UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

KOMPARATÍVNA ANALÝZA ML MODELOV V KONTEXTE KLASIFIKÁCIE SENTIMENTU V BANKOVNÍCTVE

DIPLOMOVÁ PRÁCA

2026 **BC. TAMARA HLADKÁ**

UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

KOMPARATÍVNA ANALÝZA ML MODELOV V KONTEXTE KLASIFIKÁCIE SENTIMENTU V BANKOVNÍCTVE

DIPLOMOVÁ PRÁCA

Študijný odbor:	18. Informatika
Študijný program:	Aplikovaná informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	prof. RNDr. Michal Munk, PhD.

Nitra 2026 Bc. Tamara Hladká

ZADANIE ZP

Univerzita Konštantína Filozofa v Nitre
Fakulta prírodných vied a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Tamara Hladká
Študijný program:	aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor:	informatika
Typ záverečnej práce:	Diplomová práca
Jazyk záverečnej práce:	slovenský
Sekundárny jazyk:	anglický

Názov: Komparatívna analýza ML modelov v kontexte klasifikácie sentimentu v bankovníctve

Anotácia: Analýza sentimentu v bankovníctve sa ukázala ako účinný nástroj pre pochopenie názorov, nálad a postojov klientov a trhu. Pomáha bankám monitorovať a chrániť ich reputáciu, predchádzať krízovým situáciám a optimalizovať marketingové stratégie podľa aktuálnych nálad na trhu. Tieto poznatky podporujú strategické rozhodovanie, čo zvyšuje konkurencieschopnosť a efektívnosť banky. Napriek existencii množstva predtrénovaných modelov pre účel klasifikácie sentimentu, len málo z nich, obzvlášť pri jazykoch s obmedzenými zdrojmi (low-resource languages), je vhodných pre doménu bankovníctva. Použitie generických modelov nemusí byť efektívne vzhľadom k doménovo špecifickému žargónu a kontextu. Finančné termíny a odborné výrazy patria medzi faktory, ktoré môžu ovplyvniť presnosť klasifikácie.

Cieľom diplomovej práce je identifikovať najvhodnejší model pre presné určenie sentimentu (pozitívny, negatívny, neutrálny) v textoch týkajúcich sa bankového sektora pre jazyk s obmedzenými zdrojmi. Identifikácia najvhodnejšieho modelu bude vyplývať z komparatívnej analýzy výkonu natrénovaných modelov z kategórií symbolické, subsymbolické a skupinové. Teoretická rovina práce pokryje súvisiace štúdie zaoberajúce sa problematikou klasifikácie sentimentu v bankovníctve ako aj popisom princípov činnosti jednotlivých modelov. Metodologicky sa práca oprie o metódy spracovania prirodzeného jazyka (NLP), vrátane techník prípravy a reprezentácie dát. Výsledky tejto práce poskytnú cenné informácie o efektivite rôznych modelov strojového učenia pri analýze sentimentu v bankovníctve.

Charakter práce:
Výskumný – stanovenie výskumného problému, metodika výskumu, výsledky výskumu (štatistická interpretácia), interpretácia výsledkov výskumu (vecná interpretácia).

Predmetové prerekvizity:
Úvod do spracovania prirodzeného jazyka (2.,Mgr.)
Úvod do strojového učenia (1., Mgr.)
Objavovanie znalostí (1., Mgr.)
Neurónové siete (1., Mgr)

Univerzita Konštantína Filozofa v Nitre
Fakulta prírodných vied a informatiky

Počítačová analýza dát (3., Bc.)

Najdôležitejšie kompetentnosti získané spracovaním témy:
syntetizovať výskumné publikácie
analyzovať veľké dáta (big data)
vykonať analytické matematické výpočty
vykonať hĺbkovú analýzu údajov
realizovať procesy dátovej kvality
referovať o výsledkoch analýzy

Školiteľ: prof. RNDr. Michal Munk, PhD.
Konzultant: RNDr. Lívia Kelebercová, PhD.
Oponent: Mgr. František Forgáč, PhD.
Katedra: KI - Katedra informatiky

Dátum zadania: 30.10.2024

Dátum schválenia: 20.11.2024

prof. RNDr. Michal Munk, PhD., v. r.
osoba zodpovedná za realizáciu študijného programu

POĎAKOVANIE

Rada by som sa poďakovala svojmu školiteľovi prof. RNDr. Michalovi Munkovi, PhD. a konzultantke RNDr. Lívii Kelebercovej, PhD. za odborné vedenie, ochotu a cenné rady pri písaní diplomovej práce.

ABSTRAKT

HLADKÁ, Tamara: Komparatívna analýza ml modelov v kontexte klasifikácie sentimentu v bankovníctve. [Diplomová práca]. Univerzita Konštantína Filozofa v Nitre. Fakulta prírodných vied a informatiky. Školiteľ: prof. RNDr. Michal Munk, PhD. Stupeň odbornej kvalifikácie: Magister odboru Aplikovaná informatika. Nitra: FPVaI, 2026. 68 s.

Práca porovnáva vybrané modely strojového učenia a jazykové modely pri klasifikácii sentimentu (pozitívny, neutrálny, negatívny) v textoch z bankového sektora v slovenskom jazyku, ktorý predstavuje jazyk s obmedzenými zdrojmi. Riešenie je navrhnuté a realizované v súlade s metodikou CRISP-DM (Cross-Industry Standard Process for Data Mining), od porozumenia problému a dát, cez predspracovanie textov, až po modelovanie a vyhodnotenie. V praktickej časti sa využívajú textové reprezentácie Bag-of-Words a TF-IDF, vrátane variantu s redukciou dimenzie pomocou SVD, a sú aplikované symbolické, subsymbolické, skupinové a jazykové modely na ručne anotovaných dátach. Cieľom práce je systematicky porovnať tieto prístupy a poskytnúť prehľad o vhodných riešeniach pre klasifikáciu sentimentu v slovenskom jazyku.

Kľúčové slová: Analýza sentimentu. Strojové učenie. Jazykové modely. Bankovníctvo. Spracovanie prirodzeného jazyka.

ABSTRACT

HLADKÁ, Tamara: Comparative analysis of ML models in the context of sentiment classification in banking. [Master Thesis]. Constantine the Philosopher University in Nitra. Faculty of Natural Sciences and Informatics. Supervisor: prof. RNDr. Michal Munk, PhD. Degree of Qualification: Master of Applied Informatics. Nitra: FNSaI, 2026. 68 p.

This thesis compares selected machine learning models and language models for sentiment classification (positive, neutral, negative) in texts from the banking sector in Slovak language, which is considered a low-resource language. The solution is designed and implemented in accordance with the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, ranging from problem and data understanding, through text preprocessing to modelling and evaluation. In the practical part, Bag-of-Words and TF-IDF text representations are used, including a variant with dimensionality reduction using SVD. Symbolic, subsymbolic, ensemble and language models are then applied to manually annotated datasets. The aim of the thesis is to systematically compare these approaches and provide an overview of suitable methods for sentiment classification in the Slovak language.

Keywords: Sentiment analysis. Machine learning. Language models. Banking. Natural language processing.

OBSAH

ÚVOD

Analýza sentimentu predstavuje významnú disciplínu v rámci spracovania prirodzeného jazyka, ktorej cieľom je identifikácia a interpretácia emócií alebo názorov vyjadrených v texte. V posledných rokoch sa uplatňuje v mnohých oblastiach, pričom dôležitosť nadobúda aj v doméne bankovníctva. V tejto oblasti umožňuje bankovým inštitúciám efektívnejšie monitorovať reputáciu, predchádzať krízovým situáciám, optimalizovať marketingové stratégie a prispôsobovať komunikáciu klientom na základe aktuálnych nálad na trhu.

Napriek neustálemu výskumu v oblasti strojového učenia zostáva aplikácia týchto metód v špecifických doménach obmedzená, najmä pri jazykoch s obmedzenými zdrojmi, ako je slovenčina. Väčšina existujúcich modelov je trénovaná na veľkých korpusoch v anglickom jazyku, pričom ich prenos do iných jazykových a tematických kontextov je často problematický.

Cieľom diplomovej práce je identifikovať najvhodnejší model pre klasifikáciu sentimentu v textoch z oblasti bankovníctva pre jazyk s obmedzenými zdrojmi, a to prostredníctvom komparatívnej analýzy symbolických, subsymbolických a skupinových modelov.

Teoretická časť práce sa venuje analýze sentimentu, jej princípom a metódam spracovania. Osobitný dôraz je kladený na metódy strojového učenia, ktoré predstavujú jadro moderných prístupov k automatickej klasifikácii sentimentu. Táto časť popisuje rôzne metódy, ktoré sa líšia mierou interpretovateľnosti, zložitosti a spôsobom učenia. Získané teoretické poznatky tvoria základ pre praktickú časť práce, ktorá je zameraná na spracovanie a analýzu dát z oblasti bankovníctva, vrátane prípravy dát, aplikácie rôznych modelov strojového učenia, ich hodnotenia a porovnania výkonu s cieľom identifikovať najvhodnejší model pre klasifikáciu sentimentu.

Zistenia tejto práce prispejú k lepšiemu pochopeniu efektívnosti rôznych modelov strojového učenia pri analýze sentimentu v bankovníctve a môžu poslúžiť ako podklad pre ďalší výskum a praktické aplikácie v oblasti finančnej textovej analytiky.

1 ANALÝZA SÚČASNÉHO STAVU

Metódy strojového učenia a spracovania prirodzeného jazyka (NLP) zohrávajú dôležitú rolu pri extrakcii užitočných informácií z veľkých objemov textových dát. Tieto prístupy umožňujú analyzovať a klasifikovať rôzne aspekty komunikácie, ako sú názory, zámery alebo postoje používateľov, čím poskytujú cenné poznatky, využiteľné ako podklad pre ďalšie analytické a strategické rozhodovanie.

Jednou z kľúčových oblastí spracovania textových dát, ktorá má čoraz väčší význam, je analýza sentimentu. Ako uvádzajú Xu et al. (2022), rýchly rozvoj informačných technológií a sociálnych médií výrazne ovplyvnil každodenný život ľudí. Tieto platformy umožňujú používateľom nielen jednoducho komunikovať a zdieľať obsah, ale aj vyjadrovať svoje názory a zverejňovať recenzie na rôzne produkty, služby a organizácie. Tento neustále rastúci objem používateľmi generovaného obsahu vyvolal záujem firiem, vládnych inštitúcií a organizácií o spracovanie týchto dát s cieľom sledovať verejnú mienku a podporovať rozhodovanie.

1.1 ANALÝZA SENTIMENTU

Analýza sentimentu predstavuje oblasť, ktorá sa zameriava na identifikáciu a hodnotenie subjektívnych aspektov textu. Cieľom je kategorizovať, či má daný text pozitívny, negatívny alebo neutrálny postoj, pričom je možné túto klasifikáciu rozšíriť tak, aby rozlišovala aj jemnejšie emócie ako sú radosť, smútok alebo hnev.

Sentiment možno hodnotiť na rôznych úrovniach v závislosti od typu lexikálnej jednotky, od jednotlivých slov, viet, až po celé texty, kedy výber vhodnej úrovne závisí od cieľa analýzy, pričom jemnejšie úrovne dokážu detailnejšie zachytiť emócie a vyššie úrovne poskytujú prehľadnejší, ale menej detailný výsledok analýzy.

Pojem „sentiment analysis“ sa v odborných publikáciách pravdepodobne prvýkrát objavil v práci od Nasukawa a Yi (2003). Rýchlo získal pozornosť výskumnej komunity, najmä v súvislosti s analýzou webových recenzií a viedol k rozvoju metód pre automatickú identifikáciu a členenie názorov v textoch. Dave et al. (2003) následne prvýkrát použili výraz „opinion mining“ a vytvorili nástroj, ktorý dokázal automaticky identifikovať pozitívne a negatívne hodnotenia jednotlivých produktov. Hoci predstavili prakticky využiteľný prístup, ich metódy narazili na viacero zásadných obmedzení. Výsledky klasifikácie boli nestabilné a modely trpeli problémom preučenia,

čo naznačuje nízku mieru generalizácie na nové dáta. Rovnako chýbal aj kvalitne anotovaný korpus, ktorý by umožnil lepšie trénovanie modelov a objektívnejšie porovnanie ich výkonu.

Hoci nástroje na analýzu sentimentu vznikli predovšetkým na spracovanie recenzií a spätnej väzby k produktom, ich uplatnenie sa postupne rozšírilo aj do vysoko špecializovaných oblastí, akými sú financie a bankovníctvo. Yusuf et al. (2018) uvádzajú, že veľká časť zákazníkov sa pred kúpou produktu spolieha na eWOM (elektronické odporúčania) a hodnotenia iných používateľov. Podobne ako pri spotrebiteľských produktoch, aj v bankovom sektore môžu názory a recenzie klientov významne ovplyvniť dôveru a rozhodovanie ostatných klientov. Pochopenie nálad a postojov zákazníkov prostredníctvom analýzy sentimentu umožňuje bankám monitorovať reputáciu, identifikovať možné problémy a optimalizovať marketingové stratégie alebo zákaznícke služby. Tento význam sentimentu v bankovom sektore potvrdzujú aj viaceré štúdie, ktoré skúmali jeho aplikáciu na rôznych typoch dát a jeho vplyv na rozhodovanie bánk a správanie klientov.

Correa et al. (2017) analyzujú vo svojej štúdii vzťah medzi sentimentom v správach o finančnej stabilite (FSR) a vývojom finančného cyklu. Autori vytvorili špecializovaný slovník, na základe ktorého zostavili Financial Stability Sentiment Index (FSS). Výsledky tejto štúdie ukazujú, že sentiment centrálnych bánk úzko súvisí s vývojom v bankovom sektore a má prediktívnu hodnotu pri identifikácii rôznych finančných rizík a kríz. Zároveň poukazujú na to, že včasná a efektívna komunikácia rizík zo strany centrálnych bánk môže prispieť k zníženiu pravdepodobnosti vzniku finančných kríz.

Štúdia Mohanty a Cherukuri (2023) sa zameriava na analýzu sentimentu v bankovníctve pomocou nového, inovatívneho prístupu, ktorý kombinuje tradičné NLP techniky s využitím synoným a antoným na zlepšenie presnosti klasifikácie. Autori vytvorili vlastný sentimentový slovník, v ktorom najprv priradili k jednotlivým slovám skóre sentimentu, a potom tento dataset rozšírili generovaním viet pomocou synoným a antoným. Model ELECTRA, natrénovaný na pripravených dátach, dosiahol vysokú presnosť (92-93%) pri klasifikácii bankových recenzií. Výsledky indikujú, že kombinácia lexikálnych prístupov a modelov hlbokého učenia môže prispieť k zvýšeniu presnosti analýzy sentimentu v oblasti bankovníctva. Medzi možnosti ďalšieho vývoja patrí rozšírenie slovníka o viac synoným a antoným, použitie ďalších

bankových datasetov a implementáciu pokročilých modelov strojového učenia pre detailnejšie hodnotenie sentimentu.

Na sentimentovú a tématickú analýzu používateľských recenzií mobilných bankových aplikácií sa zameriava výskum Mahmood et al. (2023). Autori použili lexikónovo-pravidlový prístup v kombinácii s viacerými modelmi strojového učenia na klasifikáciu sentimentu, a model Top2Vec na analýzu pozitívnych a negatívnych tém. Výsledky ukázali, že najlepšiu presnosť klasifikácie dosiahol skupinový model, zatiaľ čo tématická analýza odhalila kľúčové faktory, ktoré ovplyvňujú spokojnosť a nespokojnosť používateľov. Štúdia poukazuje na prínos kombinácie sentimentovej a tématickej analýzy pre zlepšenie používateľskej skúsenosti a kvality digitálnych bankových služieb.

Hajek a Munk (2023) sa zameriavajú na vplyv emocionálnych znakov z konferenčných hovorov spoločností na predikciu finančných ťažkostí. Autori navrhli model založený na hlbokom učení, ktorý kombinuje textové a zvukové dáta, konkrétne analýzu sentimentu v transkriptoch a rozpoznávanie emócií z reči. Výsledky ukázali, že aj textové, aj hlasové emócie poskytujú dôležité informácie na predpoveď finančných problémov. Negatívne emócie často symbolizovali finančné ťažkosti, zatiaľ čo pozitívne emócie signalizovali dobrý finančný výkon. Avšak autori upozorňujú na určité obmedzenia: emocionálne stavy sa môžu počas hovoru meniť, čo môže ovplyvniť presnosť analýzy, a rozdelenie firiem podľa úrovne finančných ťažkostí bolo zjednodušené.

Výskum Huang et al. (2023) rozširuje využitie sentimentu na predikciu finančnej tiesne podnikov v Číne. Autori analyzovali sentiment v audítorských správach a manažérskych diskusiách pomocou hlbokých neurónových sietí, a z výsledkov zistili, že sentiment výrazne zlepšuje presnosť predikčných modelov. Sentiment odvodený z audítorských správ mal vyššiu prediktívnu hodnotu ako z manažérskych diskusií, pričom predpokladom je, že manažéri často prezentujú optimistickejší obraz o výkonnosti podniku. Štúdia zdôrazňuje význam sentimentu ako doplňujúci zdroj informácií pre predikciu finančnej stability a pre modelovanie rizík.

Z uvedených štúdií vyplýva, že analýza sentimentu predstavuje univerzálny a flexibilný nástroj, ktorý umožňuje bankám lepšie pochopiť svojich klientov, predikovať finančné zmeny a poskytovať podporu pri rozhodovacích procesoch. Rovnako sa ukazuje, že aplikácia analýzy sentimentu nie je obmedzená len

na analýzu zákazníckych recenzií, ale zahŕňa aj správy centrálnych bánk, audítorské správy alebo manažérske diskusie. Tieto poznatky potvrdzujú, že sentiment predstavuje významný zdroj neštruktúrovaných informácií, ktorý dokáže obohatiť tradičné ukazovatele a tak prispievať k presnejšiemu hodnoteniu finančnej stability a rizika bankových inštitúcií.

