámci technické analýzy budu z burzovních dat počítat pouze nejpoužívanější [13] technické indikátory, jako jsou MA, RSI a CCI (viz 2.7.1), s různou periodou. Tyto indikátory se budou počítat pro celý den, tedy i pro hodnoty mimo obchodní dobu. Pro každý den dostaneme tolik hodnot indikátoru, kolik bude hodnot v datech. Kvůli rozjezdu výpočtu indikátorů, bude obvykle na začátku každého dne několik hodnot nedefinovaných, ale zato budou jednoduše přiřazena k výchozím datům. Přu výběru dat z obchodní doby, pak bude stačit vystřihnout z pole indikátorů úplně stejnou část jako z dat.

6.2 Analýza nálady

Obecně se nálada v textu může analyzovat několika způsoby. Nejvíce pracné je, když někdo ručně projde celý text a subjektivně ohodnotí všechna slova nebo věty, což je obvykle nejpřesnější. Tuto metodu lze částečně zautomatizovat. Ručně se ohodnotí jen část textu a na základě podobností ve struktuře vět nebo uspořádání slov se automaticky určí nálada v ostatním textu. Existuje také zcela automatický způsob, který nahradí ruční ohodnocování, ale lze s ním analyzovat jen nálady pozitivní a negativní. Použije se velká databáze recenzí například k filmům (IBDM) nebo zboží (Amazon, eBay). Každá recenze má kromě textu i výsledné hodnocení autora, z kterého vyplývá, zda je recenze spíše pozitivní nebo negativní. Hledají se slova nebo větné konstrukce, které se nejčastěji vyskytují u recenzí s velmi nízkým nebo naopak vysokým hodnocením.

Zjednodušením výsledků těchto metod a to zanedbáním struktury vět a pořadí slov, můžeme získat pouze seznam slov, který odráží danou náladu. Tato slova vytvoří náladový slovník, pomocí kterého se může určit nálada jednoho samostatného slova, bez ohledu na to v jakém je kontextu. Použití slovníků je obecně nejpoužívanější způsob analýzy nálady v textu, protože existují volně dostupné kvalitní slovníky a jejich použití je díky zanedbání kontextu slov velmi jednoduché. Pro svou práci použiji několik slovníků viz 5.4, z nichž jsou některé vytvořeny ručně a jiné zcela automaticky.

Systém bude zjišťovat, zda se v opakujících dobách s podobnou náladou chová podobně i burzovní trh. Proto nepůjde o přesnou náladu v tweetech jako takovou, ale spíše o to nalézt pomocí slovníků vhodnou hromadnou charakteristiku textu tweetů.

Nejmenší časový úsek, na kterém budu náladu hromadně analyzovat, bude jedna hodina. Kratší doba by neměla smysl, protože chci zachytit obecnou průměrnou náladu, která je v menších intervalech příliš proměnlivá a může být ovlivněna i jediným tweetem. Mezi nápadem na tweet, jeho publikací a případně dobou než si ho lidé přečtou často uběhne i několik desítek minut, proto je vhodnější seskupovat tweety z delšího časového období. Kromě nálady z každé hodiny budu zjišťovat i souhrnnou náladu za celý den, která by mohla lépe zachytit chování trhu v rámci dne.

Vstupy analýzy nálady tedy budou několik náladových slovníků a dva seznamy plnohodnotných slov v tweetech (první kompletní a druhý projekcí přes slovníky). Každé slovo bude ohodnoceno přes všechny slovníky, z čehož bude pro jedno slovo vytvořen náladový vektor. Tyto vektory budou pro všechna slova v rámci tweetu sečteny a po přidání položky s počtem slov vznikne vektor souhrnné nálady v jednom tweetu. Ve skupině tweetů se jejich náladové vektory také sečtou a k výslednému vektoru se přidá položka s počtem tweetů. Tak vznikne finální náladový vektor, který charakterizuje náladu ve skupině tweetů. Tento vektor tedy obsahuje sumy jednotlivých nálad pro všechna slova a počet slov a tweetů. Nálady se uchovávají v podobě sum, protože se tak celé vektory lehce kombinují (sčítají) a dá se z nich vypočítat průměrná nálada na jeden tweet, které se budou dál používat. Finální vektor se potom bude používat celý nebo se z něj vyberou jen prvky z určitých slovníků. Výsledkem analýzy nálady tak bude náladový vektor pro každou hodinu a jeden souhrnný vektor pro celý den a to pro každou sadu tweetů zvlášť.

Vektor nálady pro kompletní sadu slov a sadu vytvořenou projekcí se bude lišit jen v počtu slov a tweetů, protože slova, která nejsou v žádném slovníku, budou mít nulový náladový vektor. Stačí tedy analyzovat náladu jen u menší sady, která vznikla projekcí, a výsledné hodnoty použít i pro kompletní sadu, jen se musí zvlášť spočítat slova a tweety. Z jednoho vektoru sum nálad tak získám dvě různé průměrné náladové charakteristiky.

Analýza nálady se provede pro malou i pro velkou sadu tweetů. Ve finále tak bude

k dispozici pro každou hodinu (den) a akcii čtyři různé náladové charakteristiky.

6.3 Analýza pomocí rekurentní sítě

Kromě analýzy nálady existují i jiné způsoby jak charakterizovat text. Jedním z nich je použití rekurentní neuronové sítě a to konkrétně nástroje RNNLM[1].

Rekurentní síť je obecně neuronová síť, ve které se signál nešíří pouze od vstupní vrstvy k výstupní, ale dochází k zpětnovazebnímu šíření informace. To vytváří jistý vnitřní stav sítě, který umožňuje zohlednit dynamické časové chování a historický kontext vstupních dat. Odezva sítě, tak není daná jen vstupem (jako je to u klasických neuronových sítí), ale závisí také na vnitřním stavu sítě.

RNNLM obsahuje rekurentní neuronovou síť založenou na jazykových modelech. Tento nástroj slouží hlavně ke statické analýze textu, která má široké uplatnění například v automatickém rozpoznávání řeči nebo strojovém překladu.

V své práci použiji RNNLM k získání charakteristiky textu tweetů. Tato charakteristika by mohla souviset s obchodováním tím způsobem, že trh se bude chovat podobně v obdobích, které mají podobnou charakteristiku slov ve tweetech.

Systém bude přímo využívat hotového nástroje RNNLM, tak že mu předá předzpracované tweety z období, které bude chtít charakterizovat. RNNLM na základě těchto slov naučí svůj vnitřní model, což znamená, že v rekurentní síti adaptuje váhy mezi neurony a jejich aktivační hodnoty. Po naučení se z rekurentní sítě vezmou neurony poslední skryté vrstvy. Tato vrstva leží přímo před vrstvou výstupní, a proto se jedná o nejužší řez celé sítě (bottleneck), kterým procházejí všechny informace. Aktivační hodnoty neuronů této vrstvy pak odráží charakteristiku vstupního textu. V této práci použiji rekurentní síť s deseti neurony ve skryté vrstvě.

Stejně jako u analýzy nálad se při této analýze bude charakterizovat text pouze v rámci každé jedné hodiny nebo každého jednoho dne. Také se budou charakterizovat zvlášť kompletní sady tweetů a sady vytvořené projekcí přes slovníky, u kterých na rozdíl od nálad bude v charakteristice znatelný rozdíl.

Aby se charakterizoval celý text a ne jen posloupnost jednotlivých slov a tweetů, protože pořadí tweetů v rámci několika minut je nevýznamné, budou do jednoho běhu RNNLM dány všechny tweety v náhodném pořadí a navíc stokrát zduplikovány. Tím se pro jednu skupinu tweetů vytvoří čtyři různé charakteristiky, což je tedy pro obě sady tweetů celkem osm různých charakteristik pro každou hodinu (den) a akcii.

6.4 Neuronová síť

Jádrem celého systému bude vrstvená neuronová síť s jednou skrytou vrstvou, která se bude učit metodou zpětného šíření chyb. Tato síť se bude učit predikovat budoucí cenu akcií na základě technické analýzy a analýzy textu tweetů.

Volitelné parametry sítě, jejichž vhodné hodnoty pro různé timeframy budu hledat v rámci experimentů, budou velikost skryté vrstvy, učící faktor a počet vektorů, po kterých se přepočítá chyba. Síť tedy dostane vstupní vektory a příslušné výstupní vektory, které představují požadovaný výstup sítě. Samotné učení bude probíhat v iteracích, ve kterých vždy natrénuje síť na celou trénovací sadu. Aby nedošlo k přeučení sítě, tedy ke stavu, kdy síť ztratí schopnost generalizace a perfektně tak reaguje na trénovací sadu, ale špatně

[1] http://rnnlm.org/

na jiná data, bude před každou iterací trénovací sada náhodně rozdělena na trénovací a validační část v poměru 4:1. Na trénovací části se síť naučí a na validační zjistí chybu odezvy.

Po každé iteraci se vyhodnotí chyba sítě a pokud se zvětšila, což znamená, že se reakce sítě zhoršily, bude navrácen stav sítě z předchozí iterace. Výhoda návratu k předchozímu lepšímu výsledku bude předmětem jednoho z experimentů.

Pokud bude rozdíl mezi aktuální chybou sítě a chybou z předchozí iterace menší než určitá tolerance, poté bude učící faktor snížen na polovinu, aby mohla síť zpřesnit svou odezvu. Učení skončí buď po dosáhnutí maximálního množství iterací, nebo když má být učící faktor zmenšen po třetí, což znamená, že se mezi iteracemi síť skoro nemění a proto není třeba dál pokračovat.

