VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

DETEKCIA FALOŠNÝCH SPRÁV S VYUŽITÍM STROJOVÉHO UČENIA

DETECTION OF FAKE NEWS USING MACHINE LEARNING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **MATEJ KOREŇ**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. DAVID HŘÍBEK**
SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

148472

Ústav:	Ústav počítačové grafiky a multimédií (UPGM)
Student:	**Koreň Matej**
Program:	Informační technologie
Specializace:	Informační technologie
Název:	**Detekce falešných zpráv s využitím strojového učení**
Kategorie:	Umělá inteligence
Akademický rok:	2022/23

Zadání:

1. Nastudujte problematiku falešných zpráv. Nalezněte dostupné datasety pro detekci falešných zpráv, popř. vytvořte dataset vlastní.
2. Navrhněte řešení pro detekci falešných zpráv, založené na strojovém učení.
3. Navržené řešení implementujte.
4. Implementované řešení otestujte a vyhodnoťte úspěšnost na dostupných datasetech. Výsledky provnejte s aktuálními state-of-the-art přístupy.

Literatura:

- Christopher M. Bishop, ISBN: 1493938436, Pattern Recognition and Machine Learning, 2016.

Při obhajobě semestrální části projektu je požadováno:
První dva body zadání.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce:	**Hříbek David, Ing.**
Vedoucí ústavu:	Černocký Jan, prof. Dr. Ing.
Datum zadání:	1.11.2022
Termín pro odevzdání:	10.5.2023
Datum schválení:	31.10.2022

Abstrakt

Cieľom tejto práce je využitie strojového učenia na detekciu falošných správ. Na túto prácu boli vybrané štyri modely – Bayesovský, rozhodovací strom, model podporných vektorov a neurónová sieť. V rámci piatich experimentov na rôznych sadách dát boli tieto modely natrénované, otestované, vyhodnotené a porovnané s modernými prístupmi riešenia. Implementácia riešenia má formu konzolovej aplikácie, ktorá umožňuje používateľom replikovať tento postup na vlastných dátach. Nad rámec zadania bol vytvorený (a otestovaný) aj vlastný slovenský dataset **Dezinfo SK** (viď. Príloha).

Abstract

This thesis focuses on the use of machine learning in fake news detection. For this purpose, four models have been selected – Bayesian, Decision Tree, Support Vector Machine and a Neural Network. In five experiments on various datasets, these models were trained, tested, evaluated and compared with state-of-the-art methods. Final implementation is in the form of a python package, which allows it's users to replicate this procedure with their own data. Beyond the assignment, Slovak dataset **Dezinfo SK** was created (see Apendix).

Kľúčové slová

falošné správy, strojové učenie, klasifikácia textu, spracovanie prirodzeného jazyku, Python, datasety, konitngenčná matica, Naive Bayes, Podporné vektory, Rozhodovací strom, neurónové siete

Keywords

Fake news, machine learing, text classificaion, natural language processing, Python, datasets, confusion matrix, Naive Bayes, Support vectors, Decision tree, neural networks

Citácia

KOREŇ, Matej. *Detekcia falošných správ s využitím strojového učenia.* Brno, 2023. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. David Hříbek

Detekcia falošných správ s využitím strojového učenia

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Davida Hříbka. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....
Matej Koreň
7. mája 2023

Poďakovanie

Rád by som poďakoval svojmu vedúcemu práce Ing. Davidovi Hříbkovi za pomoc a ochotu pri vedení mojej bakalárskej práce. Tiež by som rád poďakoval mojej rodine a blízkym za ich podporu.

Obsah

Kapitola 1

Úvod

Internet je celosvetová sieť počítačov, ktoré sú navzájom prepojené a komunikujú spolu. Jeho história siaha do roku 1969, kedy bol vytvorený spoločnosťou ARPA za zámerom prepojiť hlavné univerzity v USA. Odvtedy sa táto sieť rozvíjala až do bodu, kde je k internetu pripojených viac ako 4 miliardy používateľov. Jeho funkcie sa postupne vylepšovali a pridávali sa nové technológie. V dnešnej dobre sa využíva na rôzne účely, avšak jeho prvotný zámer stále pretrváva – jednoduché a dostupné získavanie informácii.

Jedným z miest, ktoré ľudia používajúci internet najčastejšie navštevujú, sú sociálne siete. Sú to väčšinou verejne prístupné stránky a aplikácie, kde spolu ľudia môžu komunikovať, zdieľať novinky, zabávať sa ale aj pracovať. Prístupnosť k informáciám však so sebou nesie aj mnoho nástrah. V tradičných médiach, ako je televízia, rádio či tlač, musia prejsť články a reportáže overením poskytovaných informácii a zdrojov. Pri internete však existuje možnosť jednoducho a verejne prezentovať svoj názor a myšlienky bez žiadnej cenzúry. Preto je potrebné častokrát myslieť kriticky a tieto informácie si overiť na vlastnú päsť.

V prípade, že niekto úmyselne zdieľa nepravdivé informácie za cieľom manipulácie obecenstva, jedná sa o šírenie dezinformácii a klamlivých správ. V súčasnosti má tento jav populárny názov Fake News. Tento trend sa začal objavovať hlavne s nástupom služby Facebook, avšak je viditeľný už aj na iných platformách po celom internete. Ich odhaľovanie je častokrát náročné a vyžaduje si hlbšie zmýšľanie.

Práca je rozdelená do štyroch kapitol. V kapitole „Štúdium problematiky“ je popísaná história falošných správ, metódy ich šírenia, prístupy na odhaľovanie takýchto informácii a úvod do strojového učenia a jeho využitie pri klasifikácii. V časti „Návrh riešenia“ Je uvedený presný cieľ práce, návrh práce na projekte a jeho architektúry, popis získavania a práce s dátami a použité prostriedky na trénovanie modelov a ich vyhodnocovanie. Sekcia „Implementácia a testovanie“ pojednáva o konkrétnom rozložení projektu, zvolenom prostredí, využitých nástrojoch a o konkrétnej stavbe jednotlivých modulov. V poslednej kapitole „Testovanie a vyhodnocovanie“ je celý projekt testovaný v piatich experimentoch na rozličných dátach, či už anglických alebo aj na novovytvorenom slovenskom datasete Dezinfo SK. Jednotlivé modely sú s využitím získaných poznatkov ohodnotené a následne odskúšané na online spravodajských článkoch z niekoľkých novinových portálov.

Kapitola 2

Štúdium problematiky

V poslednom období sa téma falošných správ stala stále viac znepokojujúcou pre spoločnosť. V dobe, keď sociálne siete majú veľký vplyv na to, ako sa ľudia rozhodujú (najmä v oblasti politiky), je dôležité starostlivo zvažovať vierohodnosť informácií, ktoré sa na nich šíria. Napríklad v súvislosti s pandémiou COVID-19 sa mnohí ľudia stretli s problémami s pravdivosťou nariadení, výskumov alebo dát, ktoré boli zverejnené na sociálnych sieťach. Úrady v štátnej správe sa tiež museli vyrovnať s pochybnosťami zo strany občanov a niektorí ľudia využili túto situáciu na šírenie správ s neoverenými faktami a subjektívnymi názormi.

Je dôležité, aby sme dokázali rozpoznať pravdivé od nepravdivých informácií, pretože falošné správy môžu mať vážne dôsledky, ako ovplyvnenie verejnej mienky alebo dokonca ovplyvnenie rozhodnutí vlády. Preto je dôležité, aby sme si vytvorili kritické myslenie a boli schopní overiť si informácie, ktoré sa nám dostanú do rúk, predtým, než ich začneme šíriť ďalej.

2.1 Ako spozorovať falošnú správu?

Táto téma určite nie je novinkou, keďže každá správa môže byť spochybniteľná. Už v minulosti sa napríklad falšovali či šírili rôzne fámy v novinách. Doba sa ale zmenila a nástupom internetu sa informácie sprístupnili širšej verejnosti. Ich šírenie sa stalo rýchlejším, jednoduchším a efektívnejším. Podľa výskumu z roku 2018 vedcov Shawna Doriusa a Sorousha Vosoughiho [16], ktorý zozbierali dáta z Twitteru za posledných 12 rokov (2006-2018) sa falošné správy šíria rýchlejšie, než tie pravdivé – a to až o 70 %. Problematikou spozorovania falošných správ sa zaoberalo množstvo výskumov. Je to veľmi populárna téma, čo len prikladá k jej dôležitosti. Vzniklo mnoho stránok, aplikácii a organizácii, ktoré sa snažia odhaliť alebo aj priamo zabrániť ich šíreniu. Hlavné otázky, ktoré by sme si mali klásť, sú nasledujúce:

- Je zdroj týchto informácii dôveryhodný alebo čerpá z iných dôveryhodných zdrojov?
- Snaží sa zdroj príliš vplývať na emócie?
- Aký je zámer zdroja zdieľať túto informáciu?
- Je táto informácia zmysluplná a uveriteľná?

Tieto kategórie rozhodovania sú kľúčové pri našom rozhodovaní. Samozrejme, nie každá informácia je ľahko overiteľná a častokrát sa overiť ani nedá. Stránka Wikipedia poskytuje

zoznam webových stránok propagujúce takéto správy v anglickom jazyku a v češtine [18, 19]. Slovenské webové portály sú takýmto spôsobom zoradené na stránke konspiratori.sk. Medzi niektoré dezinformačné weby v týchto jazykoch patria:

- Zem & Vek – slovenský jazyk,
- badatel.net – slovenský jazyk,
- World Daily News – anglický jazyk,
- Infowars – anglický jazyk,
- AE News– český jazyk,
- Parlamentní listy – český jazyk.

2.2 Moderné prístupy

V reakcií na alarmujúce zvýšenie šírenia nepravdivých a zavádzajúcich správ v období pandémie koronavíru sa Európska Únia rozhodla vytvoriť projekt SOMA (Social Observatory for Disinformation and Social Media Analysis), [1] ktorý s pomocou spoločností TruthNest a Truly Media pravidelne adresujú najkontroverznejšie témy a problémy a uvádzajú ich na pravú mieru. Ich metóda využíva spojenie výkonnej technológie, založenej na strojovom učení a tímu vedcov, žurnalistov a politológov.

Spoločnosť Google taktiež vytvorila vlastnú verziu nástroju na klasifikáciu článkov s názvom Fact-Check Explorer [2]. Tento nástroj ponúka hodnotenia príspevkov založené priamo na ľudských zdrojoch, ako napríklad recenzie zo stránky CheckYourFact alebo francúzskej stránky FactCheck spadajúcej pod AFP (*Agence France-Presse*) [3].

Pre priame využitie Fact-Check Explorer-u sa dá v jeho sekcií *Explorer* zadať kľúčové slovo článku alebo priamo zadefinovať (vyššie spomínané) overovacie stránky. Výsledkom je panel tvrdení o danej téme z konkrétneho zdroja, hodnotenie a odkaz na jeho podrobnejšie vysvetlenie.

Ďalším dostupným online nástrojom je na stránke FakeNews.research.sfu.ca a je súčasťou výskumu univezity Simona Frasera v Kalifornii s názvom „Big data and quality data for fake news and misinformation detection“ od Fatemeh Torabi Asr and Maite Taboada [1]. K výskumu je dostupné aj video [4], kde jedna z autoriek popisuje vznik tohoto nástroja, zbieranie dát a testovanie modelov. Počas ich práce bolo zistené, že klasické spracovanie textu formou váhovania slov a využitia matematických modelov má na menších datasetoch vyššiu úspešnosť ako moderné transformátory – pri testovaní nimi vytvorenej sady dát MisInfoText [5] s 10,000 článkami bol rozdiel úspešností „klasických“ modelov a modelov s hĺbkovým učením takmer 15%. Vo výslednom klasifikátore je využitá metóda podporných vektorov (viď. 2.4) a jeho výstupom je skóre istoty v intervale $x \in < -2, 2 >$; $x \in \mathbb{R}$, kde spodná hranica predstavuje 100% reálnu správu a horná hranica 100% falošnú správu.

[1] www.disinfobservatory.org
[2] https://toolbox.google.com/factcheck/about#fce
[3] Pre slovenský jazyk existuje podstránka https://fakty.afp.com/list
[4] https://bigdatasoc.blogspot.com/2019/05/video-abstract-big-data-and-quality.html
[5] https://github.com/sfu-discourse-lab/MisInfoText

2.3 Strojové učenie - základné poznatky

Strojové učenie je oblasť umelej inteligencie, ktorá sa snaží napodobniť procesy učenia sa ľudí v sfére výpočtovej techniky. Ide o učenie stroja, ktoré využíva hlavnú výhodu počítačov – ich rýchlosť a schopnosť zvládať veľké množstvo úloh z nízkymi nárokmi na zdroje.

Cieľom strojového učenia je poskytnúť počítačovému algoritmu veľké množstvo dát, na základe ktorých je schopný vytvoriť hranice medzi skupinami rôznych prvkov (tie sa nazývajú „boundaries“) a potom dokáže každý nový vstup zaradiť alebo klasifikovať do príslušnej skupiny s určitou mierou istoty. Ďalším cieľom je využiť tieto dáta na vytváranie predpovedí.

Strojové učenie sa dnes používa vo viacerých oblastiach, ako napríklad v marketingu, financiách, medicíne alebo vo vývoji softvéru, kde sa využívajú rôzne algoritmy a metódy, ktoré umožňujú vytvárať modely schopné vykonávať rôzne úlohy, ako napríklad predpovedanie vývoja cien akcií na burze alebo rozpoznávanie obrázkov. Základné delenie kategórii je nasledovné:

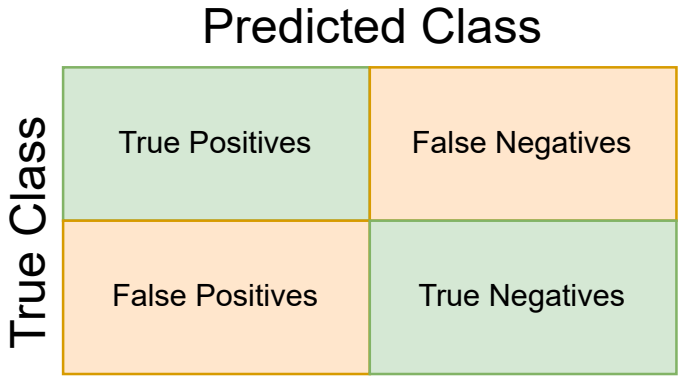
- Učenie s učiteľom,
- Učenie bez učiteľa,
- Kombinované učenie (s aj bez učiteľa),
- Posilňované učenie,
- Neurónové siete,
- Ostatné typy (Hĺbkovéučenie, Robotické učenie, ...).