1.2 PRÍSTUPY K ANALÝZE SENTIMENTU

Analýza sentimentu sa realizuje rôznymi metódami, pričom najčastejšie sa využívajú prístupy založené na lexikónoch a prístupy založené na strojovom učení. V posledných rokoch sa objavujú aj hybridné riešenia, ktoré spájajú silné stránky oboch prístupov s cieľom zvýšiť presnosť klasifikácie a umožniť hlbšie pochopenie významu textu (Verma a Thakur 2018).

1.2.1 Prístupy založené na lexikónoch

Tento prístup využíva lexikóny, teda slovníky, alebo zoznamy slov obsahujúce informácie o tom, ktoré slová a frázy majú pozitívny, a ktoré negatívny význam. Tieto slová, často označované ako hodnotiace slová, sa bežne používajú na vyjadrenie určitého postoja, napríklad „úžasný“ a „výborný“ majú pozitívne hodnotenie, zatiaľ čo „slabý“ alebo „smutný“ sú považované za negatívne. Vďaka týmto lexikónom vieme identifikovať a určiť orientáciu sentimentu v texte a metódy môžu byť aplikované na rôznych úrovniach textu, od jednotlivých slov, cez vety, až po celé dokumenty, pričom výber úrovne závisí od cieľa analýzy. (Mao et al. 2024)

Lexikóny možno vytvárať rôznymi spôsobmi. Manuálne zostavené slovníky sú časovo náročné a preto sa často používajú ako doplnková kontrola pre iné lexikóny. Slovníkové metódy vychádzajú z existujúcich lexikografických zdrojov ako sú WordNet alebo HowNet, ktoré umožňujú rozšíriť pôvodnú množinu slov o synonymá a antonymá. Korpusové metódy naopak používajú počiatočné slová, takzvané „seed words“, ktorých polaritu poznáme vopred. Vďaka týmto počiatočným slovám vieme identifikovať nové slová a ich sentiment na základe vzorov spoločného výskytu, čím sa lexikón postupne rozširuje. Pri tvorbe lexikónov sa často vychádza z prídavných mien, ktoré slúžia ako indikátory sémantickej orientácie textu. Pre každý relevantný pojem sa určí hodnota sentimentu, ktorá vyjadruje jeho polaritu (pozitívny, negatívny

alebo neutrálny). Tieto hodnoty sa následne agregujú do jedného celkového skóre, ktoré reprezentuje sentiment textu ako celku (Khoo a Johnkhan 2018).

Lexikónové prístupy čelia v praxi viacerým výzvam. Texty z online prostredia, ako sú príspevky na sociálnych sieťach, často obsahujú neštandardné skratky a výrazy, ktoré nie sú obsiahnuté v existujúcich slovníkoch. Ďalším problémom je správne spracovanie negácií a modifikátorov, ktoré môžu zmeniť význam vety (napríklad „nie je zlý“ má pozitívny význam). Lexikóny všeobecného použitia, ako je SentiWordNet, majú obmedzené pokrytie doménovo špecifických výrazov, keďže význam slova sa môže meniť na základe kontextu. Výkonnosť lexikónových metód preto závisí najmä od kvality použitého slovníka a od toho, ako dobre zodpovedá konkrétnej úlohe. Napriek týmto obmedzeniam má táto jednoduchá a transparentná povaha lexikónových metód aj svoje výhody. Takéto spracovanie textu umožňuje lepšie pochopiť výslednú klasifikáciu, ako aj odhaliť rôzne psychologické a lingvistické mechanizmy, ktoré ovplyvňujú vyjadrený sentiment (Sadia et al. 2018; Hartmann et al. 2023).

V oblasti financií a bankovníctva sa lexikónové metódy prispôsobujú doménovému jazyku, ktorý sa od bežného jazyka líši špecifickou terminológiou a interpretáciou niektorých významov. Preto vznikli špecializované slovníky, ako je Loughran-McDonaldov zoznam slov o finančnom sentimente (Loughran–McDonald Financial Sentiment Word Lists), ktorý patrí medzi najväčšie a najpoužívanejšie nástroje na analýzu sentimentu vo finančnej sfére v rámci anglického jazyka. Bol vytvorený na základe finančných správ a obsahuje veľké množstvo špecializovaných slov, ktoré sa v bežných sentimentových lexikónoch nenachádzajú, alebo majú iný význam v závislosti od domény. Tento slovník je rozdelený na viacero kategórií podľa významu slov, kde okrem negatívnych slov obsahuje aj pozitívne výrazy, slová vyjadrujúce neistotu, právne termíny, ako aj silné a slabé modálne výrazy, ktoré spolu zachytávajú mieru istoty tvrdení v textoch finančných správ (Loughran a Mcdonald 2011). Ďalším príkladom lexikónu zameraného na finančný žargón je Slovník sentimentu finančnej stability (Financial Stability Sentiment (FSS) Dictionary), ktorý bol vytvorený pre analýzu správ centrálnych bánk o finančnej stabilite. Vyvinuli ho Correa et al. (2017; 2021) na základe analýzy komunikácie centrálnych bánk z 35 krajín, s dôrazom na slová vyjadrujúci pozitívny alebo negatívny sentiment zameraný na správy o finančnej stabilite. Novším príkladom je FinSenticNet, ktorý rozširuje

SenticNet o finančný kontext. SenticNet je lexikón určený na analýzu sentimentu, ktorý priraďuje hodnoty slovám a konceptom na základe ich sémantického vzťahu. FinSenticNet sa líši od ostatných finančných lexikónov v tom, že viac ako 65% jeho záznamov sú viacslovné výrazy, vďaka čomu dokáže lepšie zachytávať významové nuansy typické pre finančný žargón (Du et al. 2023).

1.2.2 Prístupy založené na strojovom učení

Strojové učenie predstavuje oblasť umelej inteligencie, ktorá sa zameriava na automatické získavanie znalostí a vzťahov z dát namiesto ich priameho programovania. Jeho cieľom je vytvárať algoritmy schopné zlepšovania svojho výkonu prostredníctvom analýzy príkladov a získaných predošlých skúseností, pričom takto vytvorené modely dokážu generovať predikcie, odporúčania či rozhodnutia a reagovať aj na nové, doposiaľ neznáme vstupy (Janiesch et al. 2021).

Metódy strojového učenia sa vo všeobecnosti rozdeľujú na učenie s učiteľom a učenie bez učiteľa, pričom hlavným rozdielom je prítomnosť označených dát v tréningovej množine. Učenie s učiteľom je charakteristické tým, že vstupné dáta sú doplnené o známe výstupné hodnoty, zatiaľ čo učenie bez učiteľa využíva algoritmy na analýzu a objavovanie vzorov a asociácií v neoznačených dátach. Učenie s učiteľom sa používa najmä pri úlohách klasifikácie a predikcie, kde cieľom je priradiť objekt do kategórie alebo odhadnúť spojitú hodnotu. Typickými algoritmami učenia s učiteľom sú stroje s podpornými vektormi, náhodné lesy alebo Naivný Bayes (Alslaity a Orji 2022).

Naopak, učenie bez učiteľa pracuje s neoznačenými dátami, a teda algoritmus nemá k dispozícii správne výstupy. Namiesto toho sa snaží objaviť skryté štruktúry alebo vzory zo vstupných dát. Tento prístup je vhodný najmä na zhlukovanie a asociačnú analýzu, pričom medzi najčastejšie používané algoritmy patria *k*-means a Apriori algoritmus. Tento typ učenia sa často využíva ako predpríprava pre učenie s učiteľom, napríklad na vytváranie označení v dátach alebo na zníženie dimenzionality dát (Ahmad et al. 2017; Pandey et al. 2019; Alloghani et al. 2020).

1.3 METÓDY STROJOVÉHO UČENIA

Nasledujúca časť sa zameriava na hlavné typy metód využívaných pri klasifikácii sentimentu. Tieto metódy predstavujú základné prístupy, ktoré sa líšia najmä spôsobom spracovania dát, reprezentáciou informácií a interpretáciou výsledkov. Vo všeobecnosti sa delia na tri hlavné skupiny: symbolické, subsymbolické a skupinové (ensemble) metódy.

1.3.1 Symbolické metódy

Symbolické metódy strojového učenia sú prístupy, ktoré pracujú s danou reprezentáciou znalostí vo forme pravidiel alebo symbolických štruktúr. Ich hlavnou výhodou je interpretovateľnosť, teda dokážu poskytnúť jasný a zrozumiteľný dôvod, prečo model dospel k určitému rozhodnutiu alebo klasifikácii. Symbolické metódy často využívajú presne určené pravidlá na klasifikáciu dát, a sú vhodné tam, kde je dôležité pochopiť proces rozhodovania, napríklad v medicíne alebo finančnej analýze. Tieto metódy fungujú najlepšie pri jasne definovaných a statických problémoch, avšak nemajú dobrý výkon pri hodnoteniach v reálnom čase ani pri spracovaní veľkých množstiev dát. Medzi typické príklady symbolických algoritmov patria *k*-najbližších susedov a rozhodovacie stromy (Ilkou a Koutraki 2020; Banafa 2025).

***K*-najbližších susedov**

Algoritmus *k*-najbližších susedov (KNN, *k*-nearest neighbor) patrí medzi najrozšírenejšie metódy používané vo výskume strojového učenia. Je to neparametrický klasifikačný algoritmus, teda nevychádza z predpokladov o základnom rozdelení údajov, a je považovaný za jednoduchý, ale efektívny algoritmus. Hlavnou myšlienkou KNN je klasifikovať označenie (triedu) novej, neznámej inštancie na základe jej najbližších susedov v danom priestore. Predpokladá sa pritom, že dáta, ktoré sa nachádzajú blízko vedľa seba, majú tendenciu patriť do rovnakej triedy (Suyal a Goyal 2022).

Proces klasifikácie prebieha v dvoch hlavných krokoch. V prvej fáze sa určí množina *k*-najbližších susedov vzhľadom na metriku vzdialenosti, najčastejšie pomocou Euklidovskej vzdialenosti (vzorec 1)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}\,, \tag{1}$$

pričom $\boldsymbol{x}$ a $\boldsymbol{y}$ tvoria *n*-rozmerné vektory reprezentujúce vlastnosti dvoch inštancií. Okrem Euklidovskéj vzdialenosti existujú aj iné metódy na výpočet vzdialenosti ako je Manhattanská vzdialenosť. V druhej fáze sa následne priradí novej inštancii trieda, ktorá sa v jej okolí nachádza najčastejšie.

Vďaka svojej jednoduchosti je KNN algoritmus ľahko pochopiteľný, implementovateľný a prispôsobiteľný na rôzne klasifikačné úlohy. Už Cover a Hart (1967) ukázali, že chyba klasifikácie KNN je maximálne dvojnásobkom optimálnej Bayesovej chyby, pričom s rastúcim počtom dát sa k nej približuje, čo aj dokazuje spoľahlivosť algoritmu.

Napriek svojej jednoduchosti má KNN aj viacero obmedzení, ktoré môžu ovplyvniť jeho presnosť a efektivitu. Algoritmus je citlivý na vysokú dimenzionalitu dát, kedy pri väčšom počte atribútov stráca vzdialenosť medzi bodmi zmysel, keďže sú si všetky body približne rovnako vzdialené. Okrem toho je tento algoritmus výpočtovo náročný ak pracuje s veľkými datasetmi, pretože pre každú novú inštanciu je potrebné prepočítať vzdialenosť medzi všetkými bodmi v trénovacej množine, čo vedie k zhoršeniu škálovateľnosti modelu a vysokým časovým nárokom (Taunk et al. 2019; Kang 2021).

Rozhodovacie stromy

Rozhodovacie stromy patria medzi najpoužívanejšie metódy strojového učenia a využívajú sa v mnohých oblastiach ako je predikcia a klasifikácia. Ich základnou myšlienkou je rekurzívne rozdeľovanie dát do menších podmnožín na základe hodnôt vstupných premenných, až kým nie sú splnené kritériá pre zastavenie. Tento proces, známy aj ako princíp „rozdeľ a panuj“, spočíva v postupnom rozdelení dát podľa takého atribútu, ktorý najlepšie odlišuje triedy, s cieľom zvýšiť homogenitu vytvorených podmnožín. Výsledkom tohto procesu je stromová štruktúra, ktorá obsahuje uzly, vetvy a listy. Každý uzol v strome predstavuje rozhodovaciu otázku na určitý atribút, vetvy zodpovedajú možným odpovediam a listy konečným rozhodnutiam alebo začleneniam do tried. Typický rozhodovací strom obsahuje

koreňový uzol, ktorý bol zvolený na začiatku procesu, niekoľko vnútorných uzlov a viacero koncových listov (Patel a Prajapati 2018).

Jednou z hlavných výhod rozhodovacích stromov je ich vysoká interpretovateľnosť, kedy je možné ľahko vizualizovať a pochopiť výsledný model bez hlbších znalostí strojového učenia. Nevyžadujú normalizáciu dát, ani predpoklady o ich rozdelení a dokážu pracovať s číselnými aj kategorickými údajmi. Na druhej strane sú náchylné na preučenie, najmä pri veľkom množstve atribútov alebo malom množstve dát. Aj menšie zmeny v tréningovej množine môžu viesť k výrazne odlišnému výslednému stromu, čo znižuje jeho stabilitu. Tento problém sa často rieši využitím skupinových metód ako sú náhodné lesy (Bansal et al. 2022).

Najdôležitejšou časťou pri tvorbe rozhodovacieho stromu je správny výber atribútu, podľa ktorého sa strom rozvetvuje. Tento proces, označovaný ako rozdeľovanie (splitting), využíva rôzne kritériá na určenie najvhodnejších atribútov. Medzi najpoužívanejšie kritériá patria (Zhou 2021; Bansal et al. 2022):

- **Gini index** – je to jedno z najpoužívanejších kritérií na rozdelenie stromu typu CART. Meria mieru nečistoty uzla, teda čím je hodnota nižšia, tým je uzol homogénnejší. Možno ho matematicky vyjadriť takto, pričom P_i je pravdepodobnosť, že vzorka patrí do triedy i (vzorec 2)

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2. \tag{2}$$

- **Informačný zisk (Information Gain)** - vychádza z entropie a meria, ako veľmi zníži rozdelenie dát neistotu pri rozdelení podľa určitého atribútu. Vyššia hodnota informačného zisku znamená, že atribút je efektívnejší pre vytváranie homogénnych podmnožín. Entropia je daná vzorcom (vzorec 3)

$$E = -\sum_{i=1}^{n} P_i \log_2(P_i), \tag{3}$$

kde n je počet unikátnych tried a P_i pravdepodobnosť, že náhodne vybraná vzorka patrí do triedy i. Informačný zisk sa ďalej vypočíta ako rozdiel entropie pôvodnej množiny a váženého priemeru entropií jednotlivých poduzlov (vzorec 4)

$$IG = E(D) - \sum_{i=1}^{n} \frac{|D_i|}{|D|} E(D_i). \quad (4)$$

- **Pomer informačného zisku (Information Gain Ratio)** – rozšírenie informačného zisku, ktoré eliminuje sklon voči atribútom s veľkým počtom hodnôt. Normalizuje informačný zisk pomocou takzvaného „split information“, teda informácie o rozdelení, čím zabezpečuje spravodlivejšie porovnanie medzi atribútmi. Vzorec je nasledovný, pričom IG je informačný zisk a SI vyjadruje mieru rozdelenia dát (vzorec 5)

$$IGR = \frac{IG}{SI}. \quad (5)$$

Po vytvorení rozhodovacieho stromu je možné, že výsledný model bude príliš zložitý v snahe čo najlepšie klasifikovať jednotlivé dáta namiesto zachytenia všeobecných vzorov. Tento jav, známy ako preučenie (overfitting), spôsobuje, že model dosiahne vysokú presnosť na tréningovej množine, ale jeho výkon na nových a neznámych dátach sa zhoršuje. Na zníženie rizika preučenia sa používa proces prerezávania (pruning), ktorého cieľom je zjednodušiť strom odstránením nadbytočných vetiev, čím sa znižuje zložitosť modelu a zlepšuje sa jeho schopnosť generalizovať (Zhou 2021; Mienye a Jere 2024). Vo všeobecnosti sa rozlišujú dva typy prerezávania:

- **Predbežné preprezávanie (Pre-pruning)** - proces zastavenia rastu stromu v momente, keď ďalšie rozdelenie neprináša významné zlepšenie podľa zvoleného kritéria.
- **Následné prerezávanie (Post-pruning)** - spätné prehodnocovanie uzlov s minimálnym dopadom po vytvorení stromu. Ak sa po nahradení niektorého uzla listom zlepší presnosť, vetva sa odstráni.

1.3.2 Subsymbolické metódy

Na rozdiel od symbolických metód, ktoré sa spoliehajú na presné pravidlá a ľudsky interpretovateľné rozhodovacie procesy, subsymbolické metódy strojového učenia sa zameriavajú na modelovanie vzťahov medzi premennými priamo zo vzorov v dátach. Tieto vzťahy bývajú zložité a často sa tvoria pomocou náročných matematických alebo štatistických funkcií. Subsymbolické prístupy, medzi ktoré patrí

napríklad logistická regresia, viacvrstvové perceptróny, neurónové siete či stroje s podpornými vektormi, sa snažia zachytiť aj skryté závislosti v dátach bez zadania presných pravidiel.

Hlavnou výhodou týchto metód je ich schopnosť pracovať s veľkými množstvami dát a ich odolnosť voči šumu a chýbajúcim dátam. Na druhej strane, ich nízka interpretovateľnosť predstavuje významnú nevýhodu, najmä v oblastiach kde je vysvetlenie výsledkov dôležité. Tieto metódy sú ideálne na predikciu, klasifikáciu, zhlukovanie alebo spracovanie prirodzeného jazyka (Ilkou a Koutraki 2020; Banafa 2025).

Logistická regresia

Logistická regresia je algoritmus strojového učenia, ktorý predikuje pravdepodobnosť príslušnosti vstupných dát k určitej kategórii na základe nezávislých premenných. Tento model môže byť rozšírený aj na multinomickú alebo ordinálnu klasifikáciu, no v rámci tejto práce sa zameriame na binárnu klasifikáciu, pri ktorej výstup predstavuje pravdepodobnosť jednej z dvoch tried. Samotný klasifikačný proces prebieha v niekoľkých krokoch. Najprv sa vypočíta lineárna kombinácia vstupných dát *x* a príslušných regresných koeficientov *w* (vzorec 6)

$$z = w_0 + w_1 x_1 + \cdots + w_n x_n. \tag{6}$$

Táto hodnota *z* sa následne prevedie sigmoidnou funkciou (vzorec 7)

$$y = \frac{1}{1 + e^{-z}}, \tag{7}$$

kde *y* predstavuje hodnotu medzi 0 a 1. Dáta s výstupom väčším ako 0,5 sú klasifikované do triedy 1, a tie s menším výstupom ako 0,5 do triedy 0 (Zou et al. 2019).