6.4.1 Vstup a výstup

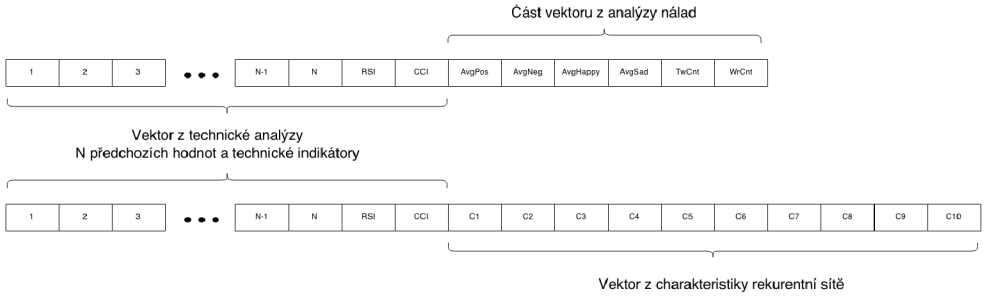
Neuronová síť nepřímo hledá vztah mezi vstupním a výstupním vektorem, proto by měly vektory pokud možno obsahovat pouze takové položky, které by mohly být teoreticky závislé.

Hlavní částí vstupního vektoru bude několik historických hodnot z bezprostředně předchozí doby, čímž jsem se inspiroval v [8]. Na rozdíl od této práce, kde se používá přímo typická cena, já budou do vektoru dávat hodnoty, které budou vyjadřovat procentuální změnu mezi historickou a aktuální typickou cenou akcie. Díky tomu se budou moci srovnávat všechny historické vzory bez ohledu na skutečnou výši ceny.

K tomuto vstupnímu vektoru se budou v rámci experimentů přidávat aktuální hodnoty indikátorů (MA, RSI, CCI), u kterých se bude sledovat, jaký budou mít vliv na učení a odezvu sítě.

Výstupním vektorem neuronové sítě bude pouze jedna hodnota a to budoucí cena akcie, která bude zase vyjádřena jako procentuální změna vůči aktuální typické ceně. Nebude se jednat o cenu akcie v následujícím časovém okamžiku, ale o průměr několika budoucích cen, díky čemuž se získá směr dlouhodobějšího pohyb trhu a vyruší se tak chvilkové výkyvy.

Do vstupního vektoru budou potom přidávány výsledky z analýzy tweetů, jak je znázorněno na obrázku 6.2 V rámci experimentů se přidá buď celý vektor z analýzy nálad nebo jen jeho určité části. Vyzkouší se například výsledky z jednotlivých slovníků nebo počet slov a tweetů. Z analýzy pomocí rekurentní sítě se bude do vstupního vektoru vždy přidávat celý vektor charakteristiky, protože jeho jednotlivé složky nemají samostatně žádný význam.



Obrázek 6.2: Ukázka skládání dvou různých vstupních vektorů.

Kromě charakteristiky z aktuální hodiny (dne) se vyzkouší charakteristika z několika předchozích hodin(dní).

Systém bude mít jeden vektor charakteristiky tweetů pro celou hodinu (den). Zatímco burzovních údajů bude mít z té doby několik. Proto se ke všem vstupním vektorům technické analýzy v rámci hodiny (dne), přidá naprosto stejný vektor nálad. Důležité bude, aby systém pohlídal časy a správně namapoval dlouhodobé nálady na příslušná burzovní data.

6.4.2 Normalizace

Aby se dosáhlo správného fungování sítě, je třeba aby hodnoty ve vstupních a výstupních vektorech byly pouze z určitého intervalu. Tento interval je dán aktivační funkcí neuronů, což bude u všech neuronů sigmoida. Její použitelná oblast z definičního oboru je zhruba mezi mínus jednou a jednou a obor hodnot sahá od nuly do jedné. Proto je třeba hodnoty vstupních vektorů transformovat do intervalu od mínus jedné do jedné a hodnoty výstupních vektorů do intervalu od nuly do jedné.

Vstupní vektory se budou normalizovat hromadně (ale každá položka zvlášť), před vstupem do sítě. Vezme se celá trénovací sada a spočítá se průměrná hodnota jednotlivých položek a směrodatná odchylka. Od hodnot se poté odečte průměr a výsledek se vydělí směrodatnou odchylkou, čímž dojde k rozprostření hodnot a co nejlepšímu využití vstupního intervalu. Tento způsob normalizace se označuje jako z-score.

Výstupní vektor, který bude udávat procentuální změnu ceny, která může teoreticky dosáhnout hodnot od mínus jedné do jedné, se lineárně transformuje přímo do intervalu nula až jedna.

6.5 Obchodní strategie

Systém bude obsahovat jednoduchou obchodní strategii, která bude vycházet převážně z výsledků predikce. Základní myšlenkou úspěšného obchodování je podle autora [7] obchodovat méně a vydělávat více. To znamená, že je výhodnější provést méně obchodů, které jsou více výnosné, než provést více obchodů, které jsou výnosné méně. Proto je třeba provádět pouze obchody, od kterých se očekává vysoký výnos.

Na základě predikce se tedy budou vybírat jen situace, kdy je predikována velká změna. Pro stanovení dostatečně velké změny bude využita dosavadní průměrná predikovaná hodnota a její směrodatná odchylka. Pokud se bude predikovaná hodnota lišit od průměrné hodnoty o více než dvě směrodatné odchylky, poté bude dán signál ke vstupu do příslušné pozice dle směru predikce. Při normálním rozdělení náhodné veličiny by se tak jednalo o méně než 5 % všech hodnot, což se dá považovat za dostatečně malý počet predikcí, které predikují velkou změnu. Výstup pozice pak bude proveden na základě malého stoplossu, nebo když se predikce otočí a dostane se do neutrální nebo dokonce opačné situace.

Testování obchodní strategie poté bude probíhat jako klasické obchodování, ale budou se používat pouze close ceny. Část systému, která bude obchodování provádět, bude postupně dostávat jednotlivé predikované hodnoty a momentální stav burzy, na základě kterých bude okamžitě vykonávat obchodní příkazy, tak jako by to bylo u skutečného obchodování.

6.6 Vizualizace

Systém bude také schopen vhodně data vizualizovat ve formě několika pod sebou paralelních aktivních grafů, ve kterých se bude moci určitá oblast přiblížit nebo posunout. Tyto grafy budou sloužit k otestování funkčnosti systému a demonstraci jeho výsledků.

Protože bude systém zaměřen na intraday obchodování, bude naráz zobrazovat pouze informace týkající se jednoho dne. Pomocí tlačítek pak půjde mezi dny přepínat. Systém také umožní kromě zobrazení aktivních grafů, uložit jednotlivé dny jako obrázky.

Stěžejním grafem bude graf svíčkový, ve kterém budou zobrazeny burzovní data. Do toho grafu se bude také zaznamenávat průběh klouzavého průměru s dvěma různými periodami. Chybějící burzovní data, které budou dopočítána jako průměry okolí, budou označena křížkem. Také se v tomto grafu budou případně zobrazovat provedené obchody a to ve formě obdélníků s výsledkem obchodu. Tyto obdélníky budou mít barvu podle typu obchodu a budou sahat od svíčky, kdy se vstoupilo do pozice, po svíčku, kdy se z pozice vystoupilo, a to mezi jejich close hodnotami.

Dále budou zobrazeny dva menší grafy jeden pro Volume a druhý pro technický indikátor RSI nebo CCI. Poslední graf bude sloužit k porovnání predikovaných hodnot a budoucích skutečných hodnot, které se měly predikovat. Bude tedy obsahovat dvě křivky, z nichž bude patrné, kde a jak moc byla predikce úspěšná či neúspěšná.

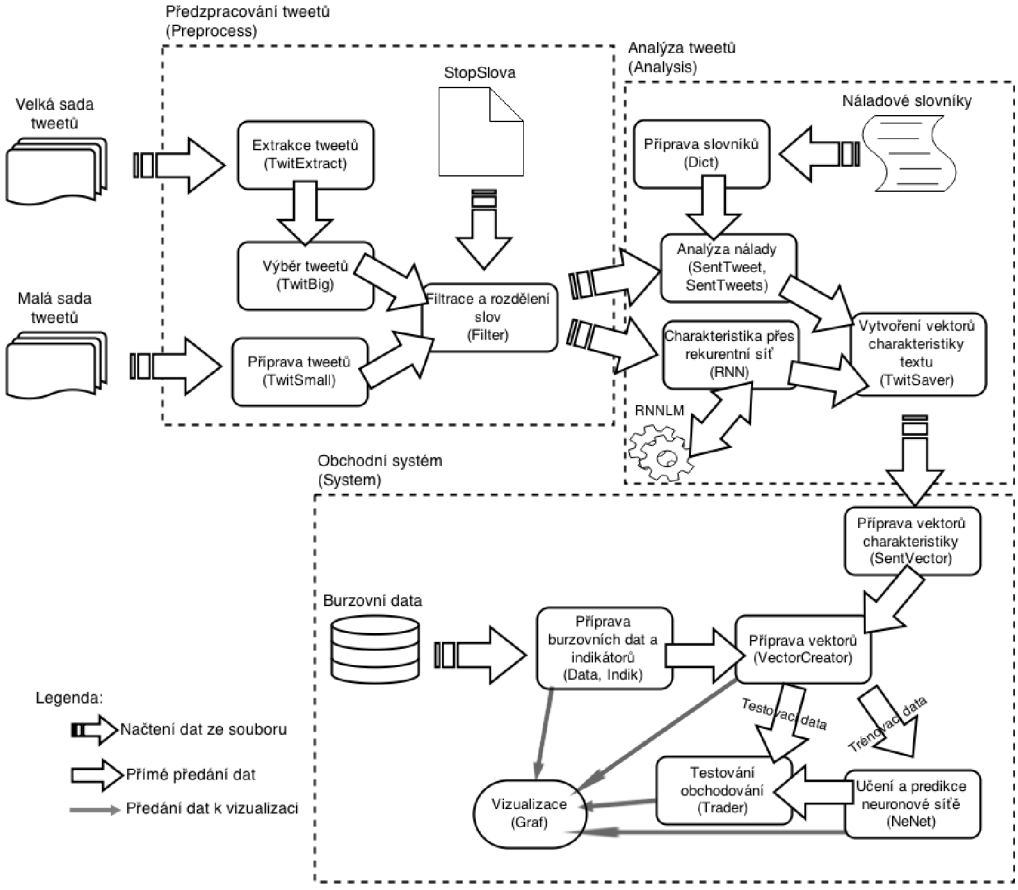
Na časové ose těchto grafů bude uveden čas v časovém pásmu ET, tedy ve vlastním čase burzy.

