Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-100241-110868

Daniela Paľuchová

Kto to napísal?

Bakalárska práca

Vedúci práce: Ing. Petrík Juraj

Máj 2024

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-100241-110868

Daniela Paľuchová

Kto to napísal?

Bakalárska práca

Študijný program: informatika
Študijný odbor: informatika
Miesto vypracovania: Ústav počítačového inžinierstva a aplikovanej informatiky, FIIT STU Bratislava
Vedúci práce: Ing. Petrík Juraj

Máj 2024

Slovenská technická univerzita v Bratislave
Ústav informatiky, informačných systémov a softvérového inžinierstva

Fakulta informatiky a informačných technológií
Akademický rok: 2023/2024
Evidenčné číslo: FIIT-100241-110868

ZADANIE BAKALÁRSKEJ PRÁCE

Študentka: **Daniela Paľuchová**
ID študenta: 110868
Študijný program: informatika
Študijný odbor: informatika
Vedúci práce: Ing. Juraj Petrík
Vedúci pracoviska:

Názov práce: **Kto to napísal?**

Jazyk, v ktorom sa práca vypracuje: slovenský jazyk

Špecifikácia zadania:

S príchodom éry internetu je čoraz ťažšie rozpoznať, s kým naozaj vzdialene komunikujeme, kto je autorom populárnych článkov alebo statusov. Alebo môžeme uvažovať o tom, či sa vôbec „na druhej strane“ nachádza človek alebo stroj. Ide napríklad o fake news, ktoré v minulosti pravdepodobne ovplyvnili napríklad voľby v USA. Tieto fake news boli šírené automatizovane napríklad Twitter botmi. Na profilovanie autora a teda aj odhalenia, či ide o stroj alebo človeka, sú často využívané štatistické metódy a metódy strojového učenia. Pri profilovaní je možné s vysokou presnosťou určiť pohlavie autora, jeho záujmy a záľuby, profesijnú oblasť a dokonca v niektorých prípadoch aj jeho vek. Cieľom práce je navrhnúť, implementovať a overiť metódu na profilovanie autora. Zamerajte sa na profilovania autora vo vybranej oblasti (text, zdrojový kód, obrázky, atď.).

Rozsah práce: 40

Termín odovzdania bakalárskej práce: 21. 05. 2024
Dátum schválenia zadania bakalárskej práce: 18. 04. 2024
Zadanie bakalárskej práce schválil: prof. Ing. Valentino Vranić, PhD. – garant študijného programu

Čestné prehlásenie

Čestne vyhlasujem, že som túto prácu vypracovala samostatne, na základe konzultácií a s použitím uvedenej literatúry.

V Bratislave, 21.05.2024 .....

Daniela Paľuchová

Poďakovanie

Vyjadrujem úprimnú vďaku svojmu vedúcemu práce, Ing. Jurajovi Petríkovi, za jeho odborné vedenie, cenné návrhy a rady, ktoré boli pre mňa neoceniteľné pri vypracovaní tejto bakalárskej práce. Rovnako by som chcela poďakovať svojej rodine a najmä partnerovi za neustálu podporu a povzbudenie, ktoré mi poskytovali počas celého procesu tvorby tejto práce.

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: informatika

Autor: Daniela Paľuchová
Bakalárska práca: Kto to napísal?
Vedúci bakalárskej práce: Ing. Petrík Juraj
Máj 2024

V ére internetu sa objavujú zásadné výzvy, ako je rozpoznávanie skutočných autorov online interakcií, identifikácia tvorcov populárnych článkov alebo príspevkov na sociálnych sieťach a rozlíšenie, či na druhej strane komunikácie je človek alebo umelá inteligencia. Táto bakalárska práca sa zaoberá týmito otázkami, sústrediac sa na komplexnú problematiku profilovania autorstva v rýchlo sa meniacej digitálnej sfére.

V práci sa najprv venujeme podrobnej analýze stylometrie, disciplíny zameranej na štúdium a analýzu písomného štýlu. Okrem toho sa práca sústreďuje na využitie metód strojového učenia v stylometrii. Zvláštny dôraz kladieme na to, ako úpravy a spracovanie dát ovplyvňujú presnosť profilovania. Cieľom tejto analýzy je identifikovať optimálne metódy a stratégie predspracovania dát a metódy strojového učenia, ktoré zvyšujú presnosť určenia veku a pohlavia autorov.

V rámci tejto práce sú predstavené rozšírené teoretické poznatky v oblasti stylometrie a vývoj vlastnej implementácie strojového učenia. Tento prístup je zameraný na profilovanie autorstva s cieľom presnejšieho určenia demografických charakteristík autorov. Výsledkom je prehĺbenie pochopenia dynamiky digitálnej komunikácie a príspevok k rozvoju nástrojov na profiláciu autorstva v digitálnej ére.

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Informatics

Author: Daniela Paľuchová
Bachelor's Thesis: Who wrote it?
Supervisor: Ing. Petrík Juraj
2024, May

In the era of the internet, significant challenges emerge, such as recognizing the real authors of online interactions, identifying the creators of popular articles or posts on social networks, and discerning whether the entity on the other side of the communication is a human or artificial intelligence. This bachelor's thesis addresses these issues, focusing on the comprehensive problem of authorship profiling within the rapidly changing digital environment.

In this work, we initially engage in a detailed analysis of stylometry, a discipline focused on the study and analysis of written style. Moreover, the thesis concentrates on the application of machine learning methods in stylometry. Special emphasis is placed on how modifications and processing of data affect the accuracy of profiling. The objective of this analysis is to identify optimal methods and strategies for data preprocessing and machine learning techniques that enhance the accuracy of determining age and gender.

Within this thesis, an expansion of theoretical knowledge in the field of stylometry and the development of a proprietary machine learning implementation are presented. This approach is aimed at profiling authorship with the aim of more precisely determining the demographic characteristics of authors. The result is an enhanced understanding of the dynamics of digital communication and a contribution to the development of tools for authorship profiling in the digital era.

Obsah

Zoznam obrázkov

Zoznam tabuliek

Zoznam skratiek

POS	Označovanie slovných druhov (angl. Part-of-Speech)
TP	Skutočne pozitívne (angl. True Positives)
TN	Skutočne negatívne (angl. True Negatives)
FP	Falošne pozitívne (angl. False Positives)
FN	Falošne negatívne (angl. False Negatives)
GBDT	Gradientovo posilnené rozhodovacie stromy (angl. Gradient Boosted Decision Trees)
BERT	Obojsmerné kódovanie reprezentácií z transformátorov (angl. Bidirectional Encoder Representations from Transformers)
NLP	Spracovanie prirodzeného jazyka (angl. Natural Language Processing)
TF-IDF	Frekvencia termínov a inverzná frekvencia dokumentov (angl. Term Frequency-Inverse Document Frequency)
SVM	Podporný vektorový stroj (angl. Support Vector Machine)
ML	Strojové učenie (angl. Machine Learning)
TF	Frekvencia termínov (angl. Term Frequency)
IDF	Inverzná frekvencia dokumentov (angl. Inverse Document Frequency)
URL	Uniformný zdrojový identifikátor (angl. Uniform Resource Locator)
AI	Umelá inteligencia (angl. Artificial Intelligence)
GBM	Gradientové posilnenie (angl. Gradient Boosting Machine)
Unicode	Unicode znakové kódovanie (angl. Unicode Character Encoding)
nltk	Nástrojová sada pre spracovanie prirodzeného jazyka (angl. Natural Language Toolkit)

Kapitola 1

Úvod

V ére internetu, kedy sa komunikácia a zdieľanie informácií deje väčšinou online, rozpoznávanie a overovanie pravých autorov digitálnych obsahov predstavuje významnú výzvu. Napriek pokrokom v oblasti informačných technológií zostáva otázka autorstva často nejasná, či už ide o identifikáciu písomných prác, multimediálny obsah, alebo softvérový kód. S rozšírením dezinformácií a falošných správ, ako to bolo vidieť počas volieb v USA, sa potreba efektívnych metód na rozpoznanie autenticity a pôvodu digitálnych textov ešte viac zvýšila. Táto bakalárska práca sa zaoberá vývojom a overením metód profilovania autorstva s využitím nástrojov strojového učenia, s cieľom zvýšiť presnosť rozpoznávania autorov textových dokumentov [1].

Prvotným cieľom tejto práce je analyzovať existujúce metódy stylometrie – disciplíny, ktorá sa venuje analýze písomného štýlu a prvkov textu. Zameriavame sa na rozpoznávanie a profilovanie autorov na základe štatistických metód a metód strojového učenia, ktoré umožňujú identifikovať osobné atribúty autora, ako sú pohlavie, vek, alebo vzdelanie. Táto práca predstavuje vlastný experimentálny model, ktorý je schopný identifikovať a profilovať autora s vysokou presnosťou vo vybranej oblasti.

Práca je štrukturovaná do niekoľkých hlavných kapitol, ktoré pokrývajú teoretický základ stylometrie, popis použitých metód a technológií, vlastný vývoj metódy na profilovanie autorstva, a detailnú analýzu výsledkov experimentov. Každá časť je navrhnutá tak, aby poskytovala postupné pochopenie problému a jeho riešení, čím sa zabezpečuje komplexný prehľad o téme a jej vedecký prínos v oblasti spracovania

prirodzeného jazyka a aplikovaných informačných technológií.

Bakalárska práca je rozdelená do nasledujúcich kapitol:

- **Kapitola 1 - Úvod:** Táto kapitola uvádza čitateľa do problematiky profilovania autorstva, zdôrazňuje význam a aktuálnosť témy, a definuje ciele bakalárskej práce.
- **Kapitola 2 - Stylometria:** Poskytuje teoretický základ disciplíny, vysvetľuje kľúčové pojmy a metódy vhodné na analýzu textu.
- **Kapitola 3 - Stylometrické črty:** Zameriava sa na detailný popis rôznych charakteristík textu, ktoré umožňujú identifikáciu a analýzu autorových štýlov.
- **Kapitola 4 - Spracovanie textu:** Popisuje techniky a metódy potrebné na spracovanie textu pred jeho analýzou, vrátane tokenizácie, stemmingu a lemmatizácie.
- **Kapitola 5 - Strojové učenie:** Predstavuje rôzne prístupy a modely strojového učenia, ktoré sú aplikované na klasifikáciu a predikciu vlastností autorov textov.
- **Kapitola 6 - Existujúce riešenia a štúdie:** Poskytuje prehľad o existujúcich štúdiách a riešeniach v oblasti profilovania autorstva, čím zdôrazňuje význam ďalšieho výskumu.
- **Kapitola 7 - Ciele práce:** Definuje konkrétne vedecké a praktické ciele, ktoré táto práce sleduje.
- **Kapitola 8 - Dataset:** Podrobne opisuje zdroje dát, proces ich prípravy a čistenia, ktoré sú nevyhnutné pre analýzu a profilovanie autorov.
- **Kapitola 9 - Extrakcia čŕt:** Venuje sa metódam a technikám extrakcie relevantných čŕt z datasetu, ktoré sú následne použité v modeloch strojového učenia.
- **Kapitola 10 - Predpovedanie demografických vlastností:** Opisuje postupy a výsledky použitia strojového učenia na predpovedanie veku a pohlavia autorov na základe ich textov.
- **Kapitola 11 - Hypotézy:** Diskutuje o formulovaných hypotézach a metódach ich overovania.

- **Kapitola 12 - Porovnanie s existujúcimi riešeniami:** Analyzuje a porovnáva efektívnosť navrhnutých metód s inými štúdiami, ktoré používajú podobné datasety.
- **Kapitola 13 - Záver:** Sumarizuje dosiahnuté výsledky, diskutuje o splnení cieľov a naznačuje možné smerovania budúceho výskumu v oblasti profilovania autorstva.

Kapitola 2

Stylometria

V dnešnom svete plnom rôznorodých textov, umelo generovaných obsahov a náročných kriminalistických výziev hrá stylometria významnú úlohu. Stylometria je veda, ktorá analyzuje štýl písania v literárnych dielach. Využíva štatistické a matematické metódy na analýzu textov s cieľom identifikovať autora a jeho charakteristické rysy.

Táto disciplína je založená na dvoch základných predpokladoch. Prvým je unikátnosť autorského štýlu, kde každý autor sa odlišuje napríklad v slovnej zásobe, štruktúre viet a používaní interpunkcie, vrátane špecifického využívania bodkočiarok, pomlčiek a ďalších znakov. Tieto aspekty sú podrobne rozoberané v nasledujúcich kapitolách. Druhým predpokladom je časová blízkosť diel, kde texty napísané v kratšom časovom rozpätí majú tendenciu byť si štýlovo podobnejšie v porovnaní s tými, ktoré vznikli v rôznych časových obdobiach. Je potrebné uvedomiť si, že autorov štýl, ako aj jeho myšlienkové a emocionálne postoje sa môžu časom meniť.

Stylometria nachádza široké uplatnenie najmä pri určovaní autorstva textov v prípadoch, keď je autor neznámy alebo sporný. Toto využitie je relevantné nielen v akademickej sfére, ale aj v právnom kontexte, kde môže slúžiť ako nástroj na identifikáciu písma alebo odhalenie plagiátu [2][3].

2.1. Rozdelenie stylometrie

Stylometria, ako interdisciplinárna veda, sa rozvíja v rôznych smeroch a nachádza široké uplatnenie v akademických, právnych a kriminalistických oblastiach. Jej rozmanitosť spočíva v schopnosti analyzovať a interpretovať texty prostredníctvom rôznych štylistických prvkov a metód. Aby sme lepšie porozumeli tejto komplexnej disciplíne, je užitočné rozčleniť ju do niekoľkých základných odvetví, z ktorých každé sa zameriava na špecifické aspekty a aplikácie v rámci stylometrie [4]. V nasledujúcom texte predstavíme päť hlavných odvetví stylometrie:

1. **Priznanie autorstva**
2. **Overovanie autorstva**
3. **Profilovanie autorstva**
4. **Stylochronometria**
5. **Kontrodiktorna stylometria**

Toto rozčlenenie poskytuje prehľad o rôznych aplikáciách stylometrie a ukazuje, ako môže táto disciplína prispieť k lepšiemu pochopeniu a analýze textov z rôznych perspektív.

2.1.1 Priznanie autorstva

Priznávanie autorstva je proces, v ktorom sa pomocou rôznych analytických metód snažíme určiť, s akou pravdepodobnosťou je určité dielo dielom konkrétneho autora. V tomto procese sa zameriavame predovšetkým na štýlové prvky, ako sú gramatické štruktúry, písací štýl a ďalšie jedinečné charakteristiky diela, pričom obsah diela nie je v centre našej pozornosti. Tento prístup je obzvlášť užitočný v prípadoch, kde je autorstvo diela anonymné alebo sporné.

Existuje niekoľko techník priznávania autorstva. Jednou z nich je metóda “closed-set attribution”, ktorá je vhodná, ak máme k dispozícii zoznam potenciálnych autorov. V tejto metóde analyzujeme a porovnávame štýly týchto autorov so štýlom skúmaného

diela. Ďalšou metódou je “open-set classification”, ktorá umožňuje možnosť neznámeho autora - autor daného diela nemusí byť v našom zozname. Táto metóda využíva prahovú hodnotu podobnosti na určenie, či sa štýl diela dostatočne podobá štýlu niektorého z autorov v našom zozname. Metóda “K-attribution”poskytuje zoznam potenciálnych autorov v poradí, kde prvý autor má najvyššiu pravdepodobnosť byť autorom diela. “Cross-domain Authorship Attribution”je metóda určená na identifikáciu autorov naprieč rôznymi žánrami a oblasťami, čo môže byť nápomocné napríklad pri určovaní, či je autor knihy tiež autorom príspevku na sociálnych sieťach. Špeciálnou technikou je “source code authorship attribution”, ktorej cieľom je identifikovať autora zdrojového kódu. Táto metóda je obzvlášť užitočná pri vyšetrovaní prípadov softvérového plagiátu, identifikácii vírusov alebo pri analýze kybernetických útokov. [4].

2.1.2 Overovanie autorstva

Overovanie autorstva je proces zamierený na potvrdenie alebo vyvrátenie autorstva konkrétneho diela. Jeho kľúčovým cieľom je analyzovať lingvistické vzorce z rôznych textov a určiť, či pochádzajú od toho istého autora.

Tradičná analýza sa vykonáva skúsenými lingvistami, ktorí hodnotia a porovnávajú rôzne jazykové aspekty, vrátane pravopisných chýb, gramatických nezrovnalostí a charakteristických štylistických prvkov textu.

S rozvojom technológií sa využívajú algoritmy strojového učenia na analýzu autorstva. Môžeme vyvinúť modely, ktoré využívajú stylometrické prvky na dosiahnutie podobných výsledkov. Avšak pri analýze rôznorodých alebo krátkych textov môže spoľahlivosť modelu klesať. Jednou z metód strojového učenia je klasifikácia jednej triedy, kde model je natrénovaný na dielach známeho autora a potom sa použije na overenie autorstva testovaného dokumentu.

Inou metódou je Many-Candidates, ktorá spočíva vo vytvorení skupiny ”podvodníkov”(falošných autorov) a porovnávaní skúmaného textu s touto skupinou. Cieľom je zistiť, či sa štýl skúmaného diela viac podobá štýlu falošných autorov alebo skutočnému autorovi. Ak je text bližší štýlu predpokladaného autora, existuje

väčšia pravdepodobnosť, že je jeho dielom; ak je bližší štýlu "podvodníkov", je menej pravdepodobné, že dané dielo napísal náš predpokladaný autor [4][5].

2.1.3 Profilovanie autorstva

Profilovanie autorstva je proces identifikácie charakteristík autora na základe jeho textov. Analyzujeme text s cieľom zistiť atribúty autora, ako sú pohlavie, veková kategória, národnosť a ďalšie. Zistilo sa, že tieto demografické údaje často korelujú s autorským štýlom; napríklad s narastajúcim vekom sa zvyšuje používanie predložiek, zatiaľ čo používanie zámen klesá.

V online prostredí je profilácia autorstva komplexnejšia. Demografické informácie o autorovi môžu byť často skryté alebo zavádzajúce. Rovnako je náročnejšie profilovať autora na základe krátkych textov, ako sú tie na sociálnych médiách, kde presnosť môže byť nižšia.

V procese profilácie je výhodné využívať obsahovo špecifické znaky. V praxi sa ukázalo, že tieto znaky môžu napomôcť k presnejšiemu určeniu hľadaných demografických charakteristík autora. Treba však mať na pamäti, že tieto obsahové znaky môžu byť občas ovplyvnené špecifickými okolnosťami písacej situácie, čo môže viesť k nesprávnym výsledkom [4][6].

2.1.4 Stylochronometria

Stylochronometria sa zaoberá detekciou a analýzou zmien v autorskom štýle cez čas. Preskúmava aspekty ako slovná zásoba a syntaktické štruktúry, a to nielen v rámci individuálnych kariér jednotlivých autorov, ale aj v kontexte štýlových trendov počas rôznych historických období. Tento prístup umožňuje lepšie pochopenie literárnych trendov a ich vývoja, napríklad ako sa štýl menil od romantizmu k modernizmu.

Stylochronometria je tiež kľúčová v procese overovania autorstva, najmä pri dielach napísaných v dlhšom časovom rozmedzí. Výskum dvoch tureckých autorov napríklad ukázal, že priemerná dĺžka slov sa zvyšuje v novších dielach oproti starším, čo naznačuje, že s rastúcim vekom autora sa môže meniť aj jeho slovník, možno v

dôsledku zlepšenia jazykových zručností.

Stylochronometria je rovnako dôležitá pri určovaní poradia, v akom autor napísal svoje diela. Táto metóda je neoceniteľná v situáciách, kedy diela neboli publikované chronologicky alebo keď sú informácie o ich vzniku neúplné či nejasné [4][7].

2.1.5 Kontrodiktorná stylonometria

Kontradiktorná stylonometria sa venuje analýze textu s cieľom zmeniť štýl písania, aby sa znížila pravdepodobnosť identifikácie autora alebo jeho charakteristických prvkov prostredníctvom stylometrickej analýzy. Tento proces zahŕňa úpravy v slovníku, slovných konštrukciách, syntaxe a gramatiky. Táto metóda je obzvlášť dôležitá v oblastiach, kde anonymita a ochrana identity majú veľký význam. Kontradiktorna stylonometria zahŕňa tri hlavné metódy:

1. **Imitácia:** Písanie textu tak, aby sa štýl približoval štýlu iného autora.
2. **Preklad:** Prekladanie textu do iného jazyka a späť pomocou strojového prekladu, čo mení štýl textu, ale môže tiež zmeniť význam niektorých častí.
3. **Zamaskovanie:** Zámerne meniť štýl písania, napríklad zmenou frekvencie používania určitých slov, viet alebo fráz.

Existujúce štúdie, ktoré skúmali zámerne zmenený štýl písania, ukázali, že tieto metódy ešte nie sú dostatočne efektívne na úplné zamaskovanie autorského štýlu. Na detekciu takýchto zmien sa využívali rôzne modely, ako napríklad BERT a GPT-2, a vo väčšine prípadov boli takto upravené texty odhalené [4][8].

Kapitola 3

Stylometrické črty

Stylometrické charakteristiky predstavujú kľúčové aspekty textu, ktoré sú využívané na jeho analýzu. Tieto charakteristiky sa zameriavajú na rozličné aspekty jazyka a písania, umožňujúc nám identifikovať unikátne vzorce, ktoré môžu byť použité na priradenie textov konkrétnym autorom, profiláciu a podobné účely. Na získanie týchto charakteristík využívame textové merania a štatistické metódy. Črty môžeme rozdeliť do troch kategórií: lexikálne(analýza slov), syntaktické (analýza štruktúry viet) a sémantické (analýza významu) črty. Okrem toho sa používajú aj textové reprezentácie, ktoré slúžia na transformáciu textu do formátu vhodného na ďalšiu analýzu a spracovanie. Niektoré črty sa môžu radiť do viacerých spomenutých kategórií [9][10].

3.1. Lexikálne črty

Lexikálne charakteristiky textu poskytujú dôležité informácie o výbere a použití slov autora v texte. Tieto charakteristiky sa zameriavajú na rôzne aspekty, ako je frekvencia znakov, rozmanitosť slovníka, dĺžka slov a podobne. Pomocou lexikálnych charakteristík môžeme získať prehľad o autorovom preferovanom slovníku a celkovej bohatosti jeho jazyka. Existuje mnoho rôznych lexikálnych charakteristík, z ktorých každá poskytuje unikátny pohľad na text a autora, vymenujeme a vysvetlíme tu niekoľko z nich:

- **Frekvencia Znakov**: Táto charakteristika zohľadňuje rozdelenie a výskyt rôznych typov znakov v texte. Pozostáva z analýzy počtu malých a veľkých písmen, cifier, nealfabetických znakov, čiastočných symbolov, ako sú emoji a medziery. Frekvencia znakov môže poskytnúť dôležité indície o štýle a formálnosti textu.

Napríklad vyšší podiel veľkých písmen môže naznačovať emocionálnejší alebo formálnejší tón, zatiaľ čo zvýšené použitie malých písmen môže odrážať neformálnejší štýl. [11][12].

- **Frekvencia Špecifických Vzorov**: Táto charakteristika analyzuje frekvenciu a distribúciu špecifických textových prvkov, napríklad hashtagy, URL adresy a užívateľské označenia (tagy). Táto metrika môže byť užitočná pre analýzu textov na sociálnych médiách alebo v obsahu z digitálnych komunikácií, kde sa často vyskytujú tieto prvky. Výskyt takýchto vzorov môže odkrývať určité vzorce správania, ako sú trendy v používaní sociálnych médií alebo preferencie odkazovania na externé zdroje [13].

- **Rozmanitosť Slovnej Zásoby**: Tento aspekt, známy tiež ako lexikálna diverzita, poukazuje na šírku a rozmanitosť autorovej slovnej zásoby v texte. Zameriava sa na počet a rozmanitosť jedinečných slov, čím poskytuje pohľad na autorovu kreativitu a štýl. Meria sa pomocou typ-token pomeru, kde vyššia hodnota indikuje širšiu slovnú zásobu. [14].

- **Priemerná Dĺžka Slov**: Táto charakteristika meria priemerný počet znakov na slovo v texte, ktorý poskytuje náhľad na štýl písania autora. Vyššia priemerná dĺžka slov často signalizuje používanie odbornejších alebo formálnejších termínov, zatiaľ čo kratšia dĺžka slov môže indikovať neformálnejší jazyk.

- **Priemerná Dĺžka Viet**: Táto metrika predstavuje pomer celkového počtu slov v texte ku celkovému počtu viet. Dlhšie vety môžu naznačovať komplexnejšie myšlienkové postupy a štruktúru, zatiaľ čo kratšie vety môžu ukazovať na priamočiary a jednoduchší štýl [15].

3.2. Syntaktické črty

Syntaktické charakteristiky textu sa zameriavajú na štruktúru a usporiadanie slov vo vetách, čím poskytujú hlbší pohľad na jazykové schopnosti autora. Tieto charakteristiky preskúmavajú rôzne aspekty, vrátane analýzy slovných druhov, funkčných slov, použitia interpunkcie, vzťahov medzi odsekmi a ďalších syntaktických prvkov. Tieto charakteristiky nám umožňujú lepšie pochopiť, ako autor konštruuje svoje vety a aké sú

jeho preferencie pri usporiadaní slov a fráz. Dovoľte nám predstaviť niekoľko príkladov syntaktických charakteristík, ktoré sa bežne používajú pri analýze textu [16]:

- **Frekvencia Funkčných Slov**: Táto charakteristika zahŕňa výskyt slov s gramatickou funkciou, ako sú členy, spojky, predložky, zámená a pomocné slovesá. Tieto slová sú fundamentom jazyka a sú nevyhnutné pre tvorbu viet a vzťahov medzi slovami. Frekvencia funkčných slov môže odhaliť špecifické vzorce písania, ktoré sú typické pre daného autora [17][18].
- **Frekvencia Časti Reči**: Táto analýza sa zameriava na výskyt rôznych slovných druhov, ako sú podstatné mená, prídavné mená, slovesá, spojky a ďalšie. Pomáha určiť, či autor uprednostňuje používanie určitých slovných druhov, čo môže byť indikátorom jeho osobitého štýlu písania [19].
- **Vzory Interpunkcie**: Táto charakteristika analyzuje, ako autor používa interpunkčné znamienka. Rôzne typy interpunkcie môžu poukazovať na gramatické a syntaktické prvky textu, ako aj na rytmus a tempo písania autora [20].
- **Prechodové Slová a Frázy**: Tieto slová a frázy slúžia ako spojnice medzi myšlienkami a odsekmi, zabezpečujúc plynulosť textu. Ich analýza môže odhaliť štýl autora a spôsob, akým sa venuje téme, ako aj jeho schopnosť vytvárať koherentné a súvislé texty [21].

3.3. Obsahové črty

Obsahové charakteristiky sa sústreďujú na tematický a sémantický obsah textu. Tento prístup sa zameriava na analýzu tém, myšlienok a emocionálneho tónu diela, čím poskytuje hlbšie pochopenie autorových zámerov a postojov. Tieto charakteristiky odhaľujú, ako autor pristupuje k rôznym témam a aké emócie a myšlienky sa snaží vyjadriť. Nižšie sú uvedené niektoré z obsahových charakteristík, ktoré nám umožňujú lepšie pochopiť a interpretovať text:

- **Pozitívne a Negatívne Slovné Triedy**: Táto charakteristika priraďuje slovám skóre sentimentu na základe ich pozitívneho alebo negatívneho významu. Slová s pozitívnym sentimentom (s hodnotením vyšším ako 0.1), ako napríklad “happy”,

"joy" alebo "love", sú zaradené do pozitívnej kategórie. Naopak, slová s negatívnym sentimentom (s hodnotením nižším ako -0.1), ako sú "sad", "pain" alebo "horrible", patria do negatívnej kategórie. Na výpočet skóre sentimentu sa často využívajú databázy ako SentiWordNet a AFINN [22].

- **Frekvencia Slovných Tried**: Tento prístup identifikuje tematické kategórie, ako sú "Rodina", "Peniaze" a "Práca", a následne analyzuje frekvenciu slov patriacich do týchto kategórií v texte. Nástroje ako RiTaWordNet pomáhajú identifikovať slová spojené s konkrétnymi témami. Táto analýza môže odhaliť autorove sklonnosti k písaniu o určitých témach a môže nám poskytnúť hlbšie porozumenie jeho záujmom a perspektívam [22].

3.4. Rozdielnosť typov čŕt vo vetách

Máme tu skupinu viet, ktoré ilustrujú rozdiely v rôznych druhoch charakteristík:

V1: Čokoláda vás môže zabiť.
V2: Čokolády, hoci sú chutné, vás môžu zabiť.
V3: Čokolády, aj keď sú chutné, vás môžu zabiť.
V4: Nastalo obdobie nepriaznivého počasia.
V5: Týždeň každý deň pršalo.

Pozrime sa na vety V2 a V3, kde pozorujeme rozdiely na lexikálnej úrovni. Veta V2, používajúca frázu "hoci sú chutné", má jazykový štýl bližší bežnej, hovorovej komunikácii. Naproti tomu, V3 s frázou äj keď sú chutné"pôsobí formálnejšie a literárnejšie, s možným nádychom akademickej alebo sofistikovanejšej reči.

Pri syntaktickej analýze viet V1 a V2 vidíme, že zatiaľ čo V1 je stručná a priama, V2 pridáva doplňujúci prvok, čím sa stáva zložitejšou a poskytuje dodatočný kontext, ktorý mení význam z absolútneho na podmienený.

Na obsahovej úrovni sú vety V4 a V5 odlišné. Veta V4, "Nastalo obdobie nepriaznivého počasia", je abstraktnejšia a menej špecifická, poskytuje všeobecný pohľad na počasie bez detailov. Naopak, V5, "Týždeň každý deň pršalo", je veľmi

konkrétna a poskytuje presný opis počasia v určitom časovom období, čím nám dáva jasné a špecifické informácie [23].

3.5. Textové reprezentácie

Textové reprezentácie sú kľúčové, pretože umožňujú počítačovým algoritmom spracovávať text. Keďže počítače pracujú efektívnejšie s číselnými údajmi, konvertujeme prirodzený jazyk do numerickej formy. Tento proces, známy ako reprezentácia textu, umožňuje algoritmom strojového učenia text efektívne analyzovať a spracovávať [24].

Textové reprezentácie zohrávajú kľúčovú úlohu v rôznych oblastiach stylometrie, kde sa používajú ako atribúty na predpovedanie rôznych aspektov, ako sú autorské štýly, demografické charakteristiky a ďalšie. Výborným príkladom je súťaž PAN, v ktorej mnohé tímy využívali práve textové reprezentácie na určenie veku účastníkov alebo na rozlíšenie, či je subjekt človek alebo počítačový program. K tomu často využívali rôzne techniky ako slovné a znakové n-gramy, metódu bag of words a n-gramy určujúce časti reči. Niektoré tímy dokonca zaviedli váhované n-gramy s využitím metódy TF-IDF na zvýšenie presnosti svojich modelov. Tieto prístupy umožňujú presnejšie a cielenejšie analyzovať a interpretovať textové dáta [25].

Reprezentácia „vreca slov“, známa ako Bag of Words (BOW), zaznamenáva frekvenciu výskytu každého slova v texte, pričom nezohľadňuje slovosled ani ďalšie lexikálne informácie. Tento model sa obvykle implementuje pomocou nástroja ako je CountVectorizer, ktorý vytvára maticu slov a textových záznamov. Každé pole tejto matice potom obsahuje numerickú frekvenciu príslušného slova v konkrétnom texte. Hlavnou nevýhodou tohto prístupu je strata informácií o pozícii slov v texte, čo môže sťažiť pochopenie ich významu. Navyše, intuitívny predpoklad, že slová s vyššou frekvenciou sú dôležitejšie alebo informatívnejšie, sa nemusí vždy potvrdiť, najmä v prípade funkčných slov ako „the“ alebo „with“. Preto je často vhodné takéto slová z textu odstrániť ešte pred vytvorením samotnej reprezentácie [24].

Reprezentácia TF-IDF (Frekvencia termínov a inverzná frekvencia dokumentov), rozšírená forma modelu ”Bag of Words”, nielenže sleduje frekvenciu výskytu slov, ale tiež hodnotí ich relevanciu v rámci dokumentu. Podstatou TF-IDF je

poskytnúť váhové skóre slovám na základe ich frekvencie v konkrétnom dokumente a ich bežnosti v celej kolekcii dokumentov. Tento prístup umožňuje identifikovať slová, ktoré sú nielen často používané, ale tiež špecificky dôležité pre daný dokument, čím zvýrazňuje jeho unikátne aspekty v kontexte celého korpusu.

Výpočet skóre TF-IDF sa realizuje ako súčin dvoch komponentov:

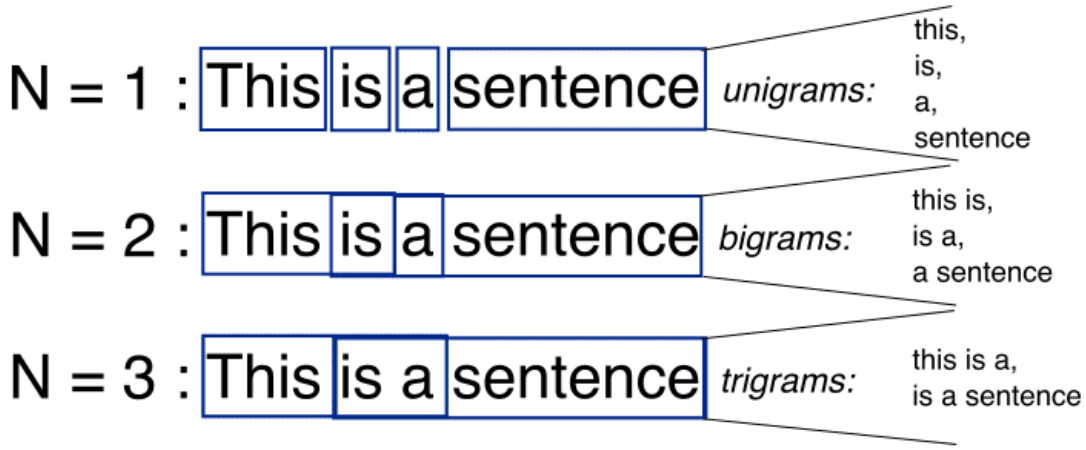
- **Term Frequency (TF):** meria, ako často sa dané slovo objaví v dokumente, vypočítané ako podiel počtu výskytov konkrétneho slova a celkového počtu slov v dokumente.
- **Inverse Document Frequency (IDF):** hodnotí unikátnosť slova v celom súbore dokumentov. IDF sa počíta ako logaritmus (so základom e) pomeru medzi celkovým počtom dokumentov a počtom dokumentov obsahujúcich dané slovo.

Takto získané TF-IDF skóre efektívne zohľadňuje nielen početnosť, ale aj sémantickú dôležitosť termínu, čím zlepšuje presnosť a účinnosť analytických modelov zameraných na spracovanie textu, ako sú vyhľadávanie informácií, klasifikácia dokumentov a extrakcia kľúčových slov.

Pre lepšie pochopenie si môžeme dať príklad. Máme 2 textové dokumenty: Dokument A: ”Rýchle hnedá líščie mláďa”, Dokument B: ”Hnedé líščie mláďa preskočilo psa”. Vypočítame si TF-IDF pre slovo “rýchle”. Ako prvé si vypočítame TF. Pre prvý dokument: TF=1/4=0.25. Pre druhý dokument TF=0/5=0. Následne si vypočítame IDF=log(2/1)=0.693. Nakoniec vypočítame TF-IDF. Pre prvý dokument TF-IDF=0.25×0.693=0.173. Pre druhý dokument TF-IDF=0×0.693=0 [24][26].

N-gram je technika v spracovaní prirodzeného jazyka, ktorá rozdeľuje text na spojité sekvencie n slov alebo znakov. Hodnota n určuje počet slov alebo znakov, ktoré tvoria jednu textovú jednotku. Napríklad unigram (1-gram) predstavuje jednotlivé slová alebo znaky, bigram (2-gram) spája dvojice za sebou idúcich slov alebo znakov, zatiaľ čo trigram (3-gram) tvorí trojice. Tento postup možno rozšíriť na akýkoľvek vyšší rád n. Podobne ako v modeli Bag of Words (BOW) sa vytvára matica frekvencií, ktorá zaznamenáva, ako často sa každý n-gram v dokumentoch vyskytuje. Navyše, n-gramy môžu byť vážené pomocou metódy TF-IDF, čo zvyšuje ich informatívnosť tým, že

poskytuje vyššiu váhu n-gramom, ktoré sú významné pre daný záznam, ale neobvyklé v celom datasete [24]. Príklady rôzdnych n-gramov si môžeme pozrieť nižšie 3.1.