Logistická regresia je relatívne jednoduchá technika binárnej klasifikácie, ktorú je ľahké pochopiť a implementovať. Medzi jej hlavné výhody patrí priame modelovanie pravdepodobnosti klasifikácie bez predpokladov o rozdelení dát, možnosť predikcie výstupov spolu s danými pravdepodobnosťami, a konvexná cieľová funkcia, ktorá umožňuje využitie numerických optimalizačných metód ako je gradientný zostup alebo Newtonova metóda na nájdenie najlepších parametrov modelu. Na druhej strane, má tento algoritmus aj svoje obmedzenia: nie je vhodná na predikciu spojitých výsledkov, jej výkon závisí od veľkosti datasetu a pri malých množinách dát môže

dochádzať k preučeniu. Okrem toho logistická regresia predpokladá nezávislosť pozorovaní, čo je kľúčové pre spoľahlivé fungovanie modelu a zabránenie skresľovania odhadov (Zhou 2021; Zaidi a Al Luhayb 2023).

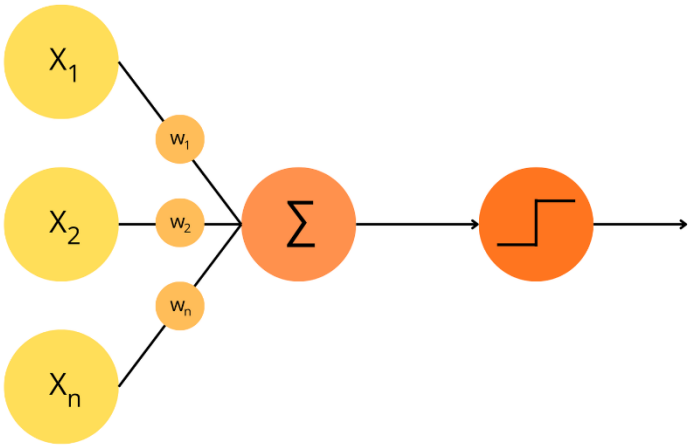
Neurónové siete

Neurónové siete sú prepojené systémy jednoduchých výpočtových jednotiek, ktoré sú inšpirované fungovaním biologických neurónov. Tieto umelé neuróny prijímajú signály, ktorým sú priradené váhy, ktoré sa sčítajú a nakoniec sa výsledok spracuje pomocou aktivačnej funkcie, čím sa rozhoduje o generovaní výstupu. Podobne ako pri biologických neurónoch, niektoré vstupy môžu pôsobiť inhibične a tak znižovať pravdepodobnosť aktivácie neurónu, zatiaľ čo iné sú excitačné a podporujú generovanie výstupu (Joshi et al. 2023).

V roku 1943 McCulloch a Pitts vytvorili prvý formálny matematický model neurónu, ktorý sa stal teoretickým základom pre neskorší vývoj perceptrónu a moderných neurónových sietí. Perceptrón, ako jeden z prvých praktických modelov umelého neurónu, dokáže kombinovať viacero vstupov na určenie výstupu (Obrázok 1). Perceptrón možno matematicky vyjadriť ako (vzorec 8)

$$y = \sum_{i=1}^{n} w_i x_i + b, \tag{8}$$

kde w je vektor váh, x predstavuje vektor vstupov pri počte n a b je bias. Každý vstup má priradenú váhu, ktorá reprezentuje jeho dôležitosť. Počas tréningu perceptrónu sa tieto váhy upravujú tak, aby sa minimalizovala chyba medzi predikovanou výslednou hodnotou a skutočným výsledkom. Ak je výstup neurónu správny, váhy sa nemenia. Naopak, ak je nesprávny, váhy aktivovaných vstupov sa upravia. Výstup neurónu je určený aktivačnou funkciou, ktorá porovná súčet vážených vstupov s prahovou hodnotou. V prípade perceptrónu sa používa jednoduchá prahová funkcia, ktorá dáva výstup 1 ak súčet prekročí prah, a 0 ak nie. Tento princíp učenia, známy ako perceptrónové učenie, predstavuje jednu z najstarších foriem učenia s učiteľom a stal sa základom pre neskorší vývoj pokročilejších algoritmov (Aggarwal 2018).

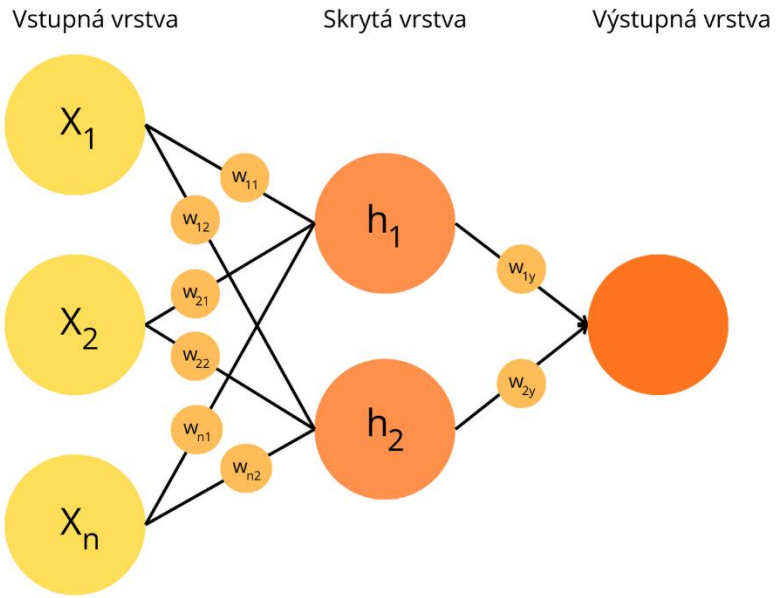


Obrázok 1 Model perceptrónu[1]

Takýto jednovrstvový perceptrón dokáže správne klasifikovať vstupy len v prípade, ak sú lineárne separovateľné. Príklady problémov zahŕňajú logické operácie AND, OR alebo NOT. Problémy ako XOR však jednovrstvový perceptrón nedokáže riešiť, pretože neexistuje lineárna hranica ktorá by takéto vstupy vedela správne rozdeliť.

Práve preto vznikla potreba viacvrstvových neurónových sietí (MLP). Ide o dopredné neurónové siete, v ktorých signál prechádza okrem vstupnej a výstupnej vrstvy aj cez tretiu, tzv. skrytú vrstvu, alebo skryté vrstvy (Obrázok 2). Vstupná vrstva prijíma externé signály, zatiaľ čo skryté a výstupné vrstvy ich spracovávajú pomocou funkčných neurónov a poskytujú spracovaný výstup (Aggarwal 2018).

[1] Zdroj Obrázok 1: autor



Obrázok 2 Model architektúry MLP[2]

Kľúčovým problémom pri trénovaní viacvrstvových neurónových sietí bolo dlhú dobu nájdenie efektívneho spôsobu, ako upravovať váhy v skrytých vrstvách. Tento problém bol vyriešený zavedením algoritmu spätného šírenia chyby (backpropagation), ktorý umožnil efektívne trénovanie neurónových sietí. Je to učiaci algoritmus, ktorý sa skladá z dvoch hlavných fáz, dopredného a spätného šírenia.

Vo fáze dopredného šírenia prechádzajú dáta cez jednotlivé vrstvy siete, pričom sa na každej vrstve aplikujú váhy a aktivačné funkcie. Výstup modelu sa následne porovnáva so skutočným výstupom a vypočíta sa chyba pomocou príslušnej stratovej funkcie, napríklad strednej kvadratickej chyby (MSE) pri regresných úlohách alebo cross-entropy loss pri klasifikačných úlohách. V spätnej fáze sa chyba potom šíri od výstupnej vrstvy až po vstupnú, pričom tento proces využíva metódu gradientného zostupu (gradient descent), na základe ktorej sa váhy menia s cieľom zníženia chyby v každom neuróne. Celý proces sa opakuje pre všetky trénovacie vzory vo viacerých epochách, až kým sa chyba modelu nezníži na alebo pod požadovanú hranicu (Aggarwal 2018; Zhou 2021).

[2] Zdroj Obrázok 2: autor

Stroje s podpornými vektormi

Stroje s podpornými vektormi (Support Vector Machine, SVM) patrí medzi klasické a často používané metódy strojového učenia, najmä na plnenie úlohy klasifikácie aj predikcie (Cervantes et al. 2020; Zhou 2021). Ide o algoritmus, ktorý sa snaží nájsť optimálnu hranicu (tzv. hyperrovinu) medzi triedami tak, aby dáta, ktoré patria do rôznych kategórií boli čo najlepšie rozdelené a tak maximalizuje vzdialenosť (margin) medzi nimi. Pre určenie ideálnej polohy nadroviny používame podporné vektory, čo sú body, ktoré ležia najbližšie k hyperrovine. Práve od týchto bodov závisí maximálna vzdialenosť medzi triedami. Čím je toto rozpätie väčšie, tým lepšie dokáže model generalizovať pri nových dátach. SVM dokáže pracovať aj v lineárnom aj v nelineárnom prostredí, pričom pri lineárne separovateľných dátach hľadá priamku v dvojrozmernom priestore, alebo hyperrovinu vo vyšších dimenziách. V prípade nelineárnych dát sa využívajú funkcie jadra, ktoré transformujú priestor do vyššieho rozmeru tak, že bude možné nájsť lineárnu hranicu medzi triedami. Výber vhodnej funkcie jadra závisí od charakteru dát. Medzi najčastejšie funkcie patrí:

- **Lineárne jadro –** najjednoduchšie jadro, vhodné na lineárne separovateľné dáta,
- **Polynómové jadro –** dokáže modelovať nelineárne vzťahy pomocou polynómu určitého stupňa,
- **Jadro funkcie radiálnej bázy (Radial Basis Function, RBF) –** je veľmi flexibilné, transformuje dáta do vyššieho rozmeru a často sa používa pri zložitejších nelineárnych dátach,
- **Sigmoidné jadro –** inšpirované aktivačnými funkciami neurónových sietí, je menej používané.

Pre lineárne separovateľné dáta hľadá SVM nadrovinu definovanú ako (vzorec 9)

$$w^T x + b = 0, \tag{9}$$

kde w je vektor váh, x vektor vstupných atribútov a b bias. Ak je výsledok väčší ako 0, bod x je klasifikovaný do jednej triedy, a ak je menší, tak je klasifikovaný do druhej.

SVM môže byť realizovaná dvoma základnými spôsobmi:

- **Tvrdá hranica (Hard margin) –** snaží sa nájsť hyperrovinu, ktorá oddelí všetky body do rôznych tried bez toho, aby sa akýkoľvek bod nachádzal v nesprávnej triede. Tento prístup je veľmi citlivý na šum a odľahlé body, ktoré dokážu hyperrovinu skresliť.
- **Mäkká hranica (Soft margin) –** umožňuje, aby sa niektoré body nachádzali na nesprávnej strane hyperroviny, alebo aby boli nesprávne klasifikované. Tento prístup zavádza penalizačný parameter, ktorý hľadá najlepší kompromis medzi najväčšou vzdialenosťou a najmenšou chybou klasifikácie. Tento spôsob je praktickejší pri reálnych dátach, pretože ponúka flexibilnejšie a robustnejšie riešenie ktoré vie pracovať aj so šumom alebo nepresnostiami.

Okrem výberu funkcie jadra SVM obsahuje aj niekoľko hyperparametrov, ktoré ovplyvňujú výkon modelu:

- ***C* –** penalizačný parameter, alebo aj parameter regulácie, určuje kompromis medzi maximálnou šírkou marginu a minimalizáciou chybne klasifikovaných bodov. Čím je *C* vyššie, tým je model nútený prísnejšie klasifikovať body,
- **Funkcie jadra –** určujú spôsob transformácie priestoru pri nelineárne separovateľných dátach,
- **Gamma –** hyperparameter rozhodovacích funkcií ako RBF alebo sigmoid. Nižšie hodnoty znamenajú, že body majú širší dosah, čo vedie k viac generalizovanej rozhodovacej hranici, zatiaľ čo vyššie hodnoty skresľujú hyperrovinu iba v okolí bodov,
- **Degree –** určuje stupeň polynomiálneho jadra.

1.3.3 Skupinové metódy

Skupinové metódy patria medzi najúčinnejšie prístupy v strojovom učení. Kombinujú viacero základných modelov (tzv. base learners) s cieľom dosiahnuť lepší prediktívny výkon, pričom základná myšlienka skupinových modelov spočíva v tom, že takáto kombinácia viacerých modelov, či už slabších alebo silnejších, môže viesť k presnejšiemu a robustnejšiemu výsledku ako jednotlivé modely samostatne. Typický proces skupinového učenia je tvorený dvoma krokmi:

1. Generovanie a tréning slabších učiteľov (weak learners),
2. Kombinácia modelov do spoločného výstupu.

Každý model sa učí rôzne aspekty problému, často na rôznych podmožinách dát, alebo na základe odlišných charakteristík. Kľúčovou podmienkou pre úspech skupinového učenia je diverzita modelov, teda, ak sú si modely veľmi podobné, budú aj pravdepodobne robiť rovnaké chyby, čím sa znižuje efektivita konečného výsledku. Z kombinácie viacerých modelov zvyčajne vzniká model s lepšou generalizačnou schopnosťou.

Existuje viacero dôvodov, prečo skupinové metódy zvyčajne zlepšujú prediktívny výkon:

- **Prevencia preučenia (overfitting) –** kombinácia viacerých nezávisle trénovaných modelov znižuje rozptyl výsledkov a tak znižuje riziko preučenia, najmä pri bagging metódach,
- **Robustnosť –** skupinové prístupy znižujú dopad náhodných chýb a šumu v dátach, čím poskytujú spoľahlivejšie predikcie,
- **Zlepšenie generalizácie -** kombináciou viacerých modelov sa znižuje rozptyl predikcií a zlepšuje schopnosť systému správne generalizovať na nové, neznáme dáta.

Skupinové metódy rozdeľujeme podľa spôsobu, akým kombinujú jednotlivé modely, pričom medzi najpoužívanejšie patria Bagging, Boosting a Stacking (Sagi a Rokach 2018; Zhou 2021; 2025).

Vrecovanie

Anglický názov pre vrecovanie, teda bagging, pochádza zo skratky Bootstrap Aggregating, čo znamená že je to technika založená na bootstrap vzorkovaní. Bagging je paralelná skupinová metóda, ktorá sa vykonáva paralelne na viacerých nezávislých modeloch, pričom sa každý model učí na inej podmnožine tréningových dát. Každá tréningová vzorka je vytvorená náhodným výberom s opakovaním (bootstrap), takže niektoré dáta sa môžu objaviť aj viackrát, zatiaľ čo iné môžu byť úplne vynechané (Ming et al. 2024).

Výstupy jednotlivých modelov sa na konci kombinujú, pričom pri klasifikácii sa používa hlasovanie, a pri regresii priemerovanie. Tento proces vedie k zníženiu celkového rozptylu predikcií a zlepšeniu generalizačnej schopnosti modelu.

Tie vzorky, ktoré neboli použité na trénovanie modelov, taktiež nazývané aj out-of-bag vzorky, sa potom dajú použiť na odhad výkonu modelu.

Náhodný les je rozšírenie baggingu, kde sa okrem vzorkovania zavádza aj náhodný výber atribútov. Pri tradičnom rozhodovacom strome sa pri každom rozdelení zvolí najlepší dostupný atribút, zatiaľ čo náhodný les vyberá náhodnú podmnožinu atribútov a znej volí ten najlepší pre dané rozdelenie. Týmto spôsobom sa zväčšuje diverzita základných modelov čo umožňuje ešte účinnejšie predikovanie model. Náhodný les tak často dosahuje stabilnejšie a presnejšie výsledky ako jednotlivé rozhodovacie stromy (Sagi a Rokach 2018; Zhou 2025).

Posilňovanie

Posilňovanie (Boosting) je sekvenčná technika, kde sa modely trénujú postupne, pričom každý nový model sa zameriava na opravu chýb svojich predchodcov. Na začiatku procesu sa natrénuje prvý slabý učiteľ na pôvodných dátach a podľa jeho chýb sa upravia váhy alebo distribúcia tréningových vzoriek tak, aby nasledujúce modely venovali väčšiu pozornosť nesprávne klasifikovaným dátam. Následne sa trénuje ďalší model na upravených dátach a proces sa opakuje, až kým sa neaktivuje preddefinovaný počet modelov. Výstupy všetkých modelov sa potom kombinujú, často váženým spôsobom, aby sa vytvoril konečný predikčný model s vyššou presnosťou, taktiež nazývaný aj silný učiteľ (Mienye a Sun 2022)

Najznámejším príkladom posilňovania je algoritmus AdaBoost, ktorý adaptívne upravuje váhy inštancií na základe presnosti predchádzajúcich klasifikácií. Váha každého slabého učiteľa sa počíta na základe jeho chyby. Po každej iterácii sa aktualizujú váhy vzoriek tak, aby nesprávne klasifikované vzorky získali vyššie váhy správne klasifikované nižšie, čím sa budú ďalšie modely viac sústrediť na ťažšie klasifikovateľné vzorky. (Mienye a Sun 2022; Zhou 2025)

Zatiaľ čo sa klasické posilňovanie zameriava na postupnú kombináciu slabých učiteľov, moderné implementácie prinášajú nové riešenia na zvýšenie presnosti a efektivity. Jedným z najvýznamnejších a najpopulárnejších nástrojov je XGBoost (Extreme Gradient Boosting), ktorý vychádza z princípov gradientného boostingu, ale kombinuje ich s pokročilými optimalizačnými technikami, regularizáciou a rozsiahlymi možnosťami prispôsobenia modelu. Ide o aditívny model, ktorý v každom kroku pridáva nový strom do celkového výsledku tak, aby korigoval chyby vzniknuté súčtom

predchádzajúcich stromov. Vďaka zahrnutiu regularizačnej zložky v stratovej funkcii sa znižuje riziko preučenia, a rôzne techniky optimalizácie pamäte a výpočtovej rýchlosti umožňujú efektívne spracovanie veľkých datasetov aj vážených vzoriek. Tieto vlastnosti robia XGBoost spoľahlivým a presným nástrojom pre predikciu na štruktúrovaných dátach, vďaka čomu je využívaný v rôznych oblastiach, od finančných analýz a marketingových predikcií až po rôzne priemyselné využitia (Sagi a Rokach 2018).