Kapitola 7

Implementace

Výsledný systém se skládá ze tří hlavních modulů, které odpovídají jednotlivým velkým částem systémů, jak byly uvedeny v návrhu. První modul předzpracuje tweety, druhý modul je analyzuje a třetí modul provádí predikci a testování obchodní strategie. Vedle toho obsahuje systém čtvrtý malý řídící modul, který je určen ke spouštění celého systému a experimentů.

Schéma na obrázku 7.1 zachycuje strukturu výsledného systému. U každé pracovní fáze je poznačena třída nebo soubor, který tuto část převážně obstarává.



Obrázek 7.1: Detailní schéma propojení jednotlivých částí výsledného systému.

Celý systém je implementován v jazyce Python, což je vysokoúrovňový hybridní jazyk, který se vyznačuje velmi krátkým a dobře čitelným kódem.

V celém systému je hojně využívaná matematická knihovna *NumPy*[1], která je určena pro práci s numerickými daty. Tato knihovna poskytuje mnoho funkcí pro pohodlnou práci s velkými a multidimenzionálními poli. Dokáže také celé pole uložit do souboru a zpátky načíst ve stejném formátu.

V této kapitole jsou dále moduly blíže popsány a jsou krátce přiblíženy funkce jejich jednotlivých souborů.

7.1 Předzpracování tweetů

Modul `Preprocess` slouží k předzpracování tweetů. Všechen výstup se ukládá do textových souborů pro další použití.

TwitExtract

Kód v tomto souboru provádí extrakci tweetů z archívů velké Twitter sady. Načítá denní soubory kompletních údajů o tweetech, které jsou ve formátu JSON. Vybírá z nich podstatné položky, které po minutách ukládá do souboru, a zvlášť po dnech ukládá unikátní uživatele.

TwitSmall a TwitBig

V těchto souborech jsou stejnojmenné třídy, které pracují s tweety uloženými v textových souborech. Poskytují tak vrstvu, které se zadá datum a čas, a ona vrátí seznam všech tweetů z daného období. `TwitSmall` si umí malou sadu připravit a rozdělit souborů po dnech. `TwitBig` načte soubory tweetů a uživatelů připravené v `TwitExtract`, které dokáže provázat a může tak filtrovat tweety podle počtu autorových followerů.

Filter

Tato knihovna obsahuje funkce pro předzpracování textu tweetů, tedy pro filtraci znaků a rozdělení slov. Jako vstup dostane pole řetězců, kde každý řetězec je text jednoho tweetu. Jako výstup vrací seznam polí s plnohodnotnými slovy.

Pro zpracování textu tweetu, jak je popsáno v kapitole 5.5.4, využívá hlavně regulárních výrazů a pro rozdělení slov funkci `word_tokenize` z knihovny pro zpracování přirozeného jazyka *NTLK*.

Helpers

Tato knihovna obsahuje pomocné funkce pro tvorbu adresářů a hlavně třídu `TimeConv`, která je určena pro práci s časem. Tato třída umožňuje jednoduchou konverzi mezi několika formáty času navzájem.

[1] http://www.numpy.org/

7.2 Analýza tweetů

Modul `Analysis`, který má za úkol analyzovat předzpracované tweety a vytvořit vektory náladové charakteristiky. K uloženým datům přistupuje prostřednictvím objektů tříd `TwitSmall` a `TwitBig`.

Dicts

Tento soubor obsahuje třídu `WordDicts`, která zapouzdří práci se všemi slovníky nálad a stopslov, které jsou popsané v 5.4. Vybrané slovníky načte, zpracuje, případně spojí a ostatním objektům je poskytuje ve formě připraveného seznamů slov.

SentiTweet a SentiTweets

Tyto soubory obsahují objekty stejnojmenných tříd, které analyzují náladu v tweetech. `SentiTweet` analyzuje a uchovává informace o jednom tweetu a jeho náladový vektor. Objekt třídy `SentiTweets` uchovává množinu objektů třídy `SentiTweet`, a poskytuje agregační funkce k výpočtu celkové nálady ve všech tweetech, které obsahuje.

S objekty se pracuje vždy jen prostřednictvím `SentiTweets`, který z výstupu funkce souboru `Filter` dostane vstupní pole polí slov. Po jednom tweetu pole rozdělí a předá do objektů třídy `SentiTweet`, kde probíhá samotné ohodnocení přes náladové slovníky.

RNN

Tato knihovna obaluje použití nástroje RNNLM, pro získání charakteristiky textu pomocí rekurentní sítě. Využívá hotovou aplikaci, kterou spouští přes systémové funkce. Nejprve připraví soubor se všemi slovy, které se mají analyzovat. Případně je ještě stokrát náhodně zopakuje a předá jako parametr do RNNLM. Počká na dokončení výpočtu a ze souboru z výstupem přečte informace o skryté vrstvě, které představují charakteristiku textu. Nakonec po sobě všechno smaže.

TweetSaver

Instance třídy `TweetSaver` řídí veškerou analýzu tweetů. Skrze `TwitSmall` a `TwitBig` dostává filtrovaný seznam slov z předzpracovaných tweetů. Slova z tweetů nechává ohodnotit, jak náladou v objektu třídy `SentiTweets`, tak charakteristikou přes rekurentní síť skrze funkce z `RNN`. Primárně ukládá souhrnné nálady za jednu hodinu, ale obsahuje i funkce pro její agregaci do jednoho dne. Výsledné vektory charakteristiky textu ukládá do textového souboru.

7.3 Systém

Modul `System` představuje samostatný predikční a obchodní systém. Jako vstup načítá vektory charakteristiky textu a burzovní data, u kterých provede technickou analýzu.

Z těchto dat vybere vhodné položky, na základě kterých predikuje budoucí pohyb trhu. Úspěšnost predikce a obchodní strategie pak otestuje a zobrazí podstatná informace.

SentVector

Objekt třídy `SentVector` načítá a připravuje vektory charakteristiky a nálady, které připravil `TwitSaver`. Podle zadaných parametrů, jako jsou typ charakteristiky, specifikace datové sady, timeframe, posun a čas, vytvoří kompletní vektor pro celý den. Na základě timeframu rozkopíruje vektory charakteristiky z jedné hodiny do příslušného počtu vektorů.

Data

Objekt `Data` slouží k práci s burzovními daty. Načte informace ze souborů, provede jejich jednoduché předzpracování a případně je přepočítá na potřebný timeframe. Připravené sady jednotlivých akcií a timeframů dokáže pro další použití uložit do souboru.

Indikator

Pomocný objekt třídy `Indikator` je určen k výpočtu indikátorů texhnické analýzy. Používá knihovnu *TA-Lib*[2], která obsahuje funkce pro výpočet asi dvou set nejrůznějších technických indikátorů. Dostane pouze hodnoty cen akcii z jednoho dne a vrátí průběh hodnot vybraných indikátorů z celého dne.

VectorCreator

`VectorCreator` je důležitý řídící objekt systému, který připravuje vektory pro neuronovou síť. Všechna data dostává prostřednictvím objektů třídy `Data`, `Indikator` a `SentVector`. Z těchto dat, tedy vektorů technické analýzy a charakteristiky textu, skládá kompletní vektory, které jsou určeny na vstup a výstup neuronové sítě.

Dostává kompletní denní data (i z doby kdy se neobchoduje), ze kterých až zde, těsně před vstupem do sítě, vyřízne použitelnou část. Na základě doby, kterou vyřezává, a timeframu získá od objektu třídy `SentVector` odpovídající vektory charakteristiky. Synchronizuje tak data z technické a fundamentální analýzy.

Pro jednodušší manipulaci vytváří vektory po jednotlivých dnech, kde je ze všech dnů vybráno stejné časové období a mají tak stejný počet vektorů. Vytvořené vektory také může uložit pro další použití.

NeNet

Instance této třídy představuje neuronovou síť, která je určena k predikci budoucího stavu trhu na základě vektorů vytvořených objektem třídy `VectorCreator`.

Neuronová síť je vytvořena prostřednictvím knihovny *Theano*[3], která slouží k efektivnímu definování, optimalizaci a vyhodnocení matematických výrazů, které využívají vícerozměrná pole. Tato knihovna je silně provázána s dříve zmíněnou knihovnou *Numpy*. Poskytuje také transparentní využití CUDA akcelerace, která umožňuje provádět výpočty na GPU.

Při vytvoření musí tento objekt dostat vstupní a výstupní trénovací vektory, podle jejichž dimenzí vytvoří vrstvenou neuronovu síť. Z jednotlivých položek vstupních vektorů také spočítá průměr a směrodatnou odchylku, které použije pro jeho normalizaci. Práce probíhá ve dvou krocích. Nejprve se síť natrénuje metodou `trainNet` na trénovacích vektorech.

[2] http://mrjbq7.github.io/ta-lib/index.html
[3] http://deeplearning.net/software/theano/

Po tomto naučení se můžou metodou `testNet` předat do sítě vstupní testovací vektory, které musí být stejného tvaru jako trénovací. Výstupem této metody jsou predikované hodnoty rozdělené po dnech tak, aby se s nimi dalo jednotlivě dál pracovat.

Objekt také umožňuje po naučení uložit (a zpětně načíst) do souboru kompletní stav sítě, jako je její struktura, váhy synapsí mezi neurony, aktivační hodnoty neuronů a udaje k normalizaci vstupního vektoru.

Trader

Objekt `Trader` obsahuje obchodní strategii, která je založená na predikci (viz 6.5). Na vstupu dostane po dnech rozděleny predikované hodnoty z neuronové sítě a close hodnoty ceny akcie z testovacích dní. Podle nich simuluje průběh obchodování a testuje tak úspěšnost obchodní strategie.

Close hodnoty má jen jako informaci o aktuální ceně, ale v rámci obchodní strategie je nevyužívá. Dává obchodní signály pouze na základě predikce a případně aktuální finanční bilance. Obchodování testuje na několika testovacích dnech, kdy obchoduje každý celý den zvlášť.