Využite pri klasifikácii

Keď chceme použiť strojové učenie na riešenie nejakého problému, potrebujeme k dispozícii dáta, ktoré môže náš algoritmus spracovať. Tieto dáta by mali byť dostatočne veľké a rôznorodé, aby sme dosiahli čo najlepšie výsledky. Ale zároveň musíme zabezpečiť, aby boli dáta súvisiace a korelované, pretože ak použijeme nezávislé alebo úplne odlišné dáta, naše rozhodovacie hranice by mohli byť príliš široké a všeobecné.

Pre úspešné vykonanie úlohy strojového učenia musíme mať kvalitné vstupné dáta. Pri trénovaní modelu sa snažíme vytvoriť algoritmus, ktorý sa učí z týchto dát a vytvára predpoveď na základe nich. Je dôležité zvoliť správny typ modelu pre danú úlohu a nastaviť jeho parametre tak, aby dosiahol čo najlepšie výsledky. Zároveň musíme zabezpečiť, aby boli dáta dostatočne relevantné pre danú úlohu, pretože inak môže náš model vytvárať široké a nepresné rozhodovacie hranice.

Pri klasifikácii sa vieme štatisticky pozerať na úspešnosť vyhodnocovania pomocou tzv. Kontingenčnej matice (viď. obrázok 2.1)



Obr. 2.1: Kontingenčná matica (tiež Matica zámen, anglicky „Confusion matrix“).

Z nej vieme výstupy utriediť nasledovne:

- True positive – (správne zaradené),
- True negative – (správne nezaradené),
- False postivie – (nesprávne zaradené),
- False negative – (nesprávne nezaradené).

Model bude najpresnejší, ak docielime nasledujúce pravidlá: „true positive“ a „true negative“ budú dosahovať čo najvyššie hodnoty a „false positive“ a „false negative“ budú dosahovať čo najnižšie hodnoty.

2.4 Metódy klasifikácie

[2] Cieľom klasifikácie v strojovom učení je schopnosť vstupnému vektoru x priradiť jednu z K diskrétnych tried C_k, kde $k = 1, ..., K$. V najčastejších prípadoch sú tieto disjunktné, takže každému vstupu je priradená práve jedna. Priestor vstupných vektorov je teda rozdelený do takzvaných *rozhodovacích oblastí*, ktoré sú oddelené *rozhodovacími hranicami*. Pre lineárne modelý sú tieto hranice lineárnymi funkciami vstupného vektoru x a teda sú definované $(D-1)$-dimenzionálnymi nadrovinami v rámci D-dimenzionálneho priestoru. Dáta, ktorých triedy vieme touto hranicouu presne rozdeliť sa nazývajú *lineárne separovateľné*.

Naive Bayes

Zaoberá sa metódou klasifikácie a je založený na tzv. Bayesovej vete (podľa štatistika Thomasa Bayesa), ktorá určuje pravdepodobnosť javu na základe vopred známych vstupných podmienok. Jeho prívlastok 'Naive' (naivný) predpokladá nezávislosť atribútov (v prípade klasifikácie textu sú to jednotlivé slová). Využíva spomínanú rovnicu

$$P(H|X) = \frac{P(X|H) * P(H)}{P(X)} \tag{2.1}$$

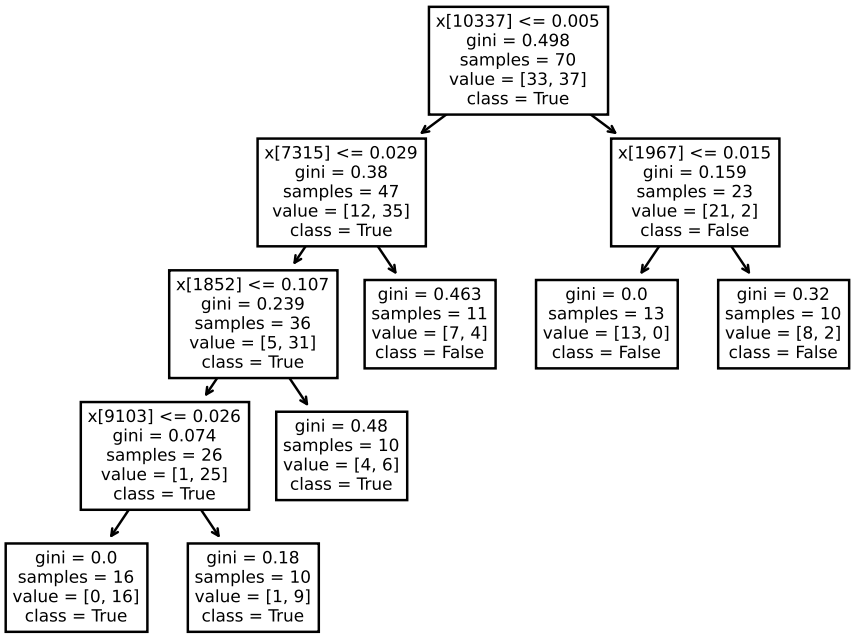
kde H je hypotéza a X je dôkaz. Potom $P(H|X)$ je výsledná pravdepodobnosť, že hypotéza H platí, $P(X|H)$ označuje pravdepodobnosť dôkazu X, keď H platí. Pre doplnenie, $P(H)$ značí pravdepodobnosť hypotézy H a $P(X)$ pravdepodobnosť dôkazu X.

Rozhodovací strom

[20] Metóda rozhodovacích stromov (anglicky. **Decision trees**) je jedným z prvých a najpoužívanejších algoritmov na klasifikáciu a predikciu. Spôsobom trénovania na množinách dát predpovedajú hodnotu cieľovej premennej. Vytvárajú štruktúru uzlov a hrán, graficky pripomínajúcu strom. Uzly sú triedy a hrany hodnoty testovacieho atribútu. Pre využite takéhoto stromu je nutné vytvoriť jeho štruktúru – zvoliť si vstupné premenné a zadefinovať ich prahové hodnoty, na základe ktorých sa bude strom ďalej rozvetvovať. Pri klasifikácií je rast stromu ovplyvňovaný metrikou jeho presnosti – informačným ziskom. Ak definujeme $p_{\tau k}$ ako podmnožinu dát množiny R_τ patriacej triede k, kde $k = 1, ..., K$, potom jednou z možností, ako informačný zisk vypočítať, je takzvaný *Gini Index*

$$Q_t(T) = \sum_{k=1}^{K} p_{\tau k}(1 - p_{\tau k}) \tag{2.2}$$

ktorého prednosťou je vysoká citlivosť na rozdelenie pravdepodobností jednotlivých tried v uzloch stromu, práve kvôli ktorej bol využitý aj v projekte.



Obr. 2.2: Vizualizácia rozhodovacieho stromu vygenerovaná po trénovaní (Experiment 5 – Dezinfo SK). Na obrázku je možné vidieť hodnoty Gini indexov v uzloch, počet vzoriek spĺňajúci kritériá zaradenia, ich hodnoty a triedy.

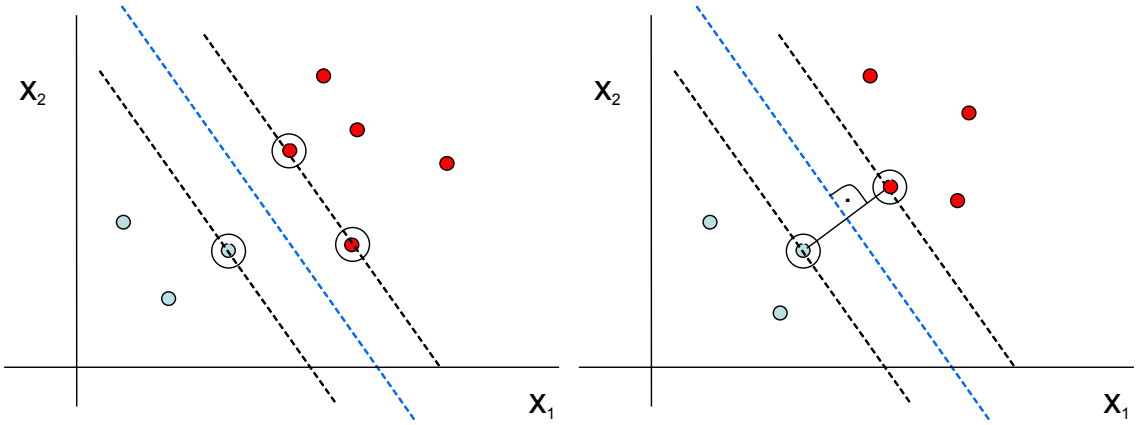
Podporné vektory

[7, 4] Ide o najčastejšie využívaný klasifikátor pre problémy s dvoma triedami. Tento algoritmus spadá do kategórie tzv. jadrových algoritmov – základným prvkom sú rozhodujúce roviny, pomocou ktorých sa hľadá tzv. nadrovina. Tá pôvodný priestor príznakov rozdeľuje tak, že trénovacie dáta z odlišných tried ležia v opačných pol priestoroch, a maximalizuje

priestor medzi danou pol rovinou a podpornými vektormi. Inak povedané, nadrovina sa snaží okolo seba vytvárať čo najväčšie pásmo bez bodov tak, aby okrajové body rozdelených oblastí boli rovnako vzdialené od stredu pásma nadroviny. Súčasťou techniky podporných vektorov je jadrová transformácia priestoru (z pravidla do vyššej dimenzie). V prípade lineárnej transformácie je vzniknutá nadrovina v dvojrozmernom priestore priamkou, v trojrozmernom priestore rovinou. Trénovanie takýchto modelov prebieha s cieľom minimalizovať vzdialenosť hraníc rozdeľúcich priestor vstupných dát x s podmienkami

$$t_n(w^T x_n + b) \geq 1,\ n \in (1, 2, ..., N) \tag{2.3}$$

kde kde w je (normálový) váhový vektor, b je explicitne definovaná odchýlka, N je počet trénovacích vektorov a $t_n \in -1, 1$ udávajú triedy pre jednotlivé vstupné dáta (v prípade binárneho klasifikátoru práve dve). Výstup modelu podporných vektorov neinterpretuje priamo pravdepodobnosť tried pre daný stup, ale takzvané „mäkké skóre“, ktoré sa dá na ňu približne previesť.

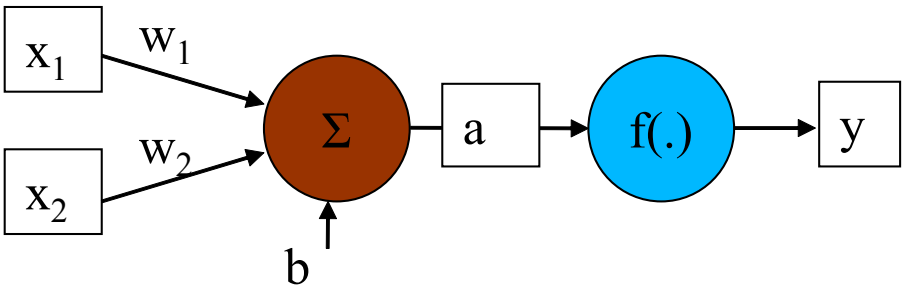


Obr. 2.3: Podporné vektory (*support vectors*) ležia na okraji prázdnej oblasti (čierna prerušovaná čiara) a ovplyvňujú riešenie. Ak by sa ostatné dáta vypustili z trénovacej sady, výsledok by se nezmenil. Prevzaté z [7].

Neurónové siete

[3, 21] Umelá neurónová sieť je jednou z najpoužívanejších metód strojového učenia. Vo svojej podstate modeluje funkcionalitu neurónov v biológií – tie reagujú na vstupné parametre, produkujú výstup a komunikujú s ostatnými neurónmi. Počet vstupov neurónu je neobmedzený, avšak výstup je len jeden. História vývoju prvých matematických modelov neurónov siaha až do štyridsiatych rokov dvadsiateho storočia. Prvý neurón s pravidlami učenia, tzv. **perceptron** bol vytvorený F. Rosenblattom o pár rokov neskôr.

Perceptron má jednoznačne určené vstupy a výstupy, ich váhy, prahovú hodnotu a aktivačnú funkciu. Jednotlivé neuróny sa spájajú do celku a vytvárajú neurónovú sieť. Predpokladom čo najsprávnejšieho výstupu takejto siete je jej schopnosť učenia sa. V biologickej sfére sú nové skúsenosti získavané pomocou zmyslov a ukladajú sa v synapsiách, resp. v spojeniach jednotlivých neurónov. V matematickom modeli je táto akcia nahradená zmenou váhovania vstupov. Podobne sa výstupy neurónov (sila elektrických impulzov) nahrádzajú rôznymi aktivačnými funkciami, ktoré modelujú rôzne chovania, ako napríklad citlivosť neurónov na vstup, prípadne takzvané „zabúdanie“.



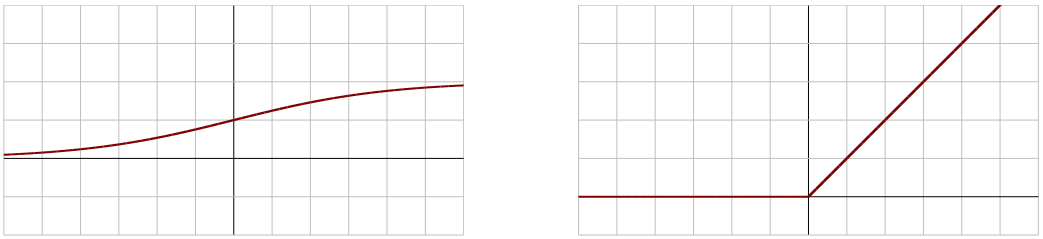
Obr. 2.4: Perceptron – $x_{1,2}$ určuje vstupy, $w_{1,2}$ váhy vstupov, b odchýlku, a výstup po súčte, $f(.)$ aktivačnú funkciu daného neurónu a y jeho finálny výstup. Prevzaté z [7]

.