2 CIELE ZÁVEREČNEJ PRÁCE

Cieľom našej práce je identifikovať najvhodnejší model strojového učenia pre klasifikáciu sentimentu (pozitívny, neutrálny, negatívny) v textoch z bankového sektora pre jazyk s obmedzenými zdrojmi. Dané texty predstavujú špecifickú doménu, ktorá sa vyznačuje používaním odborného bankového žargónu a terminológie, a tak neumožňuje jednoducho aplikovať generické modely strojového učenia trénované na všeobecných dátach.

Na základe teoretických zistení a predchádzajúcich štúdií formulujeme nasledovné hypotézy:

HA: Symbolické metódy dosiahnu najnižšiu výkonnosť pri klasifikácii sentimentu.

HB: Skupinové (ensemble) modely dosiahnu najvyššiu výkonnosť.

2.1 ČIASTKOVÉ CIELE

Na dosiahnutie hlavného cieľa boli definované tieto čiastkové ciele:

- Identifikovať existujúce prístupy k analýze sentimentu v bankovníctve a podobných oblastiach.
- Pripraviť dataset obsahujúci texty súvisiace s bankovým sektorom v slovenskom jazyku.
- Aplikovať vybrané symbolické, subsymbolické a skupinové modely na pripravený dataset.
- Porovnať výkonnosť jednotlivých modelov na základe štandardných metrík klasifikácie.
- Interpretovať výsledky a formulovať odporúčania pre budúce použitie analýzy sentimentu v tejto oblasti.

3 METODIKA VÝSKUMU

Predmetom skúmania diplomovej práce je identifikácia a porovnanie modelov strojového učenia určených na klasifikáciu sentimentu v textoch pochádzajúcich z bankového sektora. Hlavným cieľom je určiť, ktorý z testovaných modelov dosahuje najlepšie výsledky pri spracovaní textov v jazyku s obmedzenými zdrojmi, akým je slovenský jazyk, s ohľadom na bankový a odborný žargón.

Proces riešenia bol navrhnutý podľa metodológie CRISP-DM (Cross Industry Standard Process for Data Mining), ktorá predstavuje štandardný rámec pre postup pri projektoch z oblasti dolovania dát a strojového učenia. Táto metodológia poskytuje o štruktúrovaný prístup k celému životnému cyklu, od pochopenia problematiky až po nasadenie modelu v praxi. V nasledujúcich kapitolách sú podrobne aplikované jednotlivé kroky, ktoré boli aplikované pri riešení diplomovej práce.

3.1 POROZUMENIE PROBLEMATIKE

Analýza sentimentu je pre bankovníctvo mimoriadne dôležitá, pretože umožňuje organizáciám lepšie pochopiť názory, postoje a emócie svojich klientov. V modernej dobe, kedy banky medzi sebou neustále konkurujú, je dôvera zákazníkov a reputácia značky jednou z kľúčových konkurenčných výhod. Analýza sentimentu ponúka bankám pohľad na to, ako sú ich služby a produkty vnímané, čím vedia zlepšiť verejnú mienku o značke, ako aj komunikáciu s klientmi.

Vďaka tomu, že analýza sentimentu dokáže odhaliť negatívne ohlasy v reálnom čase, banky vedia rýchlo reagovať na vznikajúce reputačné riziká, ľahšie riešiť sťažnosti zákazníkov alebo predchádzať krízovým situáciám. Výsledky z analýzy vedia byť rovnako využité aj na marketingové účely, napríklad pri zlepšovaní kampaní alebo prispôsobovaní produktových ponúk podľa aktuálnych trendov v správaní zákazníkov. Okrem marketingového prínosu vedia banky získané výsledky využiť aj pri strategických rozhodovaniach na zlepšenie konkurencieschopnosti na trhu.

Výsledkom fázy porozumenia problematiky je určenie úlohy objavovania znalostí ako úlohy klasifikácie sentimentu, teda zaradenia textových vstupov do troch tried sentimentu (pozitívny, neutrálny, negatívny). Na riešenie tejto úlohy sú ďalej vybrané a porovnávané rôzne prístupy strojového učenia, a to symbolické,

subsymbolické a skupinové metódy. Nad rámec pôvodného zadania sú do riešenia zahrnuté aj jazykové modely, ktoré sú najprv analyzované ako samostatná skupina a následne porovnané s modelmi strojového učenia.

3.2 POROZUMENIE DÁTAM

Pre účely tejto práce bol použitý dataset v slovenskom jazyku, obsahujúci texty z bankového sektora, ktoré zahŕňajú typický odborný bankový žargón a terminológiu. Každý záznam predstavuje samostatnú vetu alebo krátky textový úryvok pochádzajúci z výročných správ, ktorý vyjadruje určitý postoj alebo hodnotenie aktivít banky. Každý text je manuálne označený sentimentom, a to: pozitívnym, neutrálnym alebo negatívnym. Kelebercová a Zozuk (2023), Kelebercová et al. (2025), Pilková et al. (2025) použili tento dataset vo svojich štúdiách, kde sa zaoberali analýzou sentimentu vo výročných správach. Dataset obsahuje nerovnomerné rozdelenie tried, pričom problémom sú duplicitné záznamy, neštandardné zápisy textov a odborné skratky, ktoré nie sú bežné v generických jazykových modeloch. Texty vyžadujú prípravu dát, ktorá zahŕňa čistenie textu, tokenizáciu, prípadnú normalizáciu odborných termínov a ďalšie kroky potrebné pre ideálne trénovanie modelov strojového učenia.

V rámci prieskumu dát sa plánuje vykonať aj základná štatistická analýza, ktorá pomôže lepšie pochopiť distribúciu sentimentu, dĺžku textov a frekvenciu najčastejšie používaných slov a odborných výrazov. Tieto informácie prispejú k identifikácii potenciálnych problémov v dátach a k návrhu vhodných krokov pri ich predspracovaní.

3.3 PRÍPRAVA DÁT

Po získaní a preskúmaní dát bolo potrebné pripraviť ich do podoby, ktorá je vhodná pre prácu s modelmi strojového učenia. Keďže zdrojom dát boli textové záznamy, proces prípravy zahŕňal niekoľko krokov textového spracovania a čistenia, ktorých cieľom bolo znížiť šum, odstrániť nepodstatné informácie a zabezpečiť jednotnú reprezentáciu textov.

Prvým krokom bolo čistenie textu od špeciálnych znakov, čísel, odkazov, nadbytočných medzier a iných nepotrebných prvkov. Texty boli následne prevedené na malé písmená, aby sa predišlo vzniku duplicitných záznamov v prípade iných slov s rôznou veľkosťou písmen. Po vyčistení textov nasledovala tokenizácia viet

na jednotlivé slová a lematizácia, ktorá transformovala slová na ich základný tvar, čo pomohlo zjednotiť dáta, a tak redukovať počet jedinečných slov v korpuse.

Ďalej bolo zo spracovaných textov potrebné odstrániť stop-slová, teda slová, ktoré nenesú žiaden význam pri určovaní sentimentu. Použitý bol upravený zoznam stop-slov pre slovenský jazyk, rozšírený o niektoré bežné výrazy používané v bankovom sektore, ak sa nachádzali v každej triede sentimentu bez zmeny vo význame.

Vzhľadom na prítomnosť bankového žargónu bolo potrebné doplniť aj normalizáciu doménovo špecifických termínov, ako sú skratky alebo pojmy, ktoré sa používajú v rôznych formách, ale znamenajú to isté.

3.4 MODELOVANIE

Počas fázy modelovania v CRISP-DM prebiehala aplikácia vybraných algoritmov strojového učenia na úlohu analýzy sentimentu. Cieľom tejto fázy bolo nielen natrénovať modely na pripravených dátach, ale aj nájsť optimálne hyperparametre pre jednotlivé modely. Na hľadanie vhodnej konfigurácie pri jednotlivých modeloch bola použitá metóda RandomizedSearch na efektívne preskúmanie rôznych kombinácií hyperparametrov bez potreby testovania každej možnej kombinácie. Pre jednotlivé metódy boli určené rozsahy parametrov, z ktorých sa náhodne vyberali kombinácie testované počas fázy tréningu. Optimalizácia parametrov bola realizovaná pomocou 10-násobnej krížovej validácie na trénovacej množine dát, pričom sa sledovala kvalita modelu. V tabuľke 1 sú zhrnuté parametre pre jednotlivé metódy.

Tabuľka 1 Metódy strojového učenia, parametre a experimentálne hodnoty[3]

Metóda	**Parameter**	**Experimentálne hodnoty**
K-najbližších susedov (KNN)	n_neighbors	3, 5, 7, 11, 15, 21, 31
	weights	uniform, distance
Rozhodovacie stromy	criterion	gini, entropy, log_loss
	max_depth	None, 10, 20, 40
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
Logistická regresia	C	10^{-4} - 10^{2} (logspace)
	solver	lbfgs, saga
	penalty	l1, l2, elasticinet
Viacvrstvový perceptrón	hidden_layer_sizes	(128,), (256,), (128, 64), (256, 128)
	activation	relu, tanh
	alpha	10^{-6} - 10^{-2}
	learning_rate_init	0,0001, 0,0003, 0,001
Lineárne stroje s podpornými vektormi	C	10^{-4} - 10^{2}
	loss	hinge, squared_hinge
	class_weight	None, balanced
Náhodný les	n_estimators	200, 400, 800
	max_depth	None, 10, 20, 40
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
LightGBM	n_estimators	300, 600, 900
	learning_rate	0,03, 0,05, 0,1
	min_child_samples	5, 10, 20, 40
XGBoost	n_estimators	300, 600, 900
	learning_rate	0,03, 0,05, 0,1
	gamma	0,0, 0,5, 1,0

[3] Zdroj Tabuľka 1: autor

3.5 EVALVÁCIA

Evalvácia modelov predstavuje kľúčový krok CRISP-DM, kde sa posudzuje schopnosť modelov správne predikovať sentiment na nových, neznámych dátach a tak umožňuje zistiť silné a slabé stránky jednotlivých prístupov. Pre objektívne porovnanie výkonnosti jednotlivých modelov bola použitá 10-násobná krížová validácia (10-fold cross-validation), ktorá zabezpečuje, že model bol testovaný na rôznych podmnožinách dát, čím bol minimalizovaný vplyv náhodného výberu trénovacích a testovacích vzoriek.

Na vyhodnotenie samotnej spoľahlivosti modelu boli využité viaceré metriky hodnotenia pre lepšie vyhodnotenie výkonnosti modelov, a to:

- **Správnosť (accuracy) –** podiel správne klasifikovaných vzoriek zo všetkých vzoriek,
- **Presnosť (precision) –** podiel správne klasifikovaných vzoriek danej triedy zo všetkých vzoriek, ktoré model priradil k tejto triede,
- **Citlivosť (recall) -** podiel správne identifikovaných pozitívnych prípadov zo skutočne pozitívnych prípadov,
- **F1 skóre –** harmonický priemer presnosti a citlivosti.

4 VÝSLEDKY VÝSKUMU

Táto kapitola obsahuje výsledky dosiahnuté počas realizácie výskumu zameraného na porovnanie modelov strojového učenia v textoch z oblasti bankovníctva v slovenskom jazyku. Výsledky boli získané aplikáciou vybraných algoritmov strojového učenia na pripravený dataset, pričom ich výkon bol hodnotený pomocou 10-násobnej krížovej validácie a viacerých metrík. Získavanie výsledkov bolo riadené v súlade s metodológiou CRISP-DM a jej jednotlivými fázami, od analýzy dát až po modelovanie a evalváciu, čo umožnilo systematické vyhodnotenie celého procesu riešenia.

Cieľom tejto kapitoly bolo vyhodnotiť výkonnosť testovaných modelov, identifikovať ich silné a slabé stránky, a určiť, ktorý z prístupov je najvhodnejší na spracovanie doménovo špecifických textov v jazyku s obmedzenými zdrojmi.

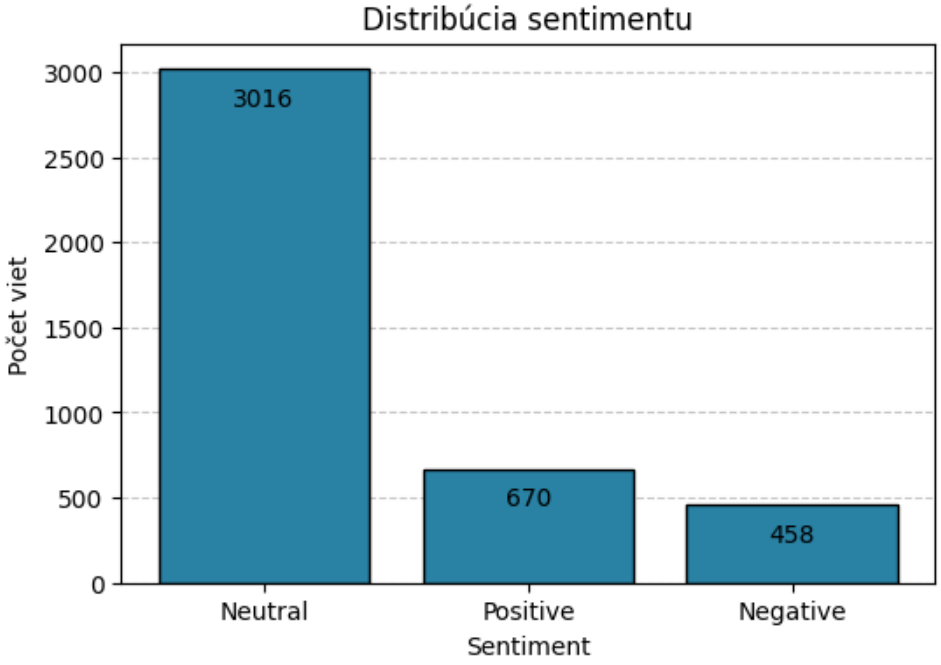
4.1 VÝSLEDKY FÁZY POROZUMENIA DÁT

Fáza porozumenia dát predstavovala dôležitý krok v procese CRISP-DM, počas ktorého sme analyzovali charakter a kvalitu súboru textov určeného na klasifikáciu textov. V rámci tejto fázy sme sa zamerali najmä na rozdelenie sentimentu, štruktúru textov, dĺžku viet a frekvenciu používaných slov, čo nám umožnilo lepšie pochopiť vlastnosti údajov a tak pripraviť vhodný postup jeho ďalšieho spracovania.

Dataset obsahuje 4144 záznamov vo forme viet alebo krátkych úryvkov z bankových dokumentov, ktoré sú označené sentimentom. Ku každému záznamu je dostupná stemmed aj lematizovaná verzia textu, ktoré redukujú slovné tvary na ich základ, čím sa znižuje riedkosť dát. Texty sú označené POS tagmi, čo umožňuje analyzovať štruktúru viet a identifikovať kľúčové slovné kategórie.

Analýza distribúcie sentimentu ukázala že triedy nie sú vyvážené, pričom neutrálny sentiment obsahuje približne 6,6-krát viac vzoriek ako negatívny sentiment. Graf 1 zobrazuje toto rozdelenie jednotlivých typov sentimentu, pričom neutrálny sentiment v datasete výrazne prevažuje nad pozitívnym a negatívnym sentimentom. Takéto rozdelenie je typické pre texty z finančného prostredia, kde správy majú často informatívny alebo neutrálny charakter. Táto nevyváženosť tried môže ovplyvniť proces trénovania klasifikačných modelov, keďže model môže mať tendenciu uprednostňovať

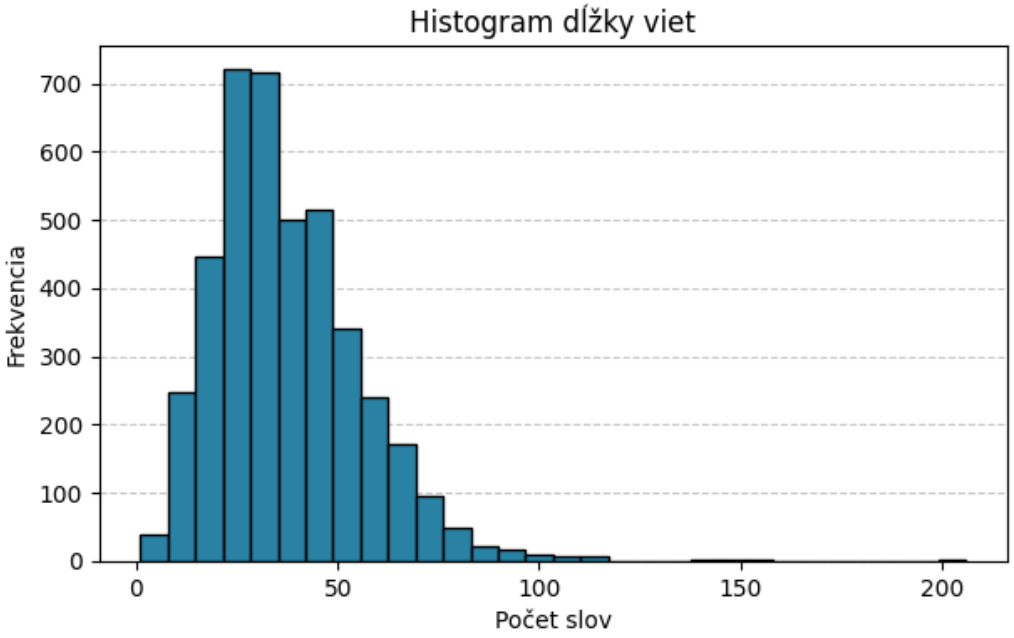
najčastejšie zastúpenú triedu. Z tohto dôvodu sme pri vyhodnocovaní modelov sledovali nielen správnosť, ale aj ďalšie metriky, a to presnosť, citlivosť a F1 skóre.