Výstupem tohoto objektu je průběh obchodování, což je obchodní list, ve kterém je pro každý den uveden seznam obchodů s informacemi od kdy do kdy obchod probíhal a jaký měl finanční výsledek.

Graf

Objekt třídy `Graf` vizualizuje podstatná data systému po jednotlivých dnech. Data dostává od objektů tříd `Data`, `Indikator`, `VectorCreator`, `NeNet` a `Trader`. Pro vykreslení dat do grafů využívá knihovnu *Matplotlib*, která umožňuje vykreslit aktivní grafy, se kterými lze dále manipulovat, nebo graf uložit jako obrázek. Jak je vidět na obrázku 7.2, zobrazuje několik paralelních grafů, tak jak jsou popsány v kapitole 6.6.

Obrázek 7.2: Grafický výstup systému.

7.4 Spouštěcí soubory

Posledním modulem systému je `Main`. Tento modul obsahuje objekty, které zapouzdřují použití systému a provádí experimenty.

Albert

Objekt třídy `Albert` zapouzdřuje práci s celým systémem pro tvorbu vektorů, predikci, obchodování a vizualizaci. Pomocí jeho metod lze zadat parametry, podle kterých nastaví jednotlivé části systému a pak řídí celý průběh chodu systému. Je to také spustitelné místo aplikace, které inicializuje a spustí celý systém dle parametrů v konfiguračním souboru.

Experiment

Funkce v tomto souboru vytváří jednotlivé experimenty. K inicializaci a spuštění systému používají objekt třídy `Albert`. Dále jsou zde také funkce, pomocí kterých jsou výsledky experimentů agregovány a vyhodnoceny.

Kapitola 8

Experimenty

Tato kapitola obsahuje popis a vyhodnocení všech experimentů. Připravil jsem několik sad experimentů, které by měly otestovat výsledný systém a pokusit se najít závislost mezi cenou akcii a zprávami z Twitteru a zhodnotit obchodní strategii. První sada experimentů má najít nejvhodnější kombinaci parametrů neuronové sítě. Druhá sada experimentů bude hledat nejlepší prvky, které se použijí z technické analýzy. Z nejlepších výsledků pak bude vycházet třetí sada experimentů, ve které se přidají výstupy z analýzy nálad Twitteru, u kterých se bude sledovat vliv na úspěšnost predikce a výnos obchodování. Stejně tak i ve čtvrté sadě experimentů, kde se místo nálady přidá vektor charakteristiky textu tweetů, který vznikl z analýzy přes rekurentní síť. V poslední sadě experimentů se vyhodnotí nejlepší obchodní systém a jeho výnos pro jednotlivé akcie.

Vstupní burzovní datovou sadu, která obsahuje 167 dní, rozdělím na trénovací část, která bude obsahovat prvních 137 dní, a testovací část se zbylými 30 dny. V těchto dnech budu provádět intraday obchodování na timeframech (TF) 1, 2, 5 a 10 minut. Počet vektorů, které budou při jednotlivých timeframech použity je uveden v tabulce 8.1.

Někteří zkušení obchodníci [16] doporučují neobchodovat na začátku a konci obchodní doby, protože je v této době trh příliš nestálý a těžko predikovatelný. Proto v rámci prvního experimentu porovnám výsledky predikce z celého obchodního dne a zkráceného o půl hodiny na začátku a na konci. Dál budu používat jen tu dobu, ze které dostanu lepší výsledek predikce.

Počet vektorů	**Celý obchodní den (9:30 - 16:00)**		**Ořezaný obchodní den (10:00 - 15:30)**	
Timeframe	**Celkem**	**Za den**	**Celkem**	**Za den**
1	53 430	390	45 210	330
2	26 715	195	22 605	165
5	10 686	78	9 042	66
10	5 343	39	4 521	33

Tabulka 8.1: Počet vektorů v závislosti na timeframu.

Úspěšnost systému budu hodnotit dvěma způsoby. První je průměrná chyba predikce v testovacích dnech. Druhým je průměrný výnos obchodování během testovacích dnů. Tyto dvě hodnoty na sobě nejsou zcela závislé, a proto může strategie postavená na horší predikci dosahovat vyššího zisku, než strategie s více přesnější predikcí. Někdy prostě horší predikce muže překvapivě dávat výhodnější obchodní signály.

Při hodnocení systému budu používat průměrné výsledky experimentů ze všech akcí, protože chci najít univerzální strategii, která není závislá na konkrétní akcii.

8.1 Experiment 1 - nastavení sítě

Cílem této sady experimentů je najít nastavení sítě, s kterým se dosáhne nejmenší chyby predikce, pro základní hledání vzorů s historickým oknem 30 minut a budoucím oknem 10 minut.

Budu sledovat vliv všech kombinací parametrů s těmito hodnotami:

- Délka obchodního dne – celý, ořezaný
- Návrat do lepšího stavu – s návratem, bez návratu
- Učící faktor – 0,05; 0,1; 0,5; 0,9
- Velikost skryté vrstvy – 10, 25, 50, 150, 500
- Velikost dávky při učení – 1, 5, 20, 50

Cílem první části experimentu bude porovnat úspěšnost predikce celého obchodního dne a ořezaného obchodního dne a zda je výhodnější po každé učící iteraci, která zhoršila chybu sítě, navrátit předchozí stav sítě.

	Délka obchodního dne		**Návrat do lepšího stavu**	
Timeframe	**Cely**	**Ořezaný**	**S návratem**	**Bez návratu**
1	0,09578	0,06533	0,08159	0,07952
2	0,09884	0,06736	0,08584	0,08036
5	0,10223	0,07486	0,09253	0,08456
10	0,11668	0,08722	0,10665	0,09725
Průměr	**0,10338**	**0,07369**	**0,09165**	**0,08542**

Tabulka 8.2: Průměrná chyba predikce na testovacích datech.

Z výsledku v tabulce 8.2, která udává průměrné chybu predikce z experimentů pro všechny kombinace parametrů, jednoznačně vyplývá, že je pro každý timeframe výhodnější použít pouze data z ořezaného obchodního dne a po zhoršení chyby při učení nevracet předchozí stav sítě. Všechny další experimenty už budou probíhat pouze s tímto nastavením.

Další částí experimentu bude hledat nejvhodnější nastavení neuronové sítě, pro které bude mít síť nejmenší průměrnou chybu predikce. Což znamená zjistit hodnoty pro velikost skryté vrstvy, velikost dávky k učení a učící faktor.

V tabulce 8.3 jsou uvedeny průměrné výsledky ze všech timeframů, všechny výsledky rozdělené po timeframech jsou v příloze D. Z těchto výsledků vyplývá, že je jen nepatrný rozdíl mezi jednotlivými nastavení, ale síť dosahuje nejmenší chyby predikce při použití 25 neuronů ve skryté vrstvě, 5 vektorů v jedné učící dávce a faktoru učení 0,5. Pro všechny další experimenty se použije síť s těmito parametry.

	Průměrná chyba predikce
Velikost dávky	
1	0,07346
5	0,07262
20	0,07401
50	0,07464
Učící faktor	
0,05	0,07427
0,1	0,07389
0,5	0,07337
0,9	0,07846
Velikost skryté vrstvy	
10	0,07320
25	0,07262
50	0,07321
150	0,07312
500	0,07522

Tabulka 8.3: Průměrná chyba predikce pro různá nastavení sítě.

8.2 Experiment 2 - technická analýza

V této sadě experimentů se bude vycházet pouze z technické analýzy, z které se budou hledat nejvhodnější prvky tak, aby se dosáhlo nejmenší chyby predikce nebo největšího zisku z obchodování.

V první části se bude hledat počet minut, z nichž se použijí historické hodnoty, ze kterých se bude vyvářet vzor do vstupního vektoru, neboli velikost historického okna. Bude se hledat také počet minut, ze kterých se bude počítat průměrná hodnota do výstupního vektoru neuronové sítě, neboli velikost budoucího okna.

Experiment jsem provedl se všemi kombinacemi těchto parametrů:

- Délka historického okna – 10, 20, 30, 40
- Délka budoucího okna – 10, 20, 30

Timeframe	**Počet minut budoucího okna**		
	10	**20**	**30**
1	0,06559	0,09510	0,11989
2	0,06707	0,09201	0,11277
5	0,07400	0,09574	0,11342
10	0,08560	0,10515	0,12087
Průměr	0,07306	0,09700	0,11674

Tabulka 8.4: Průměrná chyba predikce v závislosti na velikosti budoucího okna.

Nejprve jsem vyhodnotil vhodnou velikost budoucího okna, jako průměr přes všechny velikosti historického okna. Podle výsledků v tabulce 8.4 je pro všechny timeframy jednoznačně nejvhodnější budoucí okno s velikostí 10 minut. Což také odpovídá skutečnosti, protože je to nejmenší velikost okna, které je tak nejvíce závislé na předchozích hodnotách.

Pro vyhodnocení velikosti historického okna použiji pouze průměrné výsledky, které jsem získal v kombinaci s budoucím oknem o velikosti 10 minut.

Timeframe	**Velikost historického okna [min]**			
	10	**20**	**30**	**40**
1	0,06435	0,06379	0,06550	0,06771
2	0,06674	0,06565	0,06720	0,06728
5	0,07380	0,07396	0,07408	0,07416
10	0,08501	0,08548	0,08580	0,08581
Průměr	0,07248	0,07222	0,07315	0,07374

Tabulka 8.5: Průměrná chyba predikce v závislosti na velikosti historického okna.

V tabulce 8.5 jsou srovnány chyby predikce v závislosti na velikosti historického okna, které vycházejí pro všechny timeframy téměř stejné. Úplně nejlépe pak pro všechny timeframy vychází historické okno o velikosti 20 minut. Jako základ všech dalších experimentů se bude brát historické okno veliké 20 minut a budoucí okno o velikosti 10 minut.