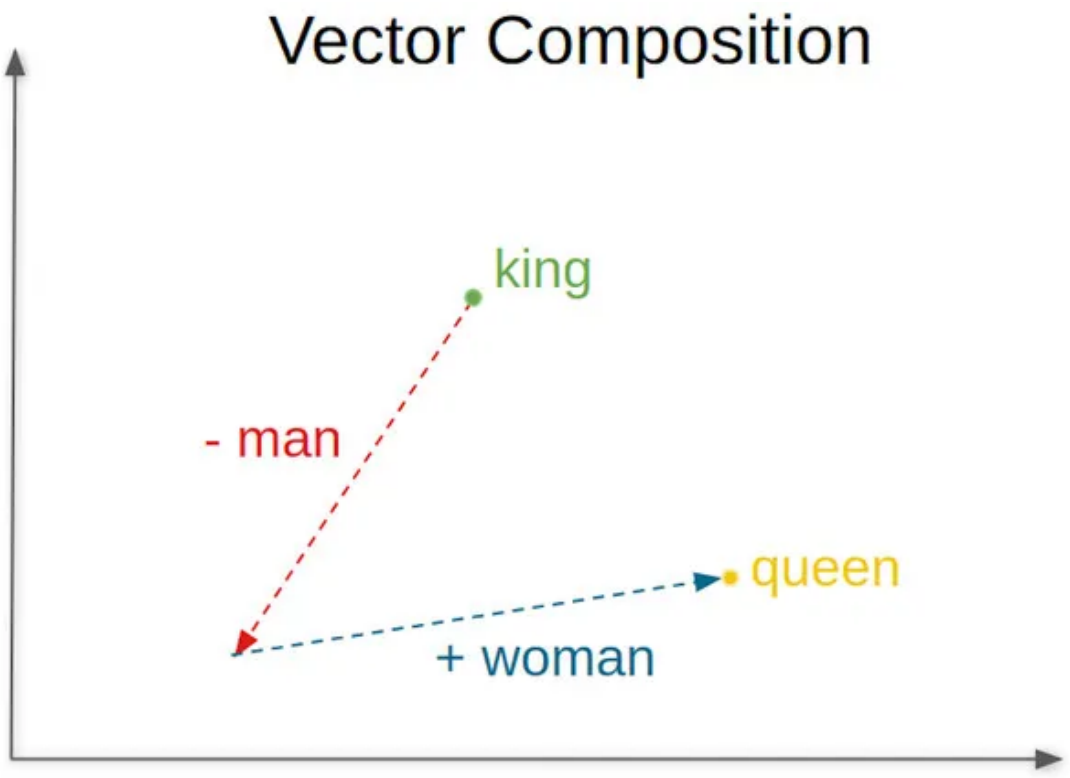


Obr. 3.1: N-gramy [27]

Word2Vec je pokročilá technika z kategórie vkladania slov (Word Embedding), ktorá transformuje slová do hustých vektorových reprezentácií v mnohorozmernom priestore, čím zachytáva sémantické vzťahy a kontextové nuansy medzi nimi. Táto metóda efektívne mapuje slová s podobnými významami tak, aby boli v priestore umiestnené blízko seba, čo umožňuje modelu lepšie pochopiť sémantické a kontextové súvislosti. Word2Vec využíva koncept, že slová vyskytujúce sa v podobných kontextoch majú tendenciu zdieľať sémantické významy. Model sa učí vektorové reprezentácie slov analyzovaním slov v ich kontextovej blízkosti. Existujú dve hlavné architektúry, ktoré Word2Vec používa: Continuous Bag of Words (CBOW) a Skip-Gram. Architektúra CBOW predpovedá cieľové slovo na základe kontextových slov, kým Skip-Gram funguje naopak, predpovedá kontextové slová na základe cieľového slova.

Pre lepšie pochopenie si môžeme uviesť známy príklad toho, ako Word2Vec zachytáva sémantické vzťahy medzi slovami a to rovnicou: „kráľ - muž + žena = kráľovná“. Tento príklad máme aj ilustrovaný nižšie 3.2. Táto rovnica ukazuje, že keď od slova „kráľ“ odčítame vektor pre „muž“ a pripočítame vektor pre „žena“, dostaneme vektor, ktorý je blízky vektoru slova „kráľovná“. Tento prístup využíva schopnosť modelu Word2Vec rozumieť kontextu a sémantike slov, čo umožňuje

vykonávať zmysluplné operácie s vektormi. Na zistenie, ktoré slovo sa najviac podobá výslednému vektoru, sa používa kosínusová podobnosť, ktorá môže potvrdiť, že „kráľovná“ je najbližším slovom k výsledku [28].



Obr. 3.2: Vizualizácia príkladu Word2Vec: kráľ - muž + žena = kráľovná [28]

Kapitola 4

Spracovanie textu

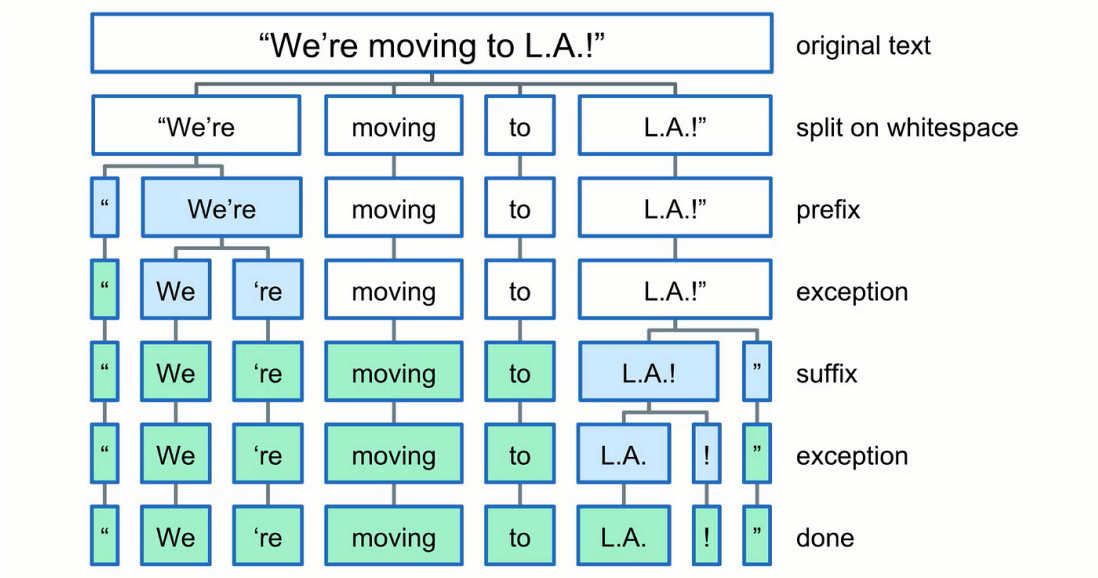
Spracovanie textu zahŕňa širokú paletu techník navrhnutých na analýzu a úpravu textových dát pomocou počítačových algoritmov. Jeho hlavným cieľom je extrahovať z neštruktúrovaného textu relevantné informácie a premeniť ich na formát vhodný pre ďalšiu analýzu a spracovanie. Tento proces je často považovaný za subdisciplínu v rámci odboru spracovanie prirodzeného jazyka (NLP), ktorý sa sústredí na vývoj algoritmov a systémov umožňujúcich počítačom rozumieť a interagovať s ľudským jazykom ako súčasť širšieho pola informatiky [29].

Spracovanie textu využíva širokú škálu techník, ako sú tokenizácia, stemming, lemmatizácia, označovanie slovných druhov (POS tagging), analýza sentimentu, identifikácia entít, sumarizácia textu, modelovanie tém a ďalšie. Podrobnejšie si vysvetlíme niekoľko z nich, ktoré považujeme za kľúčové pre implementačnú časť našej práce [29][30].

4.1. Tokenizácia

Tokenizácia je proces rozdeľovania textu na menšie jednotky, známe ako tokény. Tieto tokény môžu zahŕňať vety, slová, časti slov, znaky alebo dokonca interpunkčné znamienka a regulárne výrazy. Rozdelením textu na tieto menšie časti sa strojom uľahčuje pochopenie ľudskej reči a zefektívňuje sa ich schopnosť analyzovať text.

Tokenizácia je kľúčová pre mnohé techniky spracovania textu, ako napríklad lemmatizácia, stemming, POS tagging, pretože tieto techniky operujú práve na úrovni tokenov. V oblasti strojového učenia sa tokenizácia často využíva v analýzach štylistiky textu, napríklad pri extrakcii špecifických čŕt alebo pri vytváraní textových reprezentácií [31][32][33]. Príklad ako funguje tokenizácia, si môžeme pozrieť na obrázku nižšie 4.1.



Obr. 4.1: Ukážka tokenizácie [34]

4.2. Stemming

Stemming je metóda normalizácie textu, ktorá sa zameriava na odstránenie predpon a prípon z pojmov, čím ich transformujeme na ich základnú alebo koreňovú formu. Táto základná forma nemusí vždy odpovedať morfologickému koreňu slova, avšak je dostačujúce, ak majú príbuzné pojmy rovnaký koreň. Primárnym účelom tejto techniky je zníženie počtu unikátnych slov v texte, čo uľahčuje jeho analýzu a zlepšuje porozumenie obsahu. Stemming je v porovnaní s lemmatizáciou jednoduchší a rýchlejší, pretože používa sady pravidiel alebo algoritmy na dosiahnutie základnej formy slova. Napriek svojej efektivite sa môže stať, že výsledná forma nie je plne

validná a môže viesť k nejasnostiam. Táto technika je vhodná najmä v prípadoch, kde presný význam slov nie je pre analýzu kriticky dôležitý [35].

4.3. Lemmatizácia

Lemmatizácia je pokročilá technika normalizácie textu, ktorá premení slová na ich morfologické základné formy. Podobne ako stemming, aj jej hlavným cieľom je znížiť množstvo jedinečných slov v textových dokumentoch, čím uľahčuje ich analýzu. Oproti stemmingu, lemmatizácia je sofistikovanejší proces, ktorý využíva rozsiahlu slovnú zásobu a morfologickú analýzu na presné určenie základnej formy slova. Tento proces je však pomalší a komplexnejší, ale zato poskytuje presnejšie výsledky, pretože každá získaná základná forma slova existuje v slovníku a je gramaticky správna. Lemmatizácia je preferovanou metódou v situáciách, kde je dôležitý presný význam slov, aby analýza bola čo najpresnejšia [35]. Názorný príklad rozdielu medzi lemmatizáciou a stemmingom si môžeme pozrieť na obrázku nižšie 4.2.

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Obr. 4.2: Porovnanie lemmatizácie a stemmingu [36]

4.4. Označovanie slovných druhov

Označovanie slovných druhov známe aj ako POS tagging, je proces, pri ktorom sa každému slovu v texte priraďujú gramatické kategórie, ako sú podstatné mená, prídavné mená a slovesá. Tento proces je základným nástrojom pre identifikáciu gramatických funkcií slov, čo umožňuje pokročilú syntaktickú analýzu. Označovanie POS takisto napomáha v pochopení významu slov v ich kontexte, napríklad rozlíšenie medzi použitím slova ako podstatného mena alebo slovesa môže zásadne zmeniť význam celej vety. Táto technika poskytuje základné informácie pre široké spektrum aplikácií v oblasti spracovania prirodzeného jazyka, vrátane textovej analýzy, strojového prekladu a vyhľadávania informácií [37].

Kapitola 5

Strojové učenie

Strojové učenie, kľúčová súčasť umelej inteligencie a informatiky, využíva dáta a algoritmy na simuláciu ľudského učebného procesu, čím neustále zlepšuje svoju presnosť. Algoritmy strojového učenia identifikujú vzťahy a vzorce v dátach, čím sa zdokonaľujú v rozhodovaní a predikciách na základe nových informácií a skúseností [38]. Existujú tri základné typy strojového učenia:

- **Učenie s Učiteľom**: Pri tomto prístupe model pracuje s označenými tréningovými dátami a učí sa identifikovať vzťahy medzi vstupnými a výstupnými premennými. Ideálne pre úlohy ako klasifikácia a regresia.
- **Učenie bez Učiteľa**: Algoritmus pracuje s neoznačenými dátami a snaží sa nájsť v nich skryté vzory alebo štruktúry, ktoré následne využíva na kategorizáciu údajových bodov. Tento typ je obzvlášť užitočný pri zhlukovaní, detekcii anomálií a redukcii rozmerov dát.
- **Posilňované Učenie**: Zameriava sa na vytváranie algoritmov, ktoré optimalizujú svoje rozhodnutia v prostredí s jasnými cieľmi a pravidlami na základe odmien a trestov. Často sa využíva v robotike a v oblastiach, kde je potrebné učiť roboty hrať hry alebo riadiť zdroje.

Strojové učenie nájde svoje uplatnenie aj v stylometrii. Používa sa na analýzu textových dát, identifikáciu charakteristických znakov textu, rozlišovanie medzi autormi, profiláciu a a ďalšie účely. Medzi využívané techniky patria neurónové siete, regresia a klasifikácia, pričom každá z nich má v kontexte stylometrie svoje unikátne využitie.

Strojové učenie dokáže spracovávať široké spektrum textov, od literárnych diel cez politické prejavy až po obsah sociálnych médií. Napriek svojej rozmanitosti aplikácií však stále čelí výzvam, ako zabezpečenie presnosti, interpretovateľnosti a riešenie etických otázok pri práci s citlivými dátami [39][40].

5.1. Klasifikácia

Klasifikátor v strojovom učení je algoritmus navrhnutý na priraďovanie kategórií alebo tried k údajovým bodom podľa ich charakteristík. Tento proces využíva matematické a štatistické metódy na predpovedanie pravdepodobnosti, že vstupné údaje patria do určitej kategórie. Hlavným výstupom klasifikačného algoritmu je model schopný identifikovať a kategorizovať nové údaje na základe naučených vzorcov.

Klasifikátor trénuje model pomocou trénovacích dát, ktoré obsahujú príklady s označenými kategóriami. Trénovací proces zahŕňa analýzu týchto dát, aby algoritmus pochopil vzorce súvisiace s konkrétnymi kategóriami. Tento model sa následne používa na klasifikáciu nových, neoznačených dát, čím sa testuje jeho schopnosť správne kategorizovať údaje.

Klasifikátor má kľúčový význam v mnohých aplikáciách strojového učenia, najmä tam, kde je dôležité rozlišovať medzi rôznymi kategóriami. Takéto aplikácie zahŕňajú rozpoznávanie obrazov, spracovanie prirodzeného jazyka, medicínsku diagnostiku a mnoho ďalších oblastí, kde je kategorizácia údajov kritická pre úspešnú analýzu a rozhodovanie [41][42].

5.1.1 Algoritmus podporného vektorového stroja

Algoritmus podporného vektorového stroja (SVM) je zameraný na lineárnu aj nelineárnu klasifikáciu a regresiu a patrí medzi algoritmy učenia s učiteľom. Je vhodný pre rôzne úlohy, ako sú klasifikácia textov, obrázkov, detekcia spamu a podobne. Výhodou SVM je schopnosť efektívne spracovať veľkorozmerné údaje a nelineárne vzťahy, pričom dosahuje vysokú presnosť. SVM sú užitočné najmä tam, kde sú dáta zložité a vyžadujú sofistikované rozhodovacie hranice. Algoritmus SVM sa prevažne

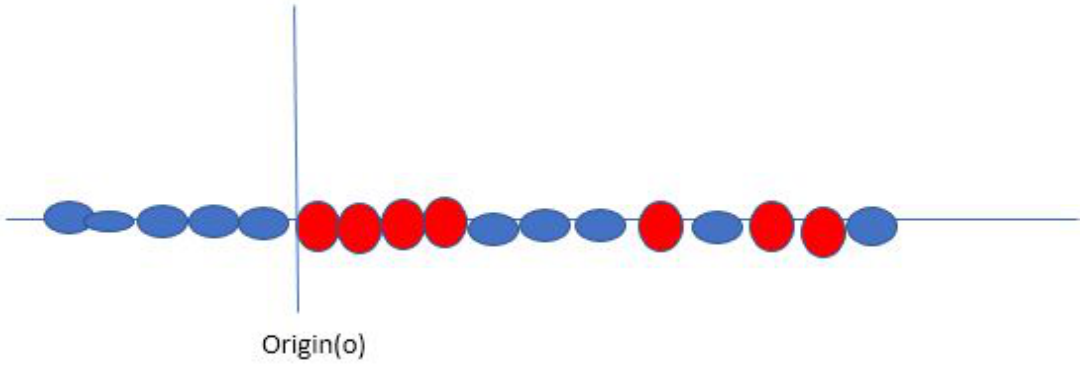
používa na binárnu klasifikáciu, kde sú dáta rozdelené do dvoch kategórií, ale zvláda aj viacrozmernú klasifikáciu.

Primárnym cieľom SVM je nájsť v N-rozmernom priestore hyperrovinu, ktorá dokáže oddeliť dátové body patriace do rôznych tried. SVM určuje hyperrovinu tak, aby bola vzdialenosť medzi bodmi rôznych tried čo najväčšia, čím sa zvyšuje presnosť. Rozmer hyperroviny závisí od počtu tried: pre 2 triedy je to čiara, pre 3 triedy 2D rovina a podobne. SVM je tiež odolný voči odľahlým hodnotám, čo znamená, že dokáže akceptovať nesprávne klasifikované body alebo odľahlé hodnoty na dosiahnutie lepšej celkovej klasifikácie. Keď je bod triedy na nesprávnej strane okraja, pridá sa penalizácia. SVM sa snaží minimalizovať stratu závesu, matematickú funkciu na výpočet penalizácie: (1/margin + (penalta)). Správne klasifikované body majú nulovú penalizáciu, ktorá rastie so vzdialenosťou od okraja. Okraje, kde sa nachádzajú odľahlé alebo nesprávne klasifikované údaje, sa nazývajú mäkké okraje, inak ide o tvrdé okraje.

V prípade nelineárnych dát, ktoré nemožno oddeliť priamkou v dvojrozmernom priestore alebo hyperrovinou vo vyššej dimenzii, ako je zobrazené na obrázku č. 5.1, SVM používa kernelové funkcie. Tieto matematické funkcie umožňujú SVM mapovať pôvodné nelineárne dáta do vyššieho rozmerového priestoru, kde ich možno lineárne separovať. Po transformácii dát do tohto priestoru pokračujeme ako pri lineárnych dátach, ako je vidieť na obrázku č. 5.2. Tento postup výrazne zlepšuje schopnosť SVM riešiť komplexné klasifikačné úlohy. Napríklad, ak máme jednorozmerné dáta pozdĺž osi x, ktoré nemožno oddeliť, kernel vytvorí novú premennú y, závislú na vzdialenosti od x. Tieto body následne zobrazíme v dvojrozmernom priestore pomocou x a y, a určíme hyperrovinu [43] [44].

Algoritmus podporných vektorových strojov (SVM) má v stylometrii široké využitie, ako ukazuje štúdia [45], ktorá sa zamerala na predpovedanie veku a pohlavia autora na základe tweetov. V rámci tejto práce sa použila technika SVM s viacerými triedami, kde sa testovali rôzne parametre jadra, ako sú gamma a C, s cieľom optimalizovať presnosť klasifikácie. Výsledky ukázali, že jemné ladenie týchto parametrov má významný vplyv na výkonnosť modelu, pričom sa používala metóda grid-search s k-fold cross-validáciou na dosiahnutie najlepších možných nastavení. Ďalšia štúdia

[46] sa zameriava na atribúciu autorstva s použitím SVM. V experimentoch s textami z nemeckých novín SVM spoľahlivo identifikovalo cieľového autora s presnosťou 60-80%. Výskum potvrdil schopnosť SVM rozlišovať medzi autormi na základe rozmanitosti slovnej zásoby a zložitých vzorov frekvencie slov, ako napríklad Poissonove distribúcie. Toto odhaľuje efektívnosť SVM v riešení klasifikačných problémov aj v prípade veľkého množstva dát.



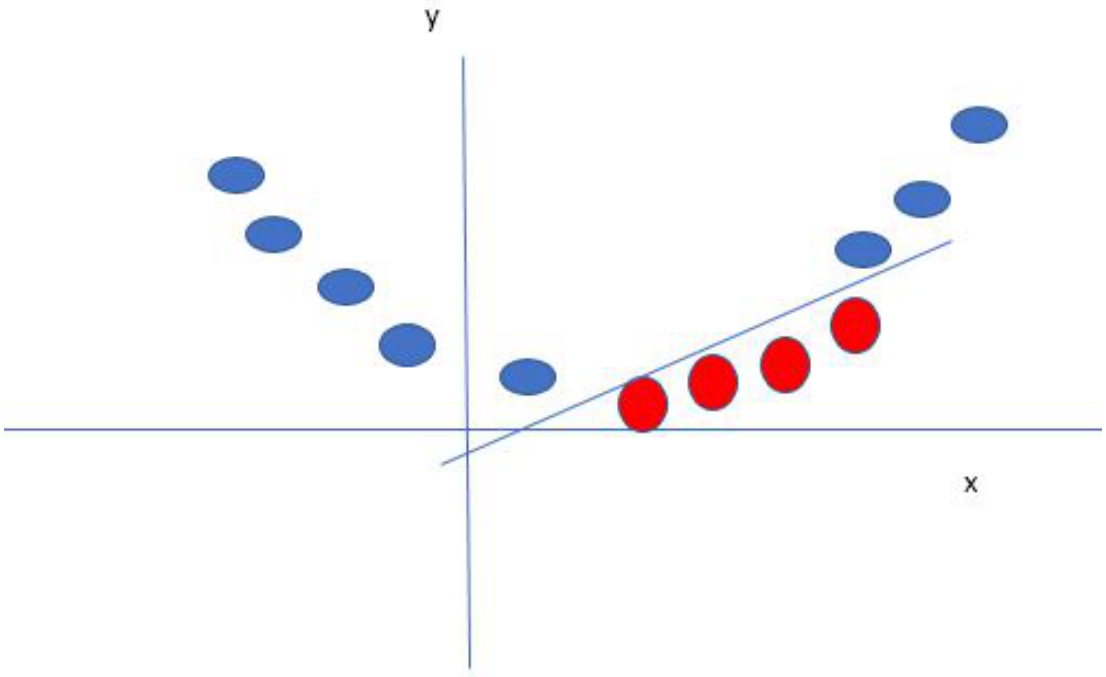
Obr. 5.1: Nelineárne dáta [43]

5.1.2 Rozhodovací strom

Rozhodovací strom je jednoduchý neparametrický algoritmus, používaný v strojovom učení pre riešenie klasifikačných aj regresných problémov. Patrí medzi metódy učenia pod dohľadom a je obľúbený vďaka svojej priamočiarosti, jednoduchej implementácii a ľahkej interpretovateľnosti. Napriek jeho prednostiam, má rozhodovací strom niekoľko obmedzení: vyžaduje úplné dáta bez chýbajúcich hodnôt, je citlivý na odľahlé hodnoty a má sklon k overfittingu, ktorému možno čeliť nastavením prísnych kritérií zastavenia.

Tento hierarchický model sa skladá z koreňového uzla, kde začína rozhodovací proces, vnútorných uzlov, ktoré reprezentujú otázky alebo testy na atribútoch, a listových uzlov, ktoré predstavujú konečné výsledky a ďalej sa nedelia. Vetvy medzi uzlami znázorňujú možné cesty rozhodnutí.

Vývoj rozhodovacieho stromu začína s celými tréningovými dátami v koreňovom uzle. Tu algoritmus vyberá atribút, ktorý optimálne rozdelí dáta podľa určitého



Obr. 5.2: Nelineárne dáta presunuté do vyššieho rozmerného priestoru [43]

kritéria (ako napríklad Giniho nečistota alebo entropia pre klasifikáciu, rozptyl alebo štandardná odchýlka pre regresiu). Týmto vybraným atribútom sa potom dáta rozdeľujú do vetiev, kde každá reprezentuje hodnotu atribútu a smeruje k ďalšiemu uzlu. Tento proces pokračuje do okamihu, kým nie sú splnené predom stanovené kritériá zastavenia, ako dosiahnutie maximálnej hĺbky stromu alebo nedosahovanie ďalšieho zlepšenia v predikciách. Po splnení týchto kritérií končí delenie a formujú sa listové uzly, ktoré obsahujú výsledný predpoklad alebo klasifikáciu [47].

5.1.3 Náhodný les

Náhodný les patrí medzi najuznávanejšie algoritmy strojového učenia, ktoré sa široko využívajú na klasifikačné a regresné úlohy. Tento algoritmus supervízovaného učenia je schopný efektívne spracovať komplexné dátové sady a účinne znižuje riziko preučenia. Náhodný les dokáže manipulovať s kontinuálnymi premennými využívanými v regresii aj s kategorickými premennými používanými v klasifikácii a má schopnosť

spracovávať dáta s chýbajúcimi hodnotami (NaN). Ako metóda zoskupovania viacerých modelov poskytuje lepšiu prediktívnu presnosť v porovnaní s použitím jedného modelu.

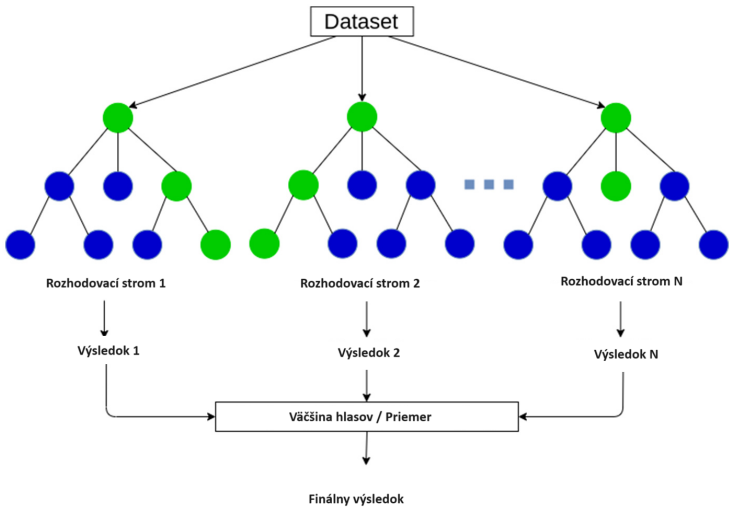
Náhodný les je založený na princípe rozhodovacích stromov. Vytvára súbor takýchto stromov, ktoré poskytujú presnejšie predpovede ako jednotlivý rozhodovací strom. Hlavnou nevýhodou tohto prístupu je však vysoká výpočtová náročnosť, keďže hodnotenie množstva stromov môže byť časovo náročné.

Algoritmus náhodného lesa funguje na princípe baggingu, známeho tiež ako Bootstrap Aggregation. Proces začína vytváraním modelov – rozhodovacích stromov – z vzoriek tréningových dát, ktoré sú vyberané metódou bootstrapping. Táto metóda umožňuje náhodný výber vzoriek z dátového súboru s možnosťou opakovaného výberu. Pri každom rozdelení uzla v strome sa vyberie náhodná podmnožina atribútov, ktoré sa použijú pre rozhodovanie, čo zvyšuje diverzitu modelov a pomáha predchádzať preučeniu. Týmto spôsobom vytvorené stromy sú schopné zvládať rôzne aspekty dát, čo umožňuje komplexnejšie modelovanie a lepšie výsledky. V klasifikácii každý strom hlasuje za jednu z možných kategórií a konečná predpoveď je určená na základe väčšinového hlasu. V regresii sa predikcie jednotlivých stromov priemerujú na určenie konečnej predpovede. Tento proces je ilustrovaný na obrázku 5.3. Náhodný les tiež dokáže určiť, ktoré atribúty sú dôležité pre presnosť predpovede, pričom dôležitosť atribútu závisí od jeho vplyvu na zlepšenie presnosti modelu [48][49].

V oblasti stylometrie sa algoritmus náhodného lesa intenzívne využíva, napríklad na profilovanie autora na základe textov z Twitteru, ako je to demonštrované v štúdii [50]. V tejto štúdii sa klasifikátory náhodného lesa používajú na určenie pohlavia a vekových skupín autorov, zatiaľ čo regresné modely hodnotia osobnostné črty ako extroverziu, stabilitu a otvorenosť. Výsledky experimentov ukázali, že metóda náhodného lesa môže dosiahnuť až 75% presnosť pri určovaní pohlavia a 70% pri klasifikácii vekových skupín. Táto štúdia ilustruje, ako môže náhodný les efektívne spracovať a klasifikovať veľké množstvá textových dát pre komplexné úlohy profilovania.

V ďalšej štúdii [51] autori skúmali využitie algoritmu náhodného lesa na atribúciu autorstva elektronických textov. Vďaka odolnosti náhodného lesa voči

nežiaducim črtám sa ukázalo, že tento prístup dosiahol vysokú klasifikačnú presnosť na viacerých dátových súboroch, s výnimkou jednej chyby klasifikácie. Tieto štúdie potvrdzujú význam a efektivitu náhodného lesa v oblasti stylometrie



Obr. 5.3: Vizualizácia procesu random forest (Preložené) [48]

5.1.4 Logická regresia

Logistická regresia predstavuje kľúčový algoritmus v strojovom učení, zaraďujúci sa do kategórie supervízovaného učenia. Je primárne aplikovaná na binárnu klasifikáciu, kde cieľom je určiť príslušnosť vstupných premenných k jednej zo dvoch možných kategórií, ako napríklad pohlavie (muž alebo žena).

Tento algoritmus využíva sigmoidnú funkciu na transformáciu reálnej hodnoty na pravdepodobnosť v rozsahu od 0 do 1. Sigmoidná funkcia je matematicky definovaná ako: $\sigma(z) = \frac{1}{1+e^{-z}}$, kde z je lineárna kombinácia vstupných premenných a váh, teda $z = \beta_0 + \beta_1 x_1$, kde β_0, β_1 sú váhy a x_1 je vstupná premenná. Ak je pravdepodobnosť vyššia ako určený prah, ktorý je vo väčšine prípadov 0.5, tak patrí do danej kategórie, ak menšia, tak patrí do inej kategórie, čo môžeme vidieť na obrázku č. 5.4

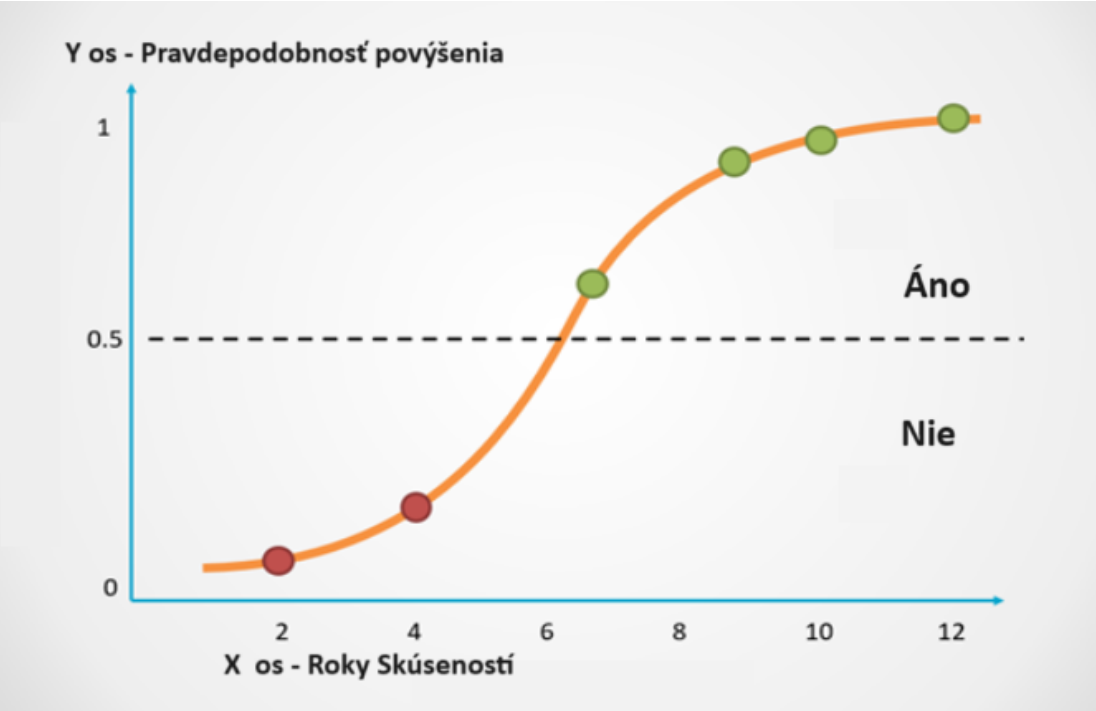
Pri trénovaní modelu sa úsilie sústreďuje na optimalizáciu váh tak, aby co najpresnejšie predpovedali kategórie. Tento proces sa obvykle realizuje dvoma hlavnými metódami: maximalizáciou logaritmickej pravdepodobnosti a metódou gradientného zostupu. Maximalizácia logaritmickej pravdepodobnosti zahŕňa hľadanie takých váh, ktoré najlepšie prispôsobia model daným dátam. Gradientný zostup začína s náhodne inicializovanými váhami a postupne ich upravuje v reakcii na rozdiely medzi skutočnými a modelom predpovedanými hodnotami [52][53].

V oblasti stylometrie sa logistická regresia využíva napríklad pri profilovaní autorstva. Štúdia [54] skúmala demografické charakteristiky autorov založené na ich tweetoch. Kombinuje technológie word2vec a logistickú regresiu na identifikáciu pohlavia autora a jazykovej variety. Text tweetu bol pretransformovaný do vektorového formátu pomocou word2vec, ktoré boli následne agregované do jedného vstupného vektora pre logistickú regresiu. Tento prístup dosiahol presnosť rozpoznávania jazykovej variety v rozmedzí od 44.88% pre arabčinu až po 97.63% pre portugalčinu, s presnosťou identifikácie pohlavia v rozmedzí od 64.25% do 74.46%. Výsledky ukazujú, že efektivita modelu závisí od počtu predikčných tried a od rozdielov medzi dialektmi. Štúdia potvrdila, že spojenie word embeddings s logistickou regresiou je efektívnym nástrojom pre profilovanie autora.

5.1.5 Naivný Bayesov algoritmus

Naivný Bayesov algoritmus je založený na aplikácii Bayesovej vety a charakteristickým predpokladom, že jednotlivé prvky v triede sú nezávislé na sebe. Hoci tento predpoklad často nie je v praxi splnený, algoritmus aj napriek tomu dosahuje vysokú úroveň presnosti. Patrí do kategórie supervízovaného strojového učenia a je často využívaný pri klasifikačných úlohách, ako je napríklad textová klasifikácia.

Naivný Bayesov algoritmus určuje pravdepodobnosti príslušnosti k triedam podľa nasledovného vzorca: $P(c \mid x) = \frac{P(x|c) \times P(c)}{P(x)}$, kde $P(c \mid x)$ je podmienená pravdepodobnosť triedy c za predpokladu vlastnosti x, $P(c)$ je pravdepodobnosť triedy c, $P(x)$ je pravdepodobnosť vlastnosti x a $P(x \mid c)$ je pravdepodobnosť vlastnosti x za predpokladu triedy c. Algoritmus následne určí pravdepodobnosti príslušnosti vstupnej hodnoty ku každej možnej triede a vyberie tú, ktorá vykazuje najvyššiu



Obr. 5.4: Vizualizácia aplikácie logistickej regresie (Preložené) [52]

pravdepodobnosť [55][56].

Využitie Naivného Bayesovho algoritmu sa rozširuje aj do oblasti stylometrie. Štúdia [57] preskúmala jeho aplikáciu pri identifikácii autorstva rozsiahlych textov, ako sú napríklad romány, kde bola klasifikačná technika algoritmu upravená pre efektívnejšie zvládanie rozsiahlych dokumentov. K týmto úpravám patrilo použitie logaritmických transformácií kondicionálnych pravdepodobností na prekonanie problémov s numerickou presnosťou, ktoré sa vyskytujú pri spracovaní veľkého množstva atribútov. Experimentálne výsledky s presnosťou až 97% ukazujú, že modifikovaný Naivný Bayesov algoritmus môže efektívne určiť, či bol konkrétny text napísaný určeným autorom, čím potvrdzujú jeho robustnosť a adaptabilitu na špecifické úlohy v analýze textu.