[5, 17] Medzi najčastejšie využívané (nelineárne) funkcie patria:

- Sigmoida,
- Gaussova krivka,
- Ricker wavelet (tiež zvaná „Mexický klobúk“),
- Hyperbolický tangens,
- Rectified linear unit (tiež „ReLU“),
- Softmax [6].

V niektorých prípadoch sa taktiež používajú aj radiálne, zložené či postupne diferencovateľné funkcie.



(a) Sigmoida (b) ReLU

Obr. 2.5: Sigmoida 2.5a (inak tiež logistická funkcia), ktorej rozsah je pevne ohraničený, čo modeluje určitú maximálnu rýchlosť odozvy neurónov. ReLU 2.5b je jedna z modernejších aktivačných funkcií, ktorej hlavné výhody sú výpočetná nenáročnosť či lepší gradientný zostup v oboch smeroch. V elektrotechnike je jej povaha analógiou k pol-vlnovému usmerňovaču striedavého prúdu. Prevzaté z [17].

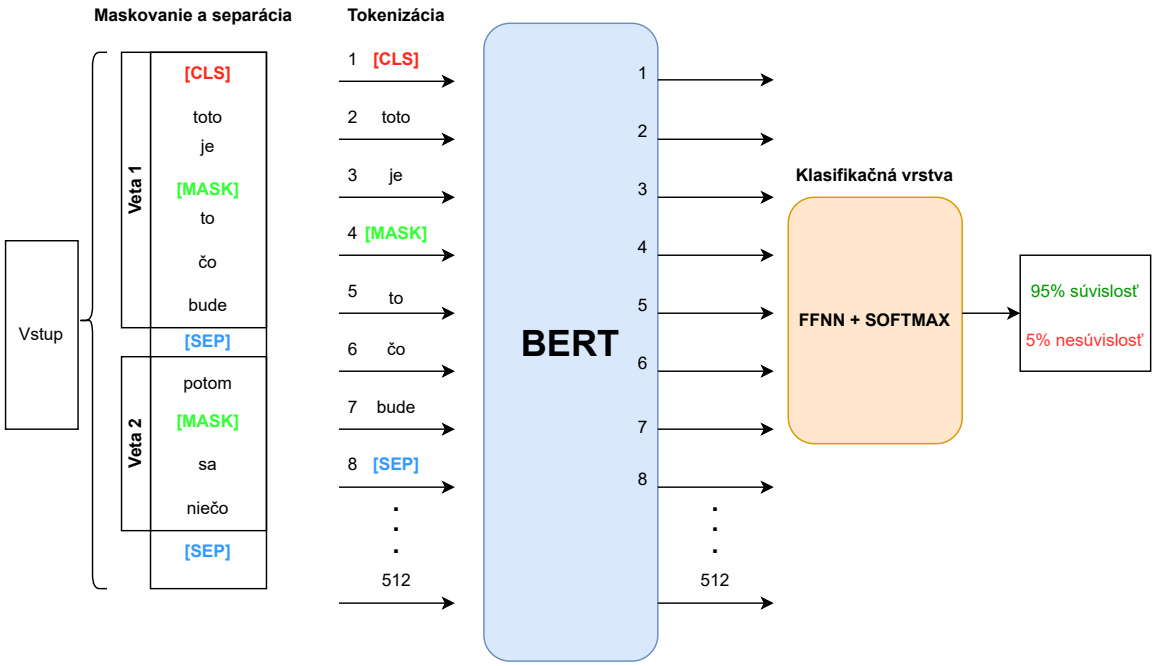
Existuje množstvo metód a algoritmov pre učenie neurónových sietí, kde medzi najpoužívanejšie patrí napríklad Hebbovské učenie (podľa paralelnej aktivácie neurónov) či Spätné šírenie chyby [(anglicky „Backpropagation“ (ako gradientný zostup zmien váhovania).

[6] Na rozdiel od vyššie spomínaných funkcii, Softmax pri výpočte používa hodnoty premenných z predošlých vrstiev. Využíva sa hlavne v posledných vrstvách sietí, kde dokáže normalizovať výstupy na rozdelenie pravdepodobností predikovaných tried. [2]

BERT

Podľa publikácie [15] z roku 2020 sa medzi najpoužívanejšie a najúspešnejšie modely zaradili predtrénované hlboké strojové učenie založené na transformátoroch, ako je **BERT**, **RoBERTa** alebo **XLNET**. **BERT**, alebo celým názvom Bidirectional Encoder Representations from Transformers bol vytvorený v roku 2017 spoločnosťou Google. Je zložený na modeli hĺbkového učenia, kde váha každého výstupu je dynamicky prerátaná ako vstup, a zároveň umožňuje prechod analyzovaným textom obojsmerne. Transformátory dovoľujú trénovať tento model aj na väčšom objeme dát než pri iných modeloch – konkrétne BERT bol trénovaný na celej anglickej Wikipédii a na dátach z Toronto BooksCorpus. Jeho hlavnou prednosťou je priraďovanie významu slovám na základe toho, aký význam majú vo vete (podľa slov, ktoré sa nachádzajú okolo nich). Existuje niekoľko voľne dostupných modelov, ktoré sú predtrénované na špecifických dátach pre klasifikáciu patentov, vedeckých alebo medicínskych textov či vizuálno-lingvistických dát.

Pri trénovaní modelov BERT sa využívajú dve techniky: maskovanie a predikcia viet [6]. Sú využívané súčasne, čo minimalizuje stratovosť oboch stratégii. Maskovanie prebieha tak, že 15% slov v určitej sekvencií je „zakrytých“, respektíve nahradených špeciálnou maskou. Tá môže byť buď to prázdny reťazec, náhodné slovo alebo pôvodné slovo vo vete. Model sa následne snaží predpovedať toto ukryté slovo na základe zvyšných slov v sekvencii. Predikcia nasledujúcich viet je proces, pri ktorom model na vstup dostáva dvojicu viet a snaží sa predpovedať, či tento pár na seba kontextovo nadväzuje, alebo nie. Počas trénovania je polovica týchto viet na seba nadväzujúca a druhá polovica nie. Taktiež sú tieto vety označené vopred označené na začiatku a na konci, čo modelu uľahčuje ich rozpoznávanie.



Obr. 2.6: Vizualizácia trénovania rozpoznávania súvislosti viet modelov BERT. Ako klasifikačná vrstva môže figurovať napríklad neurónová sieť (na obrázku Feedforward Neural Network) s aktivačnou funkciou Softmax 2.4 alebo iný klasifikátor. Obrázok inšpirovaný [7].

[7] https://www.geeksforgeeks.org/understanding-bert-nlp/

Transformátory sa v odvetví spracovania textu stávajú čoraz viac populárnejšími a ich využitie je naozaj všestranné. Na základe toho sa vynára otázka, či matematicky založené modely (ako napríklad Naive Bayes, Rozhodovacie stromy či Podporné vektory) sú stále relevantné na použitie a či je vôbec možné dosiahnuť porovnateľnú efektivitu voči neurónovým sieťam a transformátorom.

Poznámka: Kvôli náročnej implementácii je prístup ku spracovaniu prirodzeného jazyka a klasifikácii na základe modelov BERT v práci nevyužitý, avšak figuruje v porovnávaní výsledkov jednotlivých experimentov, pri ktorých ho buď samotní autori zbierok dát alebo ich používatelia využili na tento účel a slúži ako „state-of-the-art“ riešenie, respektíve najlepšie dosiahnuté skóre pre daný dataset.

Kapitola 3

Návrh riešenia

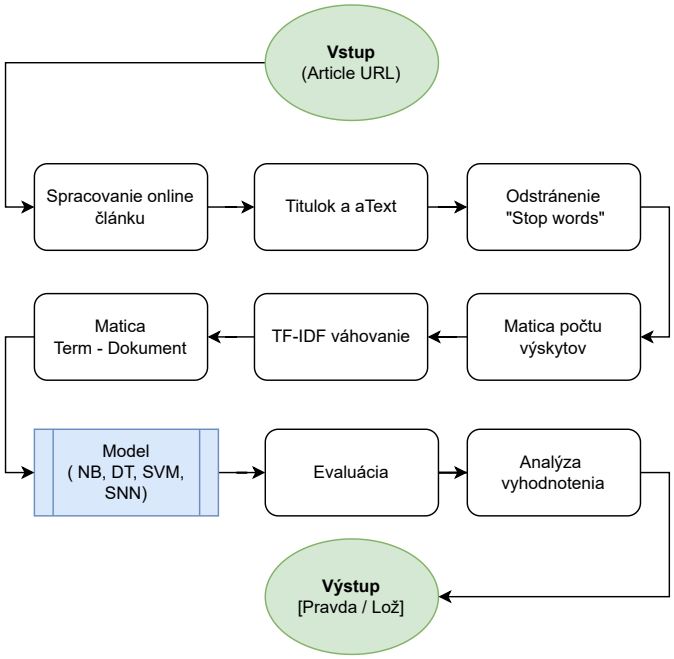
3.1 Cieľ práce

Cieľom tejto práce je vytvoriť a poskytnúť začínajúcim programátorom jednoduchú knižnicu pre klasifikáciu novinových článkov na základe ich spoľahlivosti. Táto knižnica spája jednotlivé kroky nutné pre spracovanie dát, trénovanie modelov a evalváciu textu z odkazu na konkrétny príspevok. Modul podporuje výber vlastného datasetu, zvolenie trénovaného modelu strojového učenia (*Naive Bayes*, *Rozhodovací strom*, *Podporné vektory* alebo *Neurónová sieť*) a jeho uloženie, prípadne využitie predtrénovaných modelov. Používateľ si vie taktiež určiť, či chce vidieť detailný popis vyhodnotenia a výsledky testov po trénovaní.

Ďalším cieľom je taktiež vytvorenie verejne dostupnej zbierky dát klasifikovaných článkov z prostredia slovenských novinových portálov, ktoré je možné použiť práve na vyššie spomínané účely. Zámienkou na jeho vytvorenie je nedostatok verejne prístupných dát pre klasifikáciu textu v slovenskom jazyk.u Dataset je tvorený manuálne a je pri ňom zabezpečený rovnomerný podiel pravdivých a nepravdivých príspevkov. Ich zaradenie do jednej z dvoch tried spoľahlivosti je taktiež robené manuálne na základe verejne dostupného hodnotenia slovenských internetových spravodajských stránok vytvoreného špecialistami v obore. Viac o tejto sade dát je možné sa dozvedieť v prílohe C, ktorá ho detailnejšie popisuje.

3.2 Návrh architektúry

Riešenie má tvar stiahnuteľného balíčku pre jazyk Python (prípadne pri používaní zdrojových súborov aj ako konzolová aplikácia), kde je možné zadať vstupné dáta na klasifikáciu a zvoliť preferovanú metódu klasifikácie. Dáta sa na pozadí vyhodnotia a ako výstup je vidno rozdelenie pravdepodobnosti vierohodnosti vstupu. Je taktiež možné pozrieť si iné štatistické údaje a porovnania s tréningovou sadou. Druhou možnosťou je využitie predtrénovaných modelov na ohodnotenie správy z internetu. Obe metódy vyžadujú rovnaký postup (viď. obrázok 3.1).



Obr. 3.1: Návrh architektúry - kostra programu a priebeh spracovania vstupného textu až po výstupnú klasifikáciu.

3.3 Výber vhodných dát a ich vizualizácia

Keď chceme použiť strojové učenie na riešenie nejakého problému, potrebujeme mať k dispozícii dostatočné množstvo vstupných dát, ktoré budú predstavovať rôzne prípady a situácie, ktoré budeme chcieť modelu predpovedať. Je dôležité, aby tieto dáta boli čo najrôznorodejšie, aby sme mohli modelu poskytnúť čo najviac informácií o rôznych možnostiach a scenároch, ktoré by mohol model v budúcnosti vidieť. Zároveň ale musíme dbať aj na to, aby boli tieto dáta relevantné a súvisiace s tým, na čo chceme model použiť. Pokiaľ trénujeme model na dátach, ktoré sú navzájom nesúvisiace alebo úplne odlišné od toho, na čo chceme model použiť, vytvorené rozhodovacie hranice budú príliš široké a všeobecné a model sa môže správať nepresne alebo dokonca chybne. Vždy je teda dôležité zvážiť kvalitu a objem dát, ktoré použijeme na trénovanie modelu, aby sme dosiahli čo najlepšie výsledky.

Pri výbere dát máme dve možnosti – vytvoriť si trénovaciu množinu samostatne alebo použiť niektorú z verejne dostupných a odskúšaných. Prvá možnosť je časovo náročná a vytvorenie dostatočne rozsiahleho balíčku klasifikovaných správ by bolo zložité. Druhá možnosť nám síce odoberie časť práce, avšak musíme viac zohľadniť niektoré faktory dôležité pre správny výber, ako napríklad ich relevantnosť a formát. Existuje niekoľko takýchto súborov dát, ktoré boli špeciálne vyvinuté pre klasifikáciu falošných správ a ďalšie sa neustále vyvíjajú. Pre tento účel bola využitá stránka `kaggle.com`, ktorá obsahuje voľne dostupné datasety rôznych kategórii a zameraní. Na túto úlohu boli z tejto stránky vybrané štyri subory dát a jeden vytvorený. Ich bližší popis sa nachádza v sekcií 5 Pre zjednodušenie práce boli všetky typu csv (comma separated values). Štruktúra všetkých dokumentov je taktiež podobná – obsahuje (spravidla) stĺpce **title**, **text**, **label**, **date** prípadne **subject**.

id	date	title	text	url	check	label
1	5.3.2023	Ďalšia skupina renomo..	Je veľmi dôležité, aby sa ...	zemavek.sk/dalsia-...	-	0
2	4.3.2023	Pfizer tajil poškodenia...	Počas uponáhľaných klinických ...	zemavek.sk/pfizer-...	-	0
3	3.3.2023	Vedecká štúdia preuká...	Ich povinné nosenie bolo preto ...	zemavek.sk/vedeck...	-	0
4	2.3.2023	Ukrajinci pritvrdili v ...	Ukrajinská armáda v posledných ...	zemavek.sk/ukrajin...	-	0
5	3.3.2023	Ukrajinskí novinári ...	Na doneckých frontoch sa ...	zemavek.sk/ukrajin...	-	0
6	5.3.2023	Vojna na Ukrajine: Ruskí...	Ruskí mobilizovaní záložníci ...	aktuality.sk/cl...	-	1
7	4.3.2023	Koronavírus: WHO ...	Generálny riaditeľ Svetovej ...	aktuality.sk/cl...	-	1
8	4.3.2023	Metsolová na Ukrajine ...	Predsedníčka Európskeho ...	aktuality.sk/cl...	-	1
9	4.3.2023	Rusko môže byť čoskoro ...	Rusko by sa mohlo už na ...	aktuality.sk/cl...	-	1
10	5.3.2023	Bachmut je stále v ...	Situácia v meste Bachmut v ...	aktuality.sk/cl...	-	1

Tabuľka 3.1: Ukážka štruktúry datasetov využitých v projekte. Konkrétny úsek pochádza z datasetu Dezinfo SK, ktorý bol navrhnutý podľa štandardnej štruktúry zbierok novinových článkov pre klasifikáciu. Trieda (*label*) **0** označuje nepravdivú správu a trieda **1** správu pravdivú.