V další části tohoto experimentu se budou hledat indikátory technické analýzy, které vylepší predikci založenou pouze na historickém okně. Budu testovat indikátory MA, RSI a CCI s periodou 10, 20 30, 40 a 50 minut.

	Průměrná chyba predikce				**Průměrný výsledek obchodování**			
	TF: 1	**TF: 2**	**TF: 5**	**TF: 10**	**TF: 1**	**TF: 2**	**TF: 5**	**TF: 10**
Žádný	0,06479	0,06705	0,07396	0,08578	9,74	3,714	-5,56	-5,80
RSI								
10	0,06497	0,06611	0,07265	0,08321	10,59	-9,43	-3,13	-5,25
20	0,06695	0,06719	0,07268	0,08310	17,12	5,11	-8,41	-4,18
30	0,07188	0,06822	0,07302	0,08340	**21,25**	**12,77**	-9,55	-10,15
40	0,07530	0,06887	0,07334	0,08391	**22,54**	**13,35**	-7,88	-10,63
50	0,07876	0,06968	0,07370	0,08422	**21,62**	**14,45**	-8,23	-12,64

Tabulka 8.6: Průměrná chyba predikce a výsledek obchodování při použití indikátoru RSI.

Protože je přesnost predikce pro všechny indikátory téměř stejná, budu zde hodnotit hlavně výslednou obchodní bilanci. Jako jediný indikátor, po kterém se výdělek systému výrazně zlepšil, je RSI s periodou 30, 40 a 50, což je patrné z tabulky 8.6. Platí to ale jen pro timeframy 1 a 2 minuty. Na ostatních timeframech dochází naopak ke zhoršení predikce. Ostatní indikátory nijak výrazně nezlepšili predikci ani výnos obchodování. Souhrnné výsledky pro všechny indikátory jsou v příloze D.

Pro další experimenty přidám do vstupního vektoru indikátor RSI s periodou 40 minut, pouze na timeframech 1 a 2 minuty.

8.3 Experiment 3 - analýza nálady

V tomto experimentu se vezmeme nejlepší výsledek z předchozích experimentů, tedy délka vstupního a výstupního okna a indikátory technické analýzy, které dopadly nejlépe. K těmto vstupům se přidá vektor nálad nebo jeho část, u které se bude sledovat zda se predikce zlepší nebo zhorší.

Provedou se experimenty pro všechny kombinace těchto parametrů vektoru nálad:

- Sada tweetů – malá, velká
- Období analýzy – hodina, den
- Projekce přes slovníky – bez projekce, s projekcí
- Historické okno – aktuální doba, předchozí doba, předchozí 3 doby
- Část vektoru nálad – celý, počet tweetů, počet slov, slovník LabMT, slovník Inquirer obecné nálady, slovník SentiWordNet

Kompletní výsledky experimentů jsou v tabulkách v příloze D. Největšího zlepšení obchodní bilance se dosáhlo kombinacemi parametrů uvedených v tabulce 8.7.

Sada	**Projekce**	**Období**	**Předch.**	**Vektor nálad**	**Bilance**	**Zlepšení**
Timeframe: 1					22,5	0
velká	bez projekce	hodina	0	celý	28,8	28 %
velká	bez projekce	den	3	celý	28,0	24 %
velká	bez projekce	den	3	počet tweetů	28,5	27 %
velka	s projekcí	hodina	0	LabMT	29,4	31 %
velka	s projekcí	hodina	1	nálady Inquirer	28,3	25 %
velka	s projekcí	hodina	3	nálady Inquirer	28,6	27 %
Timeframe: 2					13,4	0
malá	bez projekce	den	3	SentiWordNet	20,9	56 %
velká	bez projekce	hodina	3	celý	19,8	48 %
velká	s projekcí	hodina	3	celý	20,5	53 %
velká	bez projekce	dny	3	celý	21,2	58 %
Timeframe: 5					-7,9	0
malá	s projekcí	hodina	0	celý	0,2	profit
velká	bez projekce	hodina	3	celý	1,9	profit
velká	s projekcí	hodina	3	celý	2,1	profit
Timeframe: 10					-10.6	0
velká	s projekcí	hodina	1	nálady Inquirer	-1,8	0
velká	s projekcí	hodina	3	nálady Inquirer	-1,8	0

Tabulka 8.7: Nejvýhodnější vektory nálad.

Na timeframu 1 minuta se dosáhlo největších zlepšení pouze s velkou Twitter sadou a období analýzy, projekce ani část vektoru neměly jednoznačný vliv. Maximální zlepšení obchodní bilance se pohybovalo kolem 30 %. Na timeframu 2 minuty byly nejúspěšnějšími parametry velká Twitter sada a použití 3 předchozích celých vektorů nálady. Zlepšení bilance dosahovalo 50 %. Na timeframu 5 minut se našlo několik náladových vektorů, které vytáhly obchodní bilanci z mínusu, ale dosáhlo se jen zanedbatelného profitu. Na timeframu 10 minut žádný náladový vektor nedokázal otočit nepříznivou obchodní bilanci.

8.4 Experiment 4 - analýza přes rekurentní síť

Podobně jako v předchozím experimentu se pracuje s nejlepšími výsledky z technické analýzy, ale k nim se přidá celý vektor, který vznikl analýzou přes rekurentní síť. Bude se experimentovat s většinou kombinací těchto parametrů vektoru charakteristiky:

- Sada tweetů – malá, velká
- Období analýzy – hodina, den
- Projekce přes slovníky – bez projekce, s projekcí
- Historické okno – aktuální doba, předchozí doba, předchozí 3 doby
- Počet opakování tweetů pro RNNLM – 1, 100

Jedinou výjimkou je charakteristika po hodinách, která byla provedena jen na datech, které vznikly projekcí přes slovníky. Výsledky všech experimentů jsou uvedeny v příloze D. Největšího zlepšení výnosu dosáhly kombinace parametrů uvedené v tabulce 8.8

Sada	**Projekce**	**Období**	**Předch.**	**Opakování**	**Bilance**	**Zlepšení**
Timeframe: 1					22,5	0
velká	bez projekce	den	3	1	28,0	24 %
velká	bez projekce	den	0	100	27,8	24 %
velká	bez projekce	den	3	100	27,5	22 %
malá	bez projekce	den	0	100	27,6	23 %
malá	bez projekce	den	0	1	27,5	22 %
Timeframe: 2					13,4	0
velká	bez projekce	den	3	1	21,2	58 %
malá	bez projekce	den	1	100	18,9	41 %
malá	bez projekce	den	3	100	18,0	34 %
Timeframe: 5					-10.6	0
malá	s projekcí	hodina	1	1	1,5	profit
malá	s projekcí	hodina	1	100	2,8	profit
Timeframe: 10					-7,9	0
velká	s projekcí	hodina	3	1	-1,8	0

Tabulka 8.8: Nejvýhodnější vektory charakteristiky.

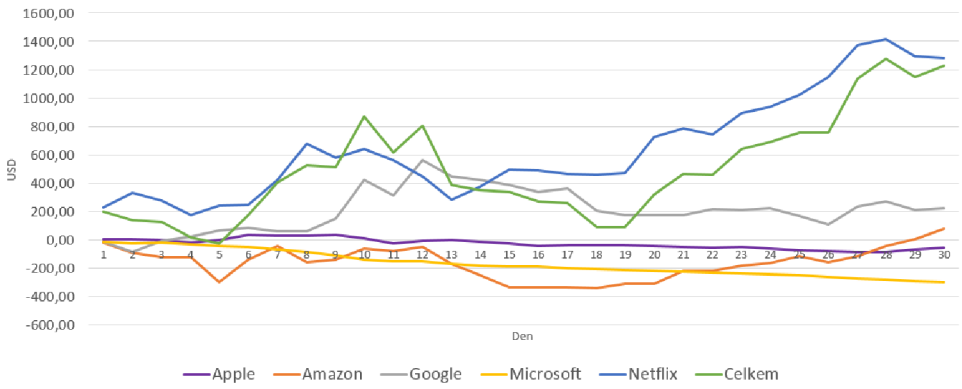
Na timeframech 1 a 2 minuty se největšího zlepšení bilance dosáhlo pouze použitím Twitter sad bez projekce, které byly analyzovány na období jednoho dne. Na timeframu 1 minuta se dosáhlo zlepšení o 24 % a na timeframu 2 minuty až 58 %. Stejně jako v předchozím experimentu se na vyšších timeframech nedosáhlo žádného výrazného zlepšení, které by bylo výnosné.

8.5 Experiment 5 - vyhodnocení strategie

Cílem poslední skupiny experimentů je vyhodnocení nejúspěšnějších strategií ze všech analýz pro jednotlivé akcie a timeframy.

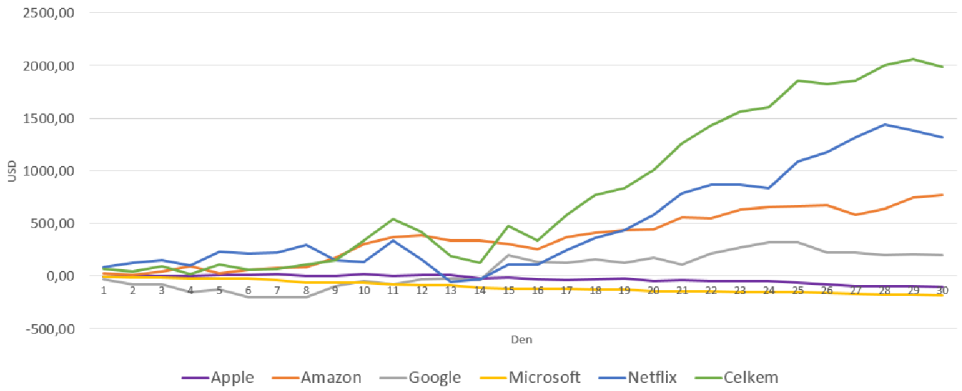
Aby byly další výsledky co nejblíže skutečnému obchodování, bude se uvažovat poplatek 5 USD za obchod a v rámci jednoho obchodu se bude počítat s objemem 50 akcií dané společnosti.