Graf 1 Distribúcia sentimentu v datasete[4]

Ďalej sme analyzovali dĺžku viet, meranú ako počet slov v jednotlivých záznamoch, vďaka čomu sme získali prehľad o štruktúre textov a identifikovali, či dáta pozostávali prevažne z kratších alebo dlhších textov. Graf 2 obsahuje histogram rozdelenia dĺžky viet v datasete, pričom priemerná dĺžka vety bola približne 37 slov a medián dosiahol hodnotu 34 slov. Väčšina viet sa pohybovala v intervale medzi 24 až 47 slovami, čo predstavuje medzikvartilové rozpätie. Tento široký rozsah dĺžky viet znamená, že texty majú rôznu zložitosť, od veľmi krátkych viet až po dlhé opisné vety, ktoré sú typické pre finančné dokumenty.

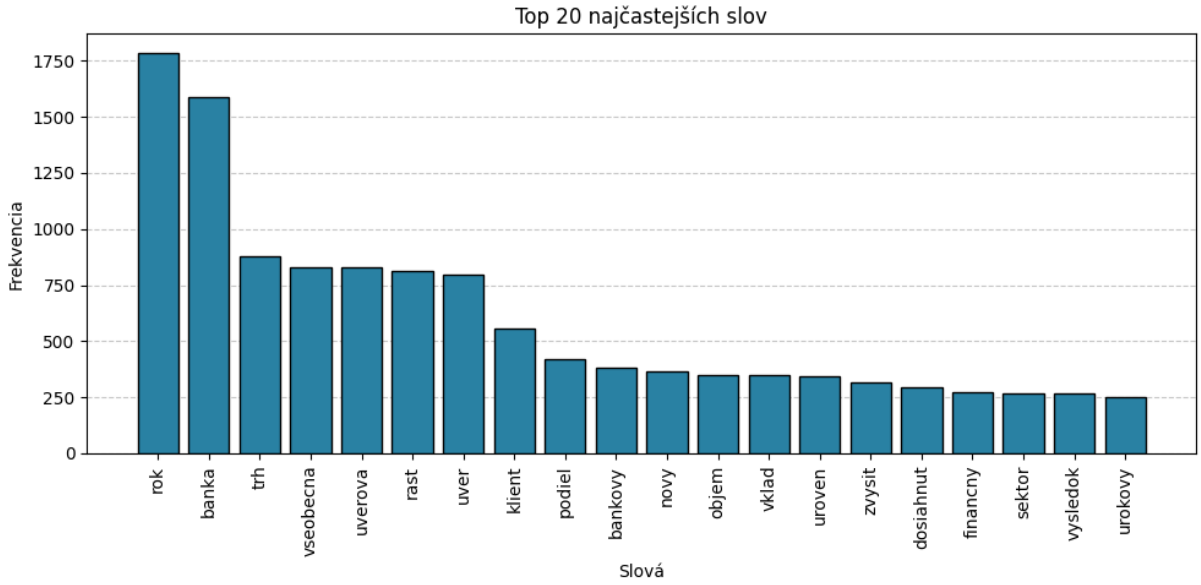
[4] Zdroj Graf 1: autor



Graf 2 Histogram rozdelenia dĺžky viet[5]

Ďalším krokom vo fáze porozumenia dát bola analýza frekvencie slov v texte. Pre tento účel sme využili lematizovanú verziu textu, pričom celkový počet slov vo všetkých záznamoch dosiahol 60 557, s počtom unikátnych slov 5 468, čo vedie k lexikálnej diverzite približne 0,09. Graf 3 zobrazuje 20 najčastejších slov v datasete, pričom medzi najpoužívanejšie slová patrili typické výrazy z finančného prostredia, ako sú rok, banka, trh, úverová alebo rast. Vďaka analýze sme identifikovali kľúčové termíny, ktoré sa v dokumentoch často opakovali, a tak sme získali lepší prehľad o obsahu textov pri príprave dát na modelovanie. Okrem individuálnych slov sme skúmali aj najčastejšie dvojice slov (bigramy), čo umožnilo zachytiť často sa vyskytujúce spojenia typické pre finančný kontext, ako napríklad „všeobecná úverová“ alebo „trhový podiel“.

[5] Zdroj Graf 2: autor



Graf 3 20 najčastejšie sa vyskytovaných slov[6]

4.2 DISKUSIA

V tejto kapitole sú prezentované interpretácie výsledkov jednotlivých modelov použitých v experimente. Cieľom nie je iba uviesť dosiahnuté hodnoty metrík, ale najmä analyzovať efektivitu jednotlivých prístupov, ich silné a slabé stránky, ako aj vplyv zvolených reprezentácií textu a optimalizovaných hyperparametrov na výsledný výkon.

Diskusia sa zameriava na porovnanie schopnosti modelov klasifikovať sentiment v kontexte nevyváženého datasetu, pričom dôraz je kladený najmä na správnosť, presnosť, citlivosť a F1 skóre uvádzané vo forme makro priemeru. Zároveň je hodnotená stabilita modelov pri rôznych rozdeleniach dát a ich schopnosť generalizácie.

4.2.1 *K*-najbližších susedov

Model *k*-najbližších susedov patril medzi výkonnejšie klasifikačné prístupy v rámci nášho experimentu. Bol použitý s reprezentáciou TF-IDF, ktorá bola

[6] Zdroj Graf 3: autor

kombinovaná s redukciou dimenzionality dát pomocou SVD a *n*-gram rozsahom (2, 2), teda s využitím bigramov.

Priemerná hodnota správnosti bola 0,82, pričom sa pohybovala v relatívne úzkom intervale medzi 0,80 a 0,84, čo na prvý pohľad hovorí o stabilnom výkone naprieč rôznymi rozdeleniami dát. Vzhľadom na nevyváženosť tried však túto metriku nemožno interpretovať ako dostatočný indikátor kvality modelu, keďže môže byť výrazne ovplyvnená najpočetnejšou triedou.

Detailnejší pohľad poskytli ostatné metriky. Presnosť dosiahla priemer približne 0,83, avšak s väčšou variabilitou (0,73 – 0,91), čím je model nestabilnejší pri identifikácii pozitívnych predikcií. Citlivosť bola výrazne nižšia, na úrovni približne 0,59, čo poukazuje na nedostatočnú schopnosť modelu zachytiť všetky triedy, najmä tie menej zastúpené. Tento nesúlad medzi presnosťou a citlivosťou sa následne odrážal aj na hodnote F1 skóre, ktoré dosiahlo priemer 0,64 so štandardnou odchýlkou 0,033. Daný výsledok poukazuje na nerovnomerný výkon medzi triedami a potvrdzuje, že model nedokáže spoľahlivo klasifikovať menej zastúpené kategórie.

Optimalizáciou parametrov sme zistili, že model dosiahol najlepšie výsledky pri počte susedov $k = 11$ a hyperparameter *weights* nastavený na *distance*, teda nastavovania váh podľa vzdialenosti kde bližšie body majú väčší vplyv na výslednú klasifikáciu. Zvolená metrika vzdialenosti p bola rovná 1, teda vieme, že model najlepšie fungoval pri využití Manhattanskej vzdialenosti.

Na základe získaných výsledkov možno konštatovať, že napriek vysokej správnosti model neposkytuje vyvážený výkon naprieč triedami. Nízka citlivosť a F1 skóre poukazujú na silný vplyv najviac zastúpenej triedy na klasifikáciu modelu. Tento efekt je typický pre KNN pri nevyvážených datasetoch, kde rozhodovanie zvýhodňuje najčastejšie sa vyskytované vzory. Okrem toho zostáva limitujúcim faktorom aj výpočtová náročnosť modelu, ktorá rastie s veľkosťou dát.

4.2.2 Rozhodovacie stromy

Rozhodovací strom predstavoval jednoduchší, ale zároveň dobre interpretovateľný prístup ku klasifikácii sentimentu. Podobne ako pri modeli KNN bola použitá reprezentácia TF-IDF s SVD redukciou a *n*-gram rozsahom (2, 2).

Priemerná hodnota správnosti dosiahla 0,74, a pohybovala sa v intervale medzi 0,71 a 0,78, čo poukazuje, že rozhodovací strom nedokáže tak efektívne spracovať dáta

ako výkonnejšie modely obsiahnuté v experimente. Presnosť aj citlivosť dosiahli približne rovnakú úroveň 0,59, teda problém modelu je nielen so správnou klasifikáciou pozitívnych predikcií, ale aj so zachytením všetkých tried. F1 skóre dosiahlo priemer 0,59 so štandardnou odchýlkou 0,035, pričom jednotlivé hodnoty sa pohybovali medzi 0,54 a 0,67. Tento rozsah odráža vyššiu citlivosť modelu na konkrétne rozdelenie dát, čo je typické pre rozhodovacie stromy bez dostatočnej regularizácie.

Optimalizácia hyperparametrov ukázala, že rozhodovací strom najlepšie klasifikoval pri maximálnej hĺbke 40, minimálnych parametroch *min_samples_split* = 2 a *min_samples_leaf* = 1, a použitím kritéria *entropy*. Parameter *class_weight = balanced* čiastočne kompenzoval nevyváženosť datasetu, avšak citlivosť pre menej zastúpené triedy zostala nižšia. Rozhodovacie stromy sú zároveň náchylné na nesprávne nastavenia parametrov, pričom príliš veľká hĺbka môže viesť k preučenému modelu a príliš malá hĺbka znižuje schopnosť modelu zachytiť zložitejšie vzory v dátach.

Celkovo model rozhodovacieho stromu poskytuje interpretovateľný, ale slabý model, vhodný skôr na rýchlu analýzu a vizualizáciu rozhodovacích pravidiel, pričom jeho použitie na plnohodnotnú klasifikáciu sentimentu je limitované vzhľadom na nižšiu presnosť a citlivosť v porovnaní s inými modelmi.

4.2.3 Logistická regresia

Logistická regresia patrila medzi najlepšie prístupy, ktoré sme testovali. Bola testovaná s dvoma reprezentáciami textu, TF-IDF a Bag-of-Words (BOW), pričom cieľom bolo zistiť, ktorá reprezentácia poskytuje lepšie vstupné hodnoty pre účel klasifikácie sentimentu.

Najlepšie výsledky dosiahol model logistickej regresie pri použití TF-IDF reprezentácie s *n*-gram rozsahom (1, 2). Priemerná hodnota správnosti dosiahla 0,88, pričom presnosť dosiahla približne 0,84 a citlivosť 0,76. Zároveň sme pozorovali stabilný výkon modelu z hľadiska F1 skóre, kedy hodnota F1 skóre bola okolo 0,80 a smerodajná odchýlka pre F1 skóre 0,031. Priemerná hodnota F1 skóre sa pohybovala v intervale od 0,75 po 0,84 pre jednotlivé iterácie. Táto hodnota svedčí o tom, že model dokáže pomerne rovnomerne klasifikovať jednotlivé triedy.

Optimalizácia hyperparametrov ukázala, že najlepšie výsledky boli dosiahnuté pri použití penalizácie typu *elasticinet*, solver typu *saga* a pri vyššej hodnote parametra

C, a to 19,68. Logistická regresia tak nepotrebuje vysokú regularizáciu a dokáže flexibilnejšie prispôsobiť hranice rozhodovania na základe dát. Parameter *use_idf = False* bol obzvlášť zaujímavým výsledkom optimalizácie, ktorý nám hovorí o tom, že model dosahoval lepšie výsledky, keď sa pri vážení nevyužila inverzná frekvencia dokumentov. Keďže sme použili doménovo špecifický korpus, ktorý obsahuje menej frekventované, ale dôležité slová, ich odstránenie by mohlo znížiť informatívnosť vstupov a tak zhoršiť klasifikáciu sentimentu.

Pri použití reprezentácie typu Bag-of-Words bola priemerná správnosť približne 0,86, presnosť 0,78 a citlivosť 0,80. F1 skóre dosiahlo priemernú hodnotu okolo 0,79. V porovnaní s TF-IDF bol výkon o niečo nižší, ale tento rozdiel nebol výrazný a tak môžeme povedať, že logistická regresia dokázala pracovať efektívne aj s jednoduchšou reprezentáciou textu založenou na počte výskytu slov. Po optimalizácii si model zachoval rovnaké nastavenia pre penalizáciu (*elasticinet*) a solver (*saga*), ako pri použití reprezentácie typu TF-IDF, ale rozdiel nastal v *n*-gram rozsahu, ktorý bol pre BOW nastavený len na unigramy (1, 1). Model v tomto prípade zachytával len jednotlivé slová bez viacslovných spojení, napriek čomu si zachoval dobrý výkon, čo svedčí o tom, že logistická regresia dokáže dosiahnuť dobré výsledky pri rôznych reprezentáciách textu.

4.2.4 Viacvrstvový perceptrón (MLP)

Viacvrstvový perceptrón predstavoval pokročilejší prístup pre klasifikáciu sentimentu, založený na neurónových sieťach. V experimente bol použitý s textovou reprezentáciou typu TF-IDF v kombinácii s redukciou dimenzionality SVD. Pre dosiahnutie optimálnych výsledkov bol zvolený *n*-gram rozsah (1, 1), teda unigramy.

Priemerná hodnota správnosti dosiahla 0,84 pričom sa pohybovala v intervale od 0,82 do 0,87, čo poukazuje na vysoký a zároveň stabilný výkon modelu naprieč iteráciami. Presnosť mala priemernú hodnotu 0,80 s relatívne nízkou variabilitou, teda model vie konzistentne, a pritom správne, klasifikovať pozitívne predikcie. Hodnota citlivosti dosahovala priemer 0,69, z čoho vieme usúdiť, že model dokáže relatívne dobre zachytiť aj menej zastúpené triedy. F1 skóre malo priemernú hodnotu 0,73 so štandardnou odchýlkou 0,025. Nízka variabilita medzi jednotlivými foldmi poukazuje na stabilitu modelu a jeho schopnosť generalizovať na rôzne rozdelenia dát,

čo podporuje záver, že MLP dokáže efektívne pracovať aj s komplexnejšími vzormi v dátach.

Optimalizácia parametrov ukázala, že najlepšie výsledky boli dosiahnuté pri použití jednej skrytej vrstvy s 256 neurónmi a aktivačnou funkciou ReLU. Rýchlosť učenia (*learning_rate_init*) bola nastavená na hodnotu 0,001 a parameter regularizácie *alpha* približne na úrovni 0,00089, čo pomohlo obmedziť preučenie modelu. Model okrem iného použil aj techniku *early stopping*, čo umožnilo zastaviť tréning v momente, kedy sa výkon prestal zlepšovať, čím sa zvýšila robustnosť modelu. Zaujímavé bolo aj využitie vyššieho počtu komponentov (400) pri SVD, ktoré indikuje, že model dokázal efektívne pracovať s väčším množstvom informácií zo vstupných dát.

Viacvrstvový perceptrón predstavoval vyvážený prístup z hľadiska presnosti, citlivosti aj stability. V porovnaní s jednoduchšími metódami, ako boli KNN alebo rozhodovacie stromy, dosiahol lepší výkon naprieč jednotlivými triedami a zároveň si udržal dobrú schopnosť generalizácie. Jeho výraznou výhodou, v porovnaní s jednoduchšími prístupmi, bola schopnosť zachytiť aj menej zastúpené triedy, čo je kľúčové pri nevyvážených datasetoch. Avšak napriek svojim výhodám nepatril medzi najvýkonnejšie modely v rámci experimentu, takže väčšia zložitosť modelovej architektúry nebola v tomto prípade rozhodujúcim faktorom. MLP predstavuje robustné riešenie, ktoré dokáže efektívne pracovať s komplexnejšími textovými reprezentáciami, hoci v tomto experimente zaostalo za najlepšími modelmi.

4.2.5 Stroje s podpornými vektormi

V prípade SVM bol analyzovaný výkon pri použití dvoch textových reprezentácií, a to konkrétne TF-IDF a BOW. Pri použití unigramovej TF-IDF reprezentácie dosiahol model správnosť 0,86, presnosť okolo 0,79 a citlivosť 0,76. Priemerná hodnota F1 skóre dosiahla 0,77 so smerodajnou odchýlkou 0,024, čo svedčí o jeho konzistentnom výkone pri rôznych rozdeleniach dát. Optimalizácia parametrov ukázala, že najlepšie výsledky boli dosiahnuté pri použití *loss* funkcie typu *squared_hinge* a pri vážení tried (*class_weight = balanced*), z čoho vieme usúdiť, že model dokázal kompenzovať nevyváženosť tried v dátach. Hodnota parametra C bola pri optimálnych nastaveniach nízka (0,30), čo znamená silnejšiu regularizáciu a dôraz na generalizačnú schopnosť namiesto presného prispôsobenia tréningovým dátam.

Pri použití reprezentácie typu Bag-of-Words došlo k miernemu zlepšeniu výsledkov. Priemerná správnosť zostala na úrovni 0,86, avšak presnosť sa zvýšila na približne 0,80 a F1 skóre na 0,78 so smerodajnou odchýlkou 0,021. Citlivosť zostala na rovnakej úrovni (0,76). Aj v tomto prípade boli zvolené rovnaké nastavenia pre *loss* a *class_weight* funkcie, pričom hodnota parametra C bola ešte nižšia (0,07), čo poukazuje na potrebu silnejšej regularizácie pri tejto reprezentácii. Okrem toho dokázalo SVM efektívne pracovať aj bez explicitného zachytávania viacslovných spojení, keďže pri oboch typoch textovej reprezentácie boli použité unigramy.

Vďaka vhodne nastavenej regularizácii si model udržal dobrú generalizačnú schopnosť, čo sa prejavilo nízkou variabilitou výsledkov medzi jednotlivými rozdeleniami dát. Tento prístup tak predstavoval spoľahlivé riešenie pre klasifikáciu sentimentu pri nevyvážených dátach, najmä v situáciách, kde je dôležitá rovnováha medzi výkonom a stabilitou modelu.

4.2.6 Náhodný les

Model náhodného lesa predstavoval robustnejší skupinový prístup založený na kombinácii viacerých rozhodovacích stromov. V experimente bol použitý s textovou reprezentáciou TF-IDF s redukciou dimenzionality SVD a *n*-gram rozsahom (1, 1).