Vizualizácia dát

Na ukážku obsahu používaných záznamov môžeme použiť takzvanú slovnú mapu – pomocou online nástroja na tvorbu „wordclouds“ [1] vieme po tokenizácii a odstránení „stop words“ vziať výsledok a jednoducho vykresliť prehľad najčastejšie vyskytujúcich sa slov v danom dokumente (najpoužívanejšie slovo je najväčšie):

Obr. 3.2: Slovná mapa - ukážka

3.4 Vektorizácia

Pred ich spracovaním je potrebné vykonať niekoľko úprav – v prvom rade boli načítané pomocou knižnice **pandas** do dátového rámca. V prípade samostatných súborov pre jednotlivé kategórie im boli manuálne priradené triedy Fake a True a datasety boli zlúčené do jedného rámca. V nasledujúcom kroku bolo vyukonané spracovanie funkciami knižnice **nltk**, kde sa odstránila interpunkcia a boli vylúčené tzv. „Stopwords“ – slová, ktoré nie sú vo vetnej syntaxi dôležité (častice, citoslovcia, spojky, predložky, zámená a niektoré slovesá). Zvyšné slová boli upravené tak, aby obsahovali len malé písmená (ang. lowercase) a stáli zvlášť (tokenizácia).

Tieto dáta sú potom zoskupované do matice podľa ich výskytov cez funkciu **Count vectorizer**. Ak máme množinu dokumentov, kde každý dokument je reprezentovaný ako reťazec znakov, tak výstupná matica má tvar $N * K$, kde N je počet dokumentov v množine a K je počet unikátnych slov vo všetkých dokumentoch. Každý riadok matice potom

[1] https://www.wordclouds.com/

reprezentuje jeden dokument a obsahuje hodnoty vektorov pre všetky slová v abecednom poradí. Pri tejto matici si vieme skontrolovať jej hustotu, resp. riedkosť – ak obsahuje príliš veľa hodnôt s nízkym alebo nulovým výskytom, musí byť upravená (stlačená) tak, aby sa obmedzili výkyvy pri klasifikácii. Napríklad pre nasledujúce dve vety:

```
Toto je ukážka matice výskytu slov.
```

a

```
Táto ukážka je maticou niekoľkých slov.
```

by výstup tejto funkcie vyzeral nasledovne:

	je	matice	maticou	niekoľkých	slov	táto	toto	ukážka	výskytu
1.	1	1	0	0	1	0	1	1	1
2.	1	0	1	1	1	1	0	1	0

Tabuľka 3.2: Ukážka výstupu funkcie Count Vectorizer bez predspracovania textu. Existuje mnoho krokov, ako zmenšiť takúto maticu – lematizácia, ostránenie 'stop words' alebo definícia minimálneho či maximálneho počtu výskytov jednotlivých slov.

V prípade experimentov ale vďaka predspracovaniu táto matica nulové hodnoty obsahovať nebude (všetky slová budú mať výskyt aspoň raz).

Ďalším dôležitým krokom je použitie výpočtu TF-IDF (Term Frequency – Inverse Document Frequency). V podstate ide o výpočet významnosti slov v korpuse dokumentov. Je to pomer výskytu konkrétneho výrazu v jednom dokumente na počet výskytu slova v celom korpuse:

$$TF - IDF(\mathbf{t}, \mathbf{d}) = TF(\mathbf{t}, \mathbf{d}) * IDF(\mathbf{t}) \tag{3.1}$$

Pričom

$$TF(\mathbf{t}, \mathbf{d}) = \frac{f_d(t)}{|d|} \tag{3.2}$$

kde $f_d(t)$ určuje počet výskytov termu $\mathbf{t}$ v dokumente $\mathbf{d}$ a $|d|$ je počet slov v danom dokumente,

$$IDF(\mathbf{t}) = log(\frac{1+n}{1+df(t)}) + 1 \tag{3.3}$$

kde n je počet dokumentov v korpuse a $df(t)$ počet dokumentov v korpuse obsahujúcich term $\mathbf{t}$. Výsledkom tejto funkcie je matica výskytov rovnakého tvaru, ako po použití *CountVectorizer*, avšak jej hodnoty sú normalizované v rozsahu $x \in < 0, 1 >$; $x \in \mathbb{R}$.

pozn.: formula popísaná vyššie použitá vo funkcií **TfidfTransformer** z knižnice sklearn sa odlišuje od klasickej rovnice – výraz +1 na konci zabezpečuje, aby slová ktoré sa vyskytujú vo všetkých dokumentoch neboli ignorované a konštanta 1 pri čitateľovi aj menovateľovi zabezpečuje prevenciu pri delení nulou.

3.5 Vyhodnocovanie modelov

Vyššie spomenuté modely boli natrénované a následne otestované na množine dát z datasetu, ktorá bola rozdelená v pomere 70:30. Po každom natrénovaní modelu je popri ich

testovaní vytvárané aj vyhodnotenie ich spoľahlivosti. Na tento úkon bola použitá funkcia **classification_report** z balíku **scikit**, ktorá produkuje textový výstup klasifikácie pre lepšiu prehľadnosť. V nej môžeme sledovať presnosť, citlivosť či spoľahlivosť daného modelu (viď. v kapitole Testovanie a vyhodnocovanie, tabuľky 5.1, 5.4, 5.7, 5.10 a 5.13.

Pri výpočte týchto hodnôt sa používajú prvky z kontingenčnej matice 2.1 (ktorá je tiež súčasťou výstupu), kde:

- precision – **TP / (TP +FP)**,
- recall – **TP / (TP +FN)**,
- accuracy – **TP + TN / počet všetkých vzoriek**,
- F1-Score (harmonický priemer citlivosti a precíznosti) – **2 * (recall * precision) / (recall + precision)**,
- support – počet reálnych výskytov danej triedy v datasete.

Hlavným smerodajným ukazateľom je práve F1-Score, pomocou ktorého dokážeme porovnávať jednotlivé modely a nie len globálnu presnosť. Jeho interval je $x \in <0, 1> \; x \in \mathbb{R}$, pričom čím vyššie skóre model po tréningu dosiahne, tým je schopnejší presnejšie klasifikovať nové vstupy. Je taktiež dôležité zohľadniť rovnováhu tried v dátach a náročnosť klasifikácie – v prípade, že triedy dát nie sú vyvážené, môže model vynikať v predikcii väčšej a/alebo častejšej triedy, ale zlyháva pri menšej a/alebo menej častej triede. V takomto prípade sa používajú techniky vyvažovania dát, ako napríklad „over-sampling“ alebo „under-sampling“ (pre-vzorkovanie a pod-vzorkovanie).

Kapitola 4

Implementácia

4.1 Technológia a prostriedky

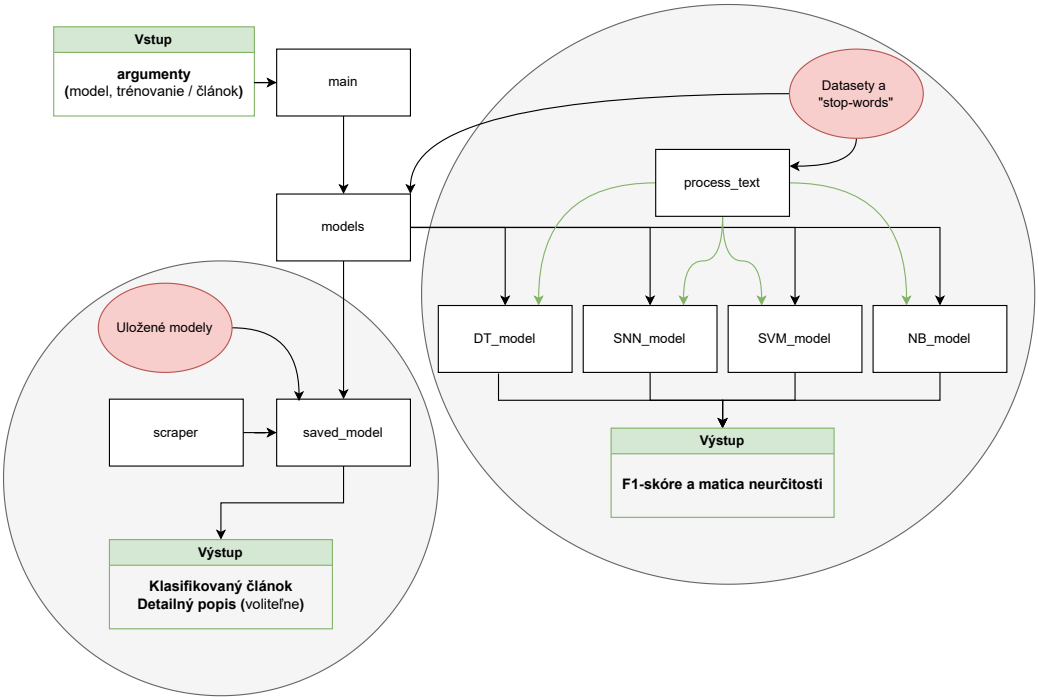
Ako vývojové prostredie bol zvolený jazyk Python verzie 3.9, ktorý je vďaka množstvu knižniciam pre strojové učenie, spracovanie textu či práce s grafickým používateľským rozhraním veľmi vhodný. Existuje k nim taktiež dostatok online dokumentácie a vzorových ukážok, takže ich použitie je jednoznačné a prácu zľahčujúce. Konkrétne sú v projekte využité nasledujúce knižnice:

- pandas – manipulácia s dátami a ich analýza[1],
- scikit-learn – prediktívna analýza a modely, vektorizéry, skórovanie[2],
- matplotlib + seaborn– grafy, vizualizácia[3],
- nltk – spracovanie prirodzeného jazyku[4],
- keras – neurónová sieť a nástroje na jej implementáciu[5],
- lime – analyzátor výstupu modelov[6],
- dill – ukladanie modelov (deserializácia objektov)[7],
- ostatné (numpy, sys, string, os).

4.2 Programová časť

Program je rozdelený na moduly, ktoré sú volané z hlavného programu – Spracovanie textu, jednotlivé modely a koncové rozhranie. Pomimo kódovej časti sú uložené predtrénované modely, datasety a k nim použité pomocné súbory.

[1] https://pandas.pydata.org/
[2] https://scikit-learn.org/stable/index.html
[3] https://matplotlib.org/, https://seaborn.pydata.org/
[4] https://www.nltk.org/
[5] https://keras.io/api/models/sequential
[6] https://lime-ml.readthedocs.io/en/latest/lime.html
[7] https://pypi.org/project/dill/



Obr. 4.1: Vizualizácia implementácie - vzájomné vzťahy medzi jednotlivými skriptami (obdĺžniky) a potrebnými dátami (elipsy). Program sa dostane do šedo vyznačených zón práve na základe argumentov zadaných pri spúšťaní. Funkcie definované v časti *process_text* sú samostatne volané modelmi pri trénovaní.

Hlavný modul

Hlavný modul (*main.py*) má za úlohu volať jednotlivé časti programu podľa argumentov zadaných užívateľom pri spúšťaní cez príkazový riadok:

`-h, --help:` zobrazenie nápovedy,

`-m, --model [názov modelu]:` výber modelov,

`-p, --pretrained:` výber medzi uložením modelu po trénovaní a použitím predtrénovaného modelu,

`-url, --url [odkaz na článok]:` zdroj online textu pre evalváciu,

`-v, --verbose:` bližší popis vyhodnotenia (html dokument a obrázky matíc zámen).

Pre účely spúšťania programu s využitím balíka FakeNewsClassifier 4.3 je vo funkcii hlavného modulu *models.py*, ktorý obsahuje definíciu triedy a jej metód zastupujúcich prácu s argumentami.

Návod na prácu obsahujúci kroky pre inštaláciu a spustenie programu je uvedený v projektovej dokumentácii v súbore *README.md* (viď. Obsah priloženého pamäťového média)

Spracovanie textu

Pre spracovanie textu bola využitá knižnica **NLTK**, ktorá zabezpečuje všetky potrebné funkcie na predspracovanie článkov pre jednoduchšiu analýzu. Z nej boli využité balíčky

punkt pre tokenizáciu, *wordnet* pre lematizáciu a *corpora* pre anglické 'stopwords'. Pre slovenské datasety bol dokument obsahujúci 418 častíc, citosloviec, zámen a iných rôznych „vylúčených slov“ prevzatý od autora *genediazjr* zo stránky GitHub [8].