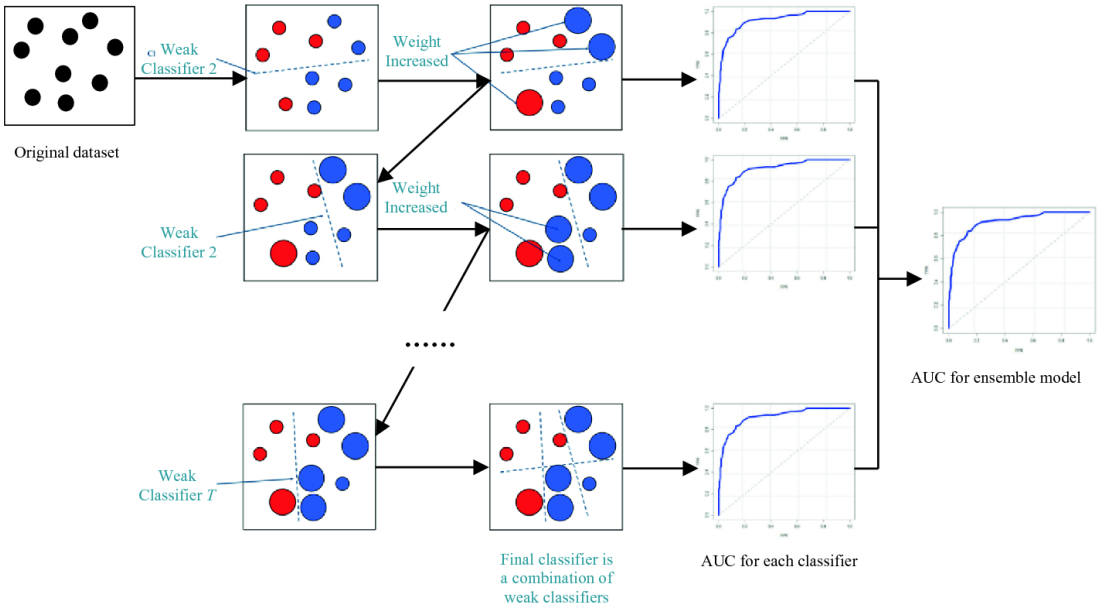
5.1.6 Gradient Boosting Maschine

Gradient Boosting Machine (GBM) je pokročilý algoritmus strojového učenia, zaradený do kategórie učenia s učiteľom. Tento algoritmus účinne kombinuje viacero základných prediktívnych modelov, najčastejšie rozhodovacích stromov, čím vytvára robustnejší a presnejší celkový model. GBM sa sústreďuje na postupné zlepšovanie predikcií pridaním nových modelov, ktoré sa zameriavajú na opravu konkrétnych chýb identifikovaných v predchádzajúcich iteráciách. Vynikajúca rýchlosť a presnosť GBM ho robia ideálnym nástrojom na prácu s rozsiahlymi a komplexnými datasetmi, kde tradičné metódy môžu zlyhávať [58][59].

Gradient Boosting Machine (GBM) začína vytvorením základného modelu, ako je napríklad rozhodovací strom, ktorý poskytuje počiatočný odhad predikcií. Tento model slúži ako základ, na ktorom sa postupne zlepšuje prediktívna schopnosť. Nasleduje výber stratovej funkcie, ktorá je kľúčová pre hodnotenie úspešnosti predikcií modelu. Typ stratovej funkcie sa líši podľa toho, či ide o klasifikačné alebo regresné úlohy. Pri klasifikácii sa často používajú funkcie ako binárna krížová entropia, viactriedna logistická strata alebo exponenciálna strata, každá so svojimi špecifickými charakteristikami a použitím. S každou iteráciou GBM pridáva ďalšie modely, ktoré sú zamerané na opravu chýb predchádzajúcich predikcií. Prostredníctvom techniky gradientného zostupu sa optimalizuje stratová funkcia tým, že sa vypočíta gradient stratovej funkcie. Tento gradient ukazuje smer, v ktorom by sa mal model zmeniť, aby sa znížila celková predikčná chyba. Aby sa znížilo riziko pretrénovania, GBM implementuje metódu zvanú shrinkage, ktorá redukuje vplyv každého nového pridaného modelu na konečnú predikciu, čím sa zvyšuje stabilita a schopnosť modelu generalizovať. Proces pridávania modelov pokračuje až do dosiahnutia stanoveného počtu iterácií alebo kým sa ďalšie zlepšovanie modelu nezastaví. Konečný model je výsledkom integrácie všetkých pridaných modelov, pričom každý z nich prispieva k predikcii podľa svojho váhového koeficientu [59][60]. Tento proces je znázornený na obrázku 5.5.

Algoritmus Gradient Boosted Decision Trees (GBDT) zohráva kľúčovú úlohu v profilácii autorov na sociálnych sieťach, ako potvrdzuje štúdia [61]. Tá sa zamerala na rozlíšenie medzi botmi a ľudskými užívateľmi a na identifikovanie pohlavia v prípade ľudských autorov. V tejto práci bol GBDT efektívne kombinovaný s hlbokým

učením, konkrétne s modelom BERT, čo výrazne zvýšilo výkonnosť celého systému. Výsledky experimentov ukazujú, že tento kombinovaný prístup dosiahol vysokú presnosť v klasifikácii, čo demonstruje robustnosť a efektivitu GBDT v rámci ensemble modelov pre profiláciu autorov. V článku [62] sa venovali určovaniu autorstva. Gradient Boosting Decision Trees (GBDT) boli súčasťou ensemble techník použitých na klasifikáciu a identifikáciu autorov, ktoré sa ukázali ako efektívne v kombinácii s pokročilými NLP modelmi ako DistilBERT. Tento kombinovaný prístup ukázal značné zlepšenie v presnosti oproti predchádzajúcim metódam. Konkrétne dosiahol zvýšenie presnosti identifikácie autorov o 3.14% a 5.25% v prvom a druhom rozsahu testovania, čo svedčí o vysokom potenciále týchto techník v oblasti určovania autorstva.



Obr. 5.5: Vizualizácia fungovania GBM [63]

5.2. Klasifikačné metriky

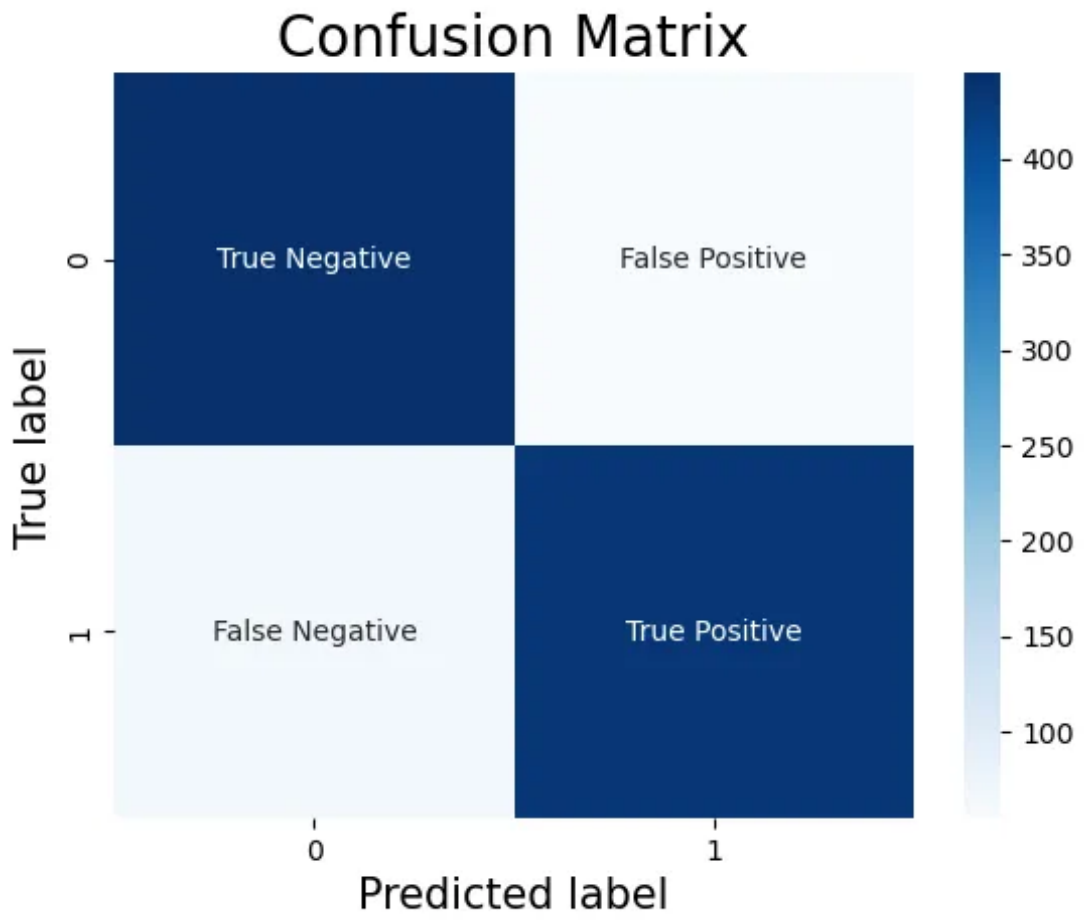
Klasifikačné metriky predstavujú nástroje na kvantitatívne vyhodnotenie výkonnosti modelov klasifikácie. Základ týchto metrík tvorí confiuson matrix, známa aj ako matica chýb, ktorá poskytuje prehľad o výkonnosti klasifikátora. Pri binárnej klasifikácii označujeme jednu triedu ako pozitívnu a druhú ako negatívnu. Nižšie je príklad matice chýb 5.6, ktorý ilustruje správne identifikované pozitívneh (TP) a negatívne (TN) prípady,

ako aj nesprávne klasifikované pozitívne (FP) a negatívne (FN) prípady. Vysvetlime si tieto komponenty podrobnejšie:

- **Skutočne pozitívne (TP)**: Počet pozitívnych prípadov, ktoré sú modelom správne klasifikované ako pozitívne.
- **Falošne pozitívne (FP)**: Počet negatívnych prípadov, ktoré sú modelom nesprávne klasifikované ako pozitívne.
- **Skutočne negatívne (TN)**: Počet negatívnych prípadov, ktoré sú modelom správne klasifikované ako negatívne.
- **Falošne negatívne (FN)**: Počet pozitívnych prípadov, ktoré sú modelom nesprávne klasifikované ako negatívne.

Klasifikačné metriky sa odvodzujú z hodnôt TP, TN, FP a FN, pričom výber adekvátnej metriky je kľúčový, keďže každá má špecifické využitie. Podrobnejšie si rozoberieme tie najčastejšie používané:

- **Presnosť (Accuracy)** : Táto metrika, často využívaná na hodnotenie klasifikátora, zisťuje pomer správne identifikovaných prípadov (TP + TN) ku celkovému počtu prípadov (TP + TN + FP + FN). Vyššia hodnota presnosti signalizuje, že model správne určuje veľkú časť prípadov, kým nižšia hodnota naznačuje opak.
- **Precíznosť (Precision)**: Táto metrika sa zameriava na pomer skutočne pozitívnych predikcií (TP) k všetkým pozitívnym predikciám modelu (TP + FP). Vyššia precíznosť ukazuje, že model efektívne identifikuje pozitívne prípady bez veľkého počtu falošne pozitívnych výsledkov.
- **Senzitivita (Recall)** : Meria schopnosť modelu správne rozpoznať pozitívne prípady (TP) z celkového počtu skutočných pozitívnych prípadov (TP + FN). Vyššie skóre v tejto metrike značí, že model efektívne zachytáva pozitívne prípady, nízke skóre naznačuje opomenutie mnohých pozitívnych prípadov.
- **Skóre F1** : Kombinuje precíznosť a senzitivitu do jedinej metriky, poskytujúc komplexný pohľad na výkonnosť modelu. Vypočíta sa ako harmonický priemer precíznosti a senzitivity, čo umožňuje vyvážené hodnotenie oboch aspektov. Vyššie skóre F1 značí dobrú výkonnosť modelu z hľadiska obidvoch metrík, zatiaľ



Obr. 5.6: Confusion matrix [64]

čo nízke skóre poukazuje na nedostatky v oblasti precíznosti alebo senzitivity [64].

Čo sme doteraz opisovali sa týkalo binárnej klasifikácie, avšak pri modeloch s viacerými triedami používame tie isté metriky, len ich aplikáciu mierne upravujeme. Presnosť a F1 skóre sa počítajú rovnakým spôsobom: presnosť ako pomer správnych predikcií ku všetkým predikciám a F1 skóre ako harmonický priemer precíznosti a senzitivity. Precíznosť a senzitivita sa však pre každú triedu počítajú osobitne, pričom každá konkrétna trieda je chápaná ako pozitívna a ostatné ako negatívne. Tieto výsledky sa potom použijú na výpočet priemernej metriky, čo možno vykonať dvoma hlavnými spôsobmi:

- **Makro-priemerovanie** : Táto metóda zahŕňa výpočet priemeru precíznosti pre

každú triedu zvlášť, pričom sa hodnoty precíznosti všetkých tried sčítajú a delia celkovým počtom tried. Rovnaký postup sa uplatní aj na senzitivitu, čím získame priemerné hodnoty pre celý model.

- **Mikro-priemerovanie** : Pri tejto technike sa najprv zosumarizujú všetky skutočne pozitívne (TP), falošne pozitívne (FP) a falošne negatívne (FN) predikcie naprieč všetkými triedami. Na výpočet precíznosti sa potom použije pomer celkového počtu TP k sume TP a FP. Pre senzitivitu sa vypočíta pomer medzi celkovým počtom TP a súčtom TP a FN, čo poskytuje celkovú mieru schopnosti modelu správne identifikovať pozitívne prípady naprieč triedami [65].

5.3. Techniky výberu čŕt pre strojové učenie

Úspech modelov strojového učenia závisí významne od efektívnosti výberu čŕt. V reálnych scenároch nie sú všetky dostupné črty prínosné, a ich zahrnutie môže negatívne ovplyvniť schopnosť modelu generalizovať na nové dáta, čím dôjde k poklesu jeho presnosti. Efektívne stratégie identifikácie a selekcie najužitočnejších čŕt zabezpečujú, že modely sú nielen presné, ale aj výkonné. Metódy výberu čŕt môžeme rozdeliť do troch základných kategórií:

- **Filter metódy:** Vykonávajú hodnotenie vlastností čŕt na základe jednoduchých štatistických meraní. Sú efektívne z hľadiska výpočtov a ideálne pre analýzu veľkých dátových súborov, kde je prioritou efektívnosť spracovania.
- **Wrapper metódy:** Zahrnujú konštrukciu modelov s rôznymi kombináciami čŕt a ich následné hodnotenie podľa výkonnosti. Obvykle poskytujú lepšiu prediktívnu presnosť, keďže sú integrované priamo do procesu modelovania.
- **Embedded metódy:** Spájajú výhody filter a wrapper metód, pretože integrujú výber čŕt priamo do procesu trénovania modelu. Udržujú výpočtové náklady na prijateľnej úrovni, zatiaľ čo efektívne identifikujú črty, ktoré sú najviac prínosné v každej iterácii trénovania.

Podrobnejšie priblíženie špecifických metód výberu čŕt v strojovom učení pre jednotlivé skupiny:

Zisk informácií – Meria úbytok entropie, ktorý je indikátorom neistoty, pri predikcii cieľovej premennej. Zisk informácií sa vyhodnocuje pre každú premennú, aby sa zistilo, ako efektívne dokáže zredukovať neistotu o cieľovej premennej.

Chi-kvadrát test – Hodnotí nezávislosť medzi kategorickými premennými. Pomáha identifikovať závislosť prítomnosti alebo absencie črty od cieľovej premennej, čím umožňuje výber najrelevantnejších čŕt pre model. Skóre chi-kvadrátu medzi každou črtou a cieľovou premennou indikuje tie najrelevantnejšie črty, pričom premenné musia byť kategorické, nezávisle vzorkované a ich očakávané frekvencie musia byť vyššie ako 5.

Postupný výber čŕt – Začína s prázdnym modelom a postupne pridáva črty, ktoré najvýraznejšie zvyšujú jeho presnosť. Črty sú vyberané na základe najväčšieho zlepšenia modelu v každom kroku, pričom proces končí, keď pridanie ďalších čŕt už neprináša zlepšenie.

Spätný výber čŕt - Začína s modelom obsahujúcim všetky dostupné črty a postupne odstraňuje tie, ktoré najmenej prispievajú k jeho efektívnosti. Proces sa zastaví, keď odstránenie ďalších čŕt by zhoršilo výkon modelu.

Rekurzívne odstraňovanie čŕt - Začína s kompletným modelom a po natrénovaní sa analyzuje dôležitosť každej črty, čo sa môže realizovať pomocou koeficientov pri lineárnych modeloch alebo metrík ako je 'feature_importances'. Črty sa postupne eliminujú na základe ich významnosti. Proces pokračuje na zredukovanom súbore čŕt, až kým nebude dosiahnutý požadovaný počet čŕt alebo ďalšie odstraňovanie by nezlepšilo výkon modelu.

Regularizácia Lasso - Ako L1 regularizácia, je embedded metóda využívaná na zabránenie nadmernému prispôsobeniu. Lasso pridáva penalizáciu k absolútnym hodnotám koeficientov v lineárnych modeloch, čím efektívne reguluje ich veľkosť a podporuje optimalizáciu modelu [66][67].

Kapitola 6

Existujúce riešenia a štúdie ohľadom profilácie autorov

V nasledujúcich podkapitolách poskytujeme systematický prehľad a kritickú analýzu súčasných výskumných prác v oblasti profilovania autorov. Preskúmame, ako rôzne prístupy využívajú strojové učenie na identifikáciu autorského štýlu, a poukážeme na ich prednosti, obmedzenia a potenciálne aplikácie. Okrem toho sa budeme venovať aj preskúmaniu štúdií, ktoré môžu prispieť k hlbšiemu pochopeniu tejto tematiky a poskytnúť podstatnú podporu pre naše teoretické a praktické závery. Tento multidisciplinárny pohľad nám umožní získať širšie pochopenie možných aplikácií profilácie autorov, ako aj identifikovať potenciálne oblasti pre ďalší výskum a rozvoj v tejto dynamicky sa rozvíjajúcej oblasti.

6.1. Profilácia autorstva z rôznych textov a jazykoch

Štúdia [68] prináša dôkladný pohľad na profilovanie autorov založené na analýze textových dokumentov, kde kľúčovým aspektom je klasifikácia veku a pohlavia autorov. Táto téma má významné uplatnenie v oblastiach ako marketing, forenzika a bezpečnostné aplikácie, pretože správne určenie týchto demografických údajov môže výrazne ovplyvniť prístupy a stratégie v týchto sektoroch.

V rámci svojej práce štúdia prezentuje rozsiahly prehľad relevantných štúdií a datasetov a skúma efektivitu systémov hlbokého učenia určených na profilovanie

autorov. Zistenia ukazujú, že väčšina datasetov v oblasti veku a pohlavia pochádza z blogov a Twitter správ, pričom dokumenty pokrývajú jazyky ako angličtina, španielčina, arabčina, holandčina, taliančina, portugalčina, turečtina a ruština. Diverzita v jazykoch a formátoch textov nás vedie k záveru, že je potrebné zohľadniť jazykové nuansy a kultúrne kontexty pri vývoji modelov.

Článok tiež zdôrazňuje, že v datasetoch existuje výrazná rozmanitosť v počte a type dokumentov, ako aj v ich rozdelení na trénovacie, vývojové a testovacie súbory. Tento fakt poukazuje na potrebnú obozretnosť pri manipulácii s datami, keďže nekonzistentnosť v dátach môže viesť k skresleným výsledkom. Výskum upozorňuje, že aj napriek zdaniu jednoduchých úloh, ako je klasifikácia pohlavia, nebola dosiahnutá očakávaná presnosť, čo poukazuje na skrytú komplexnosť profilovania veku a pohlavia.

Významným zistením je, že tradičné metódy strojového učenia, ako sú náhodné lesy alebo podporné vektory, preukázali väčšiu efektivitu v úlohách klasifikácie veku a pohlavia, ako môžeme vidieť v tabuľke č. 6.1 dané metódy sú efektívnejšie v úlohách klasifikácie veku a pohlavia v porovnaní s metódami hlbokého učenia, čo môže byť dôsledkom ich dlhodobejšieho vývoja a optimalizácie. Toto pozorovanie nás inšpiruje k integrácii a ďalšiemu rozvoju týchto osvedčených techník pri návrhu našich modelov profilovania autorov. Je dôležité, aby sme brali do úvahy, že dlhodobý vývoj a optimalizácia týchto tradičných metód môžu poskytnúť robustnejšie a spoľahlivejšie modely v porovnaní s relatívne novými prístupmi hlbokého učenia.

Diskusia na záver článku o budúcom smerovaní výskumu zdôrazňuje dôležitosť experimentovania s kombináciami rôznych parametrov v modeloch hlbokého učenia a poukazuje na nevyhnutnosť konzistentnosti v datasetoch a metodikách predspracovania. Tieto poznatky sú pre nás cenné, pretože naznačujú, že pre presnejšiu a spoľahlivejšiu klasifikáciu je kritické dôkladne zvážiť spôsob, akým sú dáta pripravené a analyzované.

Tieto zistenia a metodologické prístupy z [68] budeme využívať ako základ pre našu vlastnú metodiku v rámci tejto práce. Implementácia osvedčených strojových učebných techník spolu s inovatívnymi prístupmi z hlbokého učenia nám umožní vytvoriť robustnejší a adaptabilnejší systém na profilovanie autorov, čo je nevyhnutné

pre správnu aplikáciu v praxi v rôznych doménach. Táto integrácia zároveň otvára priestor pre budúce inovácie a zlepšenia v presnosti a efektívnosti systémov určených na profilovanie autorov.

Paper	**ML Metóda**	**Úloha a Jazyk**	**Presnosť v %**
Meina et al.	RF, NB	Pohlavie EN, Vek EN	59.21, 64.91
Santosh et al.	DT, SVM, MaxEnt	Pohlavie ES, Vek ES	64.73, 64.30
Lopez-Monroy et al.	Lib-LINEAR klasifikátor	Pohlavie EN	28.95
Alvarez-Carmona et al.	Lib-Linear SVM	Pohlavie EN	78.28
Modaresi et al.	Logistická regresia	Pohlavie EN	51.79
Deneva (no paper)	Neznáme	Pohlavie ES	73.21
Busger et al.	Lineárny SVM	Vek EN, ES	58.97, 51.79
Basile et al.	Lineárny SVM	Pohlavie EN	82.33
Daneshvar and Inkpen	Lineárny SVM	Pohlavie EN, ES	82.21, 82.00
Tellez et al.	Lineárny SVM	Pohlavie AR	81.70
Valencia et al.	Logistická Regresia	Pohlavie EN	84.32
Pizarro	Lineárny SVC	Pohlavie ES	81.72

Tabuľka 6.1: Prehľad výskumu(2022) v oblasti profilovania autorov pomocou metód strojového učenia (Preložená a upravená)[68]

6.2. Štúdia o stratégiách profilovania

Článok [69] poskytuje podrobnú analýzu súčasných techník strojového a hlbokého učenia, ktoré sa používajú na identifikáciu unikátnych charakteristík autorov na sociálnych sieťach. Táto analýza je obzvlášť relevantná pre moju bakalársku prácu, ktorej cieľom je profilovanie autorov podľa veku a pohlavia. Článok skúma rôzne demografické a psychologické aspekty, ako sú vek, pohlavie, osobnostné črty a materinský jazyk na základe analýzy publikovaných textových príspevkov.

Autori článku kategorizujú stratégie profilovania do troch hlavných skupín:

1. **Štylistické prístupy -** Zameriavajú sa na štruktúrnu, syntaktickú a lexikálnu analýzu písania. Tieto prístupy sú pre našu prácu kľúčové, pretože pomáhajú rozpoznať jemné jazykové signály, ktoré môžu naznačovať vek alebo pohlavie autora. Napríklad, mladší autori môžu používať modernejšie slangu a neformálny jazyk, zatiaľ čo používanie zložitej syntaxe môže naznačovať vyššie vzdelanie alebo staršie vekové skupiny.

2. **Obsahové prístupy -** Preskúmavajú tematickú a sémantickú štruktúru textov, čo je rozhodujúce pre identifikáciu pohlavia a veku. Určité témy alebo topiky môžu byť častejšie diskutované špecifickými pohlaviami alebo vekovými skupinami, čo pomáha pri ich identifikácii.

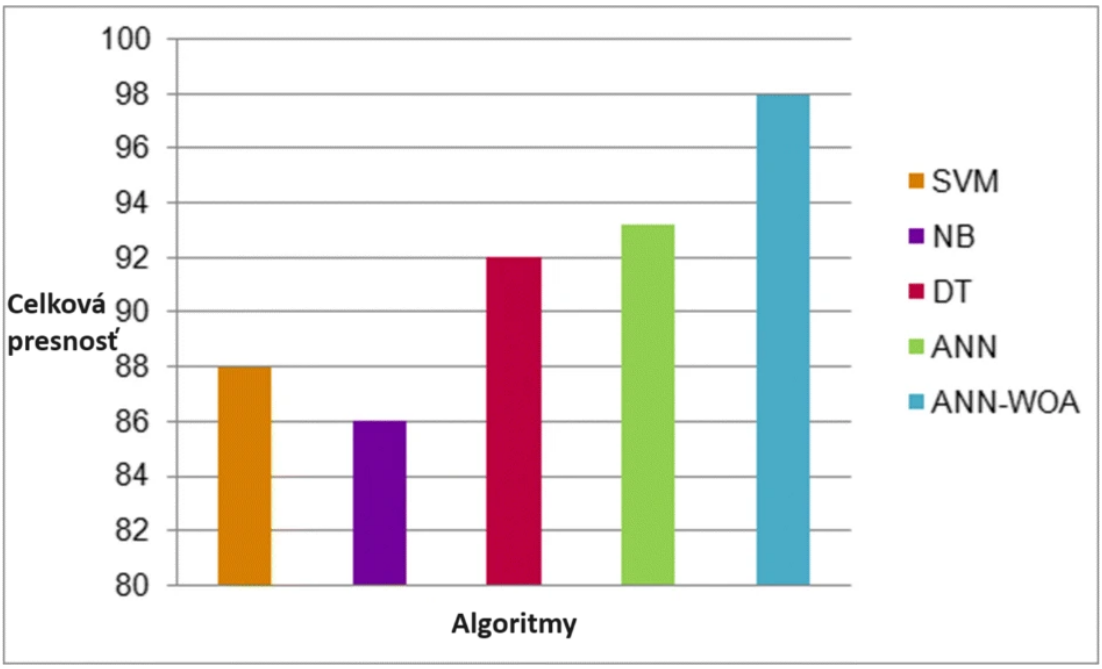
3. **Hybridné metódy -** Článok ukazuje, že kombinácia štylistických a obsahových stratégií môže výrazne zvýšiť efektivitu profilovacích systémov. Toto je mimoriadne užitočné pre našu prácu, pretože integrácia viacerých analytických techník by mohla viesť k presnejším predikciám veku a pohlavia.

Štúdia zdôrazňuje účinnosť rôznych pokročilých algoritmických techník, vrátane konvolučných neurónových sietí (CNN), rekurentných neurónových sietí (RNN) a optimalizačných algoritmov ako ANN-WOA (Artificial Neural Networks with Whale Optimization Algorithm). Tieto techniky sú nevyhnutné pre dosiahnutie vysoké presnosti pri identifikácii atribútov autorov a môžu byť priamo aplikované v našej práci na zlepšenie presnosti profilovania veku a pohlavia. Z obrázku č. 6.1 je zrejmé, že kombinácia neuronových sietí a optimalizačných algoritmov môže výrazne zlepšiť schopnosť modelu presne profilovať autora.

Diskusia v článku sa tiež zaoberá výzvami profilovania autorov v rôznorodom a dynamickom prostredí sociálnych sietí. Poukazuje na problémy, ako je nekonzistentnosť v databázach a potreba metód, ktoré dokážu spracovať multijazyčný a kultúrne rozmanitý obsah sociálnych médií. Tieto informácie môžu pomôcť pri výbere správnych prístupov pre riešenie problémov.

Záverom, článok nielen obohacuje moje pochopenie súčasného stavu technológií profilovania autorov, ale poskytuje aj priamu základňu, na ktorej môžeme stavať náš výskum. Dôraz na neustále inovácie v stratégiách profilovania by mohol inšpirovať

ďalšie skúmanie v našej práci na zlepšenie presnosti a spoľahlivosti systémov profilovania veku a pohlavia.



Obr. 6.1: Výsledok presnosť algoritmov štúdie (Preložený) [69]

6.3. Štúdia o vplyve predspracovania na profiláciu autorstva

Štúdia [70] predstavuje prísny metodologický prístup, ktorý využíva pokročilé techniky spracovania prirodzeného jazyka (NLP) a strojového učenia na vývoj multijazyčného klasifikačného systému zameraného na profiláciu autorov. Hlavným cieľom bolo presne predpovedať pohlavie a vek autorov pomocou klasifikátorov SVM, analyzujúc tweety v angličtine, holandčine, taliančine a španielčine. Metodológia zdôrazňovala zlepšenie kvality údajov prostredníctvom systematického predspracovania, čo je nevyhnutné pre efektívny tréning modelov a dosiahnutie presných predpovedí.

V štúdii bolo predspracovanie starostlivo plánované, v takej forme aby zabezpečilo integritu a užitočnosť údajov vstupujúcich do klasifikátorov SVM. Zahŕňalo

nasledujúce kroky:

1. **Tokenizácia a zmena na malé písmená -** Všetky tweety boli tokenizované rozdelením textu na medzery, čo pomáha štandardizovať textový formát naprieč rôznymi jazykmi. Jednotná zmena na malé písmená bola aplikovaná na zníženie variability rovnakých slov v rôznych prípadoch, čím sa štandardizovali vstupné údaje pre efektívnejšiu klasifikáciu.

2. **Odstránenie užívateľských mien a URL -** Tieto prvky boli odstránené, aby sa analýza sústredila čisto na jazykový obsah bez hluku zavedeného irelevantnými metadátami a hyperlinkami, ktoré nepomáhajú pri profilácii autora.

3. **Filtrovanie stop-slov -** Použitím NLTK pre angličtinu a špecifických balíčkov pre ostatné jazyky boli filtrované neinformatívne slová, ako sú predložky a spojky, čím sa model sústredil na zmysluplný obsah.

4. **Nepoužitie stemmingu a lematizácie -** Tieto techniky boli pôvodne testované, ale nakoniec neboli použité v konečnom modeli, aby sa zachovala pôvodná forma textu, ktorá je kľúčová pre zachytenie jazykových nuancií potrebných pre presné demografické predpovede.

Výber nástrojov v štúdii ako NLTK a špecifické balíčky jazyku Python bol motivovaný ich osvedčenou efektívnosťou pri riešení osobitostí jednotlivých jazykov, čo zaisťuje prispôsobený prístup k analýze textu.

Klasifikátor SVM, známy svojou schopnosťou zvládať dáta s vysokou dimenzionalitou a modelovať komplexné nelineárne hranice rozhodnutia, bol použitý. Modely boli hodnotené pomocou 3-násobnej krížovej validácie, čo zvyšuje robustnosť a zovšeobecnenie zistení. Výkonnostné metriky ukázali maximálnu priemernú presnosť 81.3 % pri klasifikácii pohlavia v španielčine a 70.3 % pri klasifikácii veku v angličtine, čo zdôrazňuje účinnosť zvolených techník predspracovania a výberu čŕt, obzvlášť použitie n-gramov slov a znakov. Bližšie výsledky si môžeme pozrieť v tabuľke č.6.2. Výsledky ukázali, že upravené modely s n-gramovými črtami dosahujú lepšie výsledky ako baseline modely s jednoduchými unigrammi. Zistilo sa, že predspracovanie dát a optimalizácia výberu čŕt môžu výrazne zlepšiť presnosť modelu. Na základe krížovej validácie bol model hodnotený ako efektívnejší o približne 10.8% pre klasifikáciu

pohlavia a 5.9% pre klasifikáciu veku v porovnaní s baseline modelmi. Hodnotenie modelov s použitím krížovej validácie ukázalo, že kombinácia rozličných techník predspracovania a selekcie čŕt môže mať pozitívny vplyv na presnosť autorovho profilovania.

Metodológie a zistenia z tejto štúdie sú priamo aplikovateľné a mimoriadne prospešné pre našu prácu v oblasti profilácie autorov, špeciálne v kontextoch predpovedania veku a pohlavia. Prísny metodologický rámec predstavený v tejto štúdii poskytuje robustnú šablónu pre vývoj sofistikovaných modelov NLP schopných spracovávať multijazyčné údaje – aspekt, ktorý je ústredný pre našu výskumnú prácu. Podrobný účet o technikách predspracovania ponúka cenné vhľady do optimalizácie prípravy dát na zvýšenie presnosti modelu.

Navyše, pozitívne výsledky získané z výberu čŕt slov a znakových n-gramov poskytujú osvedčenú cestu pre inžiniering čŕt v našej práci. Adopcia podobných čŕt by mohla zvýšiť schopnosť našich modelov rozlišovať jemné jazykové vzory, ktoré sú indikatívne pre vek a pohlavie autora, čím by sa zlepšila celková presnosť našich pokusov o profiláciu.

Táto integrácia metodológií a zistení významne obohatí našu bakalársku prácu, prispieva k hlbšiemu porozumeniu a efektívnejšiemu vykonaniu techník profilácie autorov na základe veku a pohlavia.

Jazyk	Presnosť	F1-makro	Precíznosť	Senzitivita
Angličtina	0.618	0.614	0.615	0.615
Taliančina	0.652	0.604	0.627	0.619
Holandčina	0.542	0.517	0.544	0.542
Španielčina	0.754	0.744	0.759	0.741

Tabuľka 6.2: Výsledky klasifikácie podľa jazyka (Preložená)[70]

6.4. Štúdia - nástrojov na analýzu pohlavia a nebinárnosť

V súčasnej dobe dynamického pokroku je esenciálne zahrnúť nebinárne identity do celkového spektra pohlavných identít. Mnohé inštitúcie a technologické nástroje však

túto potrebu ignorujú, čo nás podnietilo k hľadaniu výskumu, ktorý sa tejto problematike venuje. Táto kapitola poskytuje prehľad štúdie [71], ktorá sa zaoberá hodnotením troch analyzátorov pohlavia — UCLASSIFY, READABLE a HACKERFACTOR — a poukazuje na ich zaujatosť voči nebinárnym osobám. Tieto nástroje identifikujú len binárne pohlavie (muž a žena), čo vedie k častému nesprávnemu zaradeniu nebinárnych textov.

V štúdii boli použité rozsiahle datasety z Redditu a Tumblr na analýzu pohlavia, kde dataset z Redditu obsahoval 660 000 komentárov a dataset z Tumblr 2,05 milióna príspevkov. Texty boli napísané jednotlivcami, ktorí sa identifikovali ako muži, ženy a nebinárne osoby. Anotácie pohlavia boli buď priamo od používateľov, alebo boli odvodené z kontextu subredditov a blogov.

Testované nástroje v štúdii na analýzu pohlavia zahŕňali:

1. **UCLASSIFY -** Webová služba s ponukou viac ako 120 textových klasifikátorov, vrátane analýzy pohlavia
2. **READABLE -** Nástroj na hodnotenie čitateľnosti textu, ktorý zahŕňa funkcie na analýzu pohlavia
3. **HACKERFACTOR -** Open-source platforma používajúca štatistiku slovnej zásoby na určenie pohlavia

Okrem toho v spominanej študii bol použitý predtrénovaný model BERT, ktorý bol jemne upravený na predpovedanie mužských, ženských a nebinárnych pohlavných štítkov, čím predstavuje moderný prístup k riešeniu identifikovaných zaujatostí. Ďalej bol ChatGPT testovaný pomocou zero-shot promptov na simuláciu úloh klasifikácie pohlavia a na poskytnutie porovnávacej analýzy schopností AI.

Pri počiatočnom hodnotení textov pre binárnym pohlavím UCLASSIFY vykázal najvyššiu presnosť, no prejavil zaujatosť voči ženským predpovediam. READABLE mal najnižšiu celkovú presnosť, zatiaľ čo HACKERFACTOR preukázal vyváženejšie, ale stále nedostatočné výsledky.

Kľúčovým zistením bolo, že všetky platformy nesprávne klasifikovali nebinárne komentáre prevažne ako ženské. READABLE identifikoval malý podiel nebinárnych komentárov ako ”neutrálnych”. Naopak, jemne upravený model BERT dosiahol na datasete Reddit celkovú presnosť 84.4%, s výnimočnou presnosťou 93% pri nebinárnych komentároch. Jemné ladenie na zmiešanom datasete z Redditu a Tumblr zlepšilo generalizovateľnosť modelu, dosahujúc až 90% presnosť pre nebinárnych autorov. Celkovo model BERT prevýšil existujúce nástroje, najmä pri nebinárnych textoch. Audit ChatGPT ukázal, že jeho výkonnosť sa vyrovnala modelu BERT na datasete Reddit, ale na datasete Tumblr bola menej efektívna. Celková presnosť ChatGPT bola 58%, čo naznačuje potenciál, ale zároveň potrebu ďalšieho vylepšenia veľkých jazykových modelov pre analýzu pohlavia.

Táto štúdia je mimoriadne relevantná pre náš výskum profilovania autorov, najmä pokiaľ ide o vek a pohlavie. Metodológie a zistenia môžu informovať viaceré aspekty tohto odvetvia:

1. **Detekcia zaujatosti -** Techniky auditu použité v tejto štúdii môžu pomôcť identifikovať zaujatosti vo nástrojoch na klasifikáciu veku a pohlavia
2. **Spracovanie datasetu/dát -** Postrehy týkajúce sa vytvárania vyvážených datasetov pre nebinárne osoby môžu byť aplikované na zahrnutie rozmanitých vekových skupín a pohlaví v datasetoch pre profilovanie autorstva
3. **Jemné ladenie modelov: -** Úspech modelu BERT pri zlepšovaní presnosti a spravodlivosti naznačuje, že podobné prístupy môžu zvýšiť spoľahlivosť klasifikátorov veku a pohlavia
4. **Stratégie evaluácie: -** Porovnanie tradičných nástrojov a pokročilých modelov ako BERT a ChatGPT poukazuje na dôležitosť komplexných evaluačných stratégií na zabezpečenie robustnosti klasifikátorov

Začlenenie týchto zistení môže vylepšiť náš výskum v oblasti profilovania autorov, čo môže viesť k presnejšej a inkluzívnejšej identifikácii charakteristík autorov, ako sú vek a pohlavie. Ak by sa to podarilo tak by to celkovo prispelo k spravodlivejším aplikáciám AI v profilovaní pohlavia a veku.