Priemerná hodnota správnosti dosiahla približne 0,83, pričom sa pohybovala vo veľmi úzkom intervale od 0,82 do 0,84 pri jednotlivých iteráciách. Presnosť tiež dosiahla priemernú hodnotu 0,83, avšak s mierne vyššou variabilitou (od 0,78 do 0,87), a teda model bol pri niektorých prípadoch citlivejší na dané rozdelenie dát. Citlivosť modelu bola nižšia, na úrovni približne 0,61, čo poukazuje na pretrvávajúci problém so zachytávaním menej zastúpených tried, podobne ako pri niektorých iných modeloch použitých v experimente. F1 skóre dosiahlo priemernú hodnotu 0,67 so smerodajnou odchýlkou 0,020. Relatívne nízka hodnota F1 skóre poukazuje na slabšiu schopnosť zachytiť menej zastúpené triedy.

Optimalizácia parametrov ukázala, že najlepšie výsledky boli dosiahnuté pri použití väčšieho počtu stromov (*n_estimators* = 400) a maximálnej hĺbky stromov 40. Obmedzenie minimálneho počtu vzoriek pre rozdelenie (*min_samples_split* = 10) a listov (*min_samples_leaf* = 4) zároveň prispelo k redukcii preučenia. Zaujímavým zistením bolo aj nastavenie TF-IDF parametrov, a to konkrétne použitie *sublinear_tf,* ktoré logaritmicky škáluje frekvenciu výrazov, vďaka čomu znižuje vplyv veľmi

častých slov, a tak umožňuje modelu lepšie zohľadniť informatívnejšie, menej frekventované výrazy. Okrem toho nebola využitá normalizácia (*norm* = None), vďaka čomu si vstupné vektory zachovali pôvodné váhy bez škálovania, čo mohlo prispieť k lepšiemu rozlišovaniu medzi dokumentmi na základe frekvenčných charakteristík.

Napriek vyššej správnosti bol výkon modelu limitovaný nižšou presnosťou a F1 skóre, čo poukazuje na slabšiu schopnosť zachytiť minoritné triedy. Model síce dokázal modelovať nelineárne vzory a bol robustný voči šumu, avšak v tomto prípade nedosiahol tak vyvážený výkon ako lineárne prístupy.

4.2.7 LightGBM

V rámci experimentu model LightGBM predstavoval pokročilý skupinový prístup založený na gradientnom boostingu rozhodovacích stromov. Bol použitý s textovou reprezentáciou TF-IDF v kombinácii s redukciou dimenzionality pomocou SVD a *n*-gram rozsahom (2, 2).

Priemerná hodnota správnosti dosiahla približne 0,85, pričom sa pohybovala v intervale od 0,83 do 0,89. Presnosť dosiahla vysokú hodnotu, v priemere okolo 0,88 a prístup bol schopný spoľahlivo identifikovať pozitívne predikcie. Naopak, citlivosť bola nižšia (približne 0,64), čo poukazuje na problém so zachytávaním menej zastúpených tried. F1 skóre dosiahlo priemernú hodnotu 0,71 so smerodajnou odchýlkou 0,036 a pohybovalo sa v intervale od 0,67 do 0,81 čo odráža miernu citlivosť výkonu na rozdelenie dát.

Najlepšie výsledky dosiahlo nastavenie vyššieho počtu stromov (*n_estimators* = 600) a hodnoty rýchlosti učenia 0,1, čo zabezpečilo postupné učenie modelu. Nastavenie *min_child_samples* = 10 znamenalo, že jednotlivé listy stromu obsahovali dostatočný počet vzoriek. Rovnako pre nás zaujímavé bolo aj nastavenie TF-IDF parametrov, konktrétne použitie *sublinear_tf* a vypnutie *use_idf,* z čoho vieme usúdiť, že pri klasifikácii bolo informatívnejšie sledovať absolútnu frekvenciu slov v dokumentoch než upravovať ich váhu podľa celkového výskytu v korpuse.

Z hľadiska charakteru modelu je dôležité zdôrazniť, že LightGBM ako boostingový prístup postupne budoval sekvenciu stromov, pričom každý ďalší strom sa snažil opraviť chyby tých predchádzajúcich. Tento mechanizmus umožnil lepšie zachytiť nelineárne vzory v dátach, avšak zároveň viedol k vyššej náchylnosti na šum alebo nevyváženosť tried, čo sa v tomto prípade prejavilo nižšou citlivosťou. Celkovo

možno tento prístup v rámci experimentu hodnotiť ako výkonný z hľadiska presnosti, avšak menej vyvážený pri klasifikácii jednotlivých tried, čo sa prejavilo najmä nižšou citlivosťou a F1 skóre.

4.2.8 XGBoost

XGBoost predstavoval ďalší pokročilý skupinový prístup založený na princípe gradientného boostingu rozhodovacích stromov. V experimente bol použitý s textovou reprezentáciou TF-IDF v kombinácii s redukciou dimenzionality pomocou SVD, pričom bol zvolený *n*-gram rozsah (1, 1), teda unigramy.

Priemerná hodnota správnosti dosiahla približne 0,84, presnosť dosiahla priemer okolo 0,83, model teda mal dobrú schopnosť správne identifikovať pozitívne predikcie. Citlivosť bola nižšia, na úrovni približne 0,65, čo opäť odráža problém s menej zastúpenými triedami. F1 skóre dosiahlo priemernú hodnotu 0,71 so smerodajnou odchýlkou 0,030, pričom variabilita medzi jednotlivými iteráciami bola mierna, a tak môžeme povedať, že tento prístup si udržiaval stabilný výkon naprieč rôznymi rozdeleniami dát.

Optimalizácia ukázala, že plytšie, ale početnejšie stromy (*n_estimators* = 900, *max_depth* = 3) poskytli najlepší kompromis medzi zachytením vzorov a rizikom preučenia. Hodnota *learning_rate* = 0,1 zároveň poukázala na kompromis medzi rýchlosťou učenia a ustálenosťou tréningového procesu, pričom sa model postupne učil a kontroloval jednotlivé stromy a ich vplyv na výsledok klasifikácie. Parameter *gamma* = 1,0 zohrával úlohu pri regularizácii zložitosti stromov určovaním minimálneho zlepšenia potrebného na vykonanie rozdelenia daného uzla. Vyššia hodnote tohto parametra prispela k zníženiu počtu zbytočných vetvení a tak podporila jednoduchšiu štruktúru modelu.

Z hľadiska spracovania textu využil model TF-IDF váženie bez aplikácie *sublinear_tf,* na rozdiel od LightGBM, kde sa logaritmické škálovanie frekvencie slov ukázalo ako vhodnejšie nastavenie. V tomto prípade teda zostala zachovaná lineárna závislosť medzi frekvenciou výskytu slova a jeho váhou, čo znamená, že častejšie sa vyskytujúce slová mali priamo úmerne väčší vplyv na výslednú reprezentáciu dokumentu.

Model XGBoost tak dokázal efektívne zachytiť komplexnejšie vzory v dátach, avšak jeho citlivosť na menej zastúpené triedy bola nižšia, čo súviselo

s nevyváženosťou datasetu. Hodnota F1 skóre spolu s nižšou citlivosťou indikovali, že výkon modelu nie je rovnomerne rozdelený medzi triedami a model má tendenciu uprednostňovať početnejšie kategórie. V dôsledku toho, napriek celkovo dobrému výkonu zostáva jeho schopnosť spoľahlivo identifikovať všetky triedy, najmä tie menej zastúpené, obmedzená.

Na základe výsledkov experimentu na pôvodnom datasete sme identifikovali viacero výkonných prístupov, pričom za najvhodnejší kompromis medzi presnosťou, stabilitou a interpretovateľnosťou sa ukázala logistická regresia s TF-IDF reprezentáciou textu.

4.3 ANALÝZA A SPRACOVANIE NOVÉHO DATASETU

Táto kapitola sa zameriava na analýzu a spracovanie nového datasetu finančných textov od VÚB za roky 2023–2024, ktorý nebol súčasťou pôvodného tréningového korpusu. Cieľom bolo charakterizovať štruktúru a vlastnosti dát, pripraviť ich na následnú klasifikáciu sentimentu a overiť použiteľnosť najlepšieho modelu identifikovaného v predchádzajúcom experimente na nových, reálnych dátach.

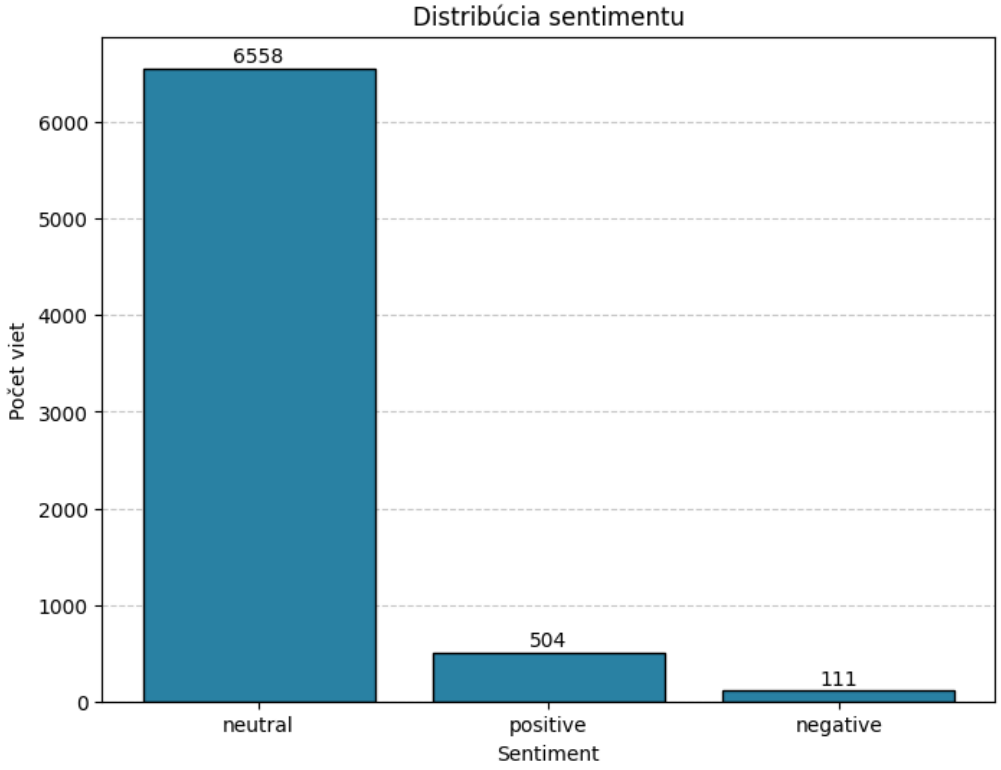
4.3.1 Popis datasetu a predspracovanie

Táto podkapitola predstavuje zozbieraný dataset a kroky jeho predspracovania. Dataset obsahuje 7177 viet z rôznych typov finančných správ a článkov, vrátane výročných správ, finančných výsledkov, kvartálnych reportov a príhovorov vedenia banky, pričom každý záznam obsahuje informácie o kapitole, roku a type správy. Pôvodné texty neobsahovali anotácie sentimentu a reprezentovali reálny finančný jazyk, ktorý často zahŕňa odbornú terminológiu, právne formulácie a opisné pasáže.

V rámci predspracovania dát boli texty najskôr segmentované a tokenizované na úrovni viet pomocou nástroja Stanza, pričom každá veta predstavovala samostatný záznam v datasete. Následne boli texty očistené od nepotrebných znakov a normalizované. Po tomto kroku bola vykonaná lemmatizácia, ktorej cieľom bolo zjednotiť slovné tvary a redukovať tak variabilitu textov, vďaka čomu sme získali konzistentnejšie dáta pripravené na následnú klasifikáciu.

Keďže dataset pôvodne neobsahoval označenie sentimentu, bola vykonaná manuálna anotácia, ktorá slúžila na vytvorenie referenčnej množiny pre evalváciu predikcií modelu. Anotácia bola vykonaná s ohľadom na finančný kontext, pričom sa zohľadňovalo, že finančný jazyk je prevažne opisný, preto boli za neutrálne považované hlavne vecné tvrdenia bez hodnotenia, zatiaľ čo pozitívny sentiment zahŕňal vyjadrenia o zlepšení, opatreniach alebo prínosoch. Negatívny sentiment bol priradený vetám, ktoré opisovali riziká, problémy alebo nepriaznivé javy.

Vzhľadom na charakter zdrojov bolo rozdelenie tried nevyvážené, pričom neutrálna trieda tvorila dominantnú časť dát (6558 viet), zatiaľ čo pozitívny sentiment bol zastúpený v 504 vetách a negatívny len v 111 vetách. Tento nepomer bol dôsledkom povahy finančných reportov, ktoré sú hlavne informatívne, pričom potenciálne negatívne hodnotenia boli zapísané miernejším spôsobom alebo boli rovno prezentované s opatreniami na ich riešenie. Rozdelenie tried je znázornené na Grafe 4, ktorý zobrazuje výraznú prevahu neutrálnej triedy.

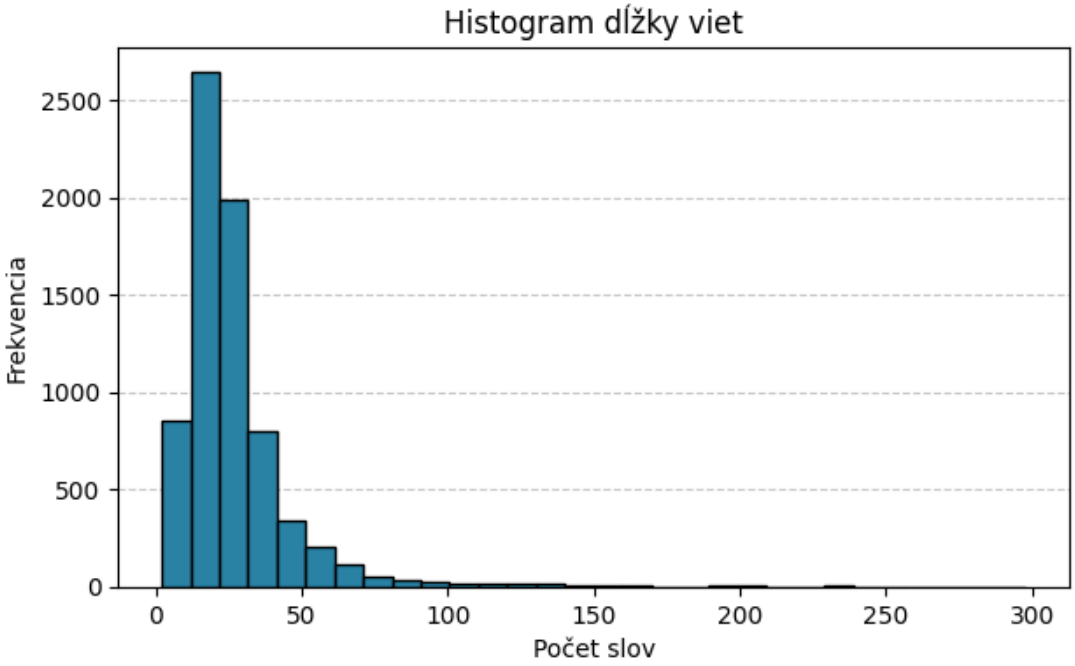


Graf 4 Distribúcia sentimentu v novom datasete[7]

[7] Zdroj Graf 4: autor

Analýza rozdelenia dát podľa typu správy ukázala, že veľkú časť tvorili najmä výročné správy, zatiaľ čo ostatné typy dokumentov boli zastúpené v obmedzenom rozsahu. Do určitej miery bol tak dataset ovplyvnený štýlom týchto výročných správ, ktoré sa hlavne vyznačujú formálnym tónom a špecifickou štruktúrou.

Väčšina viet mala dĺžku medzi 11 a 45 slovami, pričom medián dosiahol 22 slov. Graf 5 znázorňuje rozdelenie dĺžky viet a poskytuje prehľad o štruktúre textov v datasete. Tieto zistenia sú konzistentné s kvantilovou analýzou, ktorá ukázala, že 10 % najkratších viet obsahovalo maximálne 11 slov, horný kvartil (75 %) dosiahol 31 slov a 10 % najdlhších viet presiahlo 45 slov. Rozdelenie podľa rokov ukázalo, že väčšinu viet tvorili texty z roku 2024, konkrétne 4 553 viet, čo predstavovalo 63 % všetkých dát, zatiaľ čo 2 621 viet (37 %) pochádzalo z roku 2023.



Graf 5 Histogram rozdelenia dĺžky viet[8]

[8] Zdroj Graf 5: autor

4.3.2 Aplikácia modelu a výsledky klasifikácie

Po úvodnom spracovaní a analýze získaného datasetu sme použili najlepší model, ktorý sme identifikovali v predošlom experimente, a to konkrétne logistickú regresiu s TF-IDF reprezentáciou textu. Tento model bol zvolený na základe dosiahnutých výsledkov, pričom poskytol najlepší kompromis medzi presnosťou a stabilitou pri jednotlivých triedach.

Po aplikovaní modelu na anotovaný dataset boli dosiahnuté nasledovné hodnoty: správnosť na úrovni okolo 0,88, presnosť 0,84, citlivosť 0,88 a F1 skóre 0,86. Z daných výsledkov sme najprv usúdili, že model si zachováva vysoký výkon aj pri spracovaní nových dát, ktoré neboli súčasťou tréningového korpusu. Detailnejšia analýza výstupov modelu však ukázala, že rozdelenie predikovaných tried bolo výrazne nevyvážené. Model klasifikoval prevažnú väčšinu viet ako neutrálne (6908 prípadov), zatiaľ čo pozitívny sentiment bol identifikovaný len v 188 vetách a negatívny v 77 vetách. V porovnaní s referenčnou anotáciou išlo o zhoršenú klasifikáciu menej zastúpených tried, najmä z hľadiska pozitívneho sentimentu.