V module *process_text.py* je definovaná funkcia, ktorej vstupným argumentom je text v čistej podobe. Ten sa v prvom kroku očistí od interpunkcie a čísel (objekty v knižnici *string*), následne tokenizuje (vytvoria sa samostatne stojace slová) a prebehne lematizácia – hľadanie základných tvarov slov v dokumente, respektíve v množine tokenov. Z takto upravených slov je následne možné odstrániť tie nepotrebné, a to tak, že sa každé slovo porovnáva so skupinou 'stop words', v našom prípade anglických či slovenských. Po tomto kroku je sada pripravená na vektorizáciu. Skrátená ukážka kódu:

```
def process_text(text):
 #odstránenie interpunkcie a čísel
    text = ''.join(    #odstránenie interpunkcie a čísel
        [c for c in text if c not in string.punctuation and
        c not in string.digits])
 #tokenizácia
    tokens = word_tokenize(text)
 #lematizácia
    lemmatized = [WordNetLemmatizer.lemmatize(word) for word in tokens]
 #stop-words
    stopped = [word for word in lemmatized if word.lower() not in sw]
    return stopped
```

Modely

Modely Naive Bayes, Decision tree a Support Vector Machine sú prístupné v knižniciach *sklearn* a model neurónovej siete je z knižnice *keras*. Pre uľahčenie práce a zjednotenie krokov trénovania a vyhodnocovania bolo naimplementovaný takzvaný „pipelining“ – ten dovoľuje špecifikovať jednotlivé kroky a uniformne medzi nimi vykonávať vkladanie a transformáciu vstupných dát. Pre každý dizajn sa táto sekvencia skladá z spracovania textu, vektorizácie a finálneho klasifikátoru. Ukážka pre model Naive Bayes:

```
pipeline = Pipeline([
        # strings to token integer counts
        ('bow', CountVectorizer(analyzer=pt.process_text)),
        # integer counts to weighted TF-IDF scores
        ('tfidf', TfidfTransformer()),
        # train on TF-IDF vectors w/ Naive Bayes classifier
        ('classifier', MultinomialNB()),
    ])
```

Môžeme vidieť, že funkcia **CountVectorizer** (popísaná v sekcii 3.4) umožňuje explicitné využitie externej funkcie na analýzu vstupu. Výstupná matica je vstupom pre váhovaciu funkciu, po ktorej model (koncový klasifikátor – položka „classifier“) pracuje priamo s váhami jednotlivých slov. Rovnaký postup je vykonávaný aj pre ostatné modely. Trénova-

[8] https://github.com/stopwords-iso/stopwords-sk

nie je potom vykonané volaním metódy `pipeline.fit(X_train, y_train)` a hodnotenie testov a článkov metódou `pipeline.predict(X_test)` [9].

Použité modely boli väčšinou konštruované so základnými implicitnými parametrami, ktoré boli uvedené v dokumentácii. Parametre, ktoré boli zmenené sú nasledovné:

- SVM – lineárne jadro, povolený výstup pravdepodobnosti (viď. 2.4),
- DT – minimálny počet listových uzlov = 10 (pre presnejší výstup pravdepodobností).

Neurónová sieť bola implementovaná trojvrstvovou architektúrou so sekvenčným prechodom a dopredným šírením (anglicky „Feed-forward neural network“, viď. obrázok 4.2). Pri jej kompilácii je ako optimalizátor zvolený (zároveň najčastejšie využívaný) „adam“ [10], stratovosť je sledovaná metódou „sparse categorical crossentropy“ [11] a hlavnou metrikou je presnosť modelu. Aby nedochádzalo k zbytočným epochám pri trénovaní a znížila sa výpočetná náročnosť a riziko pretrénovania, sieť implementuje a predčasné ukončenie tréningu v momente, kedy sa stratovosť prestane znižovať, takzvaný „Early stopping“ [12].

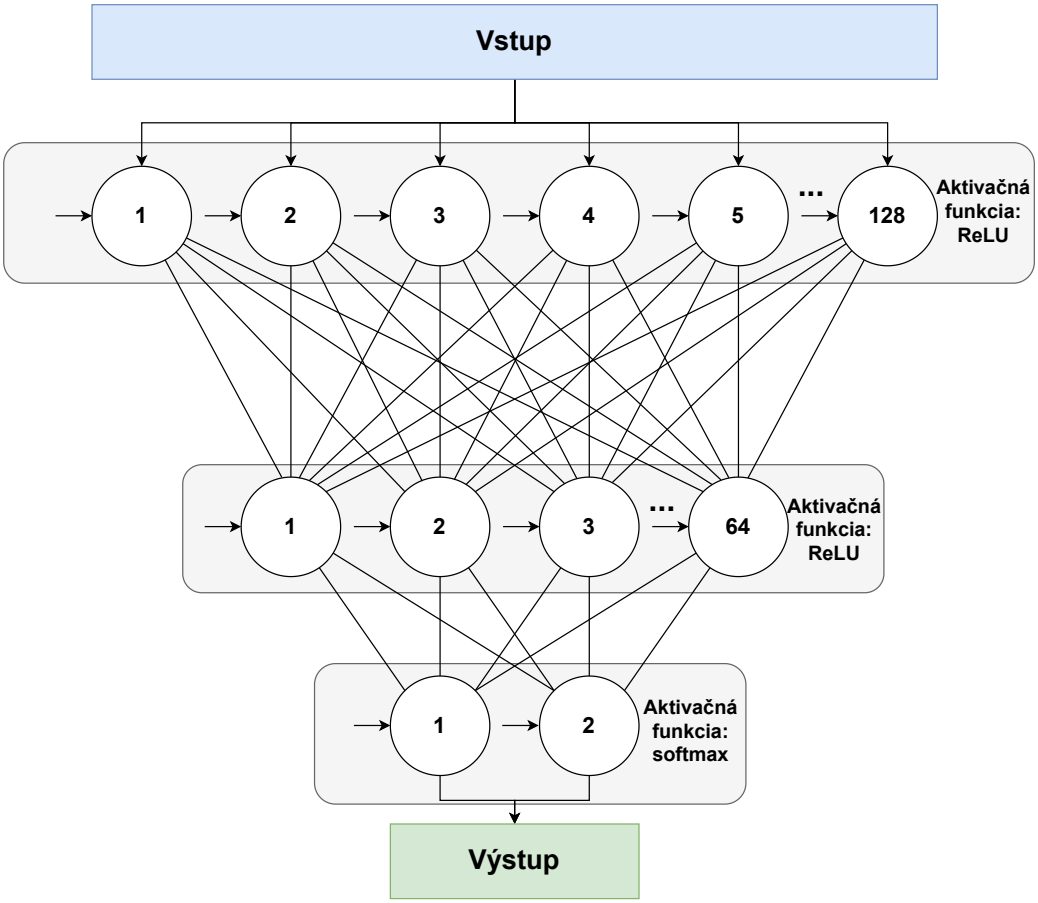
Po trénovaní program vypíše štatistiky úspešnosti a vyzýva užívateľa na potvrdenie uloženia modelu. Modely sa ukladajú ako celok vo vytvorenej „pipeline“, čo uľahčuje jej znovu načítanie pri evalvácii článku. Na túto prácu je využitá knižnica *dill*, ktorá metódou *dump* jednotlivé objekty deserializuje a inverzne ich naopak serializuje. Pre uloženie neurónovej siete je samostatne ukladané spracovanie textu (tiež *pipeline* a váhy modelu vo formáte *.h5*

[9] **X_train**, **X_test**, **y_train** a **y_test** sú množiny vstupných dát, ktoré vznikli pri rozdeľovaní na trénovaciu a testovaciu sadu funkciou **train_test_split** z balíku *scikit*.

[10] https://optimization.cbe.cornell.edu/index.php?title=Adam

[11] https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e

[12] https://paperswithcode.com/method/early-stopping



Obr. 4.2: Implementovaná neurónová sieť. Vstup tvorí matica získaná z TF-IDF váhovania. Každý neurón má konštantný vstup, takzvaný „bias“ (na obrázku vyznačený orientovanou šípkou), ktorý ovplyvňuje celkový výstup. Pre vstup skladajúci napríklad z 1000 hodnôt je počet trénovateľných parametrov, respektíve váh (na obrázku spojenia medzi neurónmi) $(1000 * 128 + 128) + (128 * 64 + 64) + (64 * 2 + 2) = \mathbf{136{,}514}$.

Webové články

Skript *scraper.py* implementuje jednoduché spracovanie článkov z poskytnutého odkazu. Využíva knižnicu newspaper3k, ktorá vytvára objekty článkov a poskytuje rôzne vstavané funkcie pre prácu s týmito objektami. Na získanie celého textu článku sú využité štyri metódy:

```
article = Article(url)                          # Vytvorenie objektu
article.download()                              # Stiahnutie článku
article.parse()                                 # Spracovanie článku
full_text = article.title + article.text        # Spojenie nadpisu a textu
```

Spracovaný text je následne využitý ako vstup na predikciu vierohodnosti v module *saved_models.py*. Ak je finálnym klasifikátorom neurónová sieť, je tento text po prevedení na maticu váh predimenzovaný tak, aby pasoval do jej vstupu.

Analýza vyhodnotenia a grafy

Pre vyhodnotenie úspešnosti testovania modelov po tréningu bola použitý modul *classification_report* z knižnice *sklearn.metrics*. Interpretuje výstupné hodnotenie modelov, respektíve celkových predikcí *pipelines* na testovacej množine pomocou f1-skórovania tak, ako je to popísané v kapitole Vyhodnocovanie modelov. Jeho výstup je možné taktiež po pretransformovaní na dátový rámec vizualizovať pomocou matice zámen. Ich vykresľovanie zabezpečuje štandardná funkcia *matplotlib* pre tvorenie grafov v jazyku *python*. Pri hodnotení internetových príspevkov sú znovu načítané uložené „pipelines“ s modelmi. Na základe vstupu od užívateľa je tiež možné nastaviť zobrazenie detailnejšieho popisu hodnoteného článku pri finálnej evalvácii. Túto možnosť zabezpečuje knižnica *Lime Explainer* od autora Marco Tulio Correia Ribeiro [13]. Umožňuje „nahliadnuť“ do rozhodovania modelov strojového učenia tak, že pozná vstupné dáta a proces ich transformácie, čím nasledovne dokáže interpretovať finálnu predpoveď klasifikátorov na originálnom vstupe [12]. Tento nástroj je vhodný pre lepšie porozumenie rozhodovaniu modelu a dokáže pridať či ubrať na jeho dôveryhodnosti vďaka hlbšiemu porozumeniu zo strany človeka. Použitie tohoto nástroju je veľmi intuitívne, pre ukážku:

```
explainer = LimeTextExplainer(                  # vytvorenie objektu
            class_names=['True', 'False'],      # triedy
            bow=True)                           # využitie "bag-of-words"
explanation = explainer.explain_instance(       # volanie metódy
            news,                               # článok
            pipeline_nb.predict_proba,          # predikcia modelu
            num_features=20)                    # najdôležitejšie slová
explanation.save_to_file('explanation.html')    # uloženie dokumentu
```

Ukážkové výstupy pre vyhodnotenie článkov z portálov **New York Times** a **WorldNetDaily** je možné vidieť v prílohe B. Všetky výstupy analyzátorov sú ukladané do zložiek `figures/..` pre obrázky matíc a `html/..` pre Lime, oboch pod unikátnymi názvami obsahujúcimi dátum a čas vytvorenia.

[13] https://github.com/marcotcr/lime

4.3 Vytvorenie balíčku

Celá implementácia tejto bakalárskej práce bola tvorená so zámerom poskytnúť jednoduchý voľne dostupný nástroj pre klasifikáciu článkov pre programovací jazyk python. Projekt bol preto štruktúrne pozmenený tak, aby vyhovoval štandardom pre vytváranie balíkov a distribúcii pre publikovanie na verejnom indexe PyPi. Sekcia popisuje a parafrázuje jednotlivé kroky práce podľa návodu zo stránok Python.org, presnejšie články [11, 10]. Vo výsledku tento úkon vyžadoval vytvorenie niekoľkých doplnkových súborov nasledovne:

```
Package/
├── LICENSE
├── pyproject.toml
├── README.md
├── MANIFEST.in
└── src/
    ├── main.py
    ├── __init__.py
    └── FakeNewsClassifier/
        ├── datasets/
        │   ├── dezinfo_sk.csv
        │   └── stopwords.txt
        ├── saved_models/
        │   └── predtrénované modely (vo formáte .pk a .hdf5) [14]
        ├── __init__.py
        ├── models.py
        ├── process_text.py
        ├── saved_model.py
        ├── scraper.py
        └── modely - nb.py, svm.py, snn.py, dt.py
```

Pričom licencia je typu MIT, *pyproject.toml* obshauje informácie o použitej platforme na zostavenie projektu, zvolený programovací jazyk, dodatočné informácie ako knižnice využité v projekte a ich verzie, popis projektu, rôzne odkazy a iné. Súbor *MANIFEST.in* dovoľuje vložiť do balíku dodatočné súbory, ktoré nie sú automaticky pridané (mimo skriptov končiacich príponou „.py“) – vďaka nemu je možné publikovať aj slovenský dataset, „stop words“ a uložené predtrénované modely. Popis inštalácie balíčku, jeho používanie či priame spustenie jeho stiahnutej verzie je popísané v dokumente *README.md* vo formáte markdown. Po úprave do tohoto rozloženia bolo spustené zostavenie balíka príkazom `python3 -m build`, ktorý vytvoril komprimované verzie jeho distribúcie vo formátoch **wheel** a **.tar.gz**. Následne bolo možné príkazom `twine upload dist/*` nahrať tieto súbory zo zložky *dist* na online index. Balík je momentálne dostupný pod menom FakeNewsClassifier a je možné ho nainštalovať cez príkazový riadok pomocou inštalátoru **pip** ako `pip install FakeNewsClassifier`, stiahnuť zdrojové súbory[15] alebo pridať do virtuálneho prostredia cez používaný editor.

[15] https://pypi.org/project/FakeNewsClassifier/

[15] Formát **.pk** je výstupom metódy dump pre uloženie „pipeline“ pre každý model. Formát **.hdf5** umožňuje uloženie váh neurónovej siete a jej štruktúry do jedného súboru.

Kapitola 5

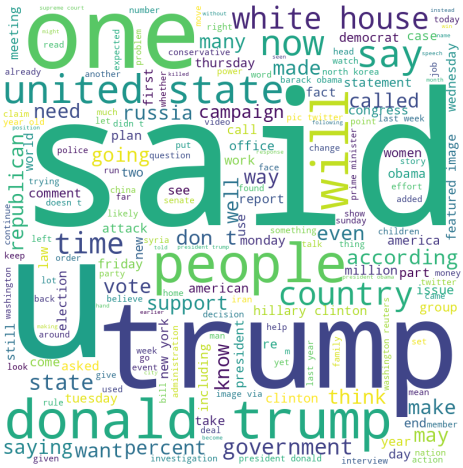
Testovanie a vyhodnocovanie

Nasledujúca kapitola popisuje trénovanie a testovanie vytvorených modelov na troch rozličných datasetoch. Každá podkapitola opisuje ich zloženie a obsah (vo forme slovnej mapy), postup práce, výsledky testovania, porovnanie s dostupnými dátami a výsledkami od iných používateľov (prevažne z využitím *BERT*, *LSTM* alebo iných moderných prístupov) a evalváciu dvoch online článkov spolu s ich referenčným hodnotením získaným z online klasifikátoru FakeNews.research.sfu.ca (jeho bližší popis sa nachádza v sekcii 2.2). Očakávané zaradenie je pri každom príspevku farebne označené.