6.5. Záver analýzy existujúcich riešení a štúdií

V tejto sekcii sme poskytli systematický prehľad a kritickú analýzu súčasných výskumných prác v oblasti profilovania autorov. Zistenia z týchto štúdií nám poskytli hodnotné poznatky o rôznych prístupoch a technikách používaných na identifikáciu autorského štýlu a ich demografických charakteristík.

Analyzované štúdie ukázali, že existujú rôzne metódy profilácie autorstva, od tradičných strojových učebných techník až po moderné hlboké učenie. Prehľad metód v [68] zdôraznil efektivitu tradičných metód, ako sú náhodné lesy a podporné vektory, pri klasifikácii veku a pohlavia. Tieto metódy, vďaka svojmu dlhodobému vývoju a optimalizácii, často dosahujú lepšie výsledky než hlboké učenie, čo nás vedie k integrácii týchto osvedčených techník do našich modelov profilovania autorov.

Na druhej strane, analýza pokročilých stratégií v [69] ukázala význam hybridných prístupov, ktoré kombinujú štylistické a obsahové analýzy textov. Tieto techniky, vrátane použitia konvolučných a rekurentných neurónových sietí, umožňujú efektívnejšie rozpoznávanie jazykových vzorcov a zvyšujú presnosť profilovacích systémov.

Štúdia [70] zdôraznila význam systematického predspracovania textov a optimalizácie výberu čŕt pre zvýšenie presnosti modelov. Ukázalo sa, že správne predspracovanie, ako je tokenizácia, odstránenie nepotrebných prvkov a filtrovanie stop-slov, môže výrazne zlepšiť výkonnosť klasifikačných modelov.

Dôležitým aspektom nášho výskumu je aj zahrnutie nebinárnych identít do profilácie pohlavia. Štúdia [71] ukázala, že existujúce nástroje majú tendenciu nesprávne klasifikovať nebinárne texty a zdôraznila potrebu moderných prístupov, ako je jemne upravený model BERT, ktorý dosahuje vyššiu presnosť a spravodlivosť.

Na základe týchto poznatkov sme schopní identifikovať silné stránky a obmedzenia existujúcich riešení a aplikovať ich pri vývoji našich modelov. Integrácia tradičných strojových učebných techník s inovatívnymi prístupmi hlbokého učenia nám umožní vytvoriť robustný a adaptabilný systém na profilovanie autorov. Zároveň

je dôležité venovať pozornosť systematickému predspracovaniu dát a zahrnutiu rôznych demografických a psychologických aspektov, aby sme zabezpečili presnosť a spoľahlivosť našich výsledkov.

Tieto zistenia nám poskytujú pevný základ pre ďalší výskum a vývoj v oblasti profilácie autorov, čo je kľúčové pre správnu aplikáciu v praxi v rôznych doménach. Zároveň otvárajú priestor pre budúce inovácie a zlepšenia v presnosti a efektívnosti systémov určených na profilovanie autorov.

Kapitola 7

Ciele práce

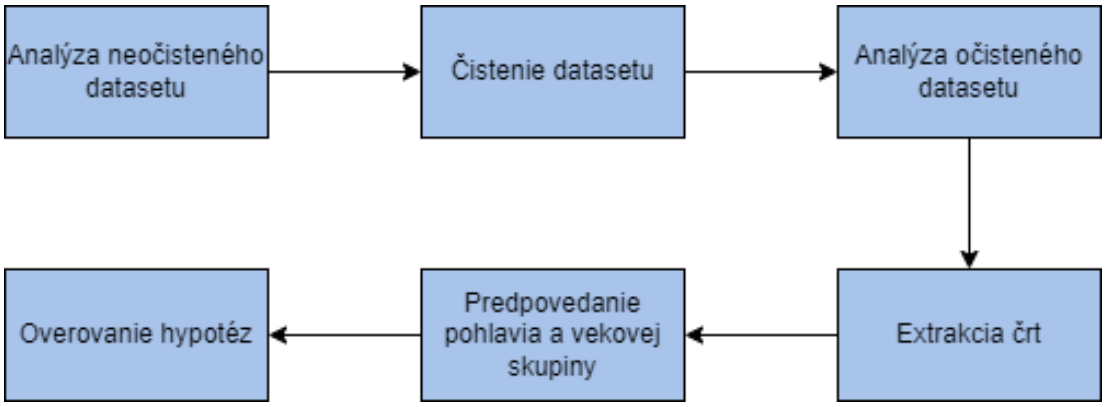
Hlavným cieľom našej bakalárskej práce je rozvíjať a aplikovať metódy profilácie autora textu na predpovedanie pohlavia a vekovej skupiny autorov. Na dosiahnutie tohto cieľa sa zameriame na sériu krokov, čo môžeme vidieť na obrázku 7.1. Zahŕňajú dôkladnú analýzu a predspracovanie dostupného datasetu, čo je opísané v kapitole 8. Úvodná analýza nám umožní identifikovať a riešiť potenciálne problémy súvisiace s kvalitou dát, ako sú nerelevantné hodnoty alebo dáta v inom ako anglickom jazyku, čím zvýšime presnosť a spoľahlivosť ďalšieho spracovania.

Po základnom predspracovaní dát bude nasledovať fáza extrakcie rozmanitých lexikálnych, syntaktických a sémantických čŕt z textov, čo opisujeme v kapitole 9. Tento proces je kľúčový pre následnú charakterizáciu autorských štýlov, ktoré sú základom pre našu prediktívnu analýzu.

V ďalšej fáze budem skúmať rôzne metódy predspracovania, vrátane prístupov bez predspracovania, s ošetrením odľahlých hodnôt, so škálovaním, a rôznych kombinácií týchto techník spolu s textovými reprezentáciami ako ngramy, tfidf ngramy, frekvencia slov a tfidf frekvencia slov. Následne aplikujeme vyvinuté metódy na predpovedanie pohlavia a vekovej skupiny v kapitole 10. Kľúčové črty, ktoré sme získali z našej extrakcie čŕt, použijeme v rôznych prediktívnych modeloch, pričom využijeme metódy výberu čŕt, ako sú forward selection a recursive feature elimination (RFE). Taktiež sa zameriame na optimalizáciu týchto modelov s ohľadom na ich hyperparametre, aby sme dosiahli čo najlepšie výsledky.

Okrem implementácie a testovania rôznych metód profilácie autora sa v rámci našej práce zameriame aj na overenie piatich konkrétnych hypotéz v kapitole 11. Tieto hypotézy sú základom pre naše experimentálne dizajny a analýzu výsledkov:

- H1: Rozmanitosť slovnej zásoby sa zvyšuje s vekom.
- H2: Priemerná dĺžka viet sa zvyšuje s vekom.
- H3: Ženy používajú viac slová s pozitívnym citovým sentimentom ako muži.
- H4: Algoritmus Random Forest za rovnakých podmienok dokáže priniesť lepšie výsledky ako Algoritmus Podporného Vekotorového Stroja (SVM) pri určovaniu pohlavia autora.
- H5: Syntaktické črty sú lepšie na zistenie vekovej skupiny autora ako lexikálne črty.



Obr. 7.1: Diagram procesu profilácie autora

Kapitola 8

Dataset

V našej práci využívame dataset používaný v rámci súťaže PAN, ktorá je zameraná na analýzu digitálnych textov a stylometriu. PAN skúma rôzne oblasti, ako sú určovanie autorstva, profilovanie autorov, detekcia plagiátov, identifikácia botov a mnoho ďalšieho v rôznych jazykoch a žánroch [72].

Využívame vzorku z datasetu zo súťaže z roku 2019, kde bol cieľom predpovedať demografické údaje celebrít, ako je pohlavie, rok narodenia, stupeň slávy a povolanie, na základe ich príspevkov z Twitteru [73]. Pre našu bakalársku prácu, ktorá sa zaoberá skúmaním vekovej skupiny a pohlavia, je z tohto dôvodu daný dataset veľmi vhodný. Celkový dataset bol veľmi rozsiahly a obsahoval profily 33 836 celebrít s až 3200 tweetmi. Avšak, kvôli obmedzeniam dostupných výpočtových prostriedkov sme pracovali iba s tridsiatimi percentami pôvodného datasetu.

Twitter je významnou platformou pre rýchle šírenie informácií a názorov. Často sa využíva na diskusie o aktuálnych udalostiach, značkové kampane a ako prostriedok pre celebrity a verejné osobnosti na udržiavanie kontaktu so svojimi sledovateľmi. Používatelia na Twitteri môžu zdieľať krátke správy, známe ako tweety, ale aj príspevky iných používateľov, čo sa označuje ako retweet. S maximálnym limitom 280 znakov podporuje Twitter stručnosť a bezprostrednosť komunikácie. Okrem textu môžu užívatelia do tweetov pridávať obrázky, videá, prieskumy, odkazy, označenia, emoji a ďalšie [74].

V nasledujúcich podkapitolách sa zameriame na analýzu vzorky, teda tridsať percent z tohto datasetu, ktorú ďalej budeme označovať iba ako náš dataset. Následne vykonáme proces čistenia datasetu od nežiaducich prvkov a analyzujeme ho opätovne.

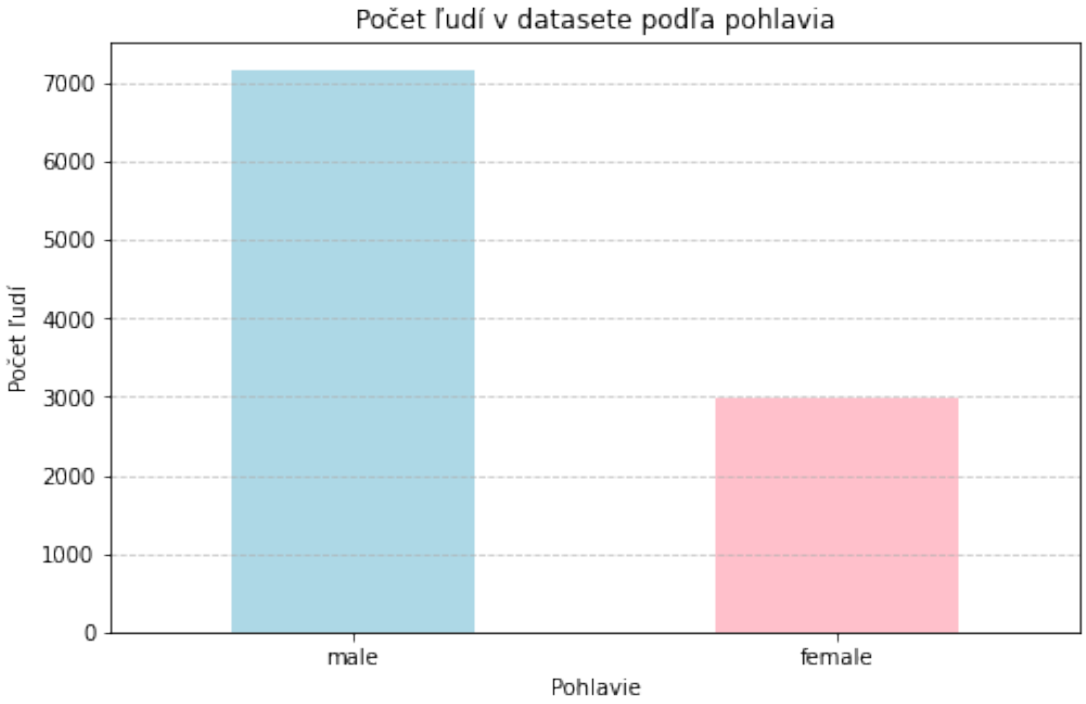
8.1. Analýza neočisteného datasetu

Našu analýzu datasetu sme začali charakteristikou jednotlivých stĺpcoch a zistením počtu záznamov. Dataset obsahuje 10 144 záznamov a žiaden z nich nevykazuje chýbajúce hodnoty (NaN) v žiadnom stĺpci. Stĺpce v datasete zahŕňajú „text“, ktorý obsahuje príspevky z Twitteru, jedinečné identifikačné čísla (id) pre každého používateľa, a „occupation“, kde sú reprezentované 8 rôznych profesii: športovci, umelci, tvorcovia, manažéri, politici, náboženské osobnosti, vedci a odborníci. Ďalej, stĺpec „gender“ klasifikuje záznamy ako muž alebo žena, zatiaľ čo „fame“ rozdeľuje osoby do troch kategórií: super hviezda, vychádzajúca hviezda a hviezda. Posledným analyzovaným stĺpcom je „birthyear“, s najnižšou hodnotou 1940 a najvyššou 2005, pričom priemerný rok narodenia činí 1977. Stĺpce „birthyear“ a „id“ sú typu int, zatiaľ čo ostatné stĺpce sú typu object. Navyše sme do datasetu pridali stĺpec „age“, vypočítaný ako aktuálny rok mínus rok narodenia, a stĺpec „age_category“, ktorý zoskupuje osoby do vekových kategórií: 0-29 rokov, 30-39 rokov, 40-49 rokov, 50-59 rokov, 60-69 rokov, 70-79 rokov a 80 rokov a viac. V nasledujúcich častiach sa budeme sústreďovať na analýzu stĺpcov „age_group“ a „gender“, keďže tieto atribúty sa budeme snažiť predpovedať pomocou algoritmov strojového učenia, ako aj na atribút text, ktorý bude slúžiť na extrakciu čŕt.

Najprv sme sa zaznamenali na počet mužov a žien v datasete, ktorý je prehľadne sumarizovaný v tabuľke č. 8.1 Zistili sme značnú nerovnováhu, keďže počet mužov výrazne prevyšuje počet žien. Tento rozdiel sme si vizuálne znázornili na grafe č. 8.1, čo nám umožňuje lepšie pochopenie tejto nerovnomernej distribúcie.

Pohlavie	**Počet**
Muž	7169
Žena	2975

Tabuľka 8.1: Počet ľudí v datasete podľa pohlavia

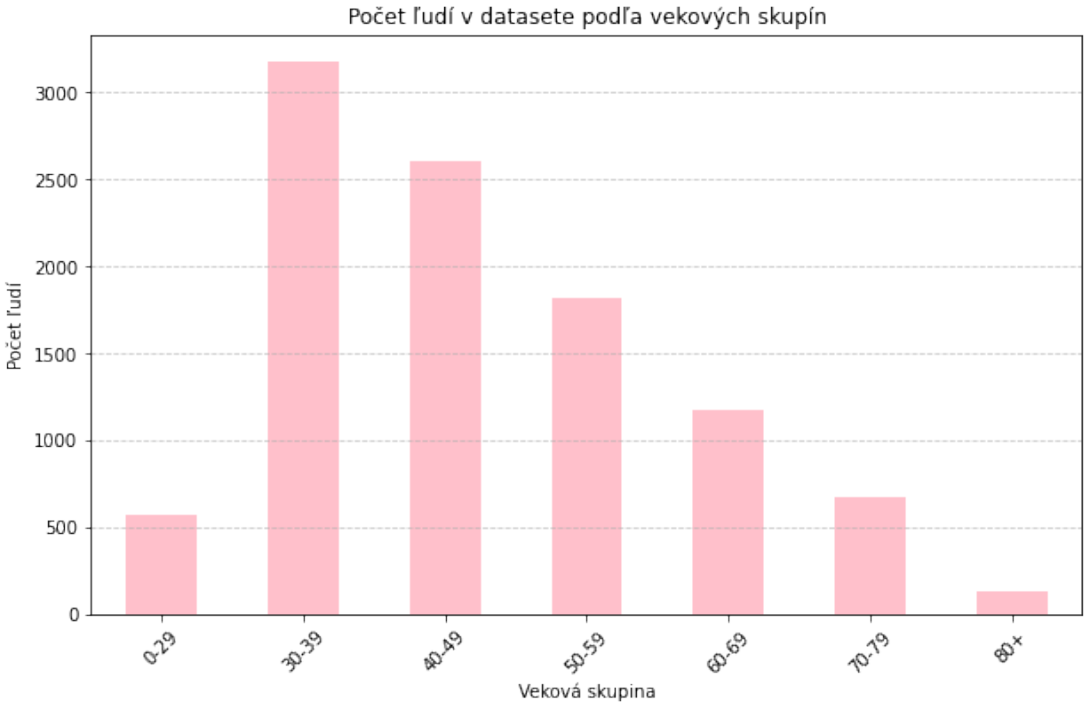


Obr. 8.1: Počet ľudí v datasete podľa pohlavia

Následne sme preskúmali rozdelenie jednotlivých vekových skupín v našom datasete, ktoré je podrobne zachytené v tabuľke č. 8.2. Podobne ako v prípade pohlavia, aj tu sme zistili, že distribúcia nie je vyvážená. Na grafe č. 8.2 je toto nerovnovážne rozloženie zreteľne viditeľné. Tento výsledok bol do istej miery očakávaný, keďže v reálnych dátach sa zriedka stretávame s perfektne vyváženým datasetom.

Veková skupina	**Počet**
0-29	571
30-39	3174
40-49	2601
50-59	1821
60-69	1172
70-79	676
80 a viac	129

Tabuľka 8.2: Počet ľudí v datasete podľa vekových skupín v neočistenom datasete)

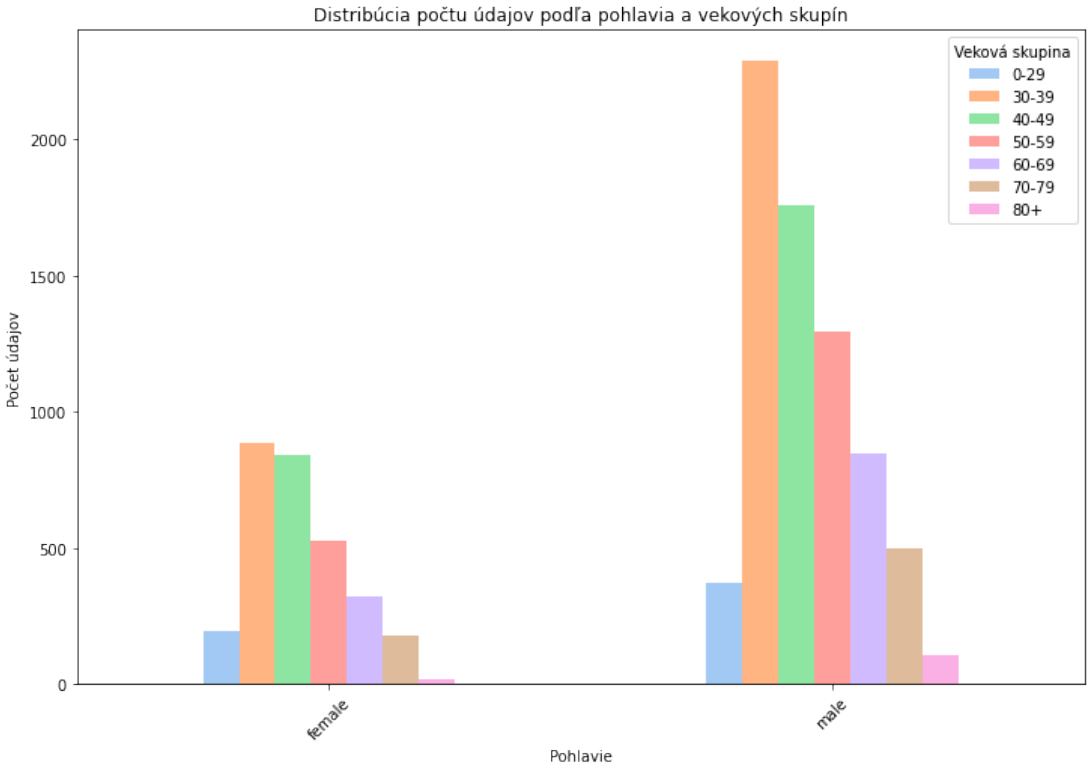


Obr. 8.2: Počet ľudí v datasete podľa vekovej skupiny

Ďalej nás zaujímali počty mužov a žien rozdelené do vekových skupín, ktoré môžeme vidieť v tabuľke č. 8.3. Z grafu č. 8.3 je zrejmé, že najväčšie zastúpenie majú muži vo vekových skupinách 30-39 a 40-49 rokov, kde ich počet výrazne prevyšuje počet žien. U žien je najviac zastúpená skupina 30-39 rokov. Tento graf 8.3 ďalej potvrdzuje naše predchádzajúce zistenia o nerovnomernom rozdelení údajov podľa pohlavia a veku v našom datasete. Dostali sme očakávané výsledky, ktoré potvrdzujú, že najväčšie zastúpenie obidvoch pohlaví je práve v najpočetnejších vekových skupinách.

	0-29	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80+**
Žena	197	886	842	529	324	176	21
Muž	374	2288	1759	1292	848	500	108

Tabuľka 8.3: Rozdelenie počtu príspevkov podľa pohlavia a vekových skupín

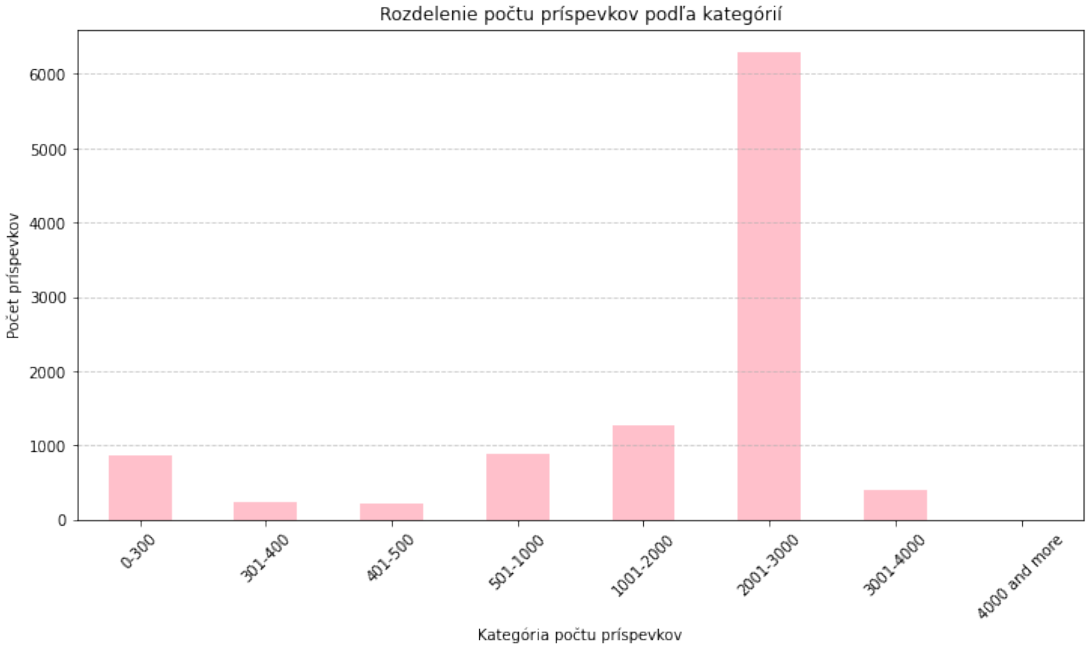


Obr. 8.3: Distribúcia počtu údajov podľa pohlavia a vekovej skupiny

Ďalším krokom v našej analýze bolo preskúmanie počtu príspevkov na osobu. Zistili sme, že priemerný počet príspevkov na osobu je 2193.95. Na jednom konci spektra je minimum s jediným príspevkom, zatiaľ čo maximum dosahuje 3000 príspevkov. Celkový počet príspevkov v našom datasete je 22 255 386. Aby sme lepšie pochopili distribúciu, rozdelili sme počty príspevkov do kategórií. Tieto kategórie sú podrobne zobrazené v tabuľke č. 8.4 a následne vizualizované na grafe č. 8.4. Analýza ukázala, že najväčší počet ľudí sa nachádza v kategórii s 2001-3000 príspevkami.

Kategoria príspevkov	Počet
0-300	867
301-400	227
401-500	208
501-1000	885
1001-2000	1262
2001-3000	6292
3001-4000	403
4000 and more	0

Tabuľka 8.4: Rozdelenie počtu príspevkov podľa kategórií



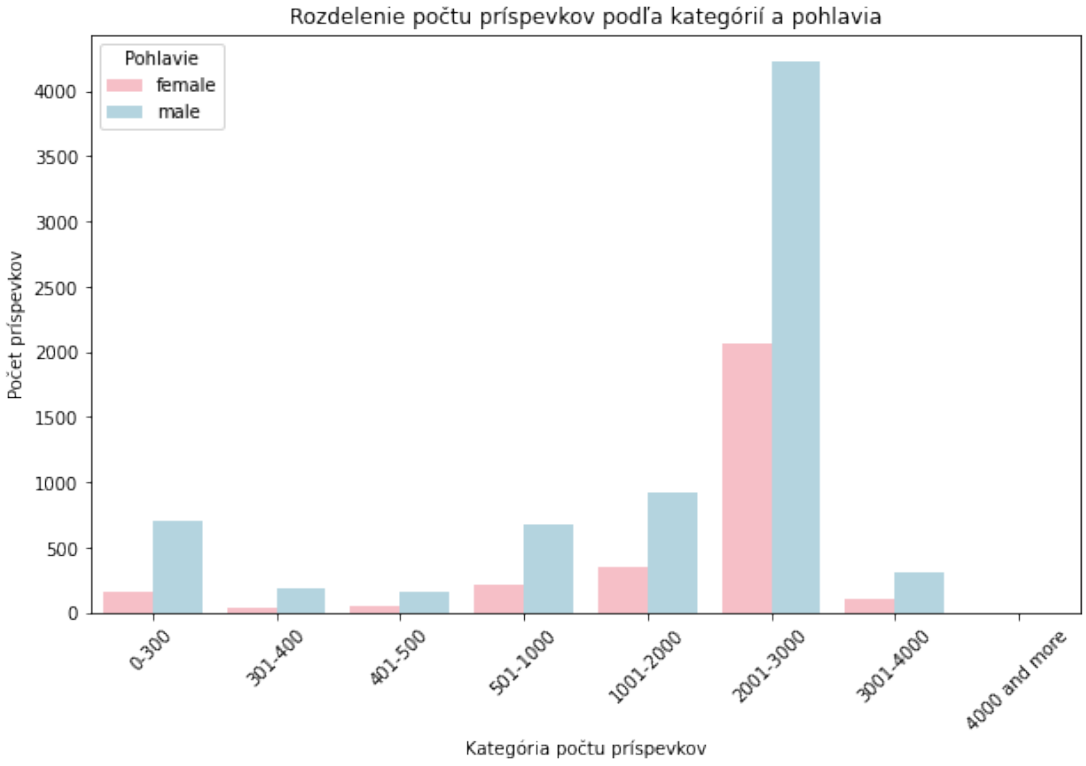
Obr. 8.4: Počet ľudí v datasete podľa počtu príspevkov

Pri analýze počtu príspevkov podľa pohlavia nás zaujímalo, či existuje výrazný rozdiel v aktivite medzi mužmi a ženami. Tieto údaje sme prehľadne zhrnuli v tabuľke č. 8.5 a vizuálne znázornili na grafe č. 8.5. Z výsledkov jasne vyplýva, že muži konzistentne prispievajú viac príspevkami než ženy. Domnievali sme sa, že tento výsledok je ovplyvnený distribúciou dát v datasete. Skúsili sme sa na to pozrieť inak, aby sme videli

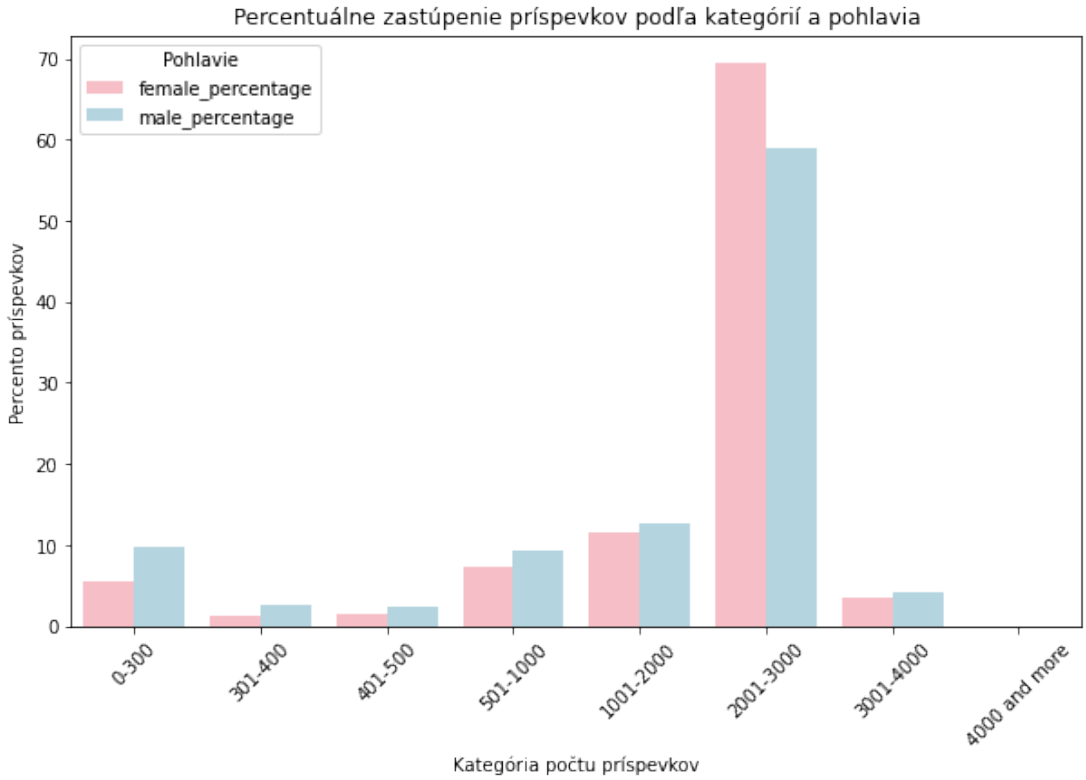
naozajstné zastúpenie pohlaví v kategóriách bez ohľadu na ich množstvo v datasete. Ako môžete vidieť na grafe č. 8.6 nižšie, tak rozdiely medzi pohlaviami v jednotlivých kategóriách sa zmenšili. Stále vo väčšine kategórií je viac mužov, však v najviac zastúpenej kategórií 2001-3000 prevažujú ženy.

	Žena	**Muž**
0-300	162	705
301-400	41	186
401-500	45	163
501-1000	214	671
1001-2000	346	916
2001-3000	2066	4226
3001-4000	101	302

Tabuľka 8.5: Počet údajov podľa pohlavia a kategórií



Obr. 8.5: Rozdelenie počtu príspevkov podľa kategorií počtu príspevkov a pohlavia

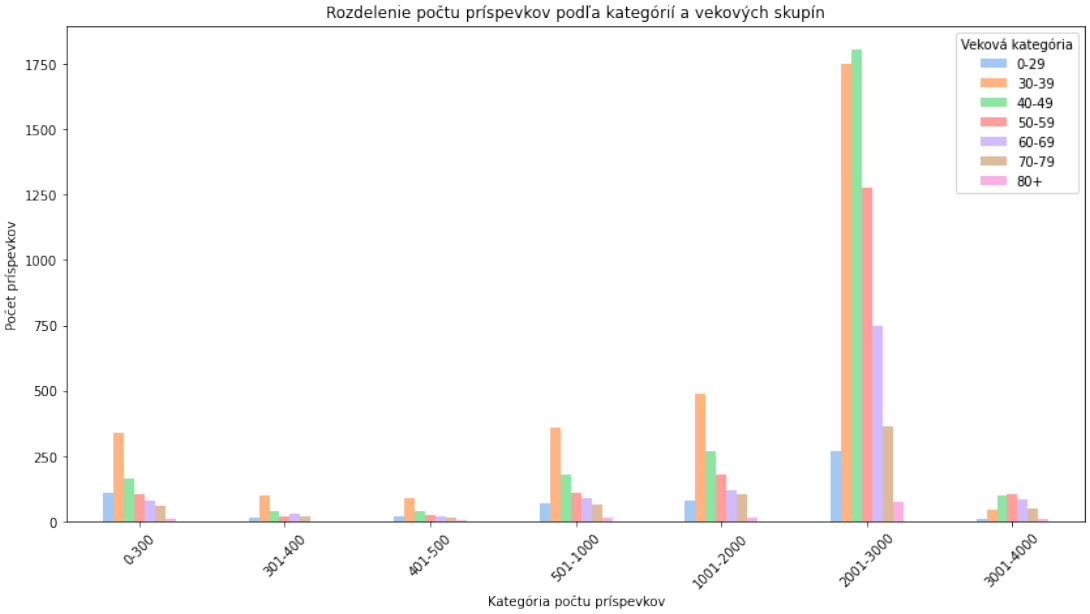


Obr. 8.6: Percentuálne zastúpenie príspevkov podľa kategorií počtu príspevkov a pohlavi

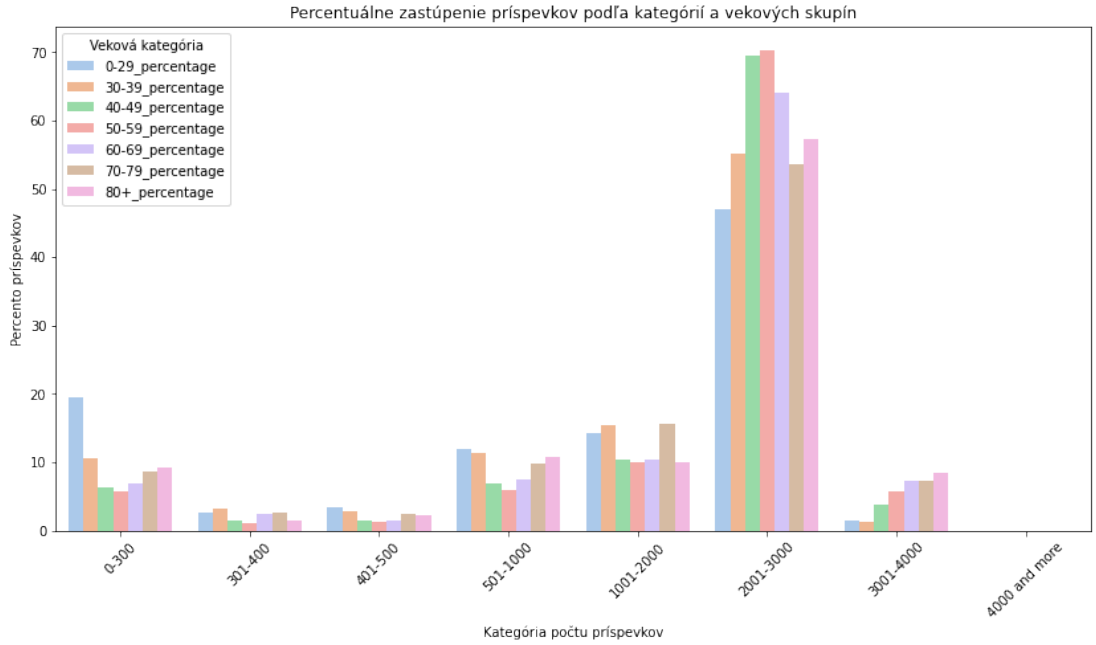
Pozreli sme sa na analýzu vzťahu medzi vekovými kategóriami a počtom príspevkov a zistili sme, že sú najviac zastúpené vekové skupiny 30-39 a 40-49 rokov. Tento nález je zrejmý z tabuľky č. 8.6 a grafu č. 8.7. Keďže tieto kategórie sú najpočetnejšie v našom datasete, rozhodli sme sa skontrolovať, či tento výsledok nie je skreslený nerovnomernou distribúciou dát. Preto sme analyzovali údaje spôsobom, ktorým sme eliminovali možné skreslenie. Výsledky z grafu č. 8.8 ukazujú, že rozdiely medzi vekovými kategóriami sa výrazne zmenšili. Zatiaľ čo mladšie a staršie vekové kategórie majú menej príspevkov, v stredných kategóriách, najmä v rozsahu 2001-3000 príspevkov, je aktivita vyššia. Najviac príspevkov, medzi 3001 a 4000, evidujeme u najstarších vekových kategórií.

	0-29	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80+**
0-300	111	339	163	104	80	58	12
301-400	15	101	41	21	29	18	2
401-500	15	89	39	24	18	16	3
501-1000	68	361	179	109	88	66	14
1001-2000	81	489	271	181	121	106	13
2001-3000	269	1751	1807	1279	750	362	74
3001-4000	8	44	101	103	86	50	11

Tabuľka 8.6: Rozdelenie počtu príspevkov podľa kategórií a vekových skupín



Obr. 8.7: Percentuálne zastúpenie príspevkov podľa kategorií počtu príspevkov a vekovej kategorie



Obr. 8.8: Percentuálne zastúpenie príspevkov podľa kategorií počtu príspevkov a vekovej kategorie

Pri detailnej manuálnej prehliadke nášho datasetu sme hľadali prítomnosť nežiaducich prvkov. Objavili sme veľké množstvo retweetov a tweetov, ktoré boli napísané v iných jazykoch než angličtine. Tieto prvky budeme riešiť v rámci procesu čistenia datasetu. Okrem toho sme narazili aj na hashtagy, URL adresy, označenia užívateľov, unicode značky, nové riadky a tabulátory. Niektoré z týchto prvkov môžu byť užitočné pre extrakciu čŕt, avšak je dôležité byť pri ich používaní opatrný a filtrovať ich, keď je to potrebné, aby nám neskreslili výsledky.