Tento jav bol dôsledkom nevyváženého datasetu, v ktorom prevládala neutrálna trieda. Model mal v takomto prípade tendenciu zvoliť najviac zastúpenú triedu, čo viedlo k vysokým hodnotám celkových metrík, najmä správnosti a citlivosti. Z tohto dôvodu bolo potrebné interpretovať dosiahnuté metriky s vyššou opatrnosťou, keďže získané hodnoty nemuseli úplne odrážať schopnosť modelu rozlišovať medzi triedami. Napriek celkovo dobrému výkonu model logistickej regresie vykazoval nižšiu citlivosť voči pozitívnym a negatívnym prípadom , čo poukazuje na jeho obmedzenú schopnosť zachytiť jemnejšie rozdiely v polarite sentimentu pri menej zastúpených triedach.

Získané výsledky tak naznačujú, že pri aplikácii modelu na reálne finančné texty je potrebné venovať zvýšenú pozornosť nevyváženosti dát, ktorá môže významne ovplyvniť výslednú klasifikáciu. Pre ďalšie zlepšenie výkonu modelu by preto bolo vhodné uvažovať o implementácii prístupov zameraných na vyrovnanie distribúcie tried, ako napríklad penalizácia dominantnej triedy, úprava váh počas trénovania alebo rozšírenie datasetu o viac vzoriek pozitívneho a negatívneho sentimentu.

4.4 ŠTATISTICKÉ VYHODNOTENIE MODELOV

V tejto časti je prezentovaná štatistická analýza výkonnosti základných klasifikačných modelov aplikovaných na rôzne textové reprezentácie (BOW, TF-IDF,

TF-IDF so znížením dimenzie pomocou SVD). Analýza bola rozšírená aj o ďalšie modely, a to konkrétne Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), Ridge klasifikátor, pasívno-agresívny klasifikátor (PA), LightGBM a klasifikátor založený na stochastickom gradientom zostupe (SGD). Cieľom analýzy bolo overiť, či existujú štatisticky významné rozdiely medzi jednotlivými modelmi a identifikovať najvýkonnejšie prístupy.

4.4.1 Vyhodnotenie modelov strojového učenia

Na úvod sme analyzovali popisné štatistiky, ktoré sa nachádzajú v Prílohe A, na základe ktorých sme zistili, že dáta boli symetrické a homogénne. Blízkosť hodnôt priemeru a mediánu vo väčšine prípadov nepoukazuje na výraznú asymetriu a zároveň bola pozorovaná nízka variabilita meraní, čo potvrdzujú aj nízke hodnoty variačného koeficientu.

Z hľadiska hodnôt výkonnostných metrík sme už na úrovni deskriptívnej analýzy spozorovali rozdiely medzi jednotlivými modelmi a ich reprezentáciami. Z popisných štatistík (Príloha A) sme spozorovali rozdiely v strednej hodnote príslušných výkonnostných mier medzi skúmanými metódami a reprezentáciami, čo naznačuje, že voľba modelu a spôsob reprezentácie vstupných dát môže mať významný vplyv na dosahovaný výkon.

Normalita rozdelenia bola overená pomocou Kolmogorov-Smirnovovho testu, ktorého výsledky sú uvedené v Prílohe B. Vo väčšine prípadov nebola zamietnutá nulová hypotéza o normálnom rozdelení ($p > 0{,}20$). V niektorých prípadoch sa však vyskytli mierne odchýlky, kde p-hodnoty klesli pod 0,20 (napr. $p < 0{,}15$ alebo $p < 0{,}10$), avšak ani v týchto prípadoch nedošlo k poklesu pod hladinu významnosti $\alpha = 0{,}05$. Na základe toho možno predpoklad normality považovať za splnený a dáta za vhodné pre použitie parametrických štatistických metód.

Na základe zistení uvedených v Prílohe A boli pre jednotlivé reprezentácie formulované globálne nulové štatistické hypotézy, ktoré predpokladajú, že medzi skúmanými modelmi neexistuje štatisticky významný rozdiel vo výkonnosti hodnotenej pomocou vybraných metrík správnosti, presnosti, citlivosti a F1 skóre. Testovanie týchto hypotéz bolo realizované samostatne pre každú textovú reprezentáciu, čím bolo možné vyhodnotiť vplyv konkrétnych spôsobov reprezentácie textu na výkonnosť jednotlivých modelov.

V prípade zamietnutia globálnej nulovej hypotézy bolo následne aplikované viacúrovňové testovanie, ktorého cieľom bolo identifikovať konkrétne rozdiely medzi modelmi a určiť tie najmenej výkonné pre danú textovú reprezentáciu. Tieto modely, ktoré sa ani v jednej z hodnotených metrík nedostali medzi tie najvýkonnejšie, boli v ďalšej analýze považované za menej vhodné a boli z nej vylúčené. Vďaka tomuto postupu sme postupne zredukovali množinu modelov a zamerali sa na tie prístupy, ktoré dosiahli vysoký výkon naprieč viacerými metrikami.

Pre lepšiu prehľadnosť boli z rozsiahlych popisných štatistík vybrané najvýkonnejšie modely na základe priemernej hodnoty F1 skóre (Tabuľka 2). Medzi najlepšie prístupy patrili najmä subsymbolické prístupy a to konkrétne logistická regresia, klasifikátor založený na stochastickom gradientom zostupe a stroje s podpornými vektormi, a to naprieč oboma reprezentáciami. Zároveň možno pozorovať že najlepšie modely dosiahli podobne dobrý výkon bez výrazných odchýlok medzi jednotlivými reprezentáciami.

Tabuľka 2 Prehľad najvýkonnejších modelov na základe priemernej hodnoty F1 skóre[9]

Model	**Reprezentácia**	**Správnosť (mean)**	**Presnosť (mean)**	**Citlivosť (mean)**	**F1 (mean)**
LogReg	TF-IDF	0,878	0,841	0,763	0,796
SGD	BOW	0,865	0,799	0,786	0,792
LogReg	BOW	0,860	0,779	0,804	0,790
Ridge	BOW	0,869	0,829	0,750	0,783
SVM	BOW	0,864	0,800	0,768	0,782
SGD	TF-IDF	0,862	0,794	0,769	0,780
Ridge	TF-IDF	0,867	0,832	0,741	0,778
SVM	TF-IDF	0,857	0,786	0,763	0,773

Na uvedené zistenia nadväzuje detailnejšia analýza pre jednotlivé textové reprezentácie. V prípade reprezentácie typu Bag-of-Words boli v Tabuľke 3 identifikované skupiny modelov s porovnateľným výkonom na základe jednotlivých metrík. Keďže premenné pochádzajú z normálneho rozdelenia a nebol porušený predpoklad sféricity kovariančnej matice ($p > 0,05$) na testovanie globálnych nulových hypotéz, ktoré predpokladajú absenciu štatisticky významných rozdielov vo výkone skúmaných metód pre danú reprezentáciu, boli použité neupravené testy pre opakované merania.

[9] Zdroj Tabuľka 2: autor

Tabuľka 3 Viacúrovňové testy pre metriky: ***a)*** *správnosť,* ***b)*** *presnosť,* ***c)*** *citlivosť,* ***d)*** *f1 skóre*[10]

Správnosť	**Mean**	**1**	**2**	**3**				
CNB_BOW_accuracy	0,781		****					
MNB_BOW_accuracy	0,802			****				
LogReg_BOW_accuracy	0,860	****						
SVM_BOW_accuracy	0,864	****						
SGD_BOW_accuracy	0,865	****						
PA_BOW_accuracy	0,867	****						
Ridge_BOW_accuracy	0,869	****						
Presnosť	**Mean**	**1**	**2**	**3**	**4**	**5**		
CNB_BOW_precision	0,665			****				
MNB_BOW_precision	0,708				****			
LogReg_BOW_precision	0,779					****		
SGD_BOW_precision	0,799	****						
SVM_BOW_precision	0,800	****						
Ridge_BOW_precision	0,829		****					
PA_BOW_precision	0,845		****					
Citlivosť	**Mean**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
MNB_BOW_recall	0,659	****						
CNB_BOW_recall	0,693		****					
PA_BOW_recall	0,719			****				
Ridge_BOW_recall	0,750				****			
SVM_BOW_recall	0,768					****		
SGD_BOW_recall	0,786						****	
LogReg_BOW_recall	0,804							****
F1	**Mean**	**1**	**2**	**3**				
CNB_BOW_f1	0,677			****				
MNB_BOW_f1	0,680			****				
PA_BOW_f1	0,767		****					
SVM_BOW_f1	0,782	****	****					
Ridge_BOW_f1	0,783	****	****					
LogReg_BOW_f1	0,790	****						
SGD_BOW_f1	0,792	****						

Poznámka: **** - $p > 0,05$, označené - vylúčené z analýzy

Pri metrike správnosti možno pozorovať, že najvyššie hodnoty dosiahli modely Ridge (0,869), pasívno-agresívny klasifikátor (0,867), SGD (0,865), SVM (0,864) a logistická regresia (0,860), ktoré zároveň tvoria homogénnu skupinu bez štatisticky významných rozdielov ($p > 0,05$). Naopak, modely založené na naivnom Bayesovskom prístupe dosahovali nižšie hodnoty správnosti a nepatrili medzi najvýkonnejšie prístupy.

[10] Zdroj Tabuľka 3: autor

Podobný trend bol pozorovaný aj pri metrike presnosti, kde najvyššie hodnoty dosiahli pasívno-agresívny klasifikátor (0,845) a Ridge klasifikátor (0,829), pričom spolu s modelmi SGD a SVM vytvárajú skupinu štatisticky porovnateľných modelov. Modely CNB a MNB opäť nedosiahli porovnateľný výkon s najlepšími prístupmi.

Z hľadiska citlivosti dosiahla najvyššie hodnoty logistická regresia (0,804) a SGD (0,786), pričom tieto modely spolu s SVM a Ridge klasifikátorom tvoria skupinu bez štatisticky významných rozdielov. Naopak, najnižšie hodnoty citlivosti boli zaznamenané pri modeloch MNB a CNB.

V prípade F1 skóre sa ako najvýkonnejšie ukázali SGD (0,792) a logistická regresia (0,790), spolu s Ridge a SVM, ktoré tvorili štatisticky porovnateľnú skupinu.

Globálne testy pre opakované merania potvrdili štatisticky významné rozdiely medzi modelmi ($df1 = 6$, $df2 = 54$, $F > 82{,}720$, $p < 0{,}001$). Následná analýza ukázala, že modely založené na naivnom Bayesovskom prístupe (MNB, CNB) dosiahli štatisticky nižší výkon a nepatria do žiadnej z homogénnych skupín najlepších modelov, a preto boli z ďalšej analýzy vylúčené.

Podobný postup bol použitý aj pri reprezentácii typu TF-IDF, ktorá bola vyhodnotená rovnakým spôsobom ako pri Bag-of-Words, a to použitím neupravených testov pre opakované merania na testovanie globálnych hypotéz, ktoré predpokladajú, že nie je štatistický významný rozdiel vo výkone skúmaných modelov. Uvedené zistenia sú vyhodnotené v Tabuľke 4.

Tabuľka 4 Viacúrovňové testy pre metriky: ***a)*** *správnosť,* ***b)*** *presnosť,* ***c)*** *citlivosť,* ***d)*** *f1 skóre*[11]

Správnosť	**Mean**	**1**	**2**	**3**	**4**	**5**
CNB_TFIDF_accuracy	0,771			****		
MNB_TFIDF_accuracy	0,803				****	
PA_TFIDF_accuracy	0,848		****			
SVM_TFIDF_accuracy	0,857	****	****			
SGD_TFIDF_accuracy	0,862	****				
Ridge_TFIDF_accuracy	0,867	****				
LogReg_TFIDF_accuracy	0,878					****
Presnosť	**Mean**	**1**	**2**	**3**	**4**	**5**
CNB_TFIDF_precision	0,652				****	
MNB_TFIDF_precision	0,701					****
PA_TFIDF_precision	0,773	****				
SVM_TFIDF_precision	0,786	****	****			
SGD_TFIDF_precision	0,794		****			
Ridge_TFIDF_precision	0,832			****		
LogReg_TFIDF_precision	0,841			****		
Citlivosť	**Mean**	**1**	**2**	**3**	**4**	**5**
MNB_TFIDF_recall	0,680				****	
CNB_TFIDF_recall	0,710					****
Ridge_TFIDF_recall	0,741			****		
PA_TFIDF_recall	0,749	****		****		
SVM_TFIDF_recall	0,763	****	****			
LogReg_TFIDF_recall	0,763	****	****			
SGD_TFIDF_recall	0,769		****			
F1	**Mean**	**1**	**2**	**3**	**4**	
CNB_TFIDF_f1	0,675		****			
MNB_TFIDF_f1	0,689		****			
PA_TFIDF_f1	0,759			****		
SVM_TFIDF_f1	0,773	****		****		
Ridge_TFIDF_f1	0,778	****				
SGD_TFIDF_f1	0,780	****				
LogReg_TFIDF_f1	0,796				****	

Poznámka: **** - $p > 0,05$, označené - vylúčené z analýzy

Pri metrike správnosti dosahovala najvyššiu hodnotu logistická regresia (0,878), ktorá tvorila samostatnú skupinu najvýkonnejších modelov. Nasledovali Ridge klasifikátor (0,867), SGD (0,862) a SVM (0,857), ktoré vytvárali homogénnu skupinu

[11] Zdroj Tabuľka 4: autor

bez štatisticky významných rozdielov. Modely MNB a CNB dosahovali najnižšie hodnoty správnosti a nepatrili medzi štatisticky porovnateľné prístupy s vyššie uvedenými modelmi.

Podobný výsledok bol zaznamenaný aj pri metrike presnosti, kde najvyššiu hodnotu dosiahla logistická regresia (0,841), ktorá spolu s Ridge klasifikátorom (0,832) tvorili homogénnu skupinu. Po nich nasledovali SGD (0,794) a SVM (0,786), ktoré tvorili ďalšiu skupinu bez štatisticky významných rozdielov. Modely MNB a CNB dosahovali nižšiu výkonnosť.

Z hľadiska citlivosti dosahoval najvyššie hodnoty SGD (0,769), logistická regresia (0,763) a SVM (0,763), ktoré vytvárali homogénnu skupinu najvýkonnejších modelov bez štatisticky významných rozdielov. Modely MNB a CNB opäť dosahovali najnižšie hodnoty citlivosti a neboli štatisticky porovnateľné s ostatnými modelmi.

Pri metrike F1 skóre dosiahla najvyššiu hodnotu logistická regresia (0,796), ktorá tvorila samostatnú skupinu najvýkonnejších metód. Nasledovala skupina modelov SGD (0,780) a Ridge (0,778), ktoré tvorili homogénnu skupinu bez štatisticky významných rozdielov. Model SVM (0,773) bol štatisticky porovnateľný so skupinou SGD a Ridge, ako aj s pasívno-agresívnym klasifikátorom (0,759), pričom pasívno-agresívny klasifikátor nepatril medzi najvýkonnejšie modely vo všetkých sledovaných metrikách.

Na základe globálnych testov pre opakované merania boli pre všetky sledované metriky zamietnuté nulové hypotézy o absencii štatisticky významných rozdielov vo výkone modelov pri reprezentácii TF-IDF ($df1 = 6$, $df2 = 54$, $F > 32{,}278$, $p < 0{,}001$), čo potvrdilo existenciu štatisticky významných rozdielov medzi modelmi.

Následná analýza ukázala, že logistická regresia, SGD, Ridge a SVM tvorili konzistentne homogénne skupiny najvýkonnejších modelov naprieč metrikami. Naopak, modely MNB, CNB a pasívno-agresívny klasifikátor systematicky dosahovali výrazne nižší alebo kolísavý výkon, a preto boli vylúčené z ďalšej analýzy.

Pri reprezentácii typu TF-IDF so znížením dimenzie pomocou SVD bola použitá rovnaká metodika štatistického vyhodnotenia ako v predchádzajúcich prípadoch. Výsledky v Tabuľke 5 poukazujú na rozdiely vo výkonnosti modelov pri použití testov pre opakované merania, pričom najväčšie rozdiely sme pozorovali najmä pri metrike F1 skóre a citlivosti.

Tabuľka 5 Viacúrovňové testy pre metriky: ***a)*** *správnosť,* ***b)*** *presnosť,* ***c)*** *citlivosť,* ***d)*** *f1 skóre*[12]

Správnosť	**Mean**	**1**	**2**	**3**		
DT_TFIDF_SVD_accuracy	0,740			****		
KNN_TFIDF_SVD_accuracy	0,819		****			
RF_TFIDF_SVD_accuracy	0,827		****			
XGB_TFIDF_SVD_accuracy	0,842	****				
MLP_TFIDF_SVD_accuracy	0,845	****				
LGBM_TFIDF_SVD_accuracy	0,847	****				
Presnosť	**Mean**	**1**	**2**	**3**	**4**	
DT_TFIDF_SVD_precision	0,597			****		
MLP_TFIDF_SVD_precision	0,801		****			
RF_TFIDF_SVD_precision	0,826	****	****			
XGB_TFIDF_SVD_precision	0,834	****				
KNN_TFIDF_SVD_precision	0,835	****				
LGBM_TFIDF_SVD_precision	0,876				****	
Citlivosť	**Mean**	**1**	**2**	**3**	**4**	
KNN_TFIDF_SVD_recall	0,580	****				
DT_TFIDF_SVD_recall	0,593	****				
RF_TFIDF_SVD_recall	0,611			****		
LGBM_TFIDF_SVD_recall	0,642		****			
XGB_TFIDF_SVD_recall	0,647		****			
MLP_TFIDF_SVD_recall	0,690				****	
F1	**Mean**	**1**	**2**	**3**	**4**	**5**
DT_TFIDF_SVD_f1	0,594		****			
KNN_TFIDF_SVD_f1	0,643			****		
RF_TFIDF_SVD_f1	0,673				****	
XGB_TFIDF_SVD_f1	0,708	****				
LGBM_TFIDF_SVD_f1	0,710	****				
MLP_TFIDF_SVD_f1	0,732					****

Poznámka: **** - $p > 0,05$, označené - vylúčené z analýzy

Po evalvácii metriky správnosti dosiahli najlepšie hodnoty modely LightGBM (0,847), viacvrstvový perceptrón (0,845) a XGBoost (0,842), ktoré tvoria homogénnu skupinu bez štatisticky významných rozdielov. Po nich nasledovali modely náhodného lesa (0,827) a KNN (0,819), ktoré dosiahli o trochu nižší, ale stále porovnateľný výkon. Na poslednom mieste bol rozhodovací strom (DT), ktorý dosiahol najnižšiu hodnotu správnosti (0,740) a nepatril medzi najvýkonnejšie modely.