5.1 Experiment 1 – Fake and Real News dataset

Pri tomto pokuse bol použitý dataset **Fake and real news dataset** od autora Clément Bisaillon [1]. Obshauje dva súbory typu .csv (Comma Separated Values) – ich štruktúra je načrtnutá v tabuľke 3.1. Celkovo balík obsahuje 38 729 unikátnych hodnôt rozdelených do dvoch súborov Fake.csv (17 903 hodnôt) a True.csv (20 826 hodnôt).

Na účely klasifikácie nám postačia stĺpce obsahujúce titulok a text správy. Takýto výcuc zo záznamov prešiel spracovaním textu. Jeho vizualizácia pomocou mapy slov bola nasledujúca:



Obr. 5.1: Slovná mapa prvého experimentu

[1] https://kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

V ďalšom kroku boli natrénované a otestované modely Multinomial Naive Bayes, Support Vectors Machine, Decision Tree a Sequential Neural Network (ďalej len **NB**, **SVM**, **DT**, **SNN**). V nasledujúcej tabuľke môžeme vidieť výstupy skórovacej funkcie:

Použitý model	**Precision (Fake/True)**	**Recall (Fake/True)**	**F1-score (Fake/True)**	**Accuracy**
NB	0.96/0.95	0.95/0.97	0.96/0.96	0.96
SVM	1.00/0.99	0.99/1.00	1.00/1.00	1.00
DT	0.98/0.99	0.99/0.99	0.99/0.99	0.99
SNN	0.98/1.00	1.00/0.98	0.99/0.99	0.99

Tabuľka 5.1: Skóre modelov v prvom experimente

Na základe nameraných hodnôt môžeme vyvodiť dva závery – buď sú naše modely nadpriemerne efektívne a dokážu klasifikovať vstupy s viac ako 96% úspešnosťou, alebo je tento dataset príliš jednoduchý, resp. obsahuje *bias* (dosl. zaujatosť). Po dôkladnom skúmaní dát a diskusií používateľov s podobnými presnosťami (dokopy 512 príspevkov, väčšinou s **BERT** a **LSTM** modelmi s úspešnosťami nad 95%) na tomto súbore sa táto hypotéza potvrdila – dáta sú aj napriek svojej rozsiahlosti príliš jednoduché na klasifikáciu a to hlavne kvôli dvom nedostatkom [9]:

- Štruktúra textu a tokenov – falošné správy obsahujú rádovo viac unikátnych slov (rôznych tokenov) než pravdivé správy, a to aj kvôli duplikácií niektorých záznamov .
- Relevantnosť slov – falošné správy obsahujú často značky používateľov Twitteru (kvôli majorite správ braných práve odtiaľ), a to hlavne účet Donalda Trumpa. Záznamy pravdivých správ obsahujú často názov spravodajskej služby Reuters.

Modely boli neskôr testované na článkoch *EU diverts road-building funds into £1.7bn Ukraine ammunition plan* (tabuľka 5.2) a *Was the Drone Attack on the Kremlin a False Flag Operation?* (tabuľka 5.3)

Použitý model	**Pravda**	**Lož**
NB	64.00%	36.00%
DT	95.00%	5.00%
SVM	60.88%	39.12%
SNN	59.25%	40.75%

Tabuľka 5.2: Evalvácia článku z portálu The Telegraph

Použitý model	**Pravda**	**Lož**
NB	21.81%	78.19%
DT	10.00%	90.00%
SVM	10.75%	89.25%
SNN	9.97%	90.03%

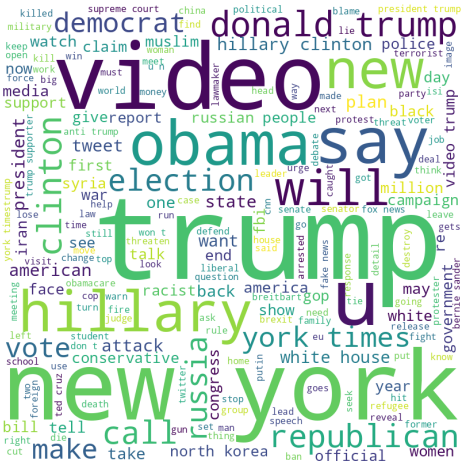
Tabuľka 5.3: Evalvácia článku z portálu Infowars.com

Online klasifikátor zaradil oba články do kategórie „False“ s hodnotou istoty 0.41 pre prvý článok a 0.29 pre druhý článok.

5.2 Experiment 2 – WELFake Dataset

Ďalší experiment bol prevedený na dátach zo súboru WELFake Dataset. Jedná sa o novší súbor, ktorý je často upravovaný a aktualizovaný rôznymi používateľmi. Popis zámeru a postupu jeho vytvorenia je dostupný v oficiálnom dokumente IEEE (Institute of Electrical and Electronics Engineers). Jedná sa o spojenie viacero dostupných datasetov do jedného celku o veľkosti 72 134 záznamov (35 028 reálnych a 37 106 falošných správ) [14].

Jeho štruktúra je podobná ako v predošlom experimente – na spracovanie sa taktiež použil titulok a obsah správy. Správy už obsahovali triedu pravdivosti ako jednu hodnotu záznamu (stĺpec 'label'), teda nebolo potrebné vykonávať spájanie rámcov. Po vykonaní spracovania záznamov vieme taktiež vizualizovať ich zloženie:



Obr. 5.2: Slovná mapa druhého experimentu

Po tréningu a testovaní boli výsledky modelov nasledujúce:

Použitý model	Precision (Fake/True)	Recall (Fake/True)	F1-score (Fake/True)	Accuracy
NB	0.93/0.86	0.86/0.93	0.90/0.89	0.90
SVM	0.97/0.98	0.98/0.97	0.97/0.98	0.98
DT	0.93/0.95	0.95/0.94	0.94/0.95	0.94
SNN	0.98/0.98	0.98/0.98	0.98/0.98	0.98

Tabuľka 5.4: Skóre modelov v druhom experimente.

Tieto hodnoty sú bližšie k očakávaným výsledkom modelov na základe všeobecnej úspešnosti, avšak modely stále nadobúdajú vysokú úspešnosť. V sekcí s kódmi ľudí používajúcich WELFake dataset nna stránke Kaggle.com [2] bola priemerná úspešnosť modelov niekde okolo 95% pre BERT modely a o čosi menej pre klasické modely. Najlepšie si viedol model podporných vektorov a model neurónovej siete.

Testovanie prebehlo na textoch s názvami *Pakistan school shooting leaves several teachers dead* (tabuľka 5.5) a *Obama officials 'met with Jeffrey Epstein,' unearthed docs reveal* (tabuľka 5.6).

[2] https://www.kaggle.com/code/anubhabkhanra/fake-news-classification, sekcia „Code“

Použitý model	Pravda	Lož
NB	70.00%	30.00%
DT	47.30%	52.70%
SVM	51.10%	48.90%
SNN	84.00%	16.00%

Tabuľka 5.5: Evalvácia článku z portálu BBC News.

Použitý model	Pravda	Lož
NB	45.81%	55.19%
DT	35.50%	64.50%
SVM	16.55%	83.45%
SNN	5.80%	94.20%

Tabuľka 5.6: Evalvácia článku z portálu WorldNetDaily.com.

Online klasifikátor zaradil oba články do kategórie „False“ s hodnotou istoty 0.20 pre prvý článok a 0.19 pre druhý článok.

5.3 Experiment 3 – Výber príspevkov z Twitter-u

Zvolený dataset je prevzatý z verejne zdieľaného repozitáru na stránke GitHub od používateľa KaiDDML [3], tiež [13]. Jedná sa o minimalistický výpis záznamov príspevkov z Twitter-u v štyroch súbory typu .csv. Rozdelené sú podľa zdroja a triedy – časť je zo stránok Politifact[4] a Gossipcop [5]

Jeho štruktúra je odlišná od predchádzajúcich datasetov, na naše účely budeme potrebovať záznamy „title“ a jednotlivé triedy musíme priradiť manuálne. Následne boli jednotlivé súbory pre zachovanie konzistencie a kompatibility spojené do jedného dátového rámca. [6]. Po týchto úpravách vieme súbory zlúčiť a pracovať s nimi ako s jedným rámcom podobným spôsobom ako v predošlých experimentoch. Vizualizácia záznamov:

Obr. 5.3: Slovná mapa tretieho experimentu.

Výsledky trénovania a následného testovania sú nasledovné:

Použitý model	Precision (Fake/True)	Recall (Fake/True)	F1-score (Fake/True)	Accuracy
NB	0.28/0.99	0.86/0.81	0.43/0.89	0.81
SVM	0.53/0.95	0.76/0.86	0.63/0.90	0.84
DT	0.53/0.86	0.56/0.85	0.54/0.86	0.78
SNN	0.63/0.89	0.66/0.88	0.64/0.88	0.82

Tabuľka 5.7: Skóre modelov v treťom experimente.

Môžeme vidieť, že najlepšie si opäť viedol model podporných vektorov. Tento krát je však možné všimnúť si veľký rozdiel medzi triedami, čo je čiastočne zapríčinené nerovnomerným rozložením hodnôt a taktiež kvôli zaujatosti vloženej do týchto ukážok, a to práve v záznamoch „true“ (politifact aj gossipcop), ktoré obsahujú viac textu než záznamy 'fake'. V porovnaní s výsledkami tejto zbierky dát uvedených priamo v publikácii [13] si autorove

[3] https://github.com/KaiDMML/FakeNewsNet/tree/master/dataset
[4] https://www.politifact.com/
[5] Stránka od septembra roku 2022 nie je viac dostupná.
[6] Tento úkon bol prevedený manuálne v externom skripte, ktorý nie je súčasťou výsledného projektu.

modely viedli horšie zhruba o 15%. Táto skutočnosť môže byť následkom či už nekompletnosťou dát (v publikácii bol použitý celý súbor príspevkov, nie len jeho časť) alebo už spomínaným nevhodným rozložením.

Online prípevky *Millions of Americans could suffer if debt showdown isn't solved in next 30 days* (tabuľka 5.8) a *Biden Troop Deployment to Border Won't Deter Illegal Migrants – Texas Governor* (tabuľka 5.9) dosiahli nasledujúce ohodnotenia:

Použitý model	**Pravda**	**Lož**
NB	44.70%	55.30%
DT	72.98%	27.02%
SVM	69.58%	30.42%
SNN	89.20%	11.80%

Tabuľka 5.8: Evalvácia článku z portálu CNN

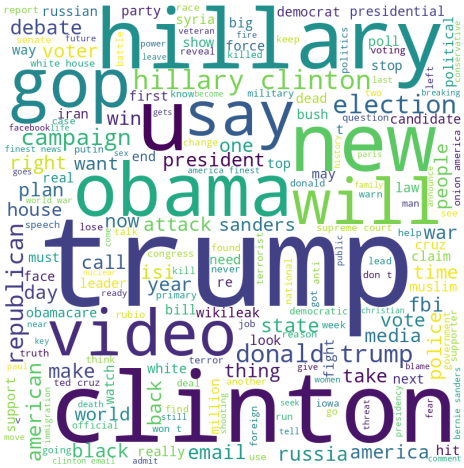
Použitý model	**Pravda**	**Lož**
NB	14.88%	85.12%
DT	20.00%	80.00%
SVM	15.05%	84.95%
SNN	2.77%	97.23%

Tabuľka 5.9: Evalvácia článku z portálu Infowars.com

Online klasifikátor zaradil článok CNN do kategórie „True“ s hodnotou istoty -0.08 a článok Infowars do kategórie „False“ s istotou 0.20.

5.4 Experiment 4 – Real&Fake News dataset

Posledný dataset je opäť zo stránky kaggle.com, tentokrát od používateľa *nop.ai* [7]. Obsahuje 6060 unikátnych záznamov, čo je celkovo menej než doposiaľ využívané súbory, avšak na rozdiel od nich má omnoho rovnomernejší pomer reálnych a falošných správ (50:50). Je štrukturovaný podobne ako väčšina datasetov z tejto oblasti – obsahuje index, titulok, text a označenie true / fake. Tak ako v druhom experimente na klasifikáciu využijeme stĺpce "label" a "text". Vizualizácia pred spracovaním:



Obr. 5.4: Slovná mapa štvrtého experimentu

Výsledky trénovania a následného testovania sú nasledovné:

Použitý model	Precision (Fake/True)	Recall (Fake/True)	F1-score (Fake/True)	Accuracy
NB	0.67/0.99	0.99/0.77	0.80/0.87	0.84
SVM	0.95/0.92	0.92/0.96	0.94/0.94	0.94
DT	0.82/0.84	0.84/0.82	0.83/0.83	0.83
SNN	0.95/0.95	0.95/0.95	0.95/0.95	0.95

Tabuľka 5.10: Skóre modelov v štvrtom experimente.

Model neurónovej siete opäť nadobudol najlepšie výsledky spomedzi všetkých modelov. Výsledky používateľov tohoto datasetu trénujúci modely transfotmátorov boli porovnateľné, s priemernou presnosťou okolo 95%. Následne bola prevedená klasifikácia online správ. Výsledky evalvácie článkov *Sudan Erupts in Chaos: Who Is Battling for Control and Why It Matters* (tabuľka 5.11) a *'Jaw dropping:' Now 2nd member of Congress warns of Biden 'prostitution'* (tabuľka 5.12):

[7] https://www.kaggle.com/datasets/nopdev/real-and-fake-news-dataset

Použitý model	Pravda	Lož
NB	78.00%	22.00%
DT	90.00%	10.00%
SVM	63.18%	36.19%
SNN	57.20%	42.80%

Tabuľka 5.11: Evalvácia článku z portálu New York Times.