V závere analýzy môžeme konštatovať, že dáta v našom datasete sú nerovnomerne distribuované. Táto skutočnosť významne ovplyvňovala naše výsledky, najmä pri hľadaní rôznych závislostí, napríklad medzi počtom príspevkov a vekovou skupinou.

8.2. Čistenie datasetu

Pred začiatkom čistiaceho procesu sme zistili celkový počet príspevkov v datasete, ktorý bol 22273611. Tento údaj sme si zaznamenali, aby sme mohli sledovať, ako sa počet príspevkov mení počas jednotlivých fáz čistenia dát. Vzhľadom na obmedzené

výpočtové a pamäťové kapacity sme sa rozhodli upraviť dataset tak, že každý používateľ mohol mať maximálne 500 príspevkov. Ak mal niekto viac príspevkov, vzali sme do úvahy len prvých 500. Zároveň sme odstránili používateľov, ktorí mali menej ako 300 príspevkov, aby sme zabezpečili, že každý z nich má dostatočný počet príspevkov pre to aby extrahované črty mali pre nás výpovednú hodnotu. Po čistiacich procesoch bol nový celkový počet príspevkov 4597403.

Nasledovalo dôkladné čistenie datasetu, aby sme zabezpečili kvalitu a relevanciu dát pre predikčné modely. Ako prvé sme odstránili všetky retweety, čo sú príspevky začínajúce na ”RT”. Tieto príspevky nie sú originálnym obsahom daného používateľa a mohli by skresliť výsledky predpovedí. Po tomto kroku nám zostalo 3205101 príspevkov, čo znamená, že sme odstránili viac ako milión príspevkov.

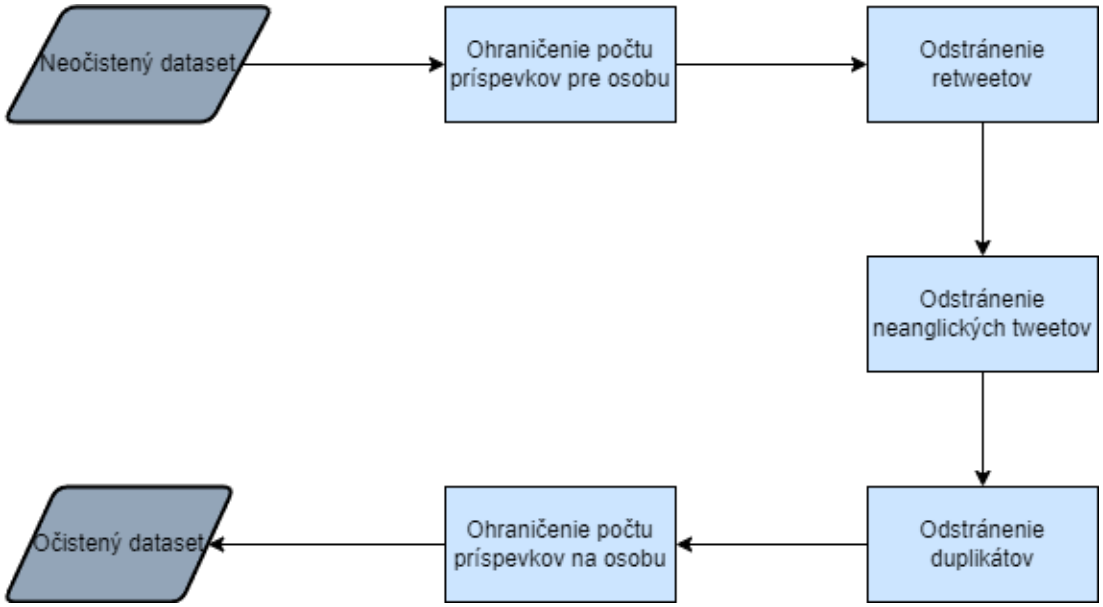
Následne sme sa zamerali na odstránenie tweetov v iných jazykoch ako angličtina, keďže štýl písania a význam môžu byť v rôznych jazykoch odlišné. Tento proces sme realizovali pomocou knižnice langdetect na detekciu jazyka každého tweetu. Ak knižnica nedokázala jazyk určiť, príslušný tweet sme označili hodnotou None. Tweet bol zachovaný len v prípade, že bol identifikovaný ako anglický. Po tomto kroku nám zostalo 2615856 príspevkov, čo znamená odstránenie viac ako pol milióna príspevkov.

Neskôr sme sa pokúsili identifikovať duplicitné riadky v datasete, avšak, ako ukázala predchádzajúca analýza, žiadne duplicity sme nenašli. Rozhodli sme sa odstrániť duplicitné príspevky na úrovni jednotlivých používateľov. Pri tejto technike sme si všetky unikátne príspevky, ktoré sme ešte neprešli, ukladali do nového zoznamu a tento nahrádzal pôvodný zoznam príspevkov. Po tomto kroku nám ostalo 2608448 príspevkov.

Záverečným krokom bolo opätovné spočítanie počtu príspevkov pre každého používateľa a odstránenie tých, ktorí mali menej ako 300 príspevkov. Po tejto úprave zostalo v datasete 1694502 príspevkov a 4536 používateľov. Celkovo sme počas čistiaceho procesu vymazali viac ako 20 miliónov príspevkov.

Celkový proces čistenia datasetu si môžeme pozrieť v diagrame č. 8.9 a záro-

veň nižšie v tabuľke č. 8.7 je zahrnutý každý krok čistenia a počet príspevkov, ktoré boli v rámci každého kroku odstránené, čo poskytuje jasný prehľad o rozsahu a vplyve jednotlivých čistiaceho krokov na našu dátovú sadu.



Obr. 8.9: Diagram procesu čistenia datasetu

Čistiaci proces	**Počet vymazaných príspevkov**
Ohraničenie počtu príspevkov na osobu	17 676 208
Odstránenie retweetov	1 392 302
Odstránenie neanglických tweetov	589 245
Odstránenie duplikátov	7 408
Ohraničenie počtu príspevkov na osobu	913 946
Celkový počet odstránených tweetov	**20 579 109**

Tabuľka 8.7: Počet vymazaných príspevkov pri čistiacom procese datasetu

8.3. Analýza očisteného datasetu

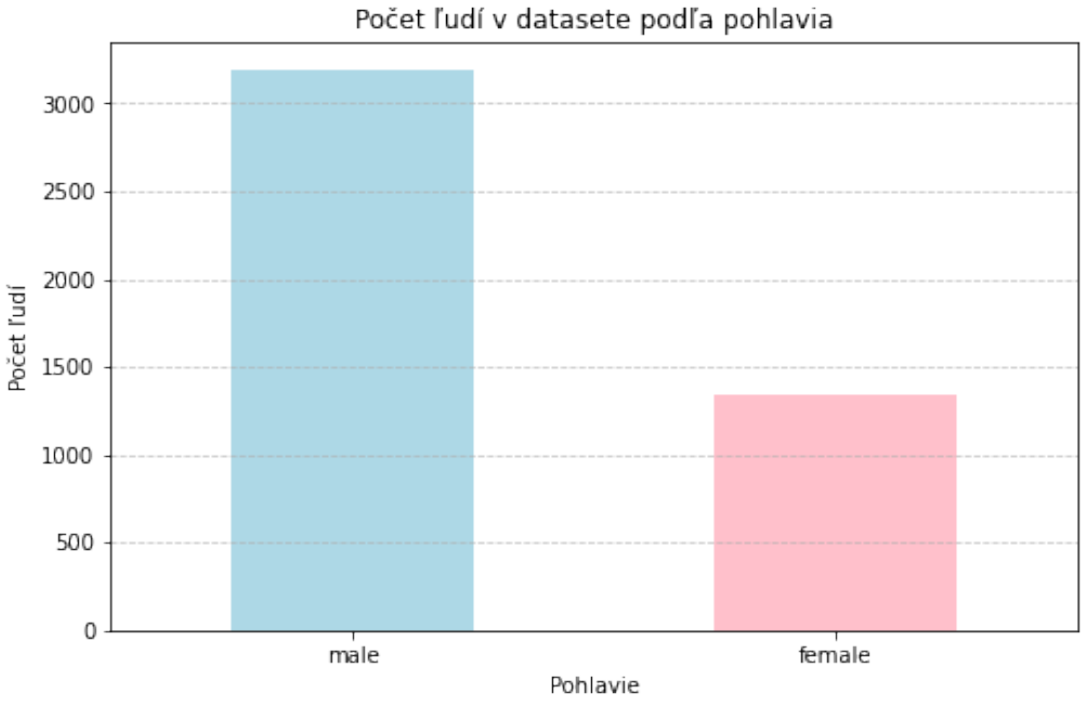
Po dôkladnom čistení datasetu je nevyhnutné znova analyzovať zmenšenú a upravenú dátovú sadu, s ktorou budeme ďalej pracovať. Počet záznamov v datasete sa znížil

na 4536, čo predstavuje výrazné zmenšenie o viac ako polovicu pôvodného počtu. Štruktúra stĺpcov v datasete zostala nezmenená oproti pôvodnému stavu, ktorý sme popisovali v predchádzajúcej analýze.

Čo sa týka distribúcie podľa pohlavia, aj keď počty mužov a žien sa znížili, nerovnomernosť v zastúpení pohlaví pretrváva. Tento nerovnovážny pomer je stále zrejmý z údajov v tabuľke č. 8.8 a na grafe č. 8.10. Hoci sa rozdiel medzi počtom mužov a žien mierne zmenšil, stále je výrazný, pričom muži predstavujú dominantnú časť datasetu.

Pohlavie	**Počet**
Muž	3191
Žena	1341

Tabuľka 8.8: Počet ľudí v datasete podľa pohlavia



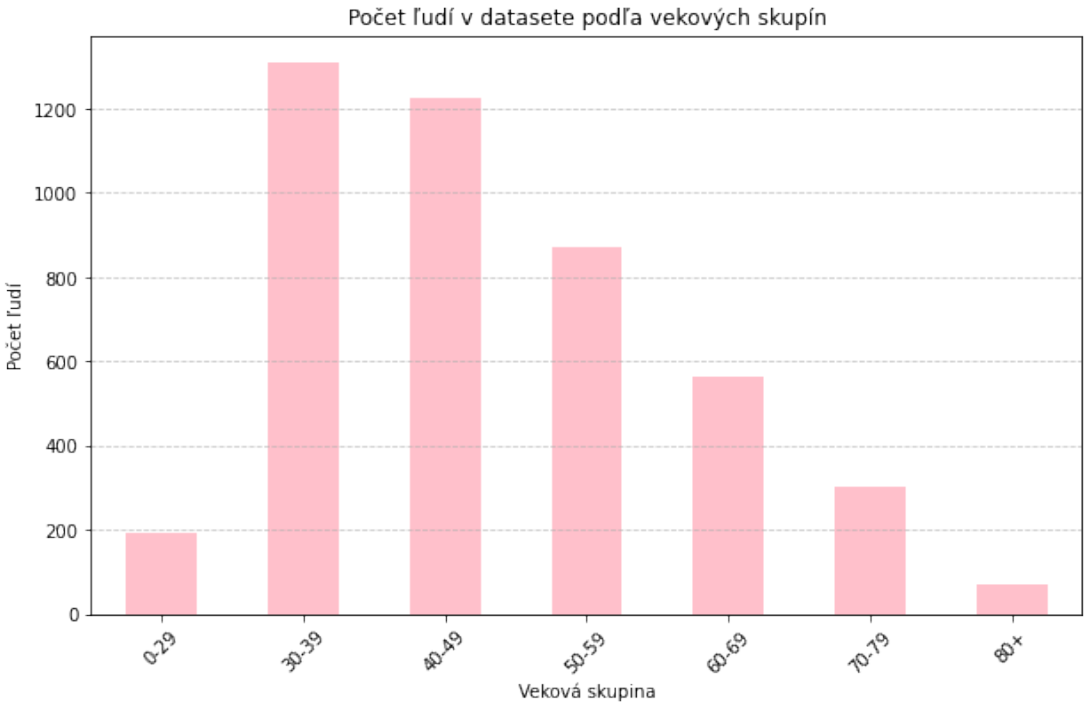
Obr. 8.10: Počet ľudí v datasete podľa pohlavia

Podobne ako v prípade pohlavia, aj počty záznamov v jednotlivých vekových kategó-

riách sa po čistení datasetu znížili. Avšak, napriek tejto zmene, distribúcia vekových kategórií ostala rovnako nerovnomerná ako pred čistením čo môžeme vidieť v tabuľke č. 8.9 a v grafe č. 8.11. Najviac záznamov sa naďalej nachádza v kategórii 30-39 rokov, čo ukazuje na pretrvávajúcu dominantnosť tejto vekovej skupiny v našom datasete.

Veková skupina	**Počet**
0-29	192
30-39	1309
40-49	1225
50-59	870
60-69	562
70-79	303
80+	71

Tabuľka 8.9: Počet ľudí v datasete podľa vekových skupín



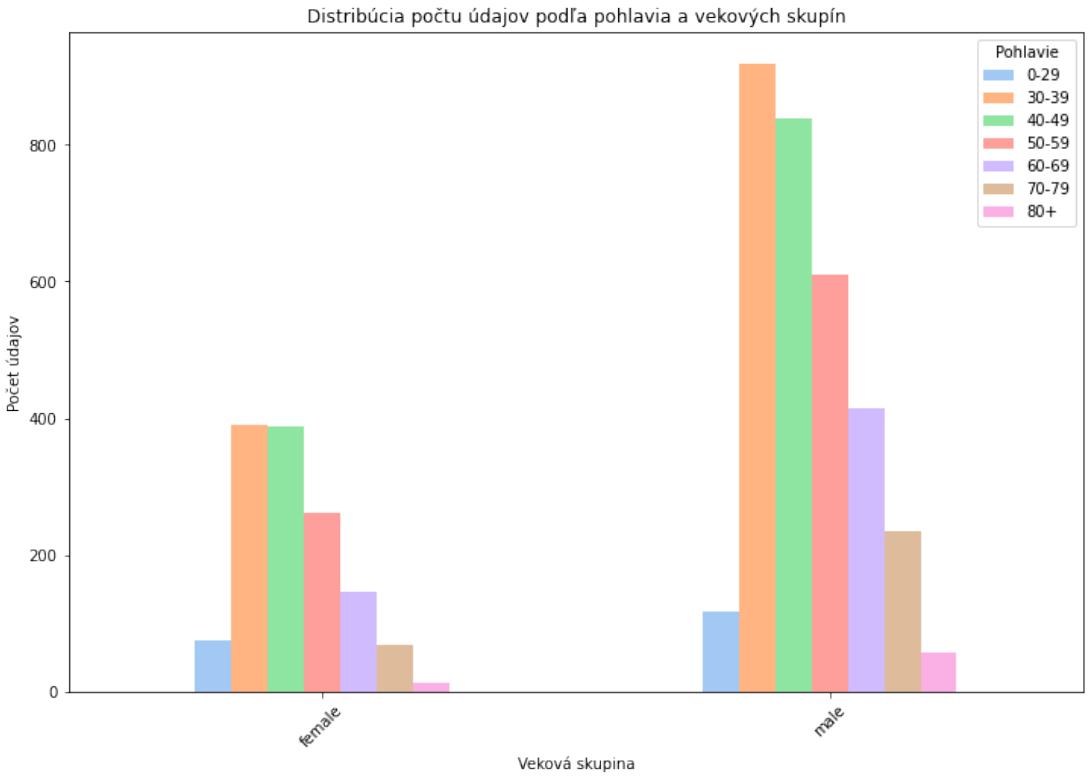
Obr. 8.11: Počet ľudí v datasete podľa vekovej skupiny

Ďalej sme skúmali, ako čistenie ovplyvnilo vzťah medzi vekovou kategóriou a pohlavím.

Zistili sme, že rozdiel v počte záznamov medzi vekovými kategóriami 30-39 rokov a 40-49 rokov sa v prípade oboch pohlaví zmenšil. Napriek tejto zmene však distribúcia pohlaví v rámci vekových kategórií ostáva stále veľmi nerovnomerná, čo je zrejmé z tabuľky č. 8.10 a z grafu č. 8.12.

Pohlavie	**0-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80+**
Žena	74	390	387	261	147	69	13
Muž	118	919	838	609	415	234	58

Tabuľka 8.10: Počet ľudí podľa pohlavia a vekových skupín



Obr. 8.12: Distribúcia počtu údajov podľa pohlavia a vekovej skupiny

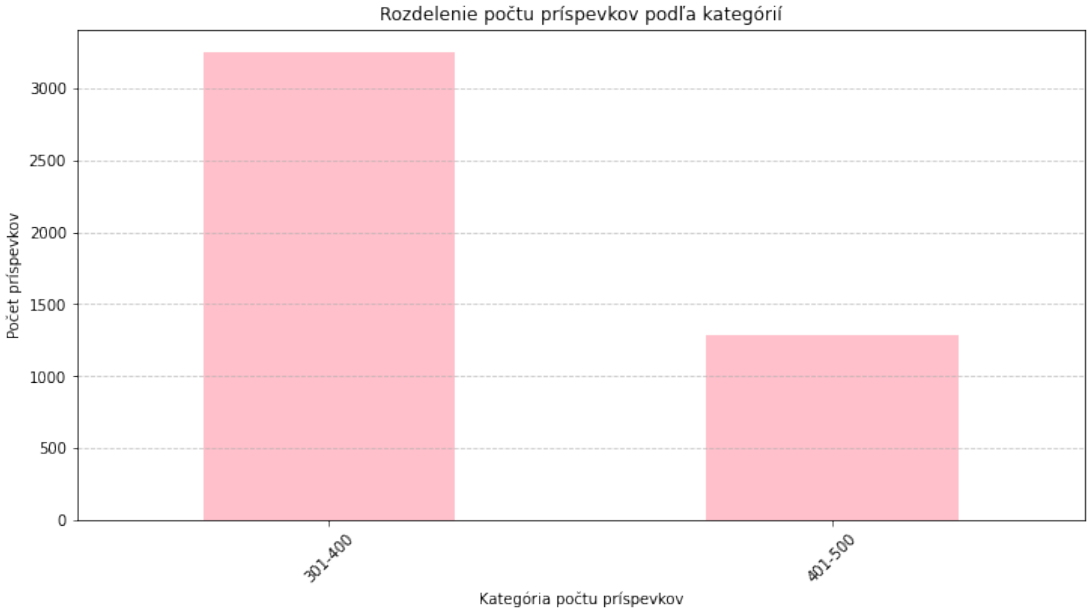
Následne sme sa zamerali na analýzu počtu príspevkov na jednotlivých používateľov. Celkový počet príspevkov sa ustálil na 1692975. Maximálny počet príspevkov, ktorý jeden používateľ zaznamenal, bol 498, zatiaľ čo minimálny počet bol 301. Priemerný počet príspevkov na osobu sa ustálil na 373.56. Všetky tieto hodnoty sa rapídne

zmenšili oproti neočistenému datasetu, však je to očakávané, keďže sme počet príspevkov každého používateľa ohraničili.

V rámci kategórií počtu príspevkov došlo k značným zmenám. Po čistení nám ostali len dve kategórie: 301-400 príspevkov a 401-500 príspevkov. Tieto kategórie, ako je vidieť v tabuľke č. 8.11 a na grafe č. 8.13, sú nerovnomerne distribuované. Kategória 301-400 príspevkov je výrazne viac zastúpená v porovnaní s kategóriou 401-500 príspevkov, čo naznačuje, že väčšina používateľov má nižší počet príspevkov.

Kategoria príspevkov	**Počet**
301-400	3245
401-500	1287

Tabuľka 8.11: Rozdelenie počtu príspevkov podľa kategórií



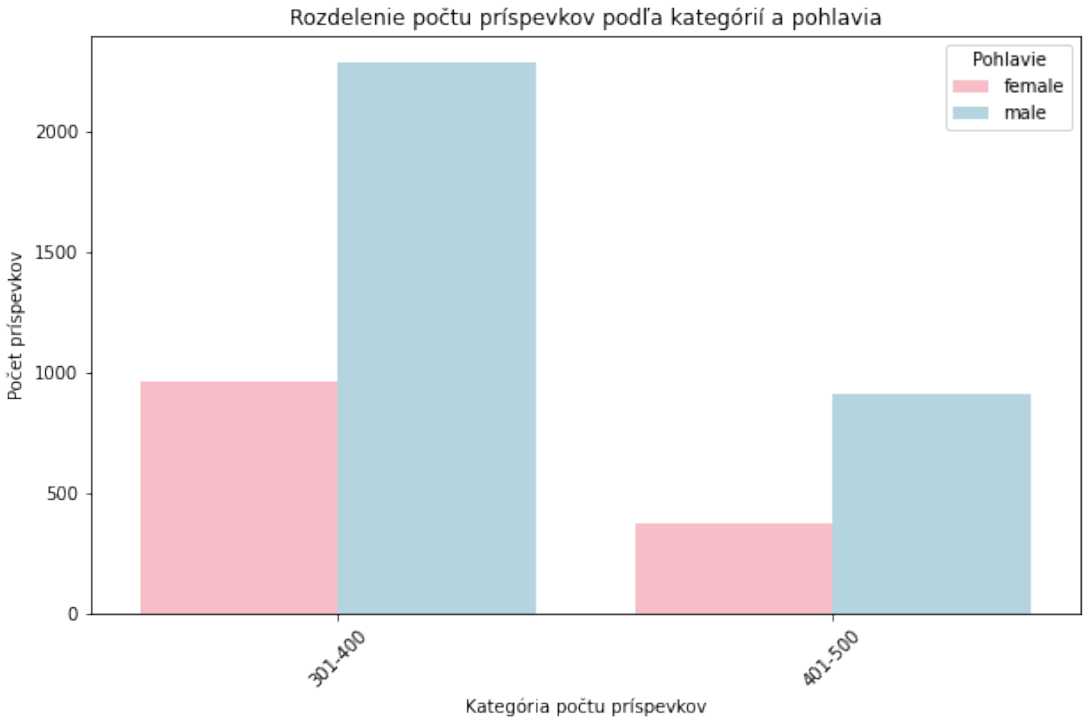
Obr. 8.13: Počet ľudí v datasete podľa počtu príspevkov

Znova sme analyzovali rozdelenie počtu príspevkov podľa kategórií a pohlavia, čo môžeme vidieť v tabuľke č. 8.12 a v grafe č. 8.14. Aj keď v oboch kategóriách prevládajú muži, zmenšili sa rozdiely medzi pohlaviami, hoci nerovnováha stále zostáva významná. Najviac zastúpená kategória je 301-400 príspevkov, čo naznačuje, že väčšina

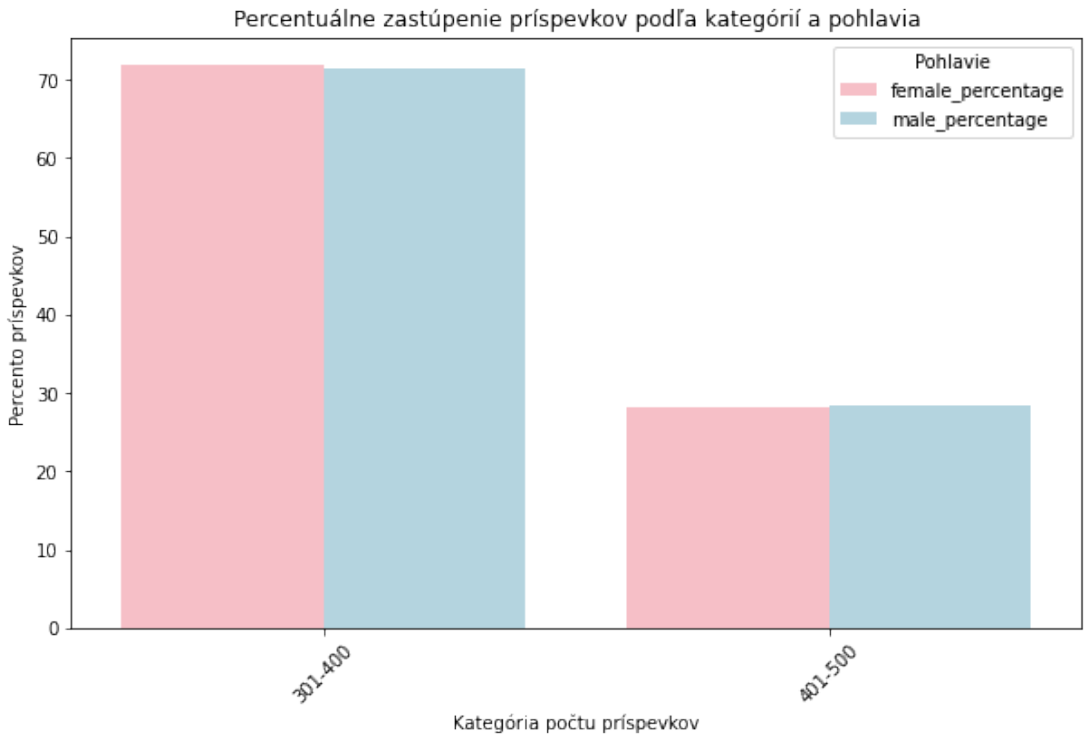
používateľov v našom očistenom datasete spadá do tejto skupiny. Ďalej sme sa rozhodli skúmať percentuálne rozdelenie príspevkov podľa kategórií a pohlavia, čo nám umožnilo eliminovať vplyv nerovnomernej distribúcie pohlaví v datasete. Výsledky tejto analýzy, ktoré vidíme na grafe č. 8.15 boli prekvapujúce: rozdiely medzi pohlaviami boli minimálne a obidve pohlavia mali podobnú početnosť v kategoriach, čo nám prezrádza to, že pohlavie nemá vplyv na počet príspevkov. Nález poukazuje na aký veľký vplyv má nerovnomerná početnosť jednotlivých skupín na výsledky.

Pohlavie	**301-400**	**401-500**
Žena	963	378
Muž	2282	909

Tabuľka 8.12: Rozdelenie počtu príspevkov podľa kategórií a pohlavia v očistenom datasete



Obr. 8.14: Rozdelenie počtu príspevkov podľa kategorií počtu príspevkov a pohlavia

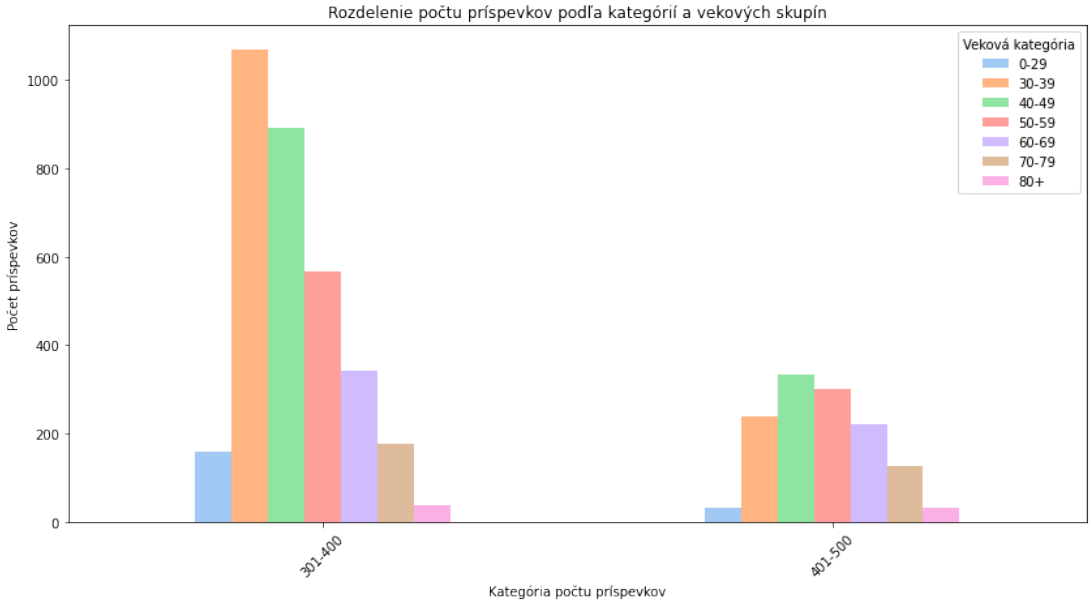


Obr. 8.15: Percentuálne zastúpenie príspevkov podľa kategorií počtu príspevkov a pohlavia

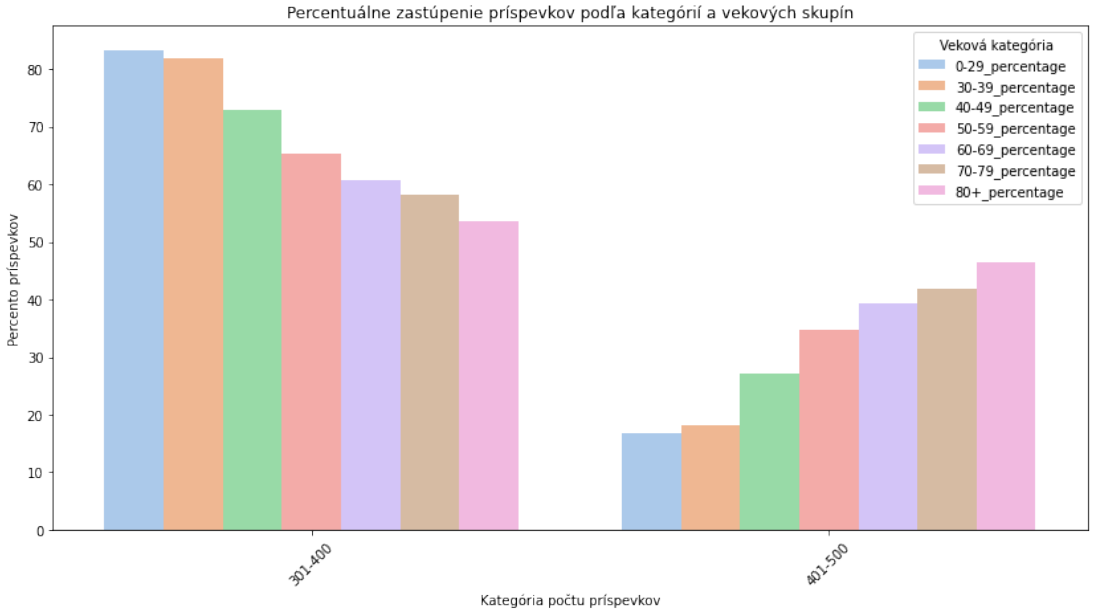
Na záver našej analýzy sme preskúmali, ako sú rozdelené vekové kategórie podľa počtu príspevkov, čo vidíme v tabuľke č. 8.13 a v grafe č. 8.16. Zistili sme, že najvyšší počet záznamov bol konzistentne v stredných vekových kategóriách v oboch kategóriách počtu príspevkov, čo zodpovedá bežnej distribúcii v našom datasete. Keď sme však zohľadnili vplyv nerovnomernej distribúcie dát a vykonali korekciu, výsledky sa značne zmenili, čo vidíme na grafe č. 8.17. Z analýzy grafu vyplynulo, že mladší ľudia majú tendenciu spadať do kategórie s menším počtom príspevkov, zatiaľ čo starší ľudia sú častejšie zastúpení v kategórii s vyšším počtom príspevkov. Tento nález môže odrážať rozdiely v používaní sociálnych médií medzi rôznymi vekovými skupinami, pričom starší ľudia môžu byť na platforme aktívnejší alebo dlhšie prítomní, čo im umožňuje nahromadiť viac príspevkov.

	0-29	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80+**
301-400	160	1070	892	568	341	176	38
401-500	32	239	333	302	221	127	33

Tabuľka 8.13: Rozdelenie počtu príspevkov podľa kategórií a vekových skupín v očistenom datasete



Obr. 8.16: Počet príspevkov podľa kategorií počtu príspevkov a vekovej kategorie



Obr. 8.17: Percentuálne zastúpenie príspevkov podľa kategorií počtu príspevkov a vekovej kategorie

V závere našej analýzy datasetu môžeme konštatovať, že aj napriek rozsiahlemu čisteniu datasetu zostala distribúcia dát vo veľkej miere rovnaká a nerovnomerná. Tento fakt zdôrazňuje, že prvotné charakteristiky našich údajov pretrvávajú aj po eliminácii mnohých záznamov a úprave dátových kategórií.

Kapitola 9

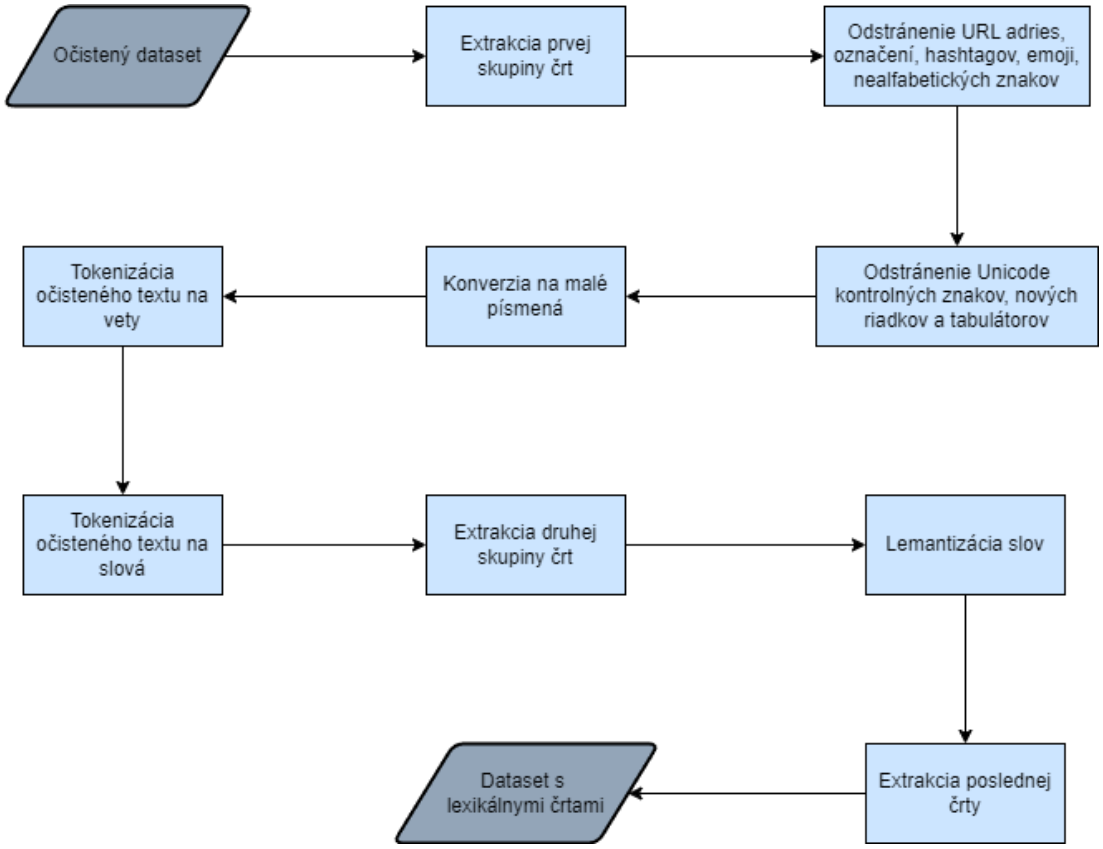
Extrakcia čŕt

Ďalším krokom bol proces extrakcie čŕt, kde sme sa zameriavali na lexikálne, syntaktické a obsahové črty textov. Na základe jednej z hypotéz, ktorá porovnáva predpovedanie vekovej kategórie pomocou lexikálnych a syntaktických čŕt, sme sa rozhodli tieto črty extrahovať samostatne. Celkovo sme identifikovali 50 čŕt: 21 lexikálnych, ktoré sa týkali primárne používania znakov a slov, 21 syntaktických, ktoré analyzovali gramatickú štruktúru textu, a 8 obsahových, zameraných na sentimenty v textoch. Celkový zoznam čŕt nájdeme v prílohách C. Po extrakcii každej skupiny čŕt sme z datasetu odstránili pomocné stĺpce, ktoré boli vytvorené počas tohto procesu, a nakoniec sme transformovali kategorické stĺpce na numerické formáty pomocou LabelEncoder, aby bolo možné s údajmi efektívne pracovať v algoritmoch strojového učenia. Detailnejšie informácie o extrahovaných črtách si rozoberieme v nasledujúcich podkapitolách.

9.1. Lexikálne črty

Extrakcia lexikálnych čŕt z očisteného datasetu prebehla v niekoľkých krokoch. Spočiatku sme sa zamerali na extrakciu tých čŕt, pri ktorých nebol potrebný zásah do obsahu tweetov. Následne sme prešli k podrobnej úprave textu, kde sme odstránili URL adresy, označenia používateľov, hashtagy, Unicode kontrolné znaky, nové riadky a tabulátory. Ďalej sme odstránili všetky nealfabetické znaky a nadbytočné medzery, a previedli text na malé písmená. Tieto úpravy sme aplikovali oddelene na slová, vety a celé tweety. Po týchto úpravách nasledovala druhá fáza extrakcie čŕt. Ako finálny krok sme realizovali lematizáciu slov, čo nám pomohlo štandardizovať rôzne tvary slov do ich základnej

formy. Extrahovali sme si poslednú črtu, ktorá sa týkala slovnej zásoby. Tento proces môžete vidieť v diagrame č. 9.1.