Podobné výsledky sme pozorovali aj pri presnosti, kde najvyššiu hodnotu dosiahol model LightGBM (0,876), ktorý tvoril samostatnú skupinu. Nasledovali

[12] Zdroj Tabuľka 5: autor

KNN (0,835), XGBoost (0,834) a náhodný les (0,826), ktoré spolu tvorili homogénnu skupinu štatisticky porovnateľných modelov. Viacvrstvový perceptrón (0,801) dosiahol o čosi nižší výkon, a rozhodovací strom (0,597) výrazne zaostával v porovnaní s ostatnými modelmi.

Pri citlivosti dosiahol najvyššiu hodnotu viacvrstvový perceptrón (0,690), ako samostatná skupina najvýkonnejších modelov. Nasledovali XGBoost (0,647) a LightGBM (0,642) ktoré spolu vytvorili homogénnu skupinu. Náhodný les (0,642) sa nachádzal v samostatnej skupine, a najnižšie hodnoty citlivosti boli zaznamenané pri KNN (0,580) a rozhodovacom strome (0,593).

Pri metrike F1 skóre sa ako najvýkonnejší opäť ukázal viacvrstvový perceptrón (0,732), ktorý bol v samostatnej skupine. Po ňom nasledovali LightGBM (0,710) a XGBoost (0,708), ktoré vytvárajú homogénnu skupinu bez štatisticky významných rozdielov. Náhodný les (0,673) dosahoval nižší výkon, zatiaľ čo KNN (0,643) a rozhodovací strom (0,594) patrili medzi najmenej výkonné prístupy.

Globálne testy pre opakované merania potvrdili existenciu štatisticky významných rozdielov medzi modelmi ($df1 = 6$, $df2 = 54$, $F > 43{,}233$, $p < 0{,}001$). Analýza ukázala, že medzi najvýkonnejšie prístupy patrili viacvrstvový perceptrón, LightGBM a XGBoost, ktoré dosiahli vysoký výkon naprieč viacerými metrikami. Naopak, jednoduchšie modely rozhodovacieho stromu a KNN mali nižšiu výkonnosť a ani v jednom prípade nepatrili do skupiny najlepších modelov. Z tohto dôvodu boli modely rozhodovacieho stromu a KNN, spolu s náhodným lesom, ktorý nemal konzistentné výsledky, vylúčené z ďalšej analýzy.

4.4.2 Vyhodnotenie jazykových modelov

Na predchádzajúcu analýzu klasických modelov strojového učenia nadväzuje vyhodnotenie jazykových modelov, ktoré predstavujú modernejší prístup založený na hlbokom učení. Cieľom tejto kapitoly je porovnať ich výkonnosť a robustnosť na základe rovnakých evalvačných metrík ako pri modeloch strojového učenia.

Z analýzy dát v Prílohe C vyplýva, že výsledky jazykových modelov boli symetrické, pričom priemer ako miera strednej hodnoty mala zmysel. Zároveň môžeme v Tabuľke 6 pozorovať, že medzi modelmi existujú rozdiely vo výkonnosti pri jednotlivých metrikách.

Tabuľka 6 Prehľad výkonnosti jazykových modelov[13]

Model	Reprezentácia	Správnosť (mean)	Presnosť (mean)	Citlivosť (mean)	F1 (mean)
slovakBERT	BPE	0,820	0,753	0,679	0,706
XLM-RoBERTa	unigram	0,745	0,434	0,440	0,417
skBLM	BPE	0,802	0,758	0,578	0,628

Najvyššiu priemernú výkonnosť dosiahol model SlovakBERT, ktorý bol v porovnaní s ostatnými modelmi lepší vo všetkých sledovaných metrikách, ale najmä v oblasti presnosti a F1 skóre. Model skBLM dosiahol mierne nižšie výsledky, zatiaľ čo model XLM-RoBERTa mal nižšiu výkonnosť pri všetkých sledovaných metrikách.

Normalita rozdelenia bola overená pomocou Kolmogorov-Smirnovovho testu (Tabuľka 7), pričom vo všetkých prípadoch nebola zamietnutá nulová hypotéza o normálnom rozdelení dát ($p > 0,20$), čo umožňuje použitie parametrických štatistických metód.

Tabuľka 7 Testy normality pre jazykové modely[14]

	N	**max *D***	**K-S *p***
slovakbert_BPE_accuracy	10	0,226	$p > 0,20$
slovakbert_BPE_precision	10	0,238	$p > 0,20$
slovakbert_BPE_recall	10	0,198	$p > 0,20$
slovakbert_BPE_f1	10	0,209	$p > 0,20$
xlmroberta_unigram_accuracy	10	0,199	$p > 0,20$
xlmroberta_unigram_precision	10	0,248	$p > 0,20$
xlmroberta_unigram_recall	10	0,254	$p > 0,20$
xlmroberta_unigram_f1	10	0,217	$p > 0,20$
skblm_BPE_accuracy	10	0,139	$p > 0,20$
skblm_BPE_precision	10	0,237	$p > 0,20$
skblm_BPE_recall	10	0,229	$p > 0,20$
skblm_BPE_f1	10	0,270	$p > 0,20$

V prípade metrík správnosti a citlivosti nebol porušený predpoklad sféricity kovariančnej matice ($p > 0,05$), zatiaľ čo pri metrikách presnosti a F1 skóre bol tento predpoklad porušený ($p < 0,05$). Z tohto dôvodu boli na testovanie globálnych nulových

[13] Zdroj Tabuľka 6: autor
[14] Zdroj Tabuľka 7: autor

hypotéz použité neupravené testy pre opakované merania pre metriky správnosti a citlivosti a upravené testy pre metriky presnosti a F1 skóre.

Na základe testov pre opakované merania boli zamietnuté globálne nulové hypotézy pre všetky sledované metriky (accuracy/recall: $df1 = 6$, $df2 = 54$, $F > 22{,}124$, $p < 0{,}001$; precision/f1: $\varepsilon = 0{,}5$, adj. $df1 = 1$, adj. $df2 = 9$, $p < 0{,}001$), čo potvrdzuje existenciu štatisticky významných rozdielov medzi skúmanými modelmi.

Viacúrovňová analýza v Tabuľke 8 upresnila vzťahy medzi jednotlivými modelmi z hľadiska štatistickej významnosti rozdielov. Ukázalo sa, že modely SlovakBERT a skBLM dosahujú v niektorých metrikách štatisticky porovnateľný výkon, keďže medzi nimi neboli identifikované významné rozdiely ($p > 0{,}05$). Naopak, model XLM-RoBERTa sa vo väčšine prípadov nachádzal mimo homogénnych skupín najvýkonnejších modelov, čo potvrdzuje jeho štatisticky nižší výkon.

Tabuľka 8 Viacúrovňové testy pre LM: ***a)*** *správnosť,* ***b)*** *presnosť,* ***c)*** *citlivosť,* ***d)*** *f1 skóre*[15]

Správnosť	**Mean**	**1**	**2**	
xlmroberta_unigram_accuracy	0,745		****	
skblm_BPE_accuracy	0,802	****		
slovakbert_BPE_accuracy	0,820	****		
Presnosť	**Mean**	**1**	**2**	
xlmroberta_unigram_precision	0,434		****	
slovakbert_BPE_precision	0,753	****		
skblm_BPE_precision	0,758	****		
Citlivosť	**Mean**	**1**	**2**	**3**
xlmroberta_unigram_recall	0,440	****		
skblm_BPE_recall	0,578		****	
slovakbert_BPE_recall	0,679			****
F1	**Mean**	**1**	**2**	
xlmroberta_unigram_f1	0,417		****	
skblm_BPE_f1	0,628	****		
slovakbert_BPE_f1	0,706	****		

Poznámka: **** - $p > 0.05$, označené - vylúčené z analýzy

Na základe získaných výsledkov bol model XLM-RoBERTa vyradený z ďalšej analýzy, keďže pri žiadnej z hodnotených metrík nepatril medzi najlepšie metódy.

[15] Zdroj Tabuľka 8: autor

4.4.3 Porovnanie najvýkonnejších modelov

Porovnanie najvýkonnejších modelov predstavuje záverečný krok štatistickej analýzy, v ktorom cieľom bolo identifikovať najefektívnejšie prístupy naprieč všetkými skúmanými reprezentáciami a typmi modelov. Do analýzy boli zahrnuté iba tie modely, ktoré v predchádzajúcich krokoch dosiahli stabilne vysoký výkon a neboli vylúčené pri viacúrovňovom testovaní.

Keďže sa potvrdilo porušenie predpokladu sféricity, na testovanie globálnych hypotéz boli použité upravené testy pre opakované merania. Na ich základe bola zamietnutá nulová hypotéza o neexistencii štatisticky významných rozdielov vo výkone medzi najvýkonnejšími modelmi (accuracy/precision/recall/f1: $\varepsilon = 0{,}077$, adj. $df1 = 1$, adj. $df2 = 9$, $p < 0{,}001$). Následné viacúrovňové porovnanie, realizované pomocou Duncanovho testu, odhalilo viaceré homogénne skupiny modelov bez štatisticky významných rozdielov ($p > 0{,}05$), ako aj dvojice, kde bol rozdiel vo výkone preukázateľný ($p < 0{,}05$).

Z výsledkov, ktoré sa nachádzajú v Prílohe D, vyplýva, že medzi najvýkonnejšie prístupy patrili najmä subsymbolické modely využívajúce textovú reprezentáciu typu TF-IDF, a to konkrétne logistická regresia, SVM, Ridge a SGD, ktoré dosiahli vysoké hodnoty pri všetkých metrikách a často sa nachádzali v rovnakých homogénnych skupinách. Dobré výsledky sme zaznamenali aj pri modeloch kombinujúcich TF-IDF so znížením dimenzie pomocou SVD, najmä viacvrstvový perceptrón, XGBoost a LightGBM. Avšak ich výhodou bola predovšetkým schopnosť efektívne využiť zredukovanú dimenziu vstupov, nie ich vyšší výkon oproti iným modelom.

Jazykové modely SlovakBERT a skBLM dosiahli stredne vysoký výkon, pričom SlovakBERT sa v niektorých ch zaradil do rovnakej homogénnej skupiny ako slabšie z najlepších modelov strojového učenia. V porovnaní s najlepšími subsymbolickými metódami však tieto modely nedosiahli štatisticky vyšší výkon, skôr dopĺňali strednú časť výkonového spektra.

Celkovo sa ako najspoľahlivejšie a najvýkonnejšie riešenie javí kombinácia TF-IDF reprezentácie so subsymbolickými lineárnymi modelmi (logistická regresia, SVM, SGD, Ridge), ktoré dosiahli relatívne vysoký výkon naprieč všetkými metrikami a poskytujú najlepší pomer medzi presnosťou, citlivosťou a F1 skóre.

4.5 VYHODNOTENIE VÝSKUMNÝCH HYPOTÉZ

V tejto časti sú vyhodnotené hypotézy stanovené na základe teoretických predpokladov a cieľov práce.

HA: Symbolické metódy dosiahnu najnižšiu výkonnosť pri klasifikácii sentimentu.

Výsledky potvrdili, že symbolické modely dosiahli vo všetkých reprezentáciách dlhodobo najnižšiu výkonnosť. Ani pri jednej metrike nedosiahli hodnoty porovnateľné s najlepšími prístupmi a vo viacúrovňových testoch sa nikdy neobjavili v skupinách najvýkonnejších modelov. Horšie dosiahnuté výsledky súvisia najmä s tým, že modely ako Naive Bayes, KNN a rozhodovacie stromy pracujú na báze pevne stanovených pravidiel, a nedokážu dostatočne zachytiť kontext a variabilitu jazyka. Zároveň sú citlivé aj na nevyvážené triedy, čo vedie k slabšiemu zachyteniu jemných rozdielov medzi triedami v porovnaní s inými modelmi. Na základe týchto zistení možno hypotézu HA považovať za potvrdenú.

HB: Skupinové (ensemble) modely dosiahnu najvyššiu výkonnosť.

Skupinové modely dosiahli pri reprezentácii TF-IDF so znížením dimenzie pomocou SVD pomerne dobrý výkon a pri niektorých metrikách, najmä pri správnosti a F1 skóre, patrili medzi lepšie prístupy. V porovnaní s ostatnými modelmi však netvorili najvýkonnejšiu skupinu naprieč všetkými metrikami ani reprezentáciami. Ich výkon nebol vyšší ako výkon subsymbolických modelov, ktoré pri reprezentácii BOW, ako aj TF-IDF opakovane dosahovali najvyššie hodnoty. Tento výsledok možno vysvetliť tým, že skupinové prístupy sú vo všeobecnosti veľmi účinné pri zložitých nelineárnych vzťahov, avšak pri nevyvážených triedach nemusia dosiahnuť lepšie výsledky ako lineárne modely so správne nastavenou regularizáciou, ktoré sa v tomto experimente ukázali ako vhodnejšie. Hypotéza HB, ktorá predpokladala najvyššiu výkonnosť skupinových modelov, preto nebola potvrdená.

ZÁVER

Cieľom diplomovej práce bolo porovnať vybrané modely strojového učenia a jazykové modely pri klasifikácii sentimentu (pozitívny, neutrálny, negatívny) v textoch z bankového sektora v slovenskom jazyku a identifikovať najvhodnejší prístup pre jazyk s obmedzenými zdrojmi. Tento cieľ bol splnený prostredníctvom návrhu a realizácie experimentov, ktoré boli v súlade s metodikou CRISP-DM, od analýzy a prípravy dát, až po vyhodnotenie modelovaných experimentov.

Výsledky experimentov ukázali, že najvyššiu výkonnosť dosahujú subsymbolické modely založené na TF-IDF reprezentácii, najmä logistická regresia, SVM, Ridge a SGD klasifikátor. Tieto modely dosahovali vyvážený výkon naprieč všetkými hodnotenými metrikami. Aplikácia redukcie dimenzie pomocou SVD na niektoré modely výkon ďalej podporila alebo udržala na porovnateľnej úrovni. Jazykové modely, najmä SlovakBERT, dosiahli konkurencieschopné výsledky, avšak nemali prevahu nad najlepšími tradičnými prístupmi strojového učenia.

Výsledky zároveň potvrdzujú, že pri menších alebo nerovnomerne zastúpených datasetoch môžu byť jednoduchšie modely strojového učenia založení na TF-IDF veľmi efektívnou voľbou. Skupinové modely naopak v tomto experimente nedosiahli svoj potenciál a neprekonali najlepšie subsymbolické klasifikátory.

Medzi hlavné obmedzenia práce patrí najmä veľkosť datasetov a ich nerovnomerné rozdelenie, ktoré komplikovalo klasifikáciu menej zastúpených tried. Priestor na ďalší výskum predstavuje rozšírenie datasetu, využitie pokročilejších techník na vyváženie tried, alebo experimentovanie s inými modelmi alebo prístupmi, najmä v oblasti hlbokého učenia.

Hlavným prínosom práce je systematické porovnanie rôznych prístupov ku klasifikácii sentimentu, ktorých výsledky poskytujú ucelený pohľad na jednotlivé prístupy a ich výkon pri klasifikácii textov v slovenskom bankovom sektore. Zistenia môžu slúžiť ako základ pre ďalší výskum aj pre praktické využitie v oblasti bankovníctva, najmä pri automatizovanom spracovaní textov, podpore hodnotenia rizík a rozhodovacích procesov.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

AGGARWAL, Charu C., 2018. An Introduction to Neural Networks. V: Charu C. AGGARWAL *Neural Networks and Deep Learning* [online]. Cham: Springer International Publishing, s. 1–52 [cit. 5.11.2025]. ISBN 978-3-319-94462-3. Dostupné na: doi:10.1007/978-3-319-94463-0_1

AHMAD, Munir, Shabib AFTAB, Syed Shah MUHAMMAD a Sarfraz AHMAD, 2017. Machine Learning Techniques for Sentiment Analysis: A Review. 2017, roč. 8, č. 3.

ALLOGHANI, Mohamed, Dhiya AL-JUMEILY, Jamila MUSTAFINA, Abir HUSSAIN a Ahmed J. ALJAAF, 2020. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. V: Michael W. BERRY, Azlinah MOHAMED a Bee Wah YAP, ed. *Supervised and Unsupervised Learning for Data Science* [online]. Cham: Springer International Publishing, Unsupervised and Semi-Supervised Learning, s. 3–21 [cit. 14.10.2025]. ISBN 978-3-030-22474-5. Dostupné na: doi:10.1007/978-3-030-22475-2_1

ALSLAITY, Alaa a Rita ORJI, 2022. Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. *Behaviour & Information Technology* [online]. 2022, roč. 43, č. 1, s. 139–164. ISSN 0144-929X, 1362-3001. Dostupné na: doi:10.1080/0144929X.2022.2156387

BANAFA, Ahmed, 2025. *Artificial intelligence in action: real-world applications and innovations*. United States: River Publishers. ISBN 978-87-7004-619-0.

BANSAL, Malti, Apoorva GOYAL a Apoorva CHOUDHARY, 2022. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal* [online]. 2022, roč. 3, s. 100071. ISSN 2772-6622. Dostupné na: doi:10.1016/j.dajour.2022.100071

CERVANTES, Jair, Farid GARCIA-LAMONT, Lisbeth RODRÍGUEZ-MAZAHUA a Asdrubal LOPEZ, 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* [online]. 2020, roč. 408, s. 189–215. ISSN 09252312. Dostupné na: doi:10.1016/j.neucom.2019.10.118

CORREA, Ricardo, Keshav GARUD, Juan M. LONDONO a Nathan MISLANG, 2017a. Sentiment in Central Bank's Financial Stability Reports. *International Finance Discussion Papers* [online]. 2017, roč. 2017.0, č. 1203, s. 1–46. ISSN 1073-2500. Dostupné na: doi:10.17016/ifdp.2017.1203

CORREA, Ricardo, Keshav GARUD, Juan M LONDONO a Nathan MISLANG, 2021. Sentiment in Central Banks' Financial Stability Reports. *Review of Finance* [online]. 2021, roč. 25, č. 1, s. 85–120. ISSN 1572-3097, 1573-692X. Dostupné na