Pro timeframe 1 minuta jsem vybral strategii, která je založena na celém náladovém hodinovém vektoru, který byl bez projekce vytvořen z velké Twitter sady. Na grafu na obrázku 8.1, kde je znázorněn stav účtu pro jednotlivé akcie, je patrné, že hlavní zisk tvoří obchodování akcií Netflixu. Obchodování akcií Microsoftu a Applu je naopak ztrátové. Celkový zisk této strategie je 1 224 USD během 30 testovacích dní.



Obrázek 8.1: Průměrná nejlepší obchodní strategie na timeframu 1 minuta.

Pro timeframe 2 minuty jsem vybral strategii, která využívá také celý náladový vektor z velké sady tweetů, ale je to denní vektor s předchozími třemi hodnotami. Z grafu na obrázku lze vyčíst, že se nejvíce na zisku podílelo obchodování akcií Netflixu a Amazonu. Obchodování akcií Applu a Microsoftu bylo opět ztrátové. Souhrnný zisk této strategie je 1 984 USD.



Obrázek 8.2: Průměrná nejlepší obchodní strategie na timeframu 2 minuta.

Zajímavé je, že strategie na minutovém timeframu má vyšší průměrný výdělek, ale nižší konečný zisk. Je to dáno tím, že se častěji obchoduje a je tak placena velká částka na poplatcích.

Kapitola 9

Závěr

Cílem této práce bylo vytvořit systém, který analyzuje historická burzovní data a tweety, na základě čehož predikuje budoucí vývoj trhu a vytvoří obchodní strategii, kterou otestuje.

Nejprve jsem nastudoval funkci burzovních trhů a principy úspěšného obchodování. Také jsem nastudoval základy analýzy burzy, jak technické, založené na indikátorech, tak fundamentální, která se snaží najít vnitřní hodnotu aktiva. Dále jsem se blíže seznámil s mikroblogovací službou Twitter a způsoby její analýzy. Poté jsem prostudoval neuronové sítě a jejich využití při obchodování na burze.

Dále jsem si obstaral datové sady, na kterých se bude obchodní systém učit a testovat. V jedné datové sadě jsou cenové hodnoty akcií pěti velkých společností za dobu osmi měsíců. Další dvě datově sady obsahují různé zprávy z Twitteru ze stejného časového období. Celkem měly obě sady více než miliardu tweetů, které jsem nejdříve předzpracoval a poté analyzoval dvěma odlišnými metodami. V rámci předzpracování jsem odstranil neznámé znaky, jednotlivá slova separoval a z nich vybral jen slova, která by mohla mít vypovídací hodnotu. První metoda analýzy spočívala v použití několika náladových slovníků, podle kterých jsem hodnotil jednotlivá slova v tweetech. Druhá metoda vytvořila charakteristiku textu jeho průchodem skrze rekurentní neuronovou síť.

Pro systém jsem vybral vrstvenou neuronovou síť, která se učí metodou zpětného šíření chyby, jenž je schopna na trhu detekovat opakující se vzory. Tato síť na základě výstupů technické a fundamentální analýzy predikuje budoucí pohyb trhu. Hlavním vstupem sítě je historické okno několika předchozích hodnot ceny a výstupem průměr několika budoucích hodnot ceny. Ke vstupům sítě jsou přidány některé technické indikátory. Tato síť je rozšířena dvěma různými způsoby. U prvního je ke vstupům sítě přidán vektor (nebo jeho část) nálad, který byl vytvořen analýzou nálady. Druhým rozšířením je přidání celého vektoru charakteristiky, který vznikl analýzou přes rekurentní síť. Pro obě rozšíření jsem hledal nejvhodnější úpravu datové sady a způsob analýzy, ze které bude patrná závislost Twitteru a burzy, a pomůže tak zlepšit obchodní bilanci. Systém také obsahuje obchodní strategii, která je založená na výsledcích predikce.

Výsledný systém jsem v rámci experimentů důkladně otestoval a našel nastavení neuronové sítě, které je nejvhodnější pro intradenní obchodování na timeframech 1, 2, 5 a 10 minut. Největších výnosů dosahovalo obchodování na timeframu 1 minuta. Ziskové bylo i obchodování na timeframu 2 minuty, ale na vyšších timeframech bylo většinou obchodování proděl̆ečné. Z výstupů technické analýzy bylo jako nejlepší vstup pro neuronovou síť vybráno historické okno o velikosti 20 minut a budoucí okno o velikosti 10 minut. A pro timeframy 1 a 2 minuty se ukázalo výhodné použít také indikátor RSI s s periodou mezi 30 a 50 minutami. V obou případech rozšíření sítě o výsledky analýzy tweetů se dosáhlo

maximálního zlepšení výnosu některých strategií o 30 - 50 %. Je to ovšem zlepšení jen na konkrétních datech, které pravděpodobně nebude fungovat na jiných akcích nebo v jiném časovém období.

Další vývoj tohoto systému by se mohl uskutečnit v několika směrech. V první řadě by se měla vylepšit jednoduchá obchodní strategie pro vstup a výstup z pozice, která je největší slabinou systému. Měla by vycházet z hodnot technických indikátorů a aktuálního stavu trhu a ne jen z predikce.

Systém by se mohl stát samostatným a fungovat v online režimu jako plnohodnotný automatický obchodní systém. Z internetu by si sám stahoval a ukládal historická burzovní data a vybrané zprávy z Twitteru. Pomocí API, které někteří brokeři poskytují, by pak sám dával skutečné obchodní signály.

Jiným směrem vývoje by mohlo být použití genetických algoritmů, pomocí kterých by se pro daný trh hledalo optimální nastavení neuronové sítě a vhodné prvky vektorů, které by se použily k predikci a obchodování.

Systém je jen tak dobrý, jak dobrá jsou data, se kterými pracuje. Proto by mohlo zlepšit výnos systému použití jiných slovníků nálad nebo vstupních datových sad. Například tweety by se mohli filtrovat dle specifické oblasti, tak aby se zacílilo na konkrétní skupinu uživatelů, kteří souvisí z vybranou burzou.

Literatura

[1] Bollen, J.; Mao, H.; Zeng, X.: Twitter mood predicts the stock market. *Journal of Computational Science*, ročník 2, č. 1, 2011: s. 1–8.

[2] Chen, R.; Lazer, M.: Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement. 2013.

[3] Fuehres, H.; Zhang, X.; Gloor, P. A.: Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, ročník 26, 2011: s. 55–62.

[4] Lawrence, R.: Using neural networks to forecast stock market prices. *University of Manitoba*, 1997.

[5] Mittal, A.; Goel, A.: Stock prediction using twitter sentiment analysis. *Standford University, CS229*, 2012.
URL http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[6] Nesnídal, T.; Podhajský, P.: *Obchodování na komoditních trzích*. Praha: Grada, druhé vydání, 2006, ISBN 8024718510.

[7] Nesnídal, T.; Podhajský, P.: *Jak se stát intradenním finančníkem*. Praha: Centrum finančního vzdělávání, 2008, ISBN 9788090387447.

[8] Putna, L.: *Predikce vývoje kurzu pomocí umělých neuronových sítí*. Diplomová práce, Brno, FIT VUT v Brně, 2011.

[9] Rejnuš, O.: *Finanční trhy*. Ostrava: Key Publishing, druhé vydání, 2010.

[10] Russell, S. J.; Norvig, P.: *Artificial Intelligence: A Modern Approach*. Pearson Education, druhé vydání, 2003, ISBN 0137903952.

[11] Turek, L.: *První kroky na burze*. Brno: Computer Press, vyd. 1. vydání, 2008, ISBN 9788025119150.

[12] Veselá, J.: Analýzy trhu cenných papírů. 2. Díl: Fundamentální analýza. Technická zpráva, ISBN 80-245-0506-1.

[13] Veselá, J.: Burzy a burzovní obchody. Technická zpráva, ISBN 80-245-0939-3.

[14] Zbořil, F. V.: Soft Computing, 2014, prezentace k předmětu SFC dostupná studentům FIT VUT v Brně. [online, 16.1.2015].
URL https://www.fit.vutbr.cz/study/courses/SFC/

[15] Zhang, L.: *Sentiment analysis on Twitter with stock price and significant keyword correlation.* Dizertační práce, 2013.

[16] Financnik.cz - komodity, akcie, burza, forex. [online, 16.1.2015].
URL http://financnik.cz/

[17] About Twitter, Inc. — About. [online, 16.1.2015].
URL https://about.twitter.com/company

[18] Twitter Developers - The Twitter platform connects your website or application with the worldwide conversation happening on Twitter. [online, 16.1.2015].
URL https://dev.twitter.com/