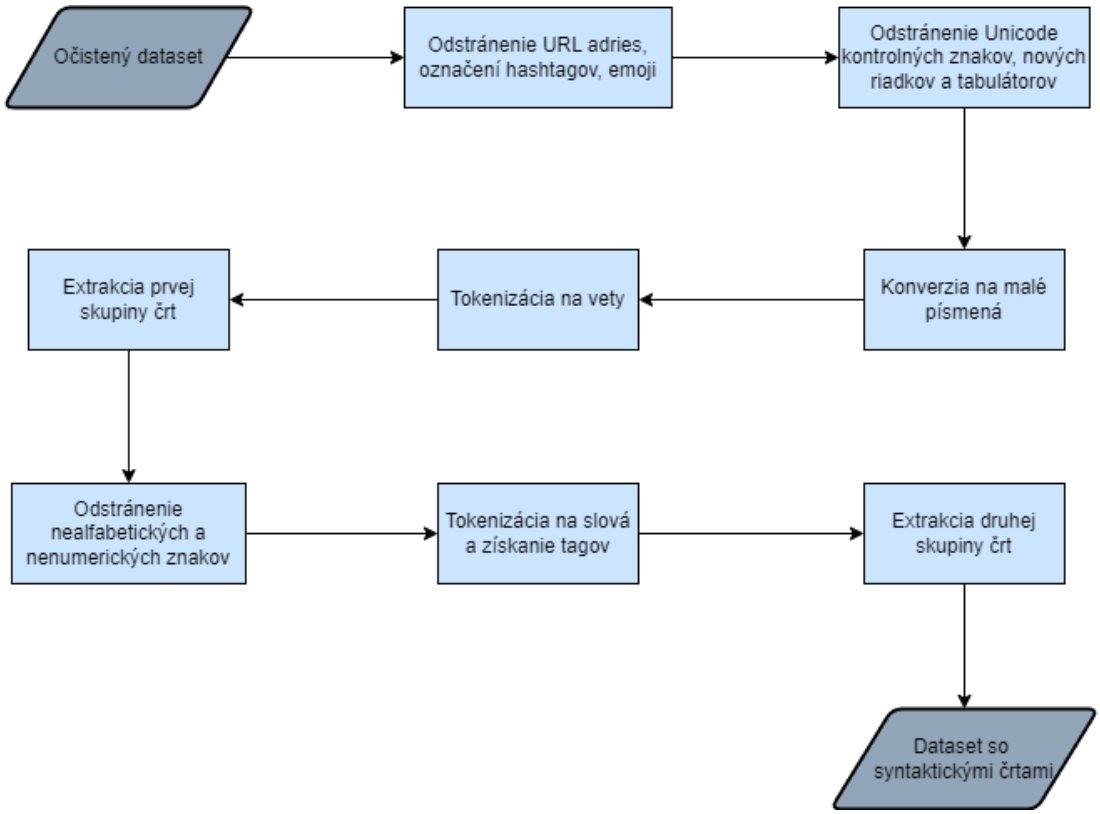
Obr. 9.1: Flowchart extrakcie lexikálnych črt

Naša prvá fáza extrakcie čŕt zahŕňala analýzu prvkov, ktoré sme plánovali v neskorších fázach odstrániť, ako sú URL adresy a emoji. Početnosť týchto prvkov sme normalizovali delením celkovým počtom tweetov, aby sme eliminovali rozdiely medzi záznamami s rozdielnym počtom tweetov. Prvým krokom bolo identifikovať počet emoji na tweet, pričom každý znak sme skontrolovali pomocou knižnice emoji. Podobne sme určili priemerný počet URL adries na tweet, kde sme vytvorili a aplikovali špecifický vzor pre URL. Tento prístup sme použili aj pre označenia a hashtagy, pričom sme definovali a hľadali špecifické vzory v tweetoch. Následne sme extrahovali črty zahŕňajúce priemerný počet číslic, veľkých písmen, malých písmen a nealfabetických znakov na tweet. Tieto kategórie sme identifikovali použitím funkcií Pythonu, ako je isalpha(). Zaujímali sme sa aj o počet uvodzoviek a apostrofov na tweet, ktoré môžu indikovať použitie

priamej reči alebo skratiek. Ďalšie analyzované črty zahŕňali počet interpunkčných znamienok a frekvenciu opakujúcich sa znakov ako ?.! v tweetoch, čo môže naznačovať špecifický štýl. Zistili sme aj počet stop slov na tweet pomocou knižnice nltk. Po ďalšej úprave sme analyzovali priemernú dĺžku slov, počet slov na vetu, počet znakov na vetu a počet viet na tweet, čím sme získali hlbšie pochopenie štruktúry textov. Zaujímavým nálezom pri manuálnom prezeraní datasetu bolo zistenie, že niektoré texty obsahovali opakované znaky v slovách (napríklad "wooow"), čo môže naznačovať špecifický štýl, tak sme si zistili frekvenciu tohto nálezu na tweet. V záverečnej fáze sme zistili slovnú zásobu každého používateľa, určenú počtom unikátnych slov použitých v tweetoch.

9.2. Syntaktické črty

Extrakciu syntaktických čŕt sme vykonali z čistej verzie datasetu. Začali sme predspracovaním tweetov, ktoré sme spojili do jedného dlhého textového reťazca. V tomto texte sme odstránili URL adresy, označenia používateľov, hashtagy, emoji, Unicode kontrolné znaky, nové riadky a tabulátory, a celý text sme previedli na malé písmená. Výsledkom bolo vytvorenie čistého textu, ktorý sme následne tokenizovali na vety. Po tokenizácii viet sme pristúpili k extrakcii prvej skupiny syntaktických čŕt, ktorá sa zamerala na nealfabetické znaky poskytujúce informácie o syntaxi textu. Po tejto fáze sme vykonali ďalšie úpravy textu, počas ktorých sme z predtým vyčisteného textu odstránili všetky nealfabetické a nenumerické znaky. V ďalšom kroku sme vytvorili nový stĺpec, v ktorom sme tokenizovali tento upravený text na jednotlivé slová a získali sme tiež gramatické kategórie (tagy) pre každé slovo. Následne sme uskutočnili ďalšiu extrakciu čŕt, ktorá bola zameraná predovšetkým na informácie získané z týchto tagov. Tento proces sme si vizuálne zobrazili v diagrame č. 9.2.



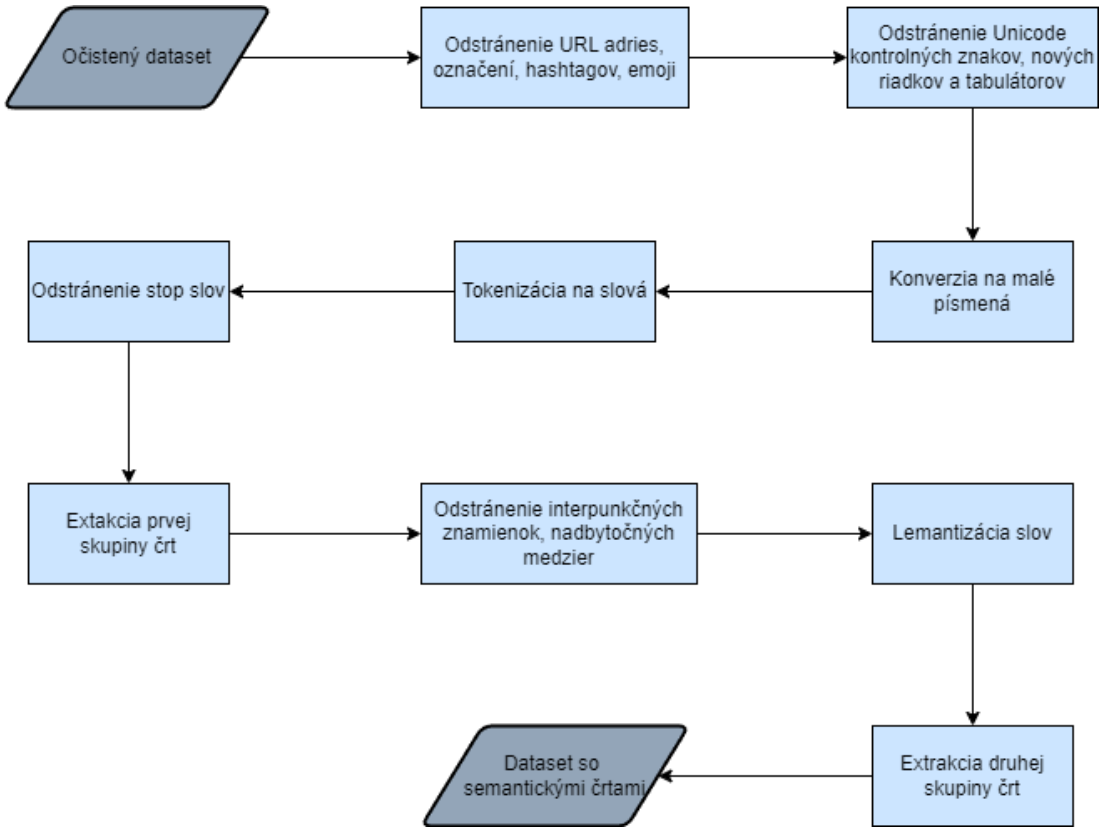
Obr. 9.2: Flowchart extrakcie syntaktickych črt

Ako som už predtým spomínala, na začiatku našej extrakcie čŕt sme sa zamerali na analyzovanie nealfabetických znakov v tweetoch. Naším cieľom bolo zistiť, ako používatelia ukončujú svoje vety – či používajú oznamovacie, opytovacie, rozkazovacie vety, alebo ich ukončujú iným spôsobom, prípadne vôbec. Pre každý typ vetného ukončenia sme zistili jeho percentuálny podiel. Ďalej sme skúmali používanie čiarok v textoch. V nasledujúcej časti extrakcie čŕt sme sa sústredili na gramatické tagy, ktoré nám poskytli informácie o slovných druhoch a gramatickom čase jednotlivých slov v tweetoch. Každý slovný druh mal pridelený zoznam tagov, napríklad podstatné mená boli identifikované tagmi ’NN’ (podstatné meno v singulári), ’NNS’ (podstatné meno v pluráli), ’NNP’ (vlastné podstatné meno v singulári), a ’NNPS’ (vlastné podstatné meno v pluráli). Z týchto tagov sme zisťovali priemerný počet výskytov jednotlivých slovných druhov na vetu. Súčasťou našej analýzy bolo aj skúmanie použitia rôznych časov v tweetoch. Zistili sme frekvenciu slovies v prítomnom čase, identifikovaných tagmi ’VBP’, ’VBZ’, ’VBG’, a minulom čase s tagmi ’VBD’ a ’VBN’. Pri analýze bu-

dúceho času sme narazili na značné výzvy, pretože slová typicky signalizujúce budúci čas, ako ”will” alebo ”going to”, nie sú v gramatických tagoch označované ako indikátory budúceho času. Napríklad, slovo ”will” je klasifikované ako modálne sloveso, čo komplikuje jeho jednoznačnú identifikáciu ako slovesa v budúcom čase. Kvôli týmto obmedzeniam a absencii spoľahlivej alternatívnej techniky na detekciu budúceho času sme sa rozhodli túto črtu neextrahovať. Následne sme sa zamerali na syntaktické závislosti, kde sme pomocou anglického jazykového modelu spaCy analyzovali lingvistické štruktúry v tweetoch. Každý tweet bol transformovaný na objekt, ktorý obsahoval informácie o lemach, gramatických kategóriách a syntaktických vzťahoch. Z týchto dát sme zistili početnosť rôznych syntaktických závislostí ako sú podmety, priame objekty, príslovkové určenia a prívlastky.

9.3. Sémantické črty

Extrakcia sémantických čŕt z nášeho očisteného datasetu prebiehala v niekoľkých krokoch. Na začiatku sme predspracovali texty spojením tweetov každého používateľa do jedného kontinuálneho textu. Počas tohto procesu sme odstránili URL adresy, označenia používateľov, hashtagy, emoji, Unicode kontrolné znaky, nové riadky a tabulátory, a text sme prekonvertovali na malé písmená. Následne sme text tokenizovali na jednotlivé slová a odstránili sme stop slová, čo sú často používané slová s nízkou informačnou hodnotou. Po týchto úpravách sme pristúpili k extrakcii hlavnej časti čŕt. Ďalší krok zahŕňal lematizáciu, pri ktorej sme previedli slová na ich základné gramatické formy. Po lematizácii sme odstránili všetky interpunkčné znamienka a nadbytočné medzery, čím sme zabezpečili čo najčistejšiu formu textu pre poslednú fázu extrakcie čŕt. Celý proces môžeme vidieť v diagrame č. 9.3.



Obr. 9.3: Flowchart extrakcie semantickych črt

Extrakcia sémantických čŕt z datasetu prebiehala v niekoľkých fázach. Ako prvé sme sa zamerali na zistenie priemerného počtu pozitívnych, negatívnych a neutrálnych slov na tweet. Na analýzu sentimentu slov sme použili nástroj SentimentIntensityAnalyzer(). Podľa výsledkov tohto analyzátora sme slová kategorizovali: slová s hodnotením sentimentu nad 0.1 boli klasifikované ako pozitívne, tie s hodnotením nižším ako -0.1 ako negatívne a slová s hodnotením medzi týmito prahmi ako neutrálne. Tieto údaje sme následne prepočítali na priemerné hodnoty na tweet. Ďalej sme skúmali priemerný sentiment jednotlivých používateľov tým, že sme sčítali sentimenty všetkých ich tweetov a vypočítali priemer pre každého používateľa. V rámci ďalšej fázy sme analyzovali počet entít, ako sú názvy organizácií a osobné mená, na tweet, k čomu sme využili jazykový model spaCy. Následne sme sa venovali extrakcii špecifických slovných skupín – negačných, kognitívnych a senzorových slov. Pre každú skupinu sme definovali zoznam príslušných slov, napríklad negačné slová zahŕňali ”not”, ”no”, ”never”, zatiaľ čo kognitívne zahŕňali ”think”, ”understand”, a senzorové slová ako ”see”, ”hear”. Zistili sme

početnosť týchto slov v tweetoch a spočítali ich priemerný výskyt na tweet.

Kapitola 10

Predpovedanie demografických vlastností pomocou strojového učenia

10.1. Pohlavie

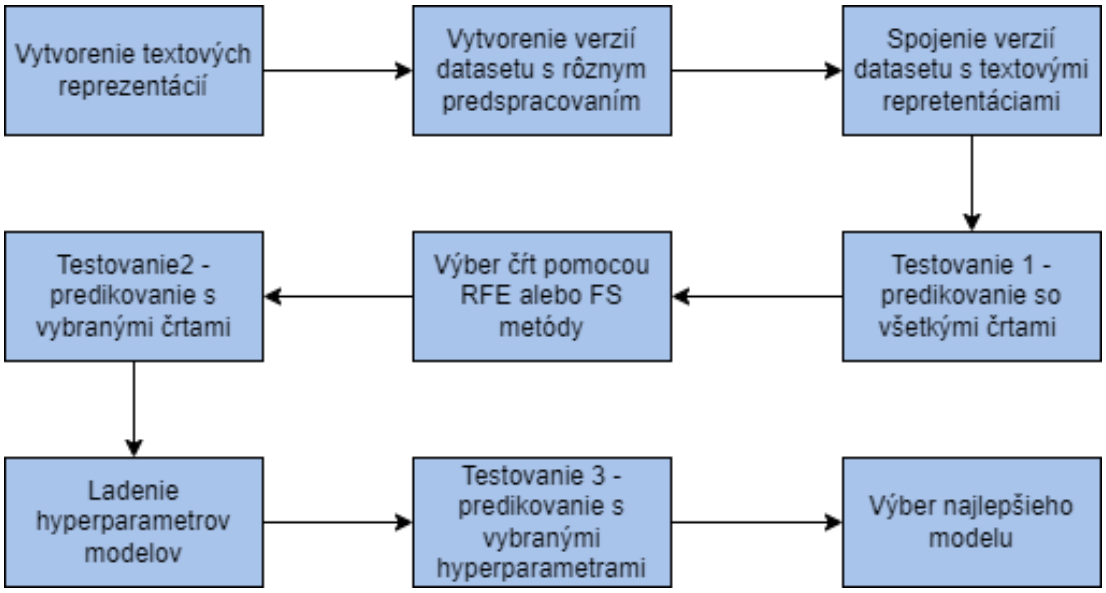
Pracovali sme s datasetom, ktorý vznikol spojením troch samostatných datasetov zameraných na extrakciu čŕt: dataset s lexikálnymi črtami, syntaktickými črtami a sémantickými črtami. Dataset obsahoval aj stĺpec s očisteným textom, ktorý bol lematizovaný a tokenizovaný na slová, čo nám umožnilo následne vytvárať textové reprezentácie. Okrem týchto čŕt dataset zahŕňal aj pohlavie, ktoré sme sa snažili predpovedať. Proces predikcie môžeme vidieť na diagrame nižšie 10.1.
Najprv sme dataset rozdelili na trénovaciu a testovaciu množinu v pomere 70 % k 30 %. Následne sme vytvorili rôzne varianty datasetu, aby sme našli optimálnu konfiguráciu pre algoritmus strojového učenia na predikciu pohlavia. Tieto varianty zahrňovali:

- Dataset bez ošetrenia outlierov a škálovania.
- Dataset bez ošetrenia outlierov a škálovania s n-gramami.
- Dataset bez ošetrenia outlierov a škálovania s TF-IDF n-gramami.
- Dataset bez ošetrenia outlierov a škálovania s word frequency.
- Dataset bez ošetrenia outlierov a škálovania s TF-IDF word frequency.
- Dataset s ošetrením outlierov.

- Dataset s ošetrením outlierov a s n-gramami.
- Dataset s ošetrením outlierov a s TF-IDF n-gramami.
- Dataset s ošetrením outlierov a s word frequency.
- Dataset s ošetrením outlierov a s TF-IDF word frequency.
- Dataset so škálovaním.
- Dataset so škálovaním a s n-gramami.
- Dataset so škálovaním a s TF-IDF n-gramami.
- Dataset so škálovaním a s word frequency.
- Dataset so škálovaním a s TF-IDF word frequency.
- Dataset s ošetrením outlierov a škálovaním.
- Dataset s ošetrením outlierov a škálovaním a s n-gramami.
- Dataset s ošetrením outlierov a škálovaním a s TF-IDF n-gramami.
- Dataset s ošetrením outlierov a škálovaním a s word frequency.
- Dataset s ošetrením outlierov a škálovaním a s TF-IDF word frequency.

Vykonali sme tri základné testovania, pričom sme skúšali všetky verzie datasetu na algoritmoch strojového učenia: Support Vector Classification, Random Forest a Gradient Boosting Machine.

- **Testovanie 1 :** Učenie a predikovanie pohlavia pomocou všetkých čŕt.
- **Testovanie 2 :** Učenie a predikovanie pohlavia s výberom čŕt. Porovnávali sme dve metódy výberu čŕt: Recursive Feature Elimination a Forward Selection.
- **Testovanie 3 :** Hyperparameter tuning pre algoritmy strojového učenia, vrátane učenia a predikovania s optimalizovanými hyperparametrami.



Obr. 10.1: Proces predikcie

10.1.1 Textové reprezentácie

Chceli sme vyskúšať viacero textových reprezentácií, aby sme zistili, s ktorou dosiahneme najväčšiu presnosť pri predpovedaní veku. Skúmali sme teda rôzne prístupy: n-gramy, tf-idf n-gramy, word frequency a tf-idf word frequency. Kvôli časovej a výpočtovej náročnosti sme sa rozhodli obmedziť maximálny počet slov/n-gramov. Toto obmedzenie zároveň slúži ako prevencia proti pretrénovaniu, pretože model neobsahuje príliš veľa riedkych čŕt, ktoré sú špecifické len pre trénovaciu sadu.

Keďže sme nevedeli presne odhadnúť, aká by bola vhodná hranica, pre každú textovú reprezentáciu sme vyskúšali hranice: 500, 1000, 2000, 3000, 4000 a 5000. Následne sme na algoritmy strojového učenia aplikovali predpovedanie veku iba na základe daných textových reprezentácií s konkrétnou hranicou. Podľa presnosti sme určili, ktorá hranica dosiahla najlepšie výsledky, a tú sme použili na vytvorenie danej textovej reprezentácie, ktorú sme neskôr kombinovali s rôznymi variantmi datasetu. Ak sa stalo, že niektoré hranice dosiahli rovnaké výsledky, zvolili sme tú menšiu, aby učenie a predpovedanie bolo rýchlejšie.

Príklad takého výstupu môžete vidieť na obrázku č. 10.2, ktorý znázorňuje vý-

stup textovej reprezentácie word frequency. Pre n-gramy, tf-idf n-gramy, word frequency a tf-idf word frequency sa najviac osvedčili hranice 3000 alebo 5000.

```
Model: SVC, Max Features: 500, Accuracy: 0.70
Model: Random Forest, Max Features: 500, Accuracy: 0.85
Model: Gradient Boosting Maschine, Max Features: 500, Accuracy: 0.88
Model: SVC, Max Features: 1000, Accuracy: 0.70
Model: Random Forest, Max Features: 1000, Accuracy: 0.84
Model: Gradient Boosting Maschine, Max Features: 1000, Accuracy: 0.89
Model: SVC, Max Features: 2000, Accuracy: 0.70
Model: Random Forest, Max Features: 2000, Accuracy: 0.85
Model: Gradient Boosting Maschine, Max Features: 2000, Accuracy: 0.89
Model: SVC, Max Features: 3000, Accuracy: 0.70
Model: Random Forest, Max Features: 3000, Accuracy: 0.85
Model: Gradient Boosting Maschine, Max Features: 3000, Accuracy: 0.90
Model: SVC, Max Features: 4000, Accuracy: 0.70
Model: Random Forest, Max Features: 4000, Accuracy: 0.84
Model: Gradient Boosting Maschine, Max Features: 4000, Accuracy: 0.89
Model: SVC, Max Features: 5000, Accuracy: 0.70
Model: Random Forest, Max Features: 5000, Accuracy: 0.85
Model: Gradient Boosting Maschine, Max Features: 5000, Accuracy: 0.90
```

Obr. 10.2: Príklad výstupu ladenia parametru max_feature pri textovej reprezentácie (word frequency)

10.1.2 Predspracovanie datasetu

Naše predspracovanie dát zahŕňalo ošetrenie extrémnych hodnôt a škálovanie. Vytvorili sme verzie datasetu s rôznymi úrovňami predspracovania: bez predspracovania, s ošetrenými extrémnymi hodnotami, so škálovaním a so škálovaním aj s ošetrenými extrémnymi hodnotami.

Rozhodli sme sa riešiť extrémne hodnoty, pretože niektoré algoritmy, ako napríklad Support Vector Classification (SVC), sú na ne citlivé. Pri vizualizácii jednotlivých stĺpcov pomocou boxplotov bolo zrejmé, že množstvo stĺpcov obsahuje extrémne hodnoty. Aby sme predišli narušeniu rozdielov medzi pohlaviami, zvolili sme netradičný prístup: riešili sme extrémne hodnoty v každom stĺpci zvlášť podľa pohlavia, teda dvakrát – raz pre ženy a raz pre mužov. Za extrémne hodnoty sme považovali horných a dolných 10 %. Tieto hodnoty sme nahradili priemerom, pretože sme nechceli odstraňovať záznamy. Nahradenie modusom sme nepovažovali za vhodné, keďže väčšina hodnôt obsahuje množstvo desatinných miest, čo znamená, že v stĺpcoch sa nevyskytuje veľa rovnakých hodnôt.

Škálovanie dát bolo nevyhnutné, pretože niektoré algoritmy strojového učenia, ako napríklad SVC, sú citlivé na rozsah dát. Naše dáta mali rôzne rozsahy – niektoré hodnoty sa pohybovali od 0 do 5, iné v percentuálnej forme od 0 do 100. Použili sme MinMaxScaler, ktorý zabezpečí, že všetky dáta budú v rozsahu od 0 do 1.

10.1.3 Testovanie 1

V testovaní jeden sme spúšťali všetky verzie datasetu na algoritmoch strojového učenia SVC, RF a GBM. Algoritmy predikovali na všetkých črtách. Ako prvý sme si spustili SVC. Vo väčšine prípadov predikoval veľmi malé množstvo záznamov ako ženy (minoritná trieda), v niektorých prípadoch dokonca predikoval, že všetky záznamy patria majoritnej triede. Predpokladali sme, že má veľký problém s tým, že triedy sú nevyvážené, avšak problém môže byť aj to, že predikuje na základe priveľa črt. Aj keď najvyššia presnosť dosiahla 0.76 v datasete s ošetrenými outliermi a n-gramami, čo sa nezdá také zlé, po zaradení všetkých záznamov do najviac zastúpenej triedy by presnosť bola 0.70, čo nie je taký veľký rozdiel. Takže vieme, že nepredpovedá úplne korektne. Budeme sa to snažiť upraviť v ďalších testovaniach.

Ďalším v poradí bol RF, ktorého predikcie boli o niečo lepšie. V niektorých verziách datasetu mal stále mierny problém s nevyváženosťou tried, avšak ošetrenie outlierov alebo pridanie textovej reprezentácie tento problém zmenšilo. Dosiahol presnosť až 0.91 na datasete s ošetrením outlierov a škálovaním a s word frequency, čo považujeme za veľmi dobré. Avšak po pozretí na presnosť trénovacej sady, ktorá bola 1, sme prišli k názoru, že model sa pretrénoval.

Posledným modelom bol GBM, ktorý nám priniesol najlepšie výsledky, s presnosťou až 0.96 na datasete, ktorý mal ošetrené outliery a textovú reprezentáciu word frequency. Vyzerá to, že nevyváženosť tried v tomto modeli nie je problém. Pri predikcii pomocou niektorých verzií datasetu sa zdá, že model by mohol byť mierne preučený, avšak rozdiel medzi presnosťou na trénovacej a testovacej sade nie je viac ako 0.1, takže to nepovažujeme za problém, ktorý by bolo nutné riešiť.

Testovanie sme spúšťali osemkrát a spriemerované tabuľky pre každý algoritmus sú uvedené v prílohe D. Tabuľky 10.1,10.2 a 10.3 zobrazujú už zjednodušenú

verziu testovania s iba zobár datasetmi a iba s metrikou presnosti.

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.759804
dataset s ošetrením outlierov	0.701716
dataset s ošetrením outlierov a s ngramami	**0.765931**
dataset s ošetrením outlierov a s word frequency	0.758946
dataset so skalovaním	0.701716
dataset s ošetrením outlierov a skalovaním	0.701716
dataset s ošetrením outlierov a skalovaním a s ngramami	**0.765931**
dataset s ošetrením outlierov a skalovaním a s word frequency	0.758946

Tabuľka 10.1: Pohlavie - Testovanie 1 SVC

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.845098
dataset s ošetrením outlierov	**0.909068833**
dataset s ošetrením outlierov a s ngramami	0.874632333
dataset s ošetrením outlierov a s word frequency	0.908333333
dataset so skalovaním	0.765318667
dataset s ošetrením outlierov a skalovaním	0.9088235
dataset s ošetrením outlierov a skalovaním a s ngramami	0.877206
dataset s ošetrením outlierov a skalovaním a s word frequency	0.906004667

Tabuľka 10.2: Pohlavie - Testovanie 1 RF

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.899387
dataset s ošetrením outlierov	0.936152
dataset s ošetrením outlierov a s ngramami	0.961152
dataset s ošetrením outlierov a s word frequency	0.960784
dataset so skalovaním	0.768505
dataset s ošetrením outlierov a skalovaním	0.936152
dataset s ošetrením outlierov a skalovaním a s ngramami	**0.961274**
dataset s ošetrením outlierov a skalovaním a s word frequency	0.960784

Tabuľka 10.3: Pohlavie - Testovanie 1 GBM

10.1.4 Pohlavie - Testovanie 2

V druhom testovaní sme aplikovali výber čŕt pomocou metód RFE (Recursive Feature Elimination) a Forward Selection, aby sme zistili, ktorá metóda dosiahne lepšie výsledky. Aplikovali sme tieto metódy na varianty datasetu: dataset bez ošetrenia outlierov a škálovania, dataset s ošetrenými outliermi, dataset so škálovaním a dataset s ošetrenými outliermi a škálovaním pre všetky algoritmy strojového učenia samostatne. Črty, ktoré sme získali pre jednotlivé varianty, sme použili aj v kombinácii s variantmi, ktoré obsahovali textové reprezentácie. Testovanie sme spustili celkovo osemkrát, štyrikrát pre RFE a štyrikrát pre Forward Selection.

Najprv sme vyskúšali metódu RFE. Ako môžete vidieť v tabuľke 1, výsledky algoritmu SVC sa zlepšili, pričom sme dosiahli presnosť až 0.84, čo je zlepšenie o 0.08 oproti prvému testu. Najvyššie skóre sme dosiahli s variantom datasetu, ktorý mal ošetrené outliery a bol spojený s textovou reprezentáciou n-gramov. Napriek tomu tento algoritmus dosiahol najnižšie výsledky aj v tomto testovaní.

Ďalším algoritmom bol RF, pri ktorom sme dosiahli približne rovnaké najvyššie skóre ako v testovaní 1. Presnosť bola stále 0.91 vo variante datasetu s ošetrením

outlierov a škálovaním. Výsledky druhého testovania pre RF sú uvedené v tabuľke 2.

Najlepšie výsledky sme dosiahli pomocou algoritmu GBM, a to s presnosťou až 0.96 vo variante datasetu s ošetrením outlierov a škálovaním a s word frequency. Táto presnosť zostala približne rovnaká ako v prvom testovaní. Výsledky druhého testovania pre GBM sú uvedené v tabuľke 3.

V tabuľkách 10.4,10.5 a 10.6 nižšie sú uvedené len niektoré z najvyšších výsledkov pre jednotlivé algoritmy. Všetky výsledky pre všetky varianty datasetu si môžeme pozrieť v prílohe D, kde nájdete aj vybrané vlastnosti pomocou RFE pre každý algoritmus a variant datasetu. Ako príklad uvádzame vybrané vlastnosti pomocou RFE (ktoré boli v tejto variante následne spojené s word frequency) pre algoritmus GBM vo variante datasetu s ošetrením outlierov a škálovaním, kde sme dosiahli najvyššie skóre:

- **GBM pre dataset s ošetrením outlierov a skalovaním:** priemerný počet emoji na tweet, priemerný počet URL adries na tweet, priemerný počet stop slov na tweet, priemerná dĺžka slova, priemerný počet znakov na vetu, priemerný počet opakovaných znakov na tweet, percentuálny podiel oznamovacích viet na tweet, percentuálny podiel otázok na tweet, percentuálny podiel rozkazovacích viet na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet číslovieok na tweet, priemerný počet prísloviek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet citosloviec na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet priamych predmetov na tweet, priemerný počet príslovkových modifikátorov na tweet, priemerný počet prívlastkov na tweet

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.840441
dataset s ošetrením outlierov	0.726716

Varianta datasetu	**Presnosť**
dataset s ošetrením outlierov a s ngramami	**0.840196**
dataset s ošetrením outlierov a s tfidf word frequency	0.747059
dataset so skalovaním	0.703922
dataset so skalovaním a s word frequency	0.830392
dataset s ošetrením outlierov a skalovaním	0.726716
dataset s ošetrením outlierov a skalovaním a s ngramami	**0.840196**
dataset s ošetrením outlierov a skalovaním a s word frequency	0.830637

Tabuľka 10.4: Pohlavie - Testovanie 2 RFE SVC

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.843137
dataset s ošetrením outlierov	0.908333
dataset s ošetrením outlierov a s ngramami	0.869853
dataset s ošetrením outlierov a s tfidf word frequency	0.890932
dataset so skalovaním	0.769118
dataset so skalovaním a s word frequency	0.843628
dataset s ošetrením outlierov a skalovaním	**0.910784**
dataset s ošetrením outlierov a skalovaním a s ngramami	0.870098
dataset s ošetrením outlierov a skalovaním a s word frequency	0.902941

Tabuľka 10.5: Pohlavie - Testovanie 2 RFE RF

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.898039

Varianta datasetu	**Presnosť**
dataset s ošetrením outlierov	0.935784
dataset s ošetrením outlierov a s ngramami	0.959314
dataset s ošetrením outlierov a s tfidf word frequency	0.957843
dataset so skalovaním	0.768872
dataset so skalovaním a s word frequency	0.880392
dataset s ošetrením outlierov a skalovaním	0.936029
dataset s ošetrením outlierov a skalovaním a s ngramami	0.959314
dataset s ošetrením outlierov a skalovaním a s word frequency	**0.959559**

Tabuľka 10.6: Pohlavie - Testovanie 2 RFE GBM

Následne sme vyskúšali výber čŕt pomocou metódy Forward Selection, v tabuľkach 10.4,10.5 a 10.6, možeme vidieť v zjednodušenú formu výsledkov tejto metódy. Pomocou algoritmu SVC sa nám podarilo dosiahnuť presnosť až 0.87 vo variante datasetu so škálovaním a s tf-idf n-gramami, čo je zlepšenie o 0.1 oproti prvému testovaniu. Tento výsledok bol tiež lepší ako pri použití metódy RFE.

Algoritmus RF dosiahol presnosť 0.91 vo variante datasetu s ošetrenými outliermi, čo je podobná presnosť ako v prvom testovaní aj v druhom testovaní pomocou metódy RFE.

Najlepšie výsledky opäť dosiahol algoritmus GBM, ktorý dosiahol presnosť 0.96 vo variante datasetu s ošetrenými outliermi a s word frequency, čo bola rovnaká presnosť ako v prvom testovaní aj v druhom testovaní pomocou metódy RFE.

Všetky spriemerované presnosti pre všetky varianty datasetu a zároveň pre všetky algoritmy strojového učenia nájdeme v prílohe D. Tam sú tiež uvedené vybrané črty pre každý algoritmus. Ako príklad uvádzame výber čŕt pomocou metódy FS pre algoritmus GBM pre variant datasetu s ošetrenými outliermi:

- **GBM pre dataset s ošetrením outlierov:** percentuálny podiel rozkazovacích viet na tweet, priemerný počet čísloviek na tweet, priemerný počet podmetov

na tweet, priemerný počet častíc na tweet, percentuálny podiel otázok na tweet, priemerný počet zámen na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet prísloviek na tweet, priemerný počet čiarok na tweet, priemerný počet apostrofov na tweet, priemerný počet predložiek na tweet, priemerný počet viet na tweet

V závere tohto testovania nemôžeme povedať, že by jedna metóda výberu čŕt bola výrazne lepšia ako druhá. V algoritmoch Random Forest a Gradient Boosting Machine dosiahli veľmi podobné presnosti, avšak v algoritme SVC sme dosiahli mierne vyššiu presnosť metódou výberu čŕt Forward Selection. Preto sme sa rozhodli v ďalšom testovaní pokračovať s črtami, ktoré nám vybrala táto metóda.

V tomto testovaní sme sa okrem zvýšenia presnosti, čo sa nám podarilo iba pri SVC, snažili aj zmenšiť preučiteľnosť, primárne v algoritme Random Forest, prípadne mierne aj v GBM. To sa nám však výberom čŕt veľmi nepodarilo. V ďalšom testovaní sa pokúsime dosiahnuť tieto ciele ladením hyperparametrov.

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.873284
dataset s ošetrením outlierov	0.730882
dataset s ošetrením outlierov a s ngramami	0.845588
dataset s ošetrením outlierov a s word frequency	0.840196
dataset so skalovaním a s tfidf ngramami	**0.875735**
dataset so skalovaním a s tfidf word frequency	0.872059
dataset s ošetrením outlierov a skalovaním	0.730882
dataset s ošetrením outlierov a skalovaním a s ngramami	0.845588
dataset s ošetrením outlierov a skalovaním a s word frequency	0.840196

Tabuľka 10.7: Pohlavie - Testovanie 2 FS SVC

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.854657
dataset s ošetrením outlierov	**0.908579**
dataset s ošetrením outlierov a s ngramami	0.866667
dataset s ošetrením outlierov a s word frequency	0.8875
dataset so skalovaním a s tfidf ngramami	0.849265
dataset so skalovaním a s tfidf word frequency	0.84853
dataset s ošetrením outlierov a skalovaním	0.901961
dataset s ošetrením outlierov a skalovaním a s ngramami	0.860049
dataset s ošetrením outlierov a skalovaním a s word frequency	0.880392

Tabuľka 10.8: Pohlavie - Testovanie 2 FS RF

Varianta datasetu	**Presnosť**
dataset bez ošetrenia outlierov a skalovania s tfidf word frequency	0.891421
dataset s ošetrením outlierov	0.93799
dataset s ošetrením outlierov a s ngramami	0.954902
dataset s ošetrením outlierov a s word frequency	**0.959559**
dataset so skalovaním a s tfidf ngramami	0.889216
dataset so skalovaním a s tfidf word frequency	0.892157
dataset s ošetrením outlierov a skalovaním	0.934314
dataset s ošetrením outlierov a skalovaním a s ngramami	0.955392
dataset s ošetrením outlierov a skalovaním a s word frequency	0.959069

Tabuľka 10.9: Pohlavie - Testovanie 2 FS GBM

10.1.5 Testovanie 3

Náš pôvodný plán pre toto testovanie zahŕňal spustenie gridsearch pre všetky varianty datasetu a algoritmy strojového učenia s vybranými črtami pomocou metódy Forward Selection (FS), pretože táto metóda priniesla lepšie výsledky v predchádzajúcich testoch. Gridsearch spočíva v skúšaní všetkých kombinácií hyperparametrov, ktoré sme si určili, s cieľom nájsť tú najlepšiu kombináciu dosahujúcu najvyššiu presnosť. Naším cieľom bolo zistiť, akú maximálnu presnosť môže každý variant datasetu dosiahnuť po optimalizácii hyperparametrov a následne nájsť celkovú najvyššiu presnosť. Tento postup sme aj vyskúšali, avšak zistili sme, že je veľmi časovo náročný. Jeden variant datasetu, ktorý obsahoval textové reprezentácie, trval viac ako štyri dni. Vzhľadom na to, že sme mali vykonať celkovo 20 gridsearchov, nebolo možné všetky varianty stihnúť včas.