Použitý model	Pravda	Lož
NB	71.81%	28.19%
DT	0.00%	100.00%
SVM	6.25%	93.75%
SNN	0.37%	99.63%

Tabuľka 5.12: Evalvácia článku z portálu WorldNetDaily.com

Vyslovene „mimo“ očakávanú hodnotu sa dostal len model Naive Bayes. Ďalšou prekvapivou hodnotou je aj 100% istota modelu rozhodovacieho stromu, čo však môže byť vysvetlené jeho nedostatočným rozvetvením v listových uzloch. Na základe výsledkov testovania nie je tento výsledok veľmi presvedčivý. Online klasifikátor zaradil článok New york Times do kategórie „True“ s hodnotou istoty -0.12 a článok WND do kategórie „False“ s istotou 0.29.

5.5 Experiment 5 – Dezinfo SK

Posledný experiment bol vykonaný na zbierke dát Dezinfo SK, ktorý bol vytvorený ako súčasť tejto baklárskej práce. Jeho štruktúra je rovnako štandardná, ako pri predošlých datasetoch. Obsahuje články z rôznych slovenských novinových portálov, ktoré boli manuálne zozbierané a ohodnotené. Viac informácii o postupe jeho vytvorenia a štruktúry je uvedených v prílohe Odborný článok – Dezinfo Dataset. Najčastejšie vyskytujúce sa slová môžeme vidieť na tomto obrázku:

Obr. 5.5: Slovná mapa piateho experimentu.

Modely natrénované a otestované na týchto dátach nadobudli nasledujúce hodnotenia:

Použitý model	Precision (Fake/True)	Recall (Fake/True)	F1-score (Fake/True)	Accuracy
NB	0.88/0.79	0.82/0.85	0.85/0.81	0.83
SVM	0.85/0.88	0.85/0.88	0.85/0.88	0.87
DT	0.86/0.81	0.80/0.87	0.83/0.84	0.83
SNN	0.94/0.85	0.89/0.92	0.91/0.88	0.90

Tabuľka 5.13: Skóre jednotlivých modelov v experimente.

Na základe týchto testov nevieme dopredu s istotou prehlásiť, že dáta sú bezchybné a nevedú k zaujatosti modelov, avšak výsledky naznačujú, že ich štruktúra a obsah je vhodne ucelený a modely sa na ňom trénujú adekvátne. Vzniknuté modely boli v ďalšom kroku boli s využitím uložených modelov z trénovacej fáze odskúšané priamo na článkoch *Kyjev kupuje od Varšavy ďalšiu techniku* (tabuľka 5.14) a *Démonom posadnutá umelá inteligencia sa učí ako použiť syntetickú biológiu, aby si vytvorila super ľudské telo* (tabuľka 5.15), ktoré sa nenachádzajú v pôvodnom datasete.

Použitý model	Pravda	Lož
NB	55.53%	44.47%
DT	68.75%	31.25%
SVM	86.09%	13.91%
SNN	76.30%	23.70%

Tabuľka 5.14: Evalvácia článku z portálu TA3.

Použitý model	Pravda	Lož
NB	36.61%	63.39%
DT	36.36%	63.64%
SVM	1.29%	98.71%
SNN	2.56%	97.44%

Tabuľka 5.15: Evalvácia článku z portálu Badatel.net.

Tieto výsledky korešpondujú s očakávaným zaradením týchto článkov. Ich hodnotenie však nie je možné porovnať s online nástrojom, ktorý bol trénovaný na sade anglických textov a cudzojazyčné výrazy by mohli byť interpretované ako chybné či zavádzajúce.

5.6 Návrh na rozšírenie tejto práce

V tejto časti sú popísané ďalšie návrhy na rozšírenie programovej časti bakalárskej práce a datasetu.

Programová časť

Ako pri každom projekte v oblasti informačných technológii, možnosti na úpravu a zlepšenie kódu je vždy možné nájsť. Keďže v práci bolo použité predspracovanie textu, existujú varianty pre jeho vylepšenie – úprava minimálneho či maximálneho počtu príznakov extrahovaných z textu, lepšia lematizácia a podobne. Tento krok priamo ovplyvňuje výslednú presnosť a výpočetnú náročnosť celého projektu. V ďalšom rade sa ponúka zväčšiť výber dostupných modelov strojového učenia. V dnešnej dobe existuje mnoho klasifikátorov založených či už to na matematických modeloch, neurónových sieťach či predtrénovaných transformátoroch, ako napríklad Passive Agressive Classifier[8], Metódy najbližších susedov [9], LSTM [10] alebo BERT modely [11]. Pre výskumné účely je taktiež možné vrámci jednotlivých krokov trénovania a vyhodnocovania ukladať či vypisovať priemerné a aktuálne presnosti modelov, odhadovaný výpočtový čas či iné štatistické údaje.

Dataset

Tvorba súboru dát implicitne poskytuje príležitosť jeho rozšírenia a zvýšenia kvality. Jeho momentálny rozsah poskytuje minimálnu funkcionalitu a svojou rozsiahlosťou nemôže konkurovať tým bežne využívaným. Napriek tomu poskytuje potrebný základ pre účely klasifikácie a eventuálnu automatizáciu pridávania ďalších článkov. Hlavným úskalím však naďalej zostáva potreba externej validácie priraďovania tried – ak má byť dataset relevantný, nepostačuje len jeho kvantita, ale aj kvalita.

[8] https://vitalflux.com/passive-aggressive-classifier-concepts-examples/
[9] https://scikit-learn.org/stable/modules/neighbors.html
[10] https://keras.io/api/layers/recurrent_layers/lstm/
[11] https://www.tensorflow.org/text/tutorials/classify_text_with_bert

Kapitola 6

Záver

Cieľom tejto práce bolo vytvoriť a poskytnúť začínajúcim programátorom jednoduchú knižnicu pre klasifikáciu novinových článkov na základe ich spoľahlivosti. Tento cieľ bol dosiahnutý – Balíček obsahujúci všetky nástroje je voľne dostupný a umožňuje jednoduchú prácu s rôznymi modelmi strojového učenia. Je možné ho nainštalovať priamo do virtuálneho prostredia pre jazyk **python** alebo stiahnuť jeho zdrojové kódy (viď. sekcia 4.3). Používateľovi stačí definovať zdroj dát na klasifikáciu v štandardnom formáte a model ktorý chce použiť. Celé trénovanie a predspracovanie dát prebieha na pozadí a nie je potrebné ho implementovať. Poskytnuté sú taktiež predtrénované modely vhodné na priamu evalváciu článkov z daného odkazu. Ďalším výstupom tejto práce je zbierka slovenských „Fake News“ spolu s pravdivými správami z online spravodajských portálov. Tento dataset je taktiež voľne prístupný a ponúka rozmanité uplatnenie nie len pre slovenský jazyk, ale taktiež pre všeobecné klasifikátory článkov v rôznych jazykoch. Hlavným zámerom jeho vytvorenia bolo to,že prístup k takýmto dátam v slovenčine je veľmi náročný – aj keď' existujú spoločnosti, ktoré sa klasifikácii textu venujú, ich zdroje sú často proprietárne a nedostupné. Vytvorenie takéhoto súboru dát bolo zložité hlavne z časového hľadiska, no vo výsledku je možné ho ľahko a intuitívne používať. Zvolené modely sa pri ich vytváraní preukázali ako dostatočne spoľahlivé a vierohodné pre túto činnosť aj napriek ich „zastaralosti“. Hlavným zistením tejto práce je celková závislosť presnosti vyhodnocovania na kvalite a kvantite vstupných dát – aj pri najmodernejších prístupoch sa zakladá práve na týchto aspektoch a aj tie najrozsiahlejšie transformátory potrebujú kvantá dát na ich natrénovanie. Avšak vďaka neustále platnému Moorovmu pravidlu [8] počet tranzistorov v procesoroch a s tým spojená výpočetná sila moderných počítačov neustále rastie, čo sa priamo odráža aj v odvetviach ako je spracovanie prirodzeného jazyka, kde sú stroje schopné efektívne spracovať enormné množstvá textu. Táto práca bola súčasťou študentskej konferencie inovácií, technológií a vedy v IT Excel@FIT2023.

Literatúra

[1] Asr, F. T. a Taboada, M. Big Data and quality data for fake news and misinformation detection. *Big Data & Society*. 1. vyd. 2019, zv. 6, č. 1. DOI: 10.1177/2053951719843310. Dostupné z: https://doi.org/10.1177/2053951719843310.

[2] Bishop, C. M. *Pattern recognition and machine learning*. 1. vyd. New York: Springer Science + Business Media, 2006. Information science and statistics. ISBN 978-1-4939-3843-8.

[3] Bíla, J. *Umělá inteligence a neuronové sítě v aplikacích*. 1. vyd. Vydavatelství ČVUT, 1996. 115 s. ISBN 8001012751.

[4] Cortes, C. a Vapnik, V. Support-vector networks. *Machine Learning*. 1. vyd. Sep 1995, zv. 20, č. 3, s. 273–297. DOI: 10.1007/BF00994018. ISSN 1573-0565. Dostupné z: https://doi.org/10.1007/BF00994018.

[5] Ding, B., Qian, H. a Zhou, J. Activation functions and their characteristics in deep neural networks. In: IEEE. *2018 Chinese Control And Decision Conference (CCDC)*. 2018, s. 1836–1841. DOI: 10.1109/CCDC.2018.8407425. ISBN 978-1-5386-1244-6.

[6] Horev, R. *BERT Explained: State of the art language model for NLP* [online]. 2018. Dostupné z: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

[7] Ing. Lukáš Burget Ph.D. doc. *Klasifikace a rozpoznávání – Umělé neuronové sítě a Support Vector Machines*. 2020. Prdnášky k magisterskému predmetu SUR. Dostupné z: https://www.fit.vutbr.cz/study/courses/SUR/public/prednasky/05_neral_networks/.en.

[8] Moore, G. E. Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. *IEEE Solid-State Circuits Society Newsletter*. 1. vyd. 2006, zv. 11, č. 3, s. 33–35. DOI: 10.1109/N-SSC.2006.4785860.

[9] Nascimento, J. *Only one word 99.2%* [online]. 2020. Dostupné z: https://www.kaggle.com/code/josutk/only-one-word-99-2/notebook.

[10] PyPA. *Including files in source distributions with MANIFEST.in* [online]. 2023. Dostupné z: https://packaging.python.org/en/latest/guides/using-manifest-in/.

[11] PyPA. *Packaging and distributing projects* [online]. 2023. Dostupné z: https://packaging.python.org/en/latest/guides/distributing-packages-using-setuptools/#packaging-your-project.

[12] Ribeiro, M. T., Singh, S. a Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Association for Computing Machinery. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: [b.n.], 2016, s. 1135–1144. KDD '16. DOI: 10.1145/2939672.2939778. ISBN 9781450342322. Dostupné z: https://doi.org/10.1145/2939672.2939778.

[13] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. a Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *ArXiv preprint arXiv:1809.01286*. 1. vyd. 2018, č. 1. Dostupné z: https://arxiv.org/abs/1809.01286.

[14] Verma, P. K., Agrawal, P., Amorim, I. a Prodan, R. WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems*. 1. vyd. 2021, zv. 8, č. 4, s. 881–893. DOI: 10.1109/TCSS.2021.3068519. Dostupné z: https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification.

[15] Vishwakarma, D. K. a Jain, C. Recent State-of-the-art of Fake News Detection: A Review. In: INCET. *2020 International Conference for Emerging Technology (INCET)*. 2020, s. 1–6. DOI: 10.1109/INCET49848.2020.9153985. ISBN 978-1-7281-6221-8. Dostupné z: https://ieeexplore.ieee.org/abstract/document/9153985.

[16] Vosoughi, S., Roy, D. a Aral, S. The spread of true and false news online. *Science*. 1. vyd. 2018, zv. 359, č. 6380, s. 1146–1151. DOI: 10.1126/science.aap9559. Dostupné z: https://www.science.org/doi/abs/10.1126/science.aap9559.

[17] Wikipedia contributors. *Activation function — Wikipedia, The Free Encyclopedia*. 2023. [Online; accessed 6-May-2023]. Dostupné z: https://en.wikipedia.org/w/index.php?title=Activation_function&oldid=1151093352.

[18] Wikipedia contributors. *List of fake news websites — Wikipedia, The Free Encyclopedia*. 2023. [Online; accessed 19-April-2023]. Dostupné z: https://en.wikipedia.org/w/index.php?title=List_of_fake_news_websites&oldid=1149631303.

[19] Wikipedie. *Seznam dezinformačních webů v češtině — Wikipedie: Otevřená encyklopedie*. 2023. [Online; navštíveno 19. 04. 2023]. Dostupné z: https://cs.wikipedia.org/w/index.php?title=Seznam_dezinforma%C4%8Dn%C3%ADch_web%C5%AF_v_%C4%8De%C5%A1tin%C4%9B&oldid=22705275.

[20] Zbořil, F. V. *Základy umělé inteligence – studijní opora*. 2023. Sekcia 5.1.1 – Rozhodovacie stromy. Dostupné z: https://www.vut.cz/www_base/priloha_fs.php?dpid=240926&skupina=dokument_priloha.

[21] Šnorek, M. *Neuronové sítě a neuropočítače*. Vyd. 1. Praha: ČVUT, 1996. ISBN 80-01-01455-X.