Příloha A

Tweet

Příklad struktury tweetu získaného skrze Twitter API ve formátu JSON (včetně všech metadat):


```
{
    "id" => 12296272736,
    "text": "test $TWTR @twitterapi
             text #hashtag http:\/\/t.co\/p5dOtmnZyu",
    "created_at": "Thu Dec 12 07:15:21 +0000 2013",
    "entities": {
        "hashtags": [{
            "text": "hashtag",
            "indices": [23, 31]
        }],
        "symbols": [{
            "text": "TWTR",
            "indices": [5, 10]
        }],
        "urls": [{
            "url": "http:\/\/t.co\/p5dOtmnZyu",
            "expanded_url": "http:\/\/dev.twitter.com",
            "display_url": "dev.twitter.com",
            "indices": [32, 54]
        }, {
        "url": "https:\/\/t.co\/ZSvIEMOPb8",
        "expanded_url": "https:\/\/ton.twitter.com\/...cOkcq9FS.jpg",
        "display_url": "pic.twitter.com\/ZSvIEMOPb8",
        "indices": [55, 78]
    }],
    "user_mentions": [{
        "screen_name": "twitterapi",
        "name": "Twitter API",
        "id": 6253282,
        "id_str": "6253282",
        "indices": [11, 22]
    }],
    "media": [{
        "id": 411031503833792512,
        "id_str": "411031503833792512",
        "indices": [55, 78],
        "media_url": "https:\/\/ton.twitter.com\/...cOkcq9FS.jpg",
        "media_url_https": "https:\/\/ton.twitter.com\/...cOkcq9FS.jpg",
        "url": "https:\/\/t.co\/ZSvIEMOPb8",
        "display_url": "pic.twitter.com\/ZSvIEMOPb8",
        "expanded_url": "https:\/\/ton.twitter.com\/...cOkcq9FS.jpg",
        "type": "photo",
        "sizes": {
            "medium": {
                "w": 600,
                "h": 450,
                "resize": "fit"
            },
            "large": {
                "w": 1024,
                "h": 768,
                "resize": "fit"
            },
            "thumb": {
                "w": 150,
                "h": 150,
                "resize": "crop"
            },
            "small": {
                "w": 340,
                "h": 255,
                "resize": "fit"
            }
        }
    }],
    "contributors": null,
    "retweet_count": null,
    "in_reply_to_status_id_str": null,
    "geo": null,
    "retweeted": false,
    "in_reply_to_user_id": null,
    "user": {
        "profile_sidebar_border_color": "C0DEED",
        "name": "Gnip, Inc.",
        "profile_sidebar_fill_color": "DDEEF6",
        "profile_background_tile": false,
        "profile_image_url": "http://a3.twimg.com/.../icon_normal.png",
        "location":  "Boulder, CO",
        "created_at": "Fri Oct 24 23:22:09 +0000 2008",
        "id_str": "16958875",
        "follow_request_sent": false,
        "profile_link_color": "0084B4",
        "favourites_count": 1,
        "url": "http://blog.gnip.com",
        "contributors_enabled": false,
        "utc_offset": -25200,
        "id": 16958875,
        "profile_use_background_image": true,
        "listed_count": 23,
        "protected": false,
        "lang": "en",
        "profile_text_color": "333333",
        "followers_count": 260,
        "time_zone": "Mountain Time (US & Canada)",
        "verified": false,
        "geo_enabled": true,
        "profile_background_color": "C0DEED",
        "notifications": false,
        "description": "Gnip makes it really easy for
                        you to collect social data for your business.",
        "friends_count": 71,
        "profile_background_image_url": "http://s.twimg.com/.../bg.png",
        "statuses_count": 302,
        "screen_name":  "gnip",
        "following": false,
        "show_all_inline_media": false
    },
    "in_reply_to_screen_name": null,
    "source": "web",
    "place": null,
    "in_reply_to_status_id": null
}
```

Příloha B

Obsah CD

Přiložené CD obsahuje tyto adresáře:

- **tex** – textová část práce ve formátu `pdf`, včetně zdrojových kódů v LaTeXu
- **src** – zdrojové kódy systému
- **burza** – historická burzovní data
- **sentimentDicts** – náladové slovníky
- **stopWords** – slovníky stopslov
- **vectors** – předzpracované vektory charakteristiky textu
- **experiment** – kompletní výsledky všech provedených experimentů
- **rnn** – soubory nástroje RNNLM
- **screen** – složka pro ukládání obrázků systému
- **cahe** – složka pro ukládání zpracovaných dat

Příloha C

Použití sytému

Ke spuštění systému je třeba mít nainstalován Python a potřebné knihovny, což jsou především NumPy, Theano, NTLK, Matplotlib, TA-Lib. Výsledný systém se spouští ve složce /src/Main příkazem `python Albert.py configFile`. Kde `configFile` je cesta k souboru s konfigurací, jehož formát a význam jednotlivých položek je popsán níže. Ve stejné složce je připraveno několik konfiguračních souborů.

Konfigurační soubor

Konfigurační soubor musí mít složen pouze z těchto položek:

- `stock NAME` – název akcie - aapl, amzn, goog, msft, nflx
- `timeframe N` – timeframe
- `testDays N` – početní posledních dní, které se budou testovat
- `lr F` – učící faktor 0 - 1
- `batchSize N` – velikost jedné učící dávky
- `betterPrev B` – návrat do lepšího stavu 0; 1
- `hidden N` – počet neuronů ve skryté vrstvě
- `normal B` – automatická normalizace 0; 1
- `in N` – délka historického okna
- `out N` – délka budoucího okna
- `rsi N` – perioda indikátoru RSI
- `cci N` – perioda indikátoru CCI
- `ma N` – perioda indikátoru MA
- `imgName NAME` – jméno souboru pro uložení obrázku systému do složky screen
- `saveLast N` – počet dní, pro které se uloží obrázky
- `show B` – zobrazit výsledek graficky 0; 1
- `logFile FILE` – jméno souboru pro uložení logů

Příloha D

Výsledky experimentů

Zde jsou zobrazeny detailnější výsledky z některých provedených experimentů.

TF	**Velikost dávky**				**Učící faktor**				**Velikost skryté vrstvy**			
	1	5	20	50	0,05	0,1	0,5	0,9	10	25	50	150
	Průměrná chyba predikce											
1	0,0672	0,0650	0,0643	0,0646	0,0680	0,0626	0,0659	0,0707	0,0652	0,0643	0,0653	0,0651
2	0,0671	0,0665	0,0670	0,0688	0,0689	0,0652	0,0678	0,0716	0,0670	0,0667	0,0671	0,0672
5	0,0741	0,0738	0,0761	0,0755	0,0750	0,0782	0,0745	0,0787	0,0744	0,0744	0,0744	0,0744
10	0,0854	0,0851	0,0886	0,0897	0,0852	0,0896	0,0852	0,0929	0,0861	0,0851	0,0859	0,0858
Prům.	0,0735	0,0726	0,0740	0,0746	0,0743	0,0739	0,0734	0,0785	0,0732	0,0726	0,0732	0,0731
	Obchodní výsledek											
1	4,3477	4,9919	0,9482	0,8867	-0,7340	4,3086	4,3117	3,2882	1,8402	2,1236	4,3778	2,5579
2	1,8661	0,5900	-0,1538	0,2635	-0,4297	0,0029	1,1780	1,8147	2,8468	1,8458	1,3127	-0,9568
5	0,2696	-1,1486	-2,1427	-2,8614	2,1445	-1,7529	-2,7432	-3,5316	0,0997	0,0952	-2,3028	-1,6494
10	-3,6896	-4,5790	-5,1622	-5,7232	-3,2015	-6,4366	-5,1174	-4,3984	-3,4602	-2,4586	-4,9674	-5,8393
Prům.	0,6984	-0,0364	-1,6276	-1,8586	-0,5552	-0,9695	-0,5927	-0,7068	0,3316	0,4015	-0,3949	-1,4719

Tabulka D.1: Detailní výsledky experimentu 1.

	Průměrná chyba predikce				Průměrný výsledek obchodování			
	TF: 1	**TF: 2**	**TF: 5**	**TF: 10**	**TF: 1**	**TF: 2**	**TF: 5**	**TF: 10**
Žádný	0,06479	0,06705	0,07396	0,08578	9,74362	3,71454	-5,56174	-5,79958
MA								
10	0,06469	0,06687	0,07357	0,08684	6,76124	-5,00268	-4,72404	-3,59872
20	0,06484	0,06688	0,07393	0,08559	4,51672	-2,44532	-3,27198	-4,18304
30	0,06480	0,06699	0,07390	0,08475	10,25462	2,71344	-6,29800	-3,29338
40	0,06590	0,06719	0,07393	0,08512	7,82888	5,82938	-5,01642	-5,09798
50	0,06626	0,06735	0,07400	0,08534	4,95084	0,34538	-3,28178	-3,97822
RSI								
10	0,06497	0,06611	0,07265	0,08321	10,59254	-9,49236	-3,13926	-5,25893
20	0,06695	0,06719	0,07268	0,08310	17,12568	5,11890	-8,41628	-4,18276
30	0,07188	0,06822	0,07302	0,08340	21,25658	12,77246	-9,55310	-10,15614
40	0,07530	0,06887	0,07334	0,08391	22,54906	13,35422	-7,88462	-10,63812
50	0,07876	0,06968	0,07370	0,08422	21,62644	14,45738	-8,23782	-12,64960
CCI								
10	0,06460	0,06654	0,07337	0,08413	4,22922	-7,06890	0,39584	-3,25480
20	0,06483	0,06660	0,07355	0,08415	9,95096	-2,77018	-3,64942	-2,82820
30	0,06496	0,06678	0,07341	0,08411	10,44022	-0,31788	-4,02404	-3,45002
40	0,06536	0,06707	0,07355	0,08399	11,88622	6,61818	-1,90152	-4,61388
50	0,06577	0,06733	0,07366	0,08408	12,56422	4,91210	-2,84560	-4,70080

Tabulka D.2: Všechny výsledky experimentu 2.

TF: 1	**Nálada po hodinách**						**Nálada po dnech**					
Typ datové sady	bez projekce			s projekcí			bez projekce			s projekcí		
Časový interval	0	1	3	0	1	3	0	1	3	0	1	3
Část vektoru nálad	Malá Twitter sada											
Celý	21,2	19,8	18,8	27,0	22,1	17,6	27,1	24,5	25,2	27,6	22,6	25,7
Počet slov	25,4	20,6	24,7	25,4	21,1	22,9	19,4	19,1	23,5	25,4	23,1	22,9
Počet tweetů	23,2	21,6	22,9	25,5	22,7	19,6	21,4	23,3	23,6	21,6	21,6	19,2
LabMT	23,0	20,8	19,7	23,2	22,0	21,9	26,3	23,5	24,6	25,6	19,5	21,9
Nálady Inquirer	24,1	22,3	16,6	25,4	24,2	22,7	25,1	24,2	25,2	24,9	22,8	24,4
SentiWordNet	25,5	23,1	19,9	22,6	23,9	24,6	25,6	23,5	24,8	23,1	22,6	23,6
	Velká Twitter sada											
Celý	28,8	23,4	23,5	18,9	23,9	27,0	25,7	20,9	28,0	27,8	24,3	27,5
Počet slov	21,1	20,8	24,9	20,4	23,0	25,6	22,7	20,9	24,7	22,8	25,2	23,0
Počet tweetů	25,1	22,2	24,5	24,3	24,9	22,0	24,4	23,4	28,5	22,6	19,3	25,4
LabMT	27,4	22,6	22,3	29,4	24,0	23,0	24,0	20,5	25,8	24,1	22,0	25,9
Nálady Inquirer	24,3	28,1	27,8	24,0	28,3	28,6	24,5	19,7	22,3	25,8	20,8	20,9
SentiWordNet	23,5	22,6	24,8	24,7	25,6	24,4	27,2	24,0	25,1	25,1	18,9	21,8

Tabulka D.3: Výsledky experimentu 3 na timeframu 1 minuta - obchodní bilance.