Preto sme sa rozhodli vykonať gridsearch iba pre varianty bez textových reprezentácií. Hyperparametre, ktoré sme týmto spôsobom získali, sme následne aplikovali aj na varianty obsahujúce textové reprezentácie. Používali sme tieto hyperparametre:

Pre Support Vector Classifier (SVC):

1. **C**: Regularizačný parameter. Kontroluje obchod medzi maximalizáciou marginu a minimalizáciou chyby. Nižšie hodnoty vedú k mäkšiemu marginu (viac chýb), vyššie hodnoty k tvrdšiemu marginu (menej chýb). Mohol nadobudnúť hodnoty: 0.1, 1, 10, 100.

2. **kernel**: Typ kernelu používaného na mapovanie vstupných dát do vyššej dimenzie. Mohol nadobudnúť hodnoty: linear, poly, rbf, sigmoid.

3. **gamma**: Koeficient pre rbf, poly a sigmoid kernely. Určuje, ako ďaleko vplyv jedného trénovacieho príkladu dosahuje. Hodnota scale znamená $1/(\text{n\_features} \times \text{X.var()})$, auto znamená $1/\text{n\_features}$. Mohol nadobudnúť hodnoty: scale, auto.

4. **coef0**: Nezávislá konštanta pridávaná do kernelu poly a sigmoid. Používa sa na riadenie oboch modelov. Mohol nadobudnúť hodnoty: 0.0, 0.5, 1.0, 1.5 [75].

Random Forest (RF) a Gradient Boosting Algorithm (GBM):

1. **n_estimators**: Počet stromov v lese. Mohol nadobudnúť hodnoty: 10, 50, 100, 200.

2. **max_depth**: Maximálna hĺbka stromu. Ak je None, uzly sa budú rozdeľovať, pokiaľ každý list neobsahuje menej ako min_samples_split vzoriek. Mohol nadobudnúť hodnoty:None, 10, 20, 30, 40.

3. **min_samples_split**: Minimálny počet vzoriek potrebný na rozdelenie uzla. Mohol nadobudnúť hodnoty: 2, 5, 10.

4. **min_samples_leaf**: Minimálny počet vzoriek potrebný v listovom uzle. Mohol nadobudnúť hodnoty: 1, 2, 4 [76][77].

Výsledky z tohto testovania však neboli uspokojivé. Prípady, u ktorých sme vykonali gridsearch, sa zvýšili v presnosti, avšak u variant bez textových reprezentácií sme často pozorovali zníženie presnosti. Aj varianta, ktorá nám v predchádzajúcom testovaní priniesla najvyššiu presnosť, konkrétne GBM dataset s ošetrením outlierov, bola jednou z tých, u ktorých sa znížila presnosť. Nedosiahli sme teda lepšiu presnosť ako pri minulom testovaní.

Ako môžeme vidieť v tabuľke 10.10, tak pri SVC sme dosiahli najvyššie skóre 0.88 vo variante dataset so skalovaním a s tfidf word frequency, čo predstavuje malé zlepšenie. Podrobné výsledky zo všetkých variant nájdeme v prílohe D. Použité hyperparametre boli nasledovné:

- **dataset bez ošetrenia outlierov a škálovania:** C: 100, coef0: 0.0, gamma: scale, kernel: rbf
- **dataset s ošetrením outlierov:** C: 100, coef0: 0.0, gamma: scale, kernel: rbf
- **dataset so skalovaním:** C: 1, coef0: 0.0, gamma: scale, kernel: rbf
- **dataset s ošetrením outlierov a skalovaním:** C: 100, coef0: 0.0, gamma: scale, kernel: rbf

Varianta datasetu	**accuracy_test**
dataset bez ošetrenia outlierov a škálovania	0.716912
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	**0.887745**
dataset s ošetrením outlierov	0.734559
dataset so skalovaním	0.722304
dataset so skalovaním a s tfidf word frequency	0.879657
dataset s ošetrením outlierov a skalovaním	0.734559
dataset s ošetrením outlierov a skalovaním a s ngramami	0.87402
dataset s ošetrením outlierov a skalovaním a s word frequency	0.86201

Tabuľka 10.10: Pohlavie - Testovanie 3 SVC

Pri Random Forest sme dosiahli najvyššiu presnosť 0.90 vo variante dataset s ošetrením outlierov, čo je zrejmé z tabuľky 10.11. Ide o mierne zníženie presnosti oproti druhému testovaniu. Podrobné výsledky zo všetkých variant nájdeme v prílohe D. Použité hyperparametre boli nasledovné:

- **dataset bez ošetrenia outlierov a škálovania:** max_depth: 10, min_samples_leaf: 4, min_samples_split: 5, n_estimators: 50
- **dataset s ošetrením outlierov:** max_depth: 20, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 200
- **dataset so skalovaním:** max_depth: 10, min_samples_leaf: 4, min_samples_split: 5, n_estimators: 100
- **dataset s ošetrením outlierov a skalovaním:** max_depth: 30, min_samples_leaf: 1, min_samples_split: 10, n_estimators: 100

Varianta datasetu	**accuracy_test**
dataset bez ošetrenia outlierov a škálovania	0.744853

Varianta datasetu	**accuracy_test**
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.846568
dataset s ošetrením outlierov	**0.903431**
dataset so skalovaním	0.714461
dataset so skalovaním a s tfidf word frequency	0.844118
dataset s ošetrením outlierov a skalovaním	0.89951
dataset s ošetrením outlierov a skalovaním a s ngramami	0.858333
dataset s ošetrením outlierov a skalovaním a s word frequency	0.875245

Tabuľka 10.11: Pohlavie - Testovanie 3 RF

Pri GBM sme dosiahli najvyššiu presnosť 0.94 vo variante dataset s ošetrením outlierov a skalovaním a s ngramami, čo môžeme vidieť v tabuľke 10.12. Presnosť sa znížila oproti druhému testovaniu. Podrobné výsledky zo všetkých variant nájdeme v prílohe D. Použité hyperparametre boli nasledovné:

- **dataset bez ošetrenia outlierov a škálovania:** max_depth: 10, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 200
- **dataset s ošetrením outlierov :** max_depth : 30, min_samples_leaf : 4, min_samples_split:10, n_estimators: 200
- **dataset so skalovaním:** max_depth: 10, min_samples_leaf: 4, min_samples_split: 10, n_estimators: 100
- **dataset s ošetrením outlierov a skalovaním:** max_depth: 10, min_samples_leaf: 4, min_samples_split: 5, n_estimators: 200

Varianta datasetu	**accuracy_test**
dataset bez ošetrenia outlierov a škálovania	0.742157
dataset bez ošetrenia outlierov a škálovania s tfidf n-gramami	0.870343
dataset s ošetrením outlierov	0.92402

Varianta datasetu	**accuracy_test**
dataset so skalovaním	0.739951
dataset so skalovaním a s tfidf word frequency	0.878186
dataset s ošetrením outlierov a skalovaním	0.919608
dataset s ošetrením outlierov a skalovaním a s ngramami	**0.943382**
dataset s ošetrením outlierov a skalovaním a s word frequency	0.93897

Tabuľka 10.12: Pohlavie - Testovanie 3 GBM

10.1.6 Finálny model

Najvyššiu presnosť sme dosiahli pomocou modelu :

- **Algoritmus :** Gradient Boosting Machine
- **Dosiahnutá presnosť :** 0.959559
- **Predspracovanie :** ošetrenie odľahlých hodnôt
- **Metóda výberu čŕt :** Postupný výber čŕt
- **Črty :** percentuálny podiel rozkazovacích viet na tweet, priemerný počet číslovieк na tweet, priemerný počet podmetov na tweet, priemerný počet častíc na tweet, percentuálny podiel otázok na tweet, priemerný počet zámen na tweet, percentuálny podiel ostatných koncov na tweet, priemerný počet prísloviek na tweet, priemerný počet čiarok na tweet, priemerný počet apostrofov na tweet, priemerný počet predložiek na tweet, priemerný počet viet na tweet
- **Textová reprezentácia :** Frekvencia slov
- **Hyperparametre:**n_estimators=100,min_samples_split=2, min_samples_leaf=1, max_depth=3

10.2. Veková kategoria

Použili sme rovnaký dataset ako pri predpovedaní pohlavia, ktorý vznikol spojením troch samostatných datasetov: datasetu s lexikálnymi črtami, datasetu so syntaktickými črtami a datasetu so sémantickými črtami. Dataset obsahoval aj očistený text, z ktorého sme extrahovali textové reprezentácie a vekovú skupinu, ktorú sa pokúsime predpovedať. Vekových skupín máme v datasete 7.

Dataset sme rozdelili na trénovaciu a testovaciu množinu v pomere 70 % ku 30 %. Následne sme vytvorili rôzne varianty datasetu, ktoré zahŕňajú rôzne predspracovanie a kombináciu s rôznymi textovými reprezentáciami. Podrobný popis jednotlivých krokov predspracovania a extrakcie textových reprezentácií nájdete v kapitole 1. Tieto varianty sú:

- Dataset bez ošetrenia outlierov a škálovania.
- Dataset bez ošetrenia outlierov a škálovania s n-gramami.
- Dataset bez ošetrenia outlierov a škálovania s TF-IDF n-gramami.
- Dataset bez ošetrenia outlierov a škálovania s word frequency.
- Dataset bez ošetrenia outlierov a škálovania s TF-IDF word frequency.
- Dataset s ošetrením outlierov.
- Dataset s ošetrením outlierov a s n-gramami.
- Dataset s ošetrením outlierov a s TF-IDF n-gramami.
- Dataset s ošetrením outlierov a s word frequency.
- Dataset s ošetrením outlierov a s TF-IDF word frequency.
- Dataset so škálovaním.
- Dataset so škálovaním a s n-gramami.
- Dataset so škálovaním a s TF-IDF n-gramami.

- Dataset so škálovaním a s word frequency.
- Dataset so škálovaním a s TF-IDF word frequency.
- Dataset s ošetrením outlierov a škálovaním.
- Dataset s ošetrením outlierov a škálovaním a s n-gramami.
- Dataset s ošetrením outlierov a škálovaním a s TF-IDF n-gramami.
- Dataset s ošetrením outlierov a škálovaním a s word frequency.
- Dataset s ošetrením outlierov a škálovaním a s TF-IDF word frequency.

Vekovú kategóriu sme predpovedali pomocou troch algoritmov strojového učenia: Support Vector Classification, Random Forest a Gradient Boosting Machine. Vykonali sme tri testovania, aby sme našli model s najvyššou presnosťou predikcie vekovej kategórie. Tieto testovania sú:

- **Testovanie 1 :** Učenie a predikovanie vekovej kategórie pomocou všetkých čŕt.
- **Testovanie 2 :** Učenie a predikovanie vekovej kategórie s výberom čŕt. Porovnávali sme dve metódy výberu čŕt: Recursive Feature Elimination a Forward Selection.
- **Testovanie 3 :** Ladenie hyperparametrov pre algoritmy strojového učenia, vrátane učenia a predikovania s optimalizovanými hyperparametrami.

10.2.1 Testovanie 1

V prvom testovaní sme spúšťali všetky verzie datasetu so všetkými črtami na troch algoritmoch strojového učenia: SVC, RF a GBM. Prvý algoritmus, ktorý sme testovali, bol SVC, pričom dosiahol presnosť 0.36. Tento algoritmus nepredikoval veľmi korektne a väčšinu záznamov zaradil do dvoch najpočetnejších skupín. Problém mohol mať s veľkým počtom čŕt alebo nerovnomerným rozdelením vekových kategórií. Model dosahoval najlepšie výsledky na variantoch datasetu s ošetrenými outliermi a spojených s textovou reprezentáciou n-gramov alebo frekvencií slov.

Ďalší v poradí bol algoritmus Random Forest (RF), s ktorým sme dosiahli presnosť

0.56, čo je omnoho vyššie ako pri SVC. Vo variantoch bez ošetrenia outlierov mal RF mierne problémy s určením, pretože často mýlil susedné kategórie. Avšak vo verziách s ošetrením outlierov tento problém výrazne klesol a model dokázal predikovať celkom presne. Model najlepšie predikoval na verziách datasetu s ošetrenými outliermi, ktoré neboli spojené s textovou reprezentáciou.

Nakoniec sme použili algoritmus Gradient Boosting Machine (GBM), s ktorým sme dosiahli najvyššiu presnosť, a to 0.81. Podobne ako pri Random Forest, mal model problémy s určením susedných kategórií vo verziách bez ošetrenia outlierov. Avšak pri variantoch s ošetrenými outliermi predikoval veľmi presne. GBM, rovnako ako RF, najlepšie predikoval na verziách datasetu s ošetrenými outliermi, ktoré neboli spojené s textovou reprezentáciou.

Testovanie sme spúšťali osemkrát a spriemerované tabuľky sú uvedené v prílohách D. Tabuľky 10.13,10.14 a 10.15 zobrazujú už zjednodušenú verziu testovania s iba zobár datasetmi a iba s metrikou presnosti.

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.354596
dataset s ošetrením outlierov	0.329228
dataset s ošetrením outlierov a s ngramami	**0.360294**
dataset s ošetrením outlierov a s word frequency	0.359559
dataset so skalovaním a s word frequency	0.353125
dataset s ošetrením outlierov a skalovaním	0.329228
dataset s ošetrením outlierov a skalovaním a s ngramami	**0.360294**
dataset s ošetrením outlierov a skalovaním a s word frequency	0.359559

Tabuľka 10.13: Vek - Testovanie 1 SVC

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.406066
dataset s ošetrením outlierov	0.709191
dataset s ošetrením outlierov a s ngramami	0.550919
dataset s ošetrením outlierov a s word frequency	0.556434
dataset so skalovaním a s word frequency	0.403309
dataset s ošetrením outlierov a skalovaním	**0.712132**
dataset s ošetrením outlierov a skalovaním a s ngramami	0.561213
dataset s ošetrením outlierov a skalovaním a s word frequency	0.565809

Tabuľka 10.14: Vek - Testovanie 1 RF

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.422427
dataset s ošetrením outlierov	0.804228
dataset s ošetrením outlierov a s ngramami	0.752758
dataset s ošetrením outlierov a s word frequency	0.745956
dataset so skalovaním a s word frequency	0.422978
dataset s ošetrením outlierov a skalovaním	**0.809191**
dataset s ošetrením outlierov a skalovaním a s ngramami	0.75239
dataset s ošetrením outlierov a skalovaním a s word frequency	0.749816

Tabuľka 10.15: Vek - Testovanie 1 GBM

10.2.2 Testovanie 2

Pri druhom testovaní sme aplikovali výber čŕt pomocou metód RFE (Recursive Feature Elimination) a Forward Selection, aby sme zistili, ktorá z nich dosahuje

lepšie výsledky. Tieto metódy sme použili na rôzne varianty datasetu: bez ošetrenia outlierov a škálovania, s ošetrenými outliermi, so škálovaním a s ošetrenými outliermi a škálovaním. Každý variant sme testovali samostatne pre všetky algoritmy strojového učenia. Črty získané pre jednotlivé varianty sme následne použili aj v kombinácii s variantmi obsahujúcimi textové reprezentácie. Testovanie sme vykonali celkovo osemkrát: štyrikrát pre RFE a štyrikrát pre Forward Selection.

Najprv sme aplikovali metódu RFE. Výsledky algoritmu SVC sa mierne zlepšili, pričom sme dosiahli presnosť 0.47, čo predstavuje nárast o 0.1. Hoci problém, že väčšina vzoriek je zaradená do dvoch najpočetnejších tried, stále pretrváva, jeho rozsah sa zmenšil. Najlepšie skóre sme dosiahli pri variante datasetu s ošetrenými outliermi a textovou reprezentáciou založenou na tf-idf frekvencii slov. Následne sme použili algoritmus Random Forest na predikciu. Dosiahol skóre 0.73, čo predstavuje mierne zlepšenie oproti prvému testovaniu. Najlepšie výsledky model dosiahol pri verzii datasetu s ošetrenými outliermi bez textových reprezentácií.

Nakoniec sme vykonali predikcie pomocou algoritmu GBM, ktorý dosiahol najlepšie výsledky. Presnosť dosiahla až 0.84, čo považujeme za vynikajúci výsledok pri siedmich triedach. V porovnaní s prvým testovaním sme zaznamenali mierne zlepšenie presnosti. Najlepšie výsledky sme dosiahli pri verziách datasetu s ošetrenými outliermi, bez použitia textových reprezentácií.

V tabuľkách 10.16,10.17 a 10.18 sú uvedené niektoré z výsledkov tohto testovania pre jednotlivé algoritmy. Kompletné výsledky pre všetky varianty datasetu sú dostupné v prílohe D, kde nájdeme aj vybrané vlastnosti pomocou metódy RFE pre každý algoritmus a variant datasetu. Ako príklad uvádzame vybrané vlastnosti pomocou RFE pre algoritmus GBM vo variante datasetu s ošetrenými outliermi a škálovaním, pri ktorom sme dosiahli najvyššie skóre:

- **GBM s ošetrením outlierov a škálovaním :** priemerný počet emoji na tweet, priemerný počet veľkých písmen na tweet, priemerný počet slov na vetu, priemerný počet slov na tweet, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, percento neštandardných zakončení viet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, prie-

merný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet prísloviek na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častic na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet, priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.408456
dataset s ošetrením outlierov	0.461765
dataset s ošetrením outlierov a s ngramami	0.415809
dataset s ošetrením outlierov a s tfidf word frequency	**0.468015**
dataset so skalovaním a s word frequency	0.408456
dataset s ošetrením outlierov a skalovaním	0.461765
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.467647
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	**0.468015**

Tabuľka 10.16: Vek - Testovanie 2 RFE SVC

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.408824
dataset s ošetrením outlierov	0.726471
dataset s ošetrením outlierov a s ngramami	0.565441
dataset s ošetrením outlierov a s tfidf word frequency	0.537868
dataset so skalovaním a s word frequency	0.396324
dataset s ošetrením outlierov a skalovaním	**0.733824**

Varianta datasetu	**presnosť**
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.549265
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.546691

Tabuľka 10.17: Vek - Testovanie 2 RFE RF

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.429044
dataset s ošetrením outlierov	0.83603
dataset s ošetrením outlierov a s ngramami	0.810662
dataset s ošetrením outlierov a s tfidf word frequency	0.796324
dataset so skalovaním a s word frequency	0.420956
dataset s ošetrením outlierov a skalovaním	**0.8375**
dataset s ošetrením outlierov a skalovaním a s tfidf ngramami	0.799633
dataset s ošetrením outlierov a skalovaním a s tfidf word frequency	0.804044

Tabuľka 10.18: Vek - Testovanie 2 RFE GBM

Následne sme použili metódu Forward Selection na výber čŕt. V tabuľkách 10.16,10.17 a 10.18 nájdeme zjednodušené výsledky tejto metódy. Pri použití algoritmu SVC sme dosiahli presnosť 0.44 vo variante s ošetrenými outliermi, čo predstavuje zlepšenie o 0.08 oproti prvému testovaniu. Tento výsledok však stále nedosiahol úroveň presnosti, ktorú sme získali použitím metódy RFE.

Algoritmus Random Forest dosiahol presnosť 0.72 pri variante datasetu s ošetrenými outliermi, čo predstavuje mierne zlepšenie oproti prvému testovaniu. Avšak, táto presnosť je stále nižšia v porovnaní s výsledkami dosiahnutými pri použití metódy RFE v druhom testovaní.

Najlepšie výsledky opäť dosiahol algoritmus GBM, ktorý dosiahol presnosť

0.83 pri variante datasetu s ošetrenými outliermi a škálovaním. Toto predstavuje mierne zlepšenie oproti prvému testovaniu, avšak presnosť bola o niečo nižšia ako v druhom testovaní s použitím metódy RFE.

V prílohe D nájdeme spriemerované presnosti pre všetky varianty datasetu a všetky algoritmy strojového učenia. Sú tam tiež uvedené vybrané črty pre každý algoritmus. Ako príklad uvádzame výber čŕt pomocou metódy RFE pre algoritmus GBM vo variante datasetu s ošetrenými outliermi a škálovaním:

- **GBM dataset s ošetrením outlierov a skalovaním :** priemerný počet priamych predmetov na tweet, priemerný počet častíc na tweet, percento oznamovacích viet, percento opytovacích viet, priemerný počet čísloviek na tweet, priemerný počet čiarok na tweet, priemerný počet príslovkových určení na tweet, percento iných zakončení viet, priemerný počet zámen na tweet, priemerný počet prídavných mien na tweet, priemerný počet spojok na tweet

Na záver tohto testovania nemôžeme jednoznačne povedať, že by jedna metóda bola výrazne lepšia než druhá, pretože výsledky boli veľmi podobné. Vo všetkých troch algoritmoch sme však dosiahli najvyššie skóre s črtami vybranými metódou RFE, hoci rozdiel oproti metóde Forward Selection bol len mierny. Preto sme sa rozhodli v ďalšom testovaní pokračovať s črtami vybranými pomocou metódy RFE.

V ďalšom testovaní sa zameriame nielen na zvyšovanie presnosti, ale aj na zníženie preučiteľnosti, predovšetkým u algoritmov Random Forest a GBM. Keďže sa nám to nepodarilo dosiahnuť výberom čŕt, pokúsime sa to dosiahnuť ladením hyperparametrov.

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.403309
dataset s ošetrením outlierov	**0.44853**
dataset s ošetrením outlierov a s ngramami	0.411765
dataset s ošetrením outlierov a s word frequency	0.409927
dataset so skalovaním a s word frequency	0.404412

Varianta datasetu	**presnosť**
dataset s ošetrením outlierov a skalovaním	**0.44853**
dataset s ošetrením outlierov a skalovaním a s ngramami	0.411765
dataset s ošetrením outlierov a skalovaním a s word frequency	0.409927

Tabuľka 10.19: Vek - Testovanie 2 FS SVC

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.405515
dataset s ošetrením outlierov	**0.726103**
dataset s ošetrením outlierov a s ngramami	0.476471
dataset s ošetrením outlierov a s word frequency	0.472794
dataset so skalovaním a s word frequency	0.404044
dataset s ošetrením outlierov a skalovaním	0.705882
dataset s ošetrením outlierov a skalovaním a s ngramami	0.492647
dataset s ošetrením outlierov a skalovaním a s word frequency	0.476471

Tabuľka 10.20: Vek - Testovanie 2 FS RF

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s word frequency	0.415809
dataset s ošetrením outlierov	0.822574
dataset s ošetrením outlierov a s ngramami	0.790441
dataset s ošetrením outlierov a s word frequency	0.784927
dataset so skalovaním a s word frequency	0.433088
dataset s ošetrením outlierov a skalovaním	**0.829559**

Varianta datasetu	**presnosť**
dataset s ošetrením outlierov a skalovaním a s ngramami	0.799633
dataset s ošetrením outlierov a skalovaním a s word frequency	0.790441

Tabuľka 10.21: Vek - Testovanie 2 FS GBM

10.2.3 Testovanie 3

Pôvodný plán pre toto testovanie zahŕňal spustenie grid search pre všetky varianty datasetu a algoritmy strojového učenia s črtami vybranými metódou RFE, pretože táto metóda priniesla lepšie výsledky v predchádzajúcich testoch. Grid search spočíva v skúšaní všetkých kombinácií vopred určených hyperparametrov s cieľom nájsť najlepšiu kombináciu dosahujúcu najvyššiu presnosť. Naším cieľom bolo zistiť, akú maximálnu presnosť môže každý variant datasetu dosiahnuť po optimalizácii hyperparametrov a následne nájsť celkovo najvyššiu presnosť. Tento postup sme vyskúšali, avšak ukázalo sa, že je veľmi časovo náročný. Jeden variant datasetu, ktorý obsahoval textové reprezentácie, trval viac ako tri dni. Vzhľadom na to, že sme mali vykonať celkovo 20 grid searchov, nebolo možné všetky varianty stihnúť včas.

Preto sme sa rozhodli vykonať grid search iba pre varianty datasetov bez textových reprezentácií. Hyperparametre získané týmto spôsobom sme následne aplikovali aj na varianty obsahujúce textové reprezentácie. Použili sme nasledujúce hyperparametre:

Pre Support Vector Classifier (SVC) :

- **C :** Regularizačný parameter. Kontroluje obchod medzi maximalizáciou marginu a minimalizáciou chyby. Nižšie hodnoty vedú k mäkšiemu marginu (viac chýb), vyššie hodnoty k tvrdšiemu marginu (menej chýb). Mohol nadobudnúť hodnoty: 0.1, 1, 10, 100.
- **kernel:** Typ kernelu používaného na mapovanie vstupných dát do vyššej dimenzie. Mohol nadobudnúť hodnoty: 'linear', 'poly', 'rbf', 'sigmoid'.
- **gamma :** Koeficient pre 'rbf', 'poly' a 'sigmoid' kernely. Určuje, ako ďaleko vplyv jedného trénovacieho príkladu dosahuje. Hodnota 'scale' znamená 1 / (n_features

* X.var()), 'auto' znamená 1 / n_features. Mohol nadobudnúť hodnoty: 'scale', 'auto'.

- **coef0 :** Nezávislá konštanta pridávaná do kernelu 'poly' a 'sigmoid'. Používa sa na riadenie oboch modelov. Mohol nadobudnúť hodnoty: 0.0, 0.5, 1.0, 1.5 [75].

Pre Random Forest (RF) a Gradient Boosting Algoritmus (GBM) :

- **n_estimators :** Počet stromov v lese. Mohol nadobudnúť hodnoty: 10, 50, 100, 200.
- **max_depth :** Maximálna hĺbka stromu. Ak je None, uzly sa budú rozdeľovať, pokiaľ každý list neobsahuje menej ako min_samples_split vzoriek. Mohol nadobudnúť hodnoty: None, 10, 20, 30, 40.
- **min_samples_split :** Minimálny počet vzoriek potrebný na rozdelenie uzla. Mohol nadobudnúť hodnoty: 2, 5, 10.
- **min_samples_leaf :** Minimálny počet vzoriek potrebný v listovom uzle. Mohol nadobudnúť hodnoty: 1, 2, 4 [76][77].

Výsledky tohto testovania však neboli uspokojivé. Hoci varianty, pri ktorých sme vykonali grid search, vo všeobecnosti vykázali zvýšenie presnosti, často sme pozorovali zníženie presnosti pri variantoch bez textových reprezentácií.

Pri použití algoritmu SVC sme dosiahli najvyššie skóre 0.50 s variantom datasetu, ktorý mal ošetrené outliery a bol škálovaný, čo je uvedené v tabuľke 10.22. Toto predstavuje zlepšenie oproti druhému testovaniu. Podrobné výsledky všetkých variantov sú uvedené v prílohe D. Použité hyperparametre boli nasledovné:

- **Dataset bez ošetrenia outlierov a škálovania :** 'C': 10, 'coef0': 0.5, 'gamma': 'auto', 'kernel': 'poly'.
- **Dataset s ošetrením outlierov :** 'C': 100, 'coef0': 0.0, 'gamma': 'auto', 'kernel': 'rbf'.
- **Dataset so škálovaním :** 'C': 100, 'coef0': 0.5, 'gamma': 'scale', 'kernel': 'poly'.

- **Dataset s ošetrením outlierov a škálovaním :** 'C': 100, 'coef0': 0.0, 'gamma': 'auto', 'kernel': 'rbf'.

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania	0.426075
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.419956
dataset s ošetrením outlierov	0.504693
dataset s ošetrením outlierov a s word frequency	0.302787
dataset so skalovaním	0.399792
dataset so skalovaním a s tfidf word frequency	0.439254
dataset s ošetrením outlierov a skalovaním	**0.507452**
dataset s ošetrením outlierov a skalovaním a s word frequency	0.298979

Tabuľka 10.22: Vek - Testovanie 3 SVC

Pri použití algoritmu Random Forest sme dosiahli najvyššiu presnosť 0.76 s variantom datasetu, ktorý mal ošetrené outliery a bol škálovaný, ako je uvedené v tabuľke 10.23. Toto predstavuje mierne zlepšenie oproti druhému testovaniu. Podrobné výsledky všetkých variantov sú uvedené v prílohe D. Použité hyperparametre boli nasledovné:

- **Dataset bez ošetrenia outlierov a škálovania :** 'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200.
- **Dataset s ošetrením outlierov :** 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200.
- **Dataset so škálovaním :** 'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200.
- **Dataset s ošetrením outlierov a škálovaním :** 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200.

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania	0.427044
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.423606
dataset s ošetrením outlierov	0.737139
dataset s ošetrením outlierov a s word frequency	0.63172
dataset so skalovaním	0.410268
dataset so skalovaním a s tfidf word frequency	0.423449
dataset s ošetrením outlierov a skalovaním	**0.757394**
dataset s ošetrením outlierov a skalovaním a s word frequency	0.615261

Tabuľka 10.23: Vek - Testovanie 3 RF

Pri použití algoritmu GBM sme dosiahli najvyššiu presnosť 0.83 s variantom datasetu, ktorý mal ošetrené outliery a využíval frekvenciu slov, ako je uvedené v tabuľke 10.24. Táto presnosť je o niečo nižšia oproti druhému testovaniu. Podrobné výsledky všetkých variantov sú uvedené v prílohe D. Použité hyperparametre boli nasledovné:

- **Dataset bez ošetrenia outlierov a škálovania :** 'max_depth': 30, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200.
- **Dataset s ošetrením outlierov :** 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 5, 'n_estimators': 200.
- **Dataset so škálovaním :** 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100.
- **Dataset s ošetrením outlierov a škálovaním :** 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 5, 'n_estimators': 200.

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania	0.404998

Varianta datasetu	**presnosť**
dataset bez ošetrenia outlierov a škálovania s n-gramami	0.424813
dataset s ošetrením outlierov	**0.836027**
dataset s ošetrením outlierov a s word frequency	0.833134
dataset so skalovaním	0.392419
dataset so skalovaním a s tfidf word frequency	0.408355
dataset s ošetrením outlierov a skalovaním	0.828871
dataset s ošetrením outlierov a skalovaním a s word frequency	0.821949

Tabuľka 10.24: Vek - Testovanie 3 GBM

10.2.4 Finálny model

Najvyššiu presnosť sme dosiahli pomocou modelu :

- **Algoritmus :** Gradient Boosting Maschine
- **Dosiahnutá presnosť :** 0.8375
- **Predspracovanie :** ošetrenie odľahlých hodnôt a škálovanie
- **Metoda výberu čŕt :** RFE
- **Črty :** priemerný počet emoji na tweet, priemerný počet veľkých písmen na tweet, priemerný počet slov na vetu, priemerný počet slov na tweet, percento oznamovacích viet, percento opytovacích viet, percento rozkazovacích viet, percento iných zakončení viet, priemerný počet čiarok na tweet, priemerný počet podstatných mien na tweet, priemerný počet prídavných mien na tweet, priemerný počet slovies na tweet, priemerný počet zámen na tweet, priemerný počet čísloviek na tweet, priemerný počet príslovie k na tweet, priemerný počet predložiek na tweet, priemerný počet spojok na tweet, priemerný počet častíc na tweet, priemerný počet minulých časov na tweet, priemerný počet prítomných časov na tweet, priemerný počet podmetov na tweet, priemerný počet predmetov na tweet,

priemerný počet príslovkových určení na tweet, priemerný počet prívlastkov na tweet, priemerný počet pomocných slovies na tweet

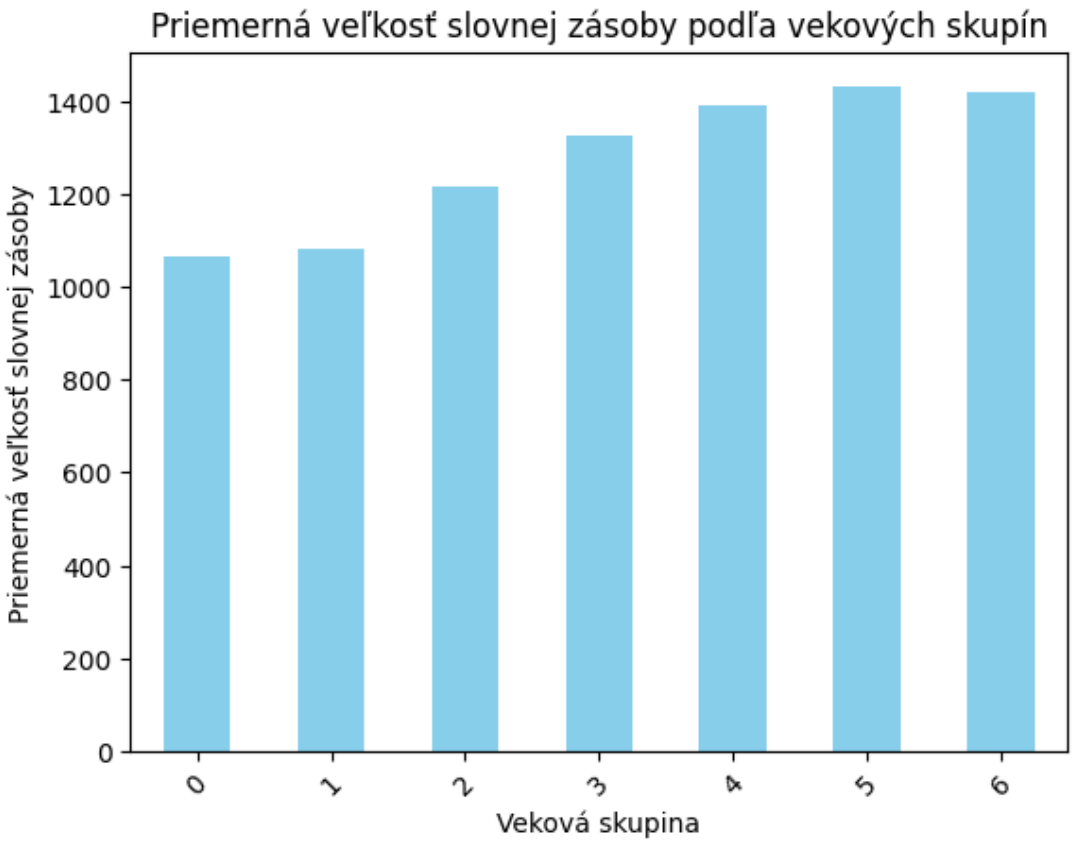
- **Textová reprezentácia :** Žiadna
- **Hyperparametre :** n_estimators=100,min_samples_split=2,min_samples_leaf=1, max_depth=3,

Kapitola 11

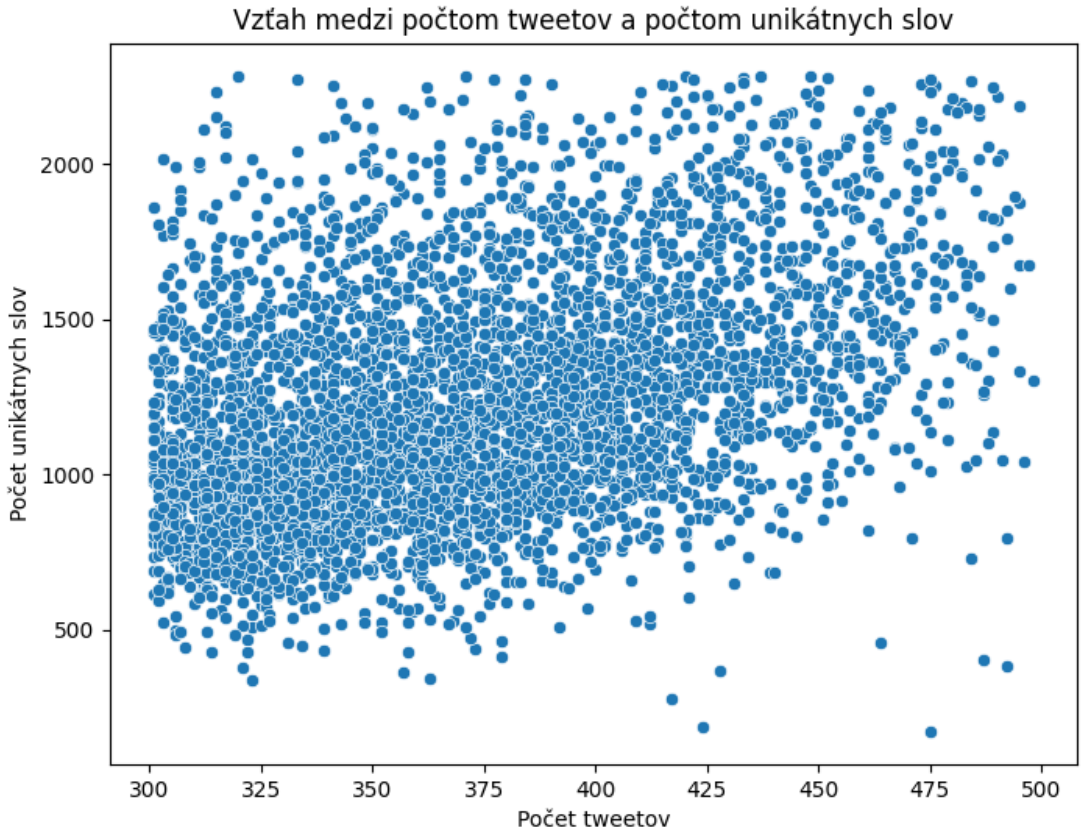
Hypotézy

11.1. H1: Rozmanitosť slovnej zásoby sa zvyšuje s vekom

Rozhodli sme sa overiť si, či hypotéza platí aj keď ošetríme vychýlené hodnoty aj keď záznamy necháme v pôvodnom stave. V prvej variante sme si vymazali vychýlené hodnoty. Za vychýlené hodnoty sme považovali dolných 25% a vrchných 25%. Následne sme si zistili priemernú slovnú zásobu pre každú vekovú skupinu a výsledky vizualizovali ako môžeme vidieť na grafe č. 11.1. Na grafe 11.1 vidíme, že sa slovná zásoba zvyšuje s vekom, však pri poslednej vekovej skupine sa mierne zníži. Rozhodli sme sa to overiť aj podľa korelačnej analýzy. Vykonali sme ako prvé Shapiro-Wilkov test aby sme zistili, či majú jednotlivé vekové skupiny normálové rozdelenie. Zistili sme, že nemali a preto sme ďalej vykonali Kruskal-Wallisov test. Jeho hodnota bola p=1.1138148016728526e-104, čo je menšie ako 0.05. Toto zistenie nám indikuje, že existujú významné rozdiely medzi vekovými skupinami. Keďže jednotlivé záznamy mohli mať 300-500 tweetov, tak nás zaujímalo, či to nemôže súvisieť s veľkosťou slovnej zásoby, teda či ľudia s viac tweetmi nemali väčšiu slovnú zásobu. Ako môžeme vidieť podľa bodového grafu č. 11.2, body nevykazujú vzor v určitom smere, skôr vyzerajú byť rozložené náhodne a zhlukované v strede. To nám naznačuje, že počet tweetov a slovná zásoba nemajú silný lineárny vzťah (koreláciu). Pri verzii s neošetrenými vychýlenými hodnotami sme dostali veľmi podobné hodnoty, taktiež priemerná slovná zásoba sa zvyšovala s vekom a pri poslednej vekovej skupine sa mierne zmenšila. Aj pri korelačnej analýze nám vyšlo, že existujú štatistické rozdiely medzi vekovými skupinami. Danú hypotézu z daných dôvodov **prijímame**.