Príloha A

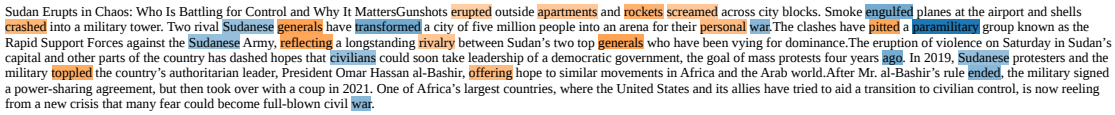
Obsah priloženého pamäťového média

- xkoren10.zip/
  - doc/
    - Zdrojové súbory pre vytvorenie technickej správy
    - xkoren10_Detection_of_Fake_News.pdf ..... Technická správa
  - code/
    - Súbory potrebné pre vytvorenie balíčku
    - src/
      - main.py ..... Hlavný skript
      - FakeNewsClassifier/
        - datasets/
          - dezinfo_sk.csv ..... Slovenský dataset
        - saved_models/
          - Predtrénované modely
        - Zdrojové súbory s kódom

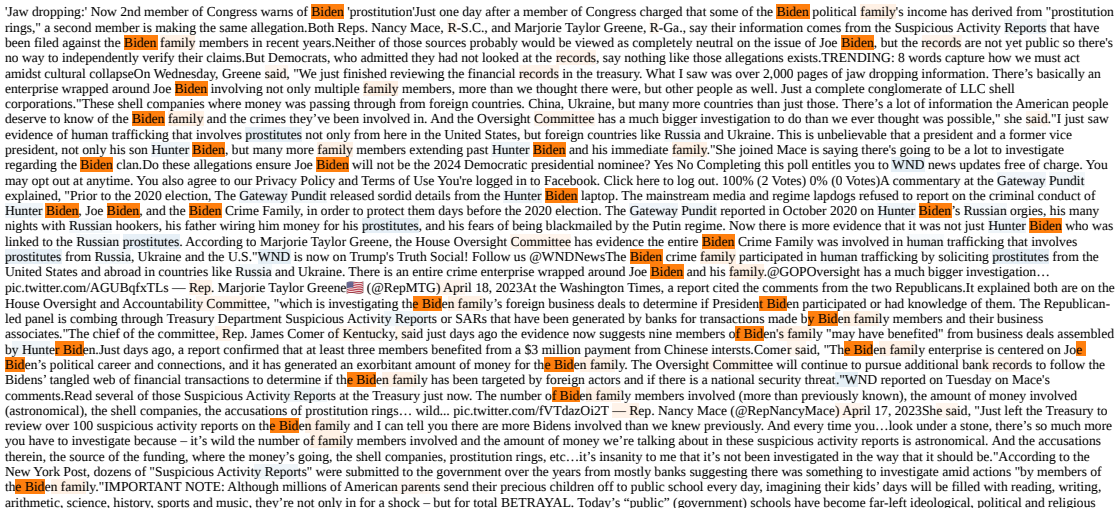
Príloha B

Ukážky analýz vyhodnotení

Priložené obrázky sú automaticky vygenerované html stránky po využití knižnice **lime**. Pre sprehľadnenie boli vybrané časti s farebne odlíšeným textom, kde odtiene modrej určujú záporné ohodnotenie a odtiene oranžovej kladné ohodnotenie, respektíve náležitosť do „Fake“ a „Real“ tried.



Obr. B.1: Analýza článku New York Times zo štvrtého experimentu vyhodnoteného modelom SNN.



Obr. B.2: Analýza článku World Net Daily zo štvrtého experimentu vyhodnoteného modelom NB.

Príloha C

Odborný článok – Dezinfo Dataset

Táto príloha obsahuje odborný článok, obsahujúci bližší popis k datasetu Dezinfo SK, ktorý bol použitý pri trénovaní a testovaní jednotlivých modelov. Obsahuje popis práce pri jeho vytváraní, jeho štruktúru a výsledky.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

http://fit.vut.cz

Odborný článok
2022/2023

Dezinfo Dataset

Matej Koreň*

Abstrakt

Dataset **Dezinfo SK - Fake News Dataset** bol vytvorený ako súčasť bakalárskej práce na tému **Detekcia falošných správ pomocou strojového učenia**. Zámienkou na jeho vytvorenie bol nedostatok verejne prístupných dát pre klasifikáciu textu v slovenskom jazyku. Taktiež slúži ako demonštrácia robustnosti modelov využívaných na tento úkon v zmysle možnosti využitia iných jazykov. Dataset svojou rozsiahlosťou nekonkuruje tým anglickým, avšak bol vytvorený manuálne so záujmom zachovania objektivity dát a taktiež ich vecnosti. Je možné ho voľne používať a trénovať na ňom vlastné modely.

Klíčová slova: Dataset — Strojové učenie — Klasifikácia textu

Přiložené materiály: Stiahnuteľný Kód

*xkoren10@fit.vut.cz, *Fakulta informačních technologií, Vysoké učení technické v Brně*

1. Úvod

[Motivácia] Existuje množstvo verejne dostupných datasetov na klasifikáciu textu, presnejšie na detekciu falošných správ, avšak majorita z nich je v anglickom jazyku. Prístup k takýmto dátam v slovenskom jazyku je veľmi náročný - aj keď existujú spoločnosti, ktoré sa klasifikácii textu venujú, ich zdroje sú často proprietárne a nedostupné. Cieľom tejto práce je teda vytvoriť voľne dostupný súbor dát pre túto špecifickú úlohu, ktorý je možné ľahko a jednoducho používať.

[Riešenie] Dataset (ku dňu 10.5.2023) obsahuje 100 unikátnych hodnôt, respektíve článkov zozbieraných manuálne. Každá hodnota sa skladá z :

- jednoznačného identifikátora - prirodzené čísla
- titulku článku - nadpis (bez podnadpisu)
- textu článku - plné znenie článku, bez odkazov,obrázkov či reklám
- zdrojového odkazu - hyertextové prepojenie
- odkazu na overenie článku - hypertextové prepojenie obsahujúce podklady k overeniu
- triedy zaradenia - dvojica tried reprezentujúca pavdivosť (1) a nepravdivosť (0)

Súbor má formát CSV, ktorý je pre tento druh dát štandardom pre jeho jednoduché spracovanie. Zber dát bol uskutočňovaný v období začiatku roka 2023 (marec). Je využitý v bakalárskej práci — , v ktorej figuruje ako príloha a zároveň na ňom pri jej vypracovávaní boli uskutočňované experimenty s rôznymi modelmi strojového učenia pre klasifikáciu a spracovanie prirodzeného jazyku.

2. Postup vytvorenia

[Zber dát] Dáta boli zozbieravané manuálne, čo je relatívne neštandardný postup pri vytváraní väčšieho množstva dát. K tejto metóde bolo pristúpené na základe toho, že na patričné zaraďovanie článkov do tried vierohodnosti by bolo nutné pri automatickom zbere použiť model, ktorý by bol predom natrénovaný na túto úlohu práve na datasete toho druhu, čo vytvára cyklickú závislosť. Ako možnosť sa taktiež ponúka využitie automatických nástrojov na získavanie obsahu z webu, ako sú rôzne web crawlery či scrapery. Tieto však články z vopred zvolených stránok vyberajú bez rozdielu na tému a obsah, čo by viedlo k nekonzistentnosti dát. Použitie takýchto nástrojov spolu s kľúčovými slovami by opäť mohlo vo výsledku viesť k takzvanej "zaujatosti" (anglicky "bias") týchto dát, nakoľko by súbor obsahoval len články týkajúce sa konkrétnych tém. Ďalším problémom je tak-

tiež ochrana niektorých stránok proti vírusom či iným škodlivým útokom, ktorá čiastočne znemožňuje prácu automatických zberačov dát.

2.1 Témy článkov

Témy jednotlivých článkov boli vyberané na základe ich relevantnosti voči momentálnej situácii vo svete, aby sa udržala celková konzistencia datasetu. Medzi ne patrí (v čase publikácie) napríklad vojna na ukrajine, Vladimír Putin, Donald Trump, Joe Biden, Čína, Koronavírus ale aj vyjadrenia politikov z rozličných krajín a iné.

[Priradzovanie tried]

Klasifikácia dát bola uskutočnená taktiež manuálne. Podkladom k hodnoteniu daných článkov a ich zaradenie do skupiny pravdivých či nepravdivých správ je založené na skóre dôveryhodnosti jednotlivých portálov, na ktorých boli dané články uverejnené. Toto skórovanie je verejne dostupné na stránkach konspiratori.sk, kde komisia žurnalistov a odborníkov udelila hodnotenie na škále (od 0 do 10, kde 10 je naviac zavádzajúce) viac ako 300 spravodajským portálom. Hodnotenia zdrojov použitých v datasete na základe tejto štúdie [1] je nasledovné:

- badatel.net - 9.6
- zemavek.sk - 9.5
- slobodnyvysielac.sk - 9.2

pričom portály ako Aktuality.sk, SME.sk, Pravda.sk či TA3.sk majú celkové hodnotenie nižšie ako 4 a sú na zozname overených stránok [2].

3. Výsledky testovania

Vzhľadom na to, že tento set bol vytvorený ako súčasť bakalárskej práce zaoberajúcej sa klasifikáciou textu s využitím modelov strojového učenia, boli tieto dáta použité ako trénovacia a testovacia sada pre modely Naive Bayes, Decision Tree, SVM a neurónovú sieť. Výsledky vyhodnotené skórovacou metódou F1-score môžeme vidieť v nasledujúcej tabuľke:

Použitý model	Precision (Fake/True)	Recall (Fake/True)	F1-score (Fake/True)	Accuracy
NB	0.88/0.79	0.82/0.85	0.85/0.81	0.83
SVM	0.85/0.88	0.85/0.88	0.85/0.88	0.87
DT	0.86/0.81	0.80/0.87	0.83/0.84	0.83
SNN	0.94/0.85	0.89/0.92	0.91/0.88	0.90

Tabulka 1. Skóre jednotlivých modelov v experimente

Na základe týchto testov nevieme dopredu s istotou prehlásiť, že dáta sú bezchybné a nevedú k zaujatosti modelov, avšak výsledky naznačujú, že ich štruktúra a obsah je vhodne ucelený a modely sa na ňom trénujú adekvátne. Vzniknuté modely boli v ďalšom kroku boli odskúšané na článkoch z internetu mimo tých, ktoré boli v pôvodnom datasete:

Used model	True	False
NB	55.53%	44.47%
DT	68.75%	31.25%
SVM	86.09%	13.91%
SNN	76.30%	23.70%

Tabulka 2. Evaluation of an article from TA3

Used model	True	False
NB	36.61%	63.39%
DT	36.36%	63.64%
SVM	1.29%	98.71%
SNN	2.56%	97.44%

Tabulka 3. Evaluation of an article from Badatel.net

Tieto výsledky korešpondujú s očakávaným zaradením týchto článkov. Pre referenciu, ChatGPT 3.5 so svojim modelom BERT klasifikoval prvý článok ako pravdivý (90%) a druhý ako nepravdivý (80%).

4. Záver

[Zhrnutie] Dataset Dezinfo SK je v počiatočnom štádiu nasadenia. Najviac obmedzujúcim faktorom jeho využitia je jeho veľkosť - aby dokázal modelom pridať na relevantnosti, musí byť pravidelne aktualizovaný kvôli neustále sa meniacim témam a trendom vo svete. Jeho manuálna tvorba je najväčším úskalím, avšak jej výhody prevyšujú nevýhody.

Získavanie ďalších dát je taktiež vďaka voľnej dostupnosti umožnené aj jeho iným používateľom, prípadne v ich záujme je možná ďalšia spolupráca na jeho rozvoji.

Poďakovanie

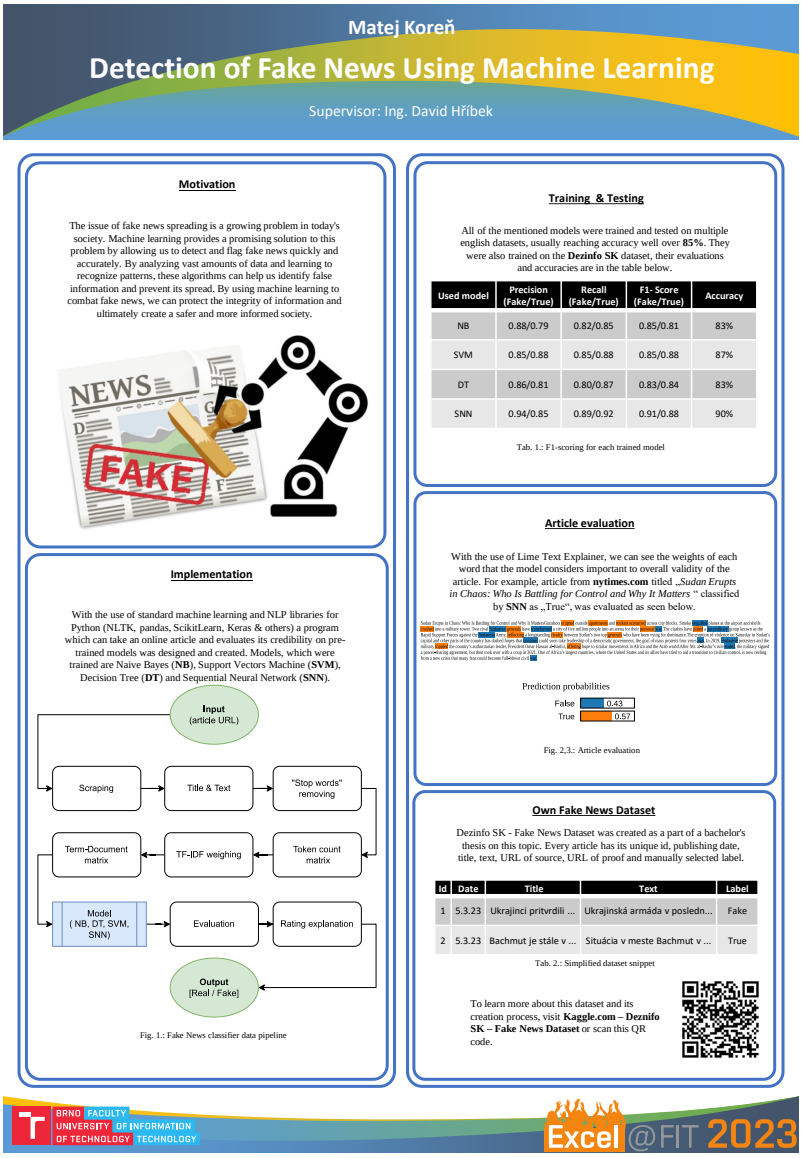
Rád by som poďakoval svojmu vedúcemu Davidovi Hříbkovi za jeho pomoc pri tvorbe tohoto datasetu a taktiež aj baklárskej práce.

[1] https://konspiratori.sk/zoznam-stranok

[2] https://konspiratori.sk/zoznam-stranok-s-overenym-obsahom

Príloha D

Plagát



Obr. D.1: Plagát práce na konferencií Excel@FIT 2023