TF: 2	**Nálada po hodinách**						**Nálada po dnech**					
Typ datové sady	bez projekce			s projekcí			bez projekce			s projekcí		
Časový interval	0	1	3	0	1	3	0	1	3	0	1	3
Část vektoru nálad	Malá Twitter sada											
Celý	14,2	10,6	16,3	10,0	14,4	16,4	14,8	16,0	16,9	16,3	18,9	18,0
Počet slov	12,9	15,2	12,3	13,0	11,5	13,0	12,8	14,0	13,9	12,3	15,8	16,4
Počet tweetů	14,7	13,8	13,5	13,4	14,8	14,2	11,5	13,2	13,8	12,0	12,7	15,5
LabMT	15,3	15,1	14,7	16,1	14,5	14,2	15,8	15,6	13,1	15,8	15,4	12,9
Nálady Inquirer	13,8	13,8	15,7	11,4	16,4	15,1	12,1	14,8	17,1	14,4	16,1	15,9
SentiWordNet	13,2	12,6	12,4	13,5	12,4	12,4	10,6	17,2	20,9	12,5	12,6	16,0
	Velká Twitter sada											
Celý	14,4	18,4	19,8	15,6	18,0	20,5	11,0	13,5	21,2	13,6	12,5	16,7
Počet slov	12,6	13,0	12,6	13,5	14,2	12,0	14,1	17,2	15,0	14,4	18,2	15,6
Počet tweetů	13,7	12,4	12,8	12,0	12,6	13,6	12,9	15,1	15,6	14,3	14,0	13,9
LabMT	14,3	12,4	14,7	15,5	14,2	15,7	15,0	14,9	14,4	14,2	16,3	15,7
Nálady Inquirer	11,6	15,3	17,6	14,3	12,6	15,1	13,6	12,7	12,3	13,3	13,8	14,1
SentiWordNet	14,6	10,2	12,4	13,7	12,3	12,7	12,8	11,8	13,5	11,7	14,3	14,0

Tabulka D.4: Výsledky experimentu 3 na timeframu 2 minuty - obchodní bilance.

TF: 5	**Nálada po hodinách**						**Nálada po dnech**					
Typ datové sady	bez projekce			s projekcí			bez projekce			s projekcí		
Časový interval	0	1	3	0	1	3	0	1	3	0	1	3
Část vektoru nálad	Malá Twitter sada											
Celý	-0,3	-2,4	-1,8	0,2	-2,3	-1,8	-2,7	-2,9	-3,3	-2,4	-3,7	-3,2
Počet slov	-2,9	-3,5	-2,3	-2,7	-3,3	-2,6	-4,1	-3,6	-3,5	-4,0	-3,9	-4,0
Počet tweetů	-3,3	-3,4	-1,8	-3,3	-3,4	-1,8	-4,6	-4,6	-3,6	-4,6	-4,6	-3,7
LabMT	-3,2	-5,1	-1,0	-3,1	-5,4	-0,9	-5,1	-4,1	-2,5	-5,3	-4,1	-2,7
Nálady Inquirer	-2,9	-3,2	-0,3	-3,0	-3,0	0,0	-5,0	-4,2	-2,9	-5,0	-4,1	-3,2
SentiWordNet	-3,5	-5,7	-2,8	-3,3	-5,8	-2,6	-2,7	-6,0	-6,4	-2,9	-5,8	-6,3
	Velká Twitter sada											
Celý	-3,2	-2,0	1,9	-3,5	-3,8	2,1	-4,6	-4,3	-3,6	-3,8	-4,6	-3,9
Počet slov	-3,7	-3,8	-3,2	-3,9	-3,9	-4,3	-3,7	-4,1	-3,2	-3,7	-4,1	-3,3
Počet tweetů	-3,5	-3,8	-3,4	-3,5	-3,7	-3,6	-3,7	-3,9	-3,2	-3,7	-3,9	-3,3
LabMT	-4,5	-4,9	-3,6	-4,5	-4,7	-3,1	-4,0	-4,0	-4,0	-4,2	-4,3	-3,7
Nálady Inquirer	-0,4	-1,1	-0,2	-0,3	-2,5	0,9	-3,0	-4,0	-3,0	-2,1	-3,4	-2,9
SentiWordNet	-3,8	-4,1	-3,7	-3,3	-4,2	-3,2	-4,0	-4,3	-5,2	-4,1	-4,0	-5,0

Tabulka D.5: Výsledky experimentu 3 na timeframu 5 minut - obchodní bilance.

TF: 10	**Nálada po hodinách**						**Nálada po dnech**					
Typ datové sady	bez projekce			s projekcí			bez projekce			s projekcí		
Časový interval	0	1	3	0	1	3	0	1	3	0	1	3
Část vektoru nálad	Malá Twitter sada											
Celý	-5,2	-4,8	-4,9	-4,5	-5,1	-5,8	-4,7	-2,9	-6,5	-4,4	-3,6	-6,8
Počet slov	-5,5	-5,7	-2,2	-5,7	-6,1	-2,6	-6,6	-5,7	-2,6	-6,6	-5,7	-2,6
Počet tweetů	-4,9	-5,5	-2,4	-4,9	-5,5	-2,4	-5,7	-5,5	-2,6	-5,7	-5,5	-2,6
LabMT	-7,2	-6,1	-4,5	-7,2	-6,1	-4,7	-6,2	-5,2	-5,3	-6,2	-5,5	-5,4
Nálady Inquirer	-5,0	-3,1	-6,8	-5,1	-3,4	-5,4	-4,7	-5,8	-5,0	-4,7	-5,7	-5,1
SentiWordNet	-7,4	-4,6	-3,2	-6,8	-4,7	-3,3	-3,2	-3,9	-5,6	-3,6	-3,8	-5,0
	Velká Twitter sada											
Celý	-4,5	-4,3	-4,9	-4,0	-4,0	-5,1	-5,0	-4,0	-3,6	-4,6	-3,2	-3,2
Počet slov	-5,7	-6,2	-4,0	-5,4	-5,5	-3,8	-6,0	-6,5	-2,9	-6,0	-6,6	-3,1
Počet tweetů	-5,3	-6,0	-3,9	-5,4	-6,0	-3,8	-6,0	-6,6	-3,3	-6,0	-6,6	-3,3
LabMT	-6,2	-6,2	-3,5	-6,3	-6,2	-3,2	-6,7	-6,0	-3,1	-6,7	-6,5	-4,8
Nálady Inquirer	-3,5	-2,0	-3,0	-3,7	-1,8	-1,8	-6,1	-2,9	-4,9	-5,8	-3,1	-5,2
SentiWordNet	-5,2	-5,1	-5,3	-5,1	-4,9	-5,7	-6,2	-6,9	-5,2	-6,2	-5,8	-4,0

Tabulka D.6: Výsledky experimentu 3 na timeframu 10 minut - obchodní bilance.

Charakteristika	**po hodinách**			**po dnech**					
Typ datové sady	s projekcí			bez projekce			s projekcí		
Časový interval	0	1	3	0	1	3	0	1	3
Timeframe: 1									
Opakování tweetů	Malá Twitter sada								
1	17,6	13,8	19,8	27,1	24,5	25,2	26,3	23,5	24,6
100	19,8	18,2	14,6	27,6	22,6	25,7	25,6	19,5	21,9
	Velká Twitter sada								
1	23,3	25,6	17,3	25,7	20,9	28,0	24,0	20,5	25,8
100	20,1	21,4	14,7	27,8	24,3	27,5	24,1	22,0	25,9
Timeframe: 2									
Opakování tweetů	Malá Twitter sada								
1	6,2	10,2	11,8	14,8	16,0	16,9	15,8	15,6	13,1
100	9,1	11,4	16,8	16,3	18,9	18,0	15,8	15,4	12,9
	Velká Twitter sada								
1	16,7	13,3	8,2	11,0	13,5	21,2	15,0	14,9	14,4
100	14,7	10,6	7,9	13,6	12,5	16,7	14,2	16,3	15,7

Tabulka D.7: Výsledky experimentu 4 na timeframu 1 a 2 minut - obchodní bilance.

Charakteristika	**po hodinách**			**po dnech**					
Typ datové sady	s projekcí			bez projekce			s projekcí		
Časový interval	0	1	3	0	1	3	0	1	3
Timeframe: 5									
Opakování tweetů	Malá Twitter sada								
1	-2,2	1,5	-2,2	-2,7	-2,9	-3,3	-5,1	-4,1	-2,5
100	-3,4	2,7	0,1	-2,4	-3,7	-3,2	-5,3	-4,1	-2,7
	Velká Twitter sada								
1	-1,8	-2,3	-1,7	-4,6	-4,3	-3,6	-4,0	-4,0	-4,0
100	-0,6	-1,4	-1,9	-3,8	-4,6	-3,9	-4,2	-4,3	-3,7
Timeframe: 10									
Opakování tweetů	Malá Twitter sada								
1	-3,5	-4,1	-2,3	-4,7	-2,9	-6,5	-6,2	-5,2	-5,3
100	-4,0	-5,1	-2,5	-4,4	-3,6	-6,8	-6,2	-5,5	-5,4
	Velká Twitter sada								
1	-3,6	-3,7	-1,8	-5,0	-4,0	-3,6	-6,7	-6,0	-3,1
100	-2,1	-3,0	-4,5	-4,6	-3,2	-3,2	-6,7	-6,5	-4,8

Tabulka D.8: Výsledky experimentu 4 na timeframu 5 a 10 minut - obchodní bilance.