Obr. 11.1: Veľkosť slovnej zásoby podľa vekových skupín

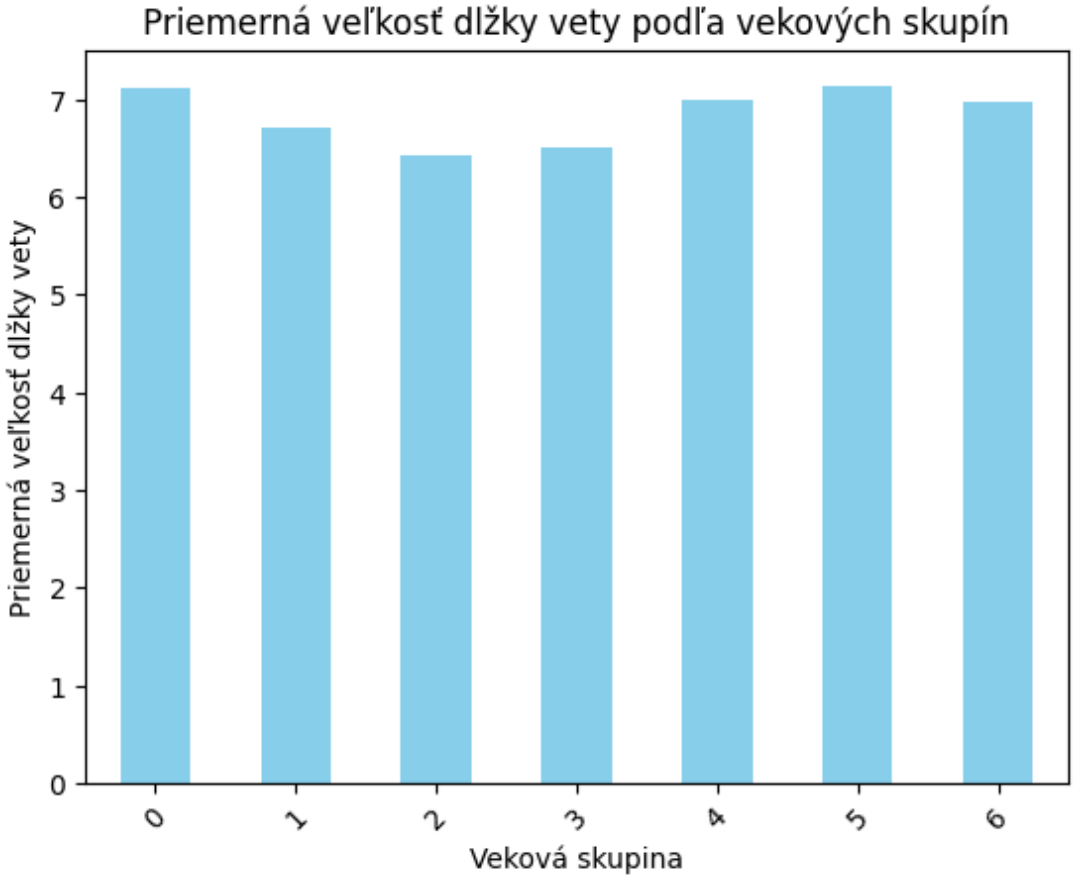


Obr. 11.2: Vzťah medzi počtom tweetov a počtom unikátnych slov

11.2. H2: Priemerná dĺžka viet sa zväčšuje s vekom

Ako aj v prvej hypotéze sa testovali hypotézu aj na variante s ošetrenými vychýlenými hodnotami aj bez ošetrenia. Ošetrili sme vychýlené hodnoty rovnakým spôsobom ako v prvej hypotéze. Zistili sme si priemer dĺžky vety pre jednotlivé skupiny a vizualizovali sme si to. Ako môžeme vidieť na grafe č. 11.3, tak na začiatku dĺžka vety sa vekom zmenšuje, potom zase stúpa a na koniec zase sa zmenšuje. Vyskúšali sme si urobiť aj štatistické testy. Testovali sme normálnosť pre každú vekovú skupinu. Keďže nemali normálové rozdelenie, tak sme vykonali Kruskal-Wallidov test, ktorým výsledkom bolo, že p je menšie ako 0.05, čo nám hovorí, že existujú významné rozdiely medzi vekovými skupinami. Pri verzii bez ošetrenia outlierov nám dĺžka vety na začiatku stúpala, potom klesala, zase stúpala, potom klesala a následne už iba stúpala. Štatistické testy znova ukázali, že existujú významné rozdiely medzi vekovými skupina. Aj keď

podľa štatistických testov sme zistili, že existujú významné rozdiely medzi vekovými skupina v obidvoch prípadoch, tak podľa grafov sme usúdili, že hypotéza by mohla platiť iba na malých častiach grafu, nie pre všetky vekové skupiny. Preto hypotézu **zamietame**.

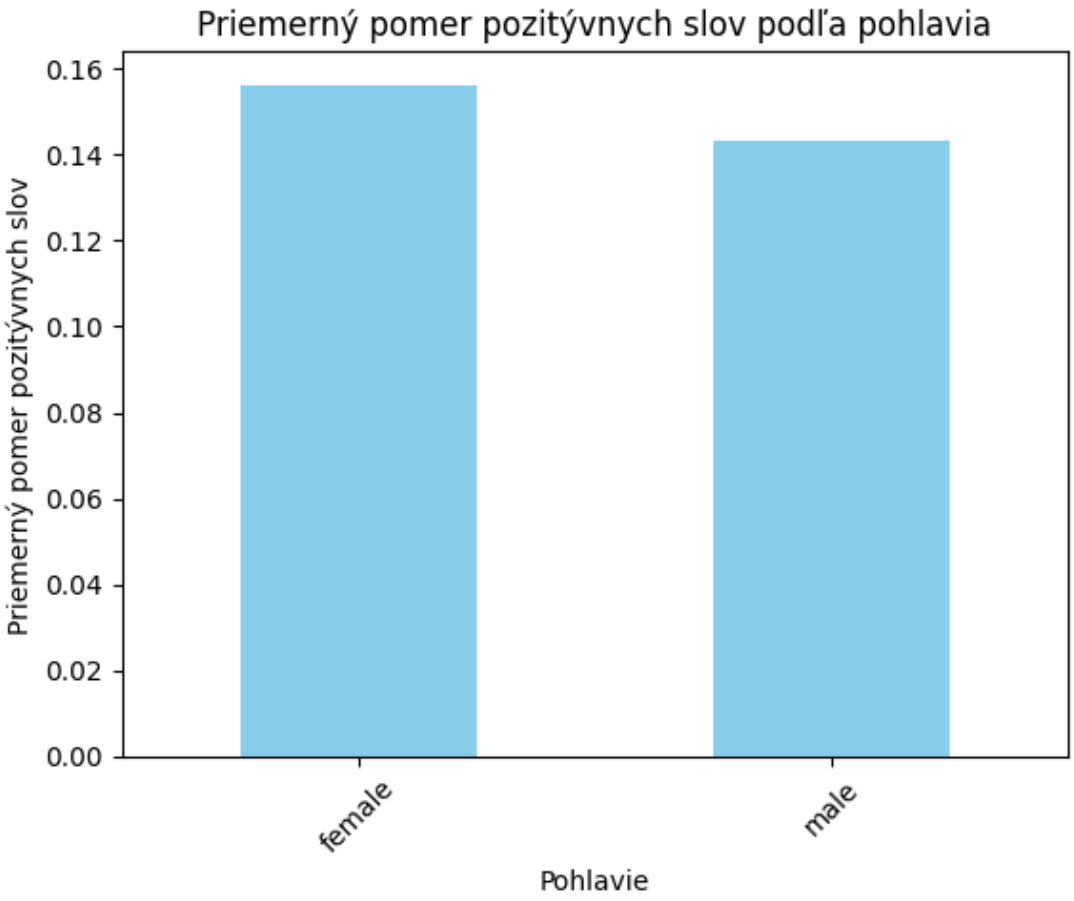


Obr. 11.3: Veľkosť dĺžky vety podľa vekových skupín

11.3. H3: Ženy používajú viac slová s pozitívnym citovým sentimentom ako muži.

Znova sme riešili hypotézu na 2 variantoch a to s ošetrenými vychýlenými hodnotami a bez ošetrenia. Vychýlené hodnoty sme riešili rovnako ako pri prvej hypotéze. Zistili sme si priemerný pomer pozitívnych slov vzhľadom ku všetkým slovám pre obidve pohlavia a vizualizovali sme si to. Ako môžeme vidieť na grafe č. 11.4, tak väčší

pomer pozitívnych slov používajú ženy. Však nie je to až taký veľký rozdiel, preto sme sa rozhodli urobiť štatistické testy aby sme zistili, či je štatisticky významný rozdiel medzi pohlaviami. Otestovali sme pomocou Shapiro-Wilkovho testu normálnosť, však hodnota p nám vyšla menšia ako 0.05, takže nemali normálne rozdelenie. Preto sme vykonali Kruskal-Wallisov test a vyšlo nám, že p=3.75974841896529e-21, čo je menšie ako 0.05. To nám hovorí, že existujú významné rozdiely medzi pohlaviami. Pre verziu s neošetrenými pohlaviami nám vyšli podobné výsledky a to, že ženy použili viac pozitívnych slov ako muži a zároveň tento rozdiel bol štatisticky významný. Preto hypotézu **prijímame**.



Obr. 11.4: Pozitívne slová podľa pohlavia

11.4. H4: Algoritmus RF dokáže priniesť lepšie výsledky ako algoritmus SVM pri určovaniu pohlavia autora

V kapitole 10 sme predpovedali pohlavie pomocou algoritmov Random Forest (RF) a Support Vector Machine (SVM). Tieto algoritmy sme testovali na datasete, ktorý obsahoval syntaktické, lexikálne a obsahové črty, spojené s rôznymi textovými reprezentáciami. Vytvorili sme tiež rôzne varianty datasetu, ktoré zahŕňali predspracovanie, ako je ošetrenie outlierov a škálovanie.

Výsledky testovania na rôznych variantách datasetu ukázali, že algoritmus Random Forest (RF) dosahuje výrazne lepšie výsledky ako Support Vector Machine (SVM). Pri použití RF sme dosiahli najvyššiu presnosť 0.91 na datasete s ošetrením outlierov, škálovaním a word frequency, zatiaľ čo najvyššia presnosť pre SVM bola 0.76 na datasete s ošetrenými outliermi a n-gramami. Tieto výsledky nám umožňujú **potvrdiť našu hypotézu**, že algoritmus Random Forest dokáže priniesť lepšie výsledky ako algoritmus Support Vector Machine pri určovaní pohlavia autora na našom datasete.

11.5. H5: Syntaktické črty sú lepšie na zistenie vekovej skupiny autora ako lexikálne črty

Rozhodli sme sa overiť túto hypotézu pomocou troch testov: test1, test2 a test3, ktoré sme vykonávali pri predpovedaní vekovej kategórie autora. Pri všetkých testoch sme použili výber čŕt pomocou metódy Recursive Feature Elimination (RFE), keďže táto metóda nám priniesla lepšie výsledky pri predchádzajúcom predpovedaní vekovej skupiny.

Každý test sme spustili samostatne pre syntaktické a lexikálne črty. Pre syntaktické črty sme používali textové reprezentácie n-gram a tf-idf n-gram. Pre lexikálne črty sme používali frekvenciu slov a tf-idf frekvenciu slov. Mali sme tiež rôzne varianty predspracovania čŕt: bez ošetrenia outlierov a škálovania, s ošetrením outlierov, so škálovaním a s ošetrením outlierov a so škálovaním.

Výsledky testov ukázali, že algoritmus Gradient Boosting Machine (GBM) dosahuje najlepšie výsledky vo všetkých testoch. Pri syntaktických črtách sme dosiahli najvyššiu presnosť 0.834559 na datasete s ošetrenými outliermi a bez textových reprezentácií. Pri lexikálnych črtách sme dosiahli najvyššiu presnosť 0.844485 na datasete s ošetrenými outliermi a škálovaním, tiež bez textových reprezentácií. Na základe týchto výsledkov môžeme zhodnotiť, že lexikálne črty poskytli lepšie výsledky na našom datasete než syntaktické črty. Preto musíme hypotézu **zamietnuť**.

Kapitola 12

Porovnanie prác s rovnakým PAN datasetom a zameraním profilácie

V tejto kapitole sa venujeme porovnaniu prístupov k profilácii autorstva, ktoré používajú rozličné metódy na rozpoznávanie demografických charakteristík autorov z textov. Naša analýza sa sústreďuje na porovnanie metodík opísaných v našej bakalárskej práci a prístupov použitých účastníkmi súťaže PAN, ktorí využívali rovnaký dataset. Zameriame sa konkrétne na dva tímy, ktoré dosiahli najlepšie výsledky v predikcii veku a pohlavia. Toto porovnanie nám umožní identifikovať kľúčové rozdiely v predspracovaní, výbere čŕt, použitých modeloch a vyhodnocovacích metódkach, ktoré viedli k ich úspechu.

V štúdii [78] sa tím sústredil na predikciu pohlavia, slávy, povolania a vekového rozpätia užívateľov Twitteru. Predspracovanie dát zahŕňalo niekoľko krokov: odstránenie retweetov, vymazanie nealfabetických znakov, nahradenie URL adries a užívateľských zmienok špeciálnymi tokenmi, odstránenie nadbytočných medzier a pridanie tokenov na konci tweetov na rozlíšenie ich začiatkov a koncov. Tieto kroky boli zamerané na čistenie textu a prípravu na efektívnu analýzu.

Naša metodológia predspracovania sa v niektorých aspektoch líšila. Nepridávali sme špeciálne tokeny a ani sme nezastupovali žiadne prvky tokenmi. Neodstraňovali sme tiež viacnásobné medzery. Avšak, ako doplnok k predspracovaniu sme odstraňovali tweety písané v iných ako anglických jazykoch a obmedzovali sme počet príspevkov

na osobu kvôli limitovaným výpočtovým a časovým kapacitám.

Pokiaľ ide o extrakciu čŕt, Radivchev a jeho tím použili TF-IDF založené na slovných a znakových n-gramoch. Na rozdiel od toho sme my extrahovali širšie spektrum čŕt vrátane lexikálnych, syntaktických a obsahových čŕt, ktoré sme kombinovali s rôznymi textovými reprezentáciami, ako sú slovné n-gramy, slovné tfidf n-gramy, frekvencie slov a tfidf frekvencie slov.

V oblasti modelovania Radivchev a jeho tím používali logistickú regresiu a SVM. My sme sa rozhodli pre širšie spektrum modelov, vrátane SVM, Random Forest a Gradient Boosting Machines (GBM). Radivchev a jeho tím hodnotili úspešnosť svojich modelov pomocou metriky F1 skóre, pričom Radivchev et al. dosiahli F1 skóre 0.92 pri predikcii pohlavia a 0.8 pri predikcii vekových skupín. Naše hodnotenie sa zameralo primárne na presnosť, kde sme dosiahli presnosť 0.95 pre pohlavie a 0.83 pre vek. Pri porovnaní výsledkov je dôležité poznamenať, že štúdia Radivcheva et al. zahrnula aj nebinárne pohlavie a delila vek do 8 skupín, zatiaľ čo naša štúdia rozlišovala len medzi mužmi a ženami a používala 7 vekových kategórií. Tento rozdiel v metodológii a metrikách robí priame porovnanie výsledkov náročnejším a nevieme povedať, ktoré riešenie bolo lepšie [78].

V štúdii [79] sa tím sústredil na profilovanie celebrít na Twitteri, pričom predikoval pohlavie, rok narodenia, povolanie a slávu na základe ich tweetov. Predspracovanie dát zahŕňalo zoskupenie tweetov od jednotlivých celebrít, obmedzenie na prvých 100 tweetov pre každú celebritu a tri úrovne čistenia textu vrátane nahradenia URL adries, zmienok a hashtagov špeciálnymi tokenmi. Tento prístup bol zameraný na zjednodušenie textov pre analýzu. V našej bakalárskej práci sme tiež implementovali predspracovanie dát, ale nepridávali sme špeciálne tokeny. Hashtagy, URL adriesy a zmienky sme použili na extrakciu niektorých čŕt a následne sme ich vymazali. Odstraňovali sme taktiež tweety písané v iných jazykoch ako angličtina. Rovnako ako ich tím, aj my sme obmedzovali počet príspevkov na osobu kvôli výpočtovým a časovým obmedzeniam, avšak naša hranica bola 500 príspevkov.

V oblasti extrakcie čŕt sa Martinc a jeho tím zameriavali predovšetkým na jed-

noduché n-gramové funkcie ako unigramy a charakterové tetragramy, ktoré boli normalizované pomocou MinMaxScaler. V našej bakalárskej práci sme použili širšie spektrum čŕt vrátane lexikálnych, syntaktických a obsahových čŕt, kombinovaných s rôznymi textovými reprezentáciami ako n-gramy, tfidf n-gramy a frekvencie slov.

V oblasti modelovania testovali rôzne klasifikátory vrátane lineárneho SVM, SVM s RBF jadrom, logistickej regresie, náhodného lesa a gradientného boosting. Napriek širokému spektru skúmaných modelov, najúspešnejšie výsledky boli dosiahnuté pomocou logistickej regresie s nastavením C=1e2 a fit_intercept=False, ktorá sa ukázala ako najefektívnejší klasifikátor pre ich úlohu V našej bakalárskej práci sme taktiež uprednostnili použitie širšieho spektra modelov, vrátane SVM, Random Forest a GBM, aby sme mohli porovnať výkonnosť rôznych prístupov. Najviac sa nám osvedčil model GBM. Martinc a jeho tím dosiahli F1 skóre 0.90 pre predikciu pohlavia a 0.06 pre predikciu veku. Pohlavie je porovnateľné s našimi výsledkami, kde naša presnosť pohlavia bola 0.95.

Pri hodnotení výsledkov štúdia [79] zahrnula predikciu nebinárneho pohlavia a rozdelila vek do 70 skupín. V našej práci sme rozdelili vek do 7 kategórií a pohlavie bolo klasifikované binárne (muž a žena). Tieto rozdiely v metodológii a metrikách komplikujú priame porovnanie efektívnosti oboch prístupov a nevieme s istotou povedať, ktoré riešenie bolo efektívnejšie.

Kapitola 13

Záver

V tejto bakalárskej práci sme sa zaoberali profilovaním autorov textov prostredníctvom analýzy stylometrie. Štúdia sa sústreďovala na vývoj a implementáciu modelov strojového učenia na určovanie demografických atribútov ako pohlavie a vek, pričom sme sa opierali o detailnú analýzu štýlu písania a jeho lingvistických charakteristík.

Naša práca začala podrobným prehľadom stylometrie, kde sme definovali kľúčové pojmy a metódy používané pri identifikácii autorského štýlu. Táto úvodná kapitola poskytla teoretické základy, ktoré boli nevyhnutné pre pochopenie následných fáz práce a zdôraznila význam stylometrie v modernej digitálnej analýze textu.

Následovala metodologická časť, kde sme sa venovali príprave a predspracovaniu dát, vyčleneniu relevantných lexikálnych, syntaktických a sémantických čŕt, ktoré sú esenciálne pre správne fungovanie klasifikačných modelov. Tento proces bol zásadný pre zabezpečenie, že dáta budú pripravené efektívne a vhodne pre strojové učenie.

Experimentálna fáza zahrnula aplikáciu viacerých algoritmov strojového učenia, kde každý z modelov bol vyhodnotený na základe jeho schopnosti predpovedať pohlavie a vek z textových dát. Diskusia o výsledkoch týchto experimentov nám umožnila porovnať účinnosť jednotlivých modelov a identifikovať tie, ktoré ponúkali najpresnejšie výsledky. Gradient Boosting Machine (GBM) sa ukázal byť najúčinnejší, pri v predpovedaní pohlavia mal presnosť až 0.959 a pri predpovedaní kategórie veku až 0.83.

V kapitole porovnávania sme naše metodiky a výsledky porovnávali s existujúcimi výskumami v oblasti stylometrie. Toto porovnanie nám umožnilo pozrieť sa na naše práce v rámci širšieho vedeckého kontextu, identifikovať príležitosti pre zlepšenia a potvrdiť, kde naše metódy excelovali oproti iným prístupom.

S ohľadom na budúci výskum navrhujeme rozšírenie modelov o ďalšiu funkcionalitu, ktorá by umožnila zahrnutie non-binárneho pohlavia. Navrhujeme taktiež ďalšie skúmanie integrácie pokročilých techník hlbokého učenia a rozšírenie implementácie o schopnosť predikovať na viacjazyčných textoch, čo by mohlo pomôcť zvýšiť schopnosti generalizácie našich modelov.

Zoznam použitej literatúry

[1] Renee Diresta et al. *DigitalCommons@University of Nebraska-Lincoln The Tactics Tropes of the Internet Research Agency*. 2019. URL: https://digitalcommons.unl.edu/senatedocs.

[2] *News - Jagiellonian University - Jagiellonian University*. URL: https://en.uj.edu.pl/en˙GB/news/-/journal˙content/56˙INSTANCE˙SxA5QO0R5BDs/81541894/141225486.

[3] François Dominic Laramée. *Introduction to stylometry with Python*. Apr. 2018. DOI: 10.46430/PHEN0078.

[4] Tempestt Neal et al. *Surveying Stylometry Techniques and Applications*. Nov. 2017. DOI: 10.1145/3132039. URL: https://dl.acm.org/doi/10.1145/3132039.

[5] *Explainable Authorship Verification in Social Media via Attention-based Similarity Learning — Papers With Code*. URL: https://paperswithcode.com/paper/explainable-authorship-verification-in-social.

[6] Michael P. Oakes. *Author Profiling and Related Applications*. Jan. 2014. DOI: 10.1093/OXFORDHB/9780199573691.013.53. URL: https://academic.oup.com/edited-volume/42643/chapter/358154668.

[7] Constantina Stamou. *Stylochronometry: Stylistic development, sequence of composition, and relative dating*. Jún 2008. DOI: 10.1093/llc/fqm029.

[8] Asad Mahmood, Zubair Shafiq a Padmini Srinivasan. *A girl has a name: Detecting authorship obfuscation*. 2020. DOI: 10.18653/v1/2020.acl-main.203.

[9] *[1909.08349] A Lexical, Syntactic, and Semantic Perspective for Understanding Style in Text*. URL: https://ar5iv.labs.arxiv.org/html/1909.08349.

[10] Fernanda López-Escobedo et al. *Analysis of Stylometric Variables in Long and Short Texts*. Okt. 2013. DOI: 10.1016/J.SBSPRO.2013.10.688.

[11] B A B Perancangan et al. *Concepts , Models ,* 2012. DOI: 10.1109/DEXA.2010.24.

[12] P G Demidov et al. *A Survey on Stylometric Text Features Ksenia Lagutina, Nadezhda Lagutina Ilya Paramonov*.

[13] Francisco Rangel a Paolo Rosso. *Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter*. URL: http://pan.webis.de.

[14] *View of Lexical diversity and lexical density in speech and writing: a developmental perspective*. URL: https://journals.lub.lu.se/LWPL/article/view/2273/1848.

[15] *N-grams in NLP — Dremio*. URL: https://www.dremio.com/wiki/n-grams-in-nlp/.

[16] Songphan Choemprayong, Fabio Crestani a Sally Jo Cunningham, ed. *Digital Libraries: Data, Information, and Knowledge for Digital Lives*. 2017. DOI: 10.1007/978-3-319-70232-2. URL: http://link.springer.com/10.1007/978-3-319-70232-2.

[17] *List of function words • Yoast*. URL: https://yoast.com/help/list-of-function-words/.

[18] *The Use of Function Word Frequencies as Indicators of Style on JSTOR*. URL: https://www.jstor.org/stable/30204237.

[19] Randy Goebel, Wolfgang Wahlster a Joerg Siekmann. *Lecture Notes in Artificial Intelligence 9811 Subseries of Lecture Notes in Computer Science LNAI Series Editors LNAI Founding Series Editor*. URL: http://www.springer.com/series/1244.

[20] Nick Moore. *Implications of various phosphoenolpyruvate-carbohydrate phosphotransferase system mutations on glycerol utilization and poly(3-hydroxybutyrate) accumulation in Ralstonia eutropha H16*. 2011. DOI: 10.1186/s40554-016-0029-x. URL: http://shura.shu.ac.uk/12219/.

[21] *Transition Words and Phrases to Improve Your Writing — Grammarly Blog*. URL: https://www.grammarly.com/blog/transition-words-phrases/.

[22] Braja Gopal Patra et al. *Automatic Author Profiling Based on Linguistic and Stylistic Features Notebook for PAN at CLEF 2013*. URL: http://weka.wikispaces.com/Use+WEKA+in+your+Java+code.

[23] *[1909.08349] A Lexical, Syntactic, and Semantic Perspective for Understanding Style in Text*. URL: https://ar5iv.labs.arxiv.org/html/1909.08349.

[24] *Gentle Introduction To Text Representation - Part - 1 — by Sundaresh Chandran — Towards Data Science*. URL: https://towardsdatascience.com/introduction-to-text-representations-for-language-processing-part-1-dc6e8068b8a4.

[25] Francisco Rangel a Paolo Rosso. *Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter*. URL: http://pan.webis.de.

[26] *From Traditional to Modern: A Comprehensive Guide to Text Representation Techniques in NLP — by Susovan Dey — Medium*. URL: https://deysusovan93.medium.com/from-traditional-to-modern-a-comprehensive-guide-to-text-representation-techniques-in-nlp-369946f67497.

[27] *N-gram Language Modeling in Natural Language Processing - KDnuggets*. URL: https://www.kdnuggets.com/2022/06/ngram-language-modeling-natural-language-processing.html.

[28] *Text Representation Techniques. The Complete NLP Guide: Text to Context. . . — by Merve Bayram Durna — Medium*. URL: https://medium.com/@mervebdurna/text-representation-techniques-d40741eb0916.

[29] *Text Processing: What Is It?* URL: https://monkeylearn.com/blog/text-processing/.

[30] *5 Natural Language Processing Techniques for Extracting Information*. URL: https://blog.aureusanalytics.com/blog/5-natural-language-processing-techniques-for-extracting-information.

[31] *What is Tokenization? Types, Use Cases, Implementation — DataCamp*. URL: https://www.datacamp.com/blog/what-is-tokenization#rdl.

[32] Elias Lundeqvist a Maria Svensson. *Examensarbete 30 hp Author profiling: A machine learning approach towards detecting gender, age and native language of users in social media*. 2017. URL: http://www.teknat.uu.se/student.

[33] *Tokenization — A complete guide. Natural Language Processing—NLP From... — by Utkarsh Kant — Medium*. URL: https://medium.com/@utkarsh.kant/tokenization-a-complete-guide-3f2dd56c0682.

[34] *Tokenization — A complete guide. Natural Language Processing—NLP From... — by Utkarsh Kant — Medium*. URL: https://medium.com/@utkarsh.kant/tokenization-a-complete-guide-3f2dd56c0682.

[35] *2— Stemming Lemmatization in NLP: Text Preprocessing Techniques — by Aysel Aydin — Medium*. URL: https://ayselaydin.medium.com/2-stemming-lemmatization-in-nlp-text-preprocessing-techniques-adfe4d84ceee.

[36] *Stemming and Lemmatization – Study Machine Learning*. URL: https://studymachinelearning.com/stemming-and-lemmatization/.

[37] *What is Parts of Speech (POS) Tagging Natural Language Processing?In what kind of applications we can use Parts of Speech (POS) Tagging in Natural Language Processing. — by Sujatha Mudadla — Medium*. URL: https://medium.com/@sujathamudadla1213/what-is-parts-of-speech-pos-tagging-natural-language-processing-in-2b8f4b07b186.

[38] *What is Machine Learning? — IBM*. URL: https://www.ibm.com/topics/machine-learning.

[39] *What is Machine Learning and How Does It Work? In-Depth Guide*. URL: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML.

[40] Jacques Savoy. *Machine Learning Methods for Stylometry*.

[41] *Machine Learning Classifiers: Definition and 5 Types — Indeed.com*. URL: https://www.indeed.com/career-advice/career-development/classifiers-in-machine-learning.

[42] *Classifier*. URL: https://c3.ai/glossary/data-science/classifier/.

[43] *Support Vector Machine (SVM) Algorithm - GeeksforGeeks*. URL: https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

[44] *Guide on Support Vector Machine (SVM) Algorithm*. URL: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/#h-how-does-support-vector-machine-work.

[45] Rodwan Bakkar Deyab, José Duarte a Teresa Gonçalves. *Author Profiling Using Support Vector Machines Notebook for PAN at CLEF 2016*. URL: http://pan.webis.de/clef16/pan16-web/author-profiling.html.

[46] Joachim Diederich et al. "Authorship attribution with support vector machines". In: *Applied Intelligence* 19 (1-2 júl 2003), s. 109–123. ISSN: 0924669X. DOI: 10.1023/A:1023824908771/METRICS. URL: https://link.springer.com/article/10.1023/A:1023824908771.

[47] *Decision Tree Algorithm - A Complete Guide - Analytics Vidhya*. URL: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/.

[48] *Building a Random Forest Model: A Step-by-Step Guide*. URL: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

[49] Alonso Palomino-Garibay et al. *A Random Forest Approach for Authorship Profiling*. URL: http://pan.webis.de/.

[50] Alonso Palomino-Garibay et al. *A Random Forest Approach for Authorship Profiling*. URL: http://pan.webis.de/.

[51] Mahmoud Khonji, Youssef Iraqi a Andrew Jones. *An evaluation of authorship attribution using random forests*. Júl 2015. DOI: 10.1109/ICTRC.2015.7156423.

[52] *Logistic Regression in Machine Learning - GeeksforGeeks*. URL: https://www.geeksforgeeks.org/understanding-logistic-regression/.

[53] *What is Logistic Regression and How Does it Work?* URL: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/.

[54] Liliya Akhtyamova, John Cardiff a Andrey Ignatov. *Twitter Author Profiling Using Word Embeddings and Logistic Regression Notebook for PAN at CLEF 2017*.

[55] *Learn Naive Bayes Algorithm : Naive Bayes Classifier Examples*. URL: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[56] *Naive Bayes Classifiers - GeeksforGeeks*. URL: https://www.geeksforgeeks.org/naive-bayes-classifiers/.

[57] James Conigliaro. *Author Identification Using Naïve Bayes Classification*.

[58] *Gradient Boosting in ML - GeeksforGeeks*. URL: https://www.geeksforgeeks.org/ml-gradient-boosting/.

[59] *Gradient Boosting: A Step-by-Step Guide*. URL: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/.

[60] *Gradient Boosting for Classification — Paperspace Blog*. URL: https://blog.paperspace.com/gradient-boosting-for-classification/.

[61] Youngjun Joo a Inchon Hwang. *Author Profiling on Social Media: An Ensemble Learning Model using Various Features Notebook for PAN at CLEF 2019*. URL: https://github.com/grantjenks/python-wordsegment.

[62] Ahmed Abbasi et al. *Authorship identification using ensemble learning*. Jún 2022. DOI: 10.1038/s41598-022-13690-4. URL: https://www.nature.com/articles/s41598-022-13690-4.

[63] *Gradient Boosting – What You Need to Know — Machine Learning — DATA SCIENCE*. URL: https://datascience.eu/machine-learning/gradient-boosting-what-you-need-to-know/.

[64] *Evaluation Metrics for Classification — by Python Programmer — Medium*. URL: https://medium.com/@impythonprogrammer/evaluation-metrics-for-classification-fc770511052d.

[65] *Accuracy, precision, and recall in multi-class classification*. URL: https://www.evidentlyai.com/classification-metrics/multi-class-metrics.

[66] *Feature Selection Techniques in Machine Learning (Updated 2024)*. URL: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/.

[67] *Feature Selection Techniques in Machine Learning - GeeksforGeeks*. URL: https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/.

[68] Yaakov HaCohen-Kerner. *Survey on profiling age and gender of text authors*. Aug. 2022. DOI: 10.1016/J.ESWA.2022.117140.

[69] Sarra Ouni, Fethi Fkih a Mohamed Nazih Omri. *A survey of machine learning-based author profiling from texts analysis in social networks*. Okt. 2023. DOI: 10.1007/S11042-023-14711-8/TABLES/9. URL: https://link.springer.com/article/10.1007/s11042-023-14711-8.

[70] Md Ataur Rahman a Yeasmin Ara Akter. *Multi-lingual author profiling: predicting gender and age from tweets!* 2021. DOI: 10.1007/978-3-030-51859-2˙46/TABLES/9. URL: https://link.springer.com/chapter/10.1007/978-3-030-51859-2˙46.

[71] Siddharth D Jaiswal, Ankit Verma a Animesh Mukherjee. "Auditing Gender Analyzers on Text Data". In: (). URL: https://forms.gle/VVSrTxoYKPW16vCYA.

[72] *PAN at CLEF 2019 - Celebrity Profiling*. URL: https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html.

[73] *CLEF 2024 — Conference and Labs of the Evaluation Forum*. URL: https://clef2024.imag.fr/index.php?page=Pages/lab˙pages/pan.html.

[74] *What is Twitter?* URL: https://www.techtarget.com/whatis/definition/Twitter.

[75] *sklearn.svm.SVC — scikit-learn 1.4.2 documentation*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[76] *sklearn.ensemble.RandomForestClassifier — scikit-learn 1.4.2 documentation*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[77] *sklearn.ensemble.GradientBoostingClassifier — scikit-learn 1.4.2 documentation*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.

[78] Victor Radivchev, Alex Nikolov a Alexandrina Lambova. *Celebrity Profiling using TF-IDF, Logistic Regression, and SVM Notebook for PAN at CLEF 2019*. 2019.

[79] Matej Martinc, Blaž Škrlj a Senja Pollak. *Who Is Hot and Who Is Not? Profiling Celebs on Twitter Notebook for PAN at CLEF 2019*. URL: http://pan.webis.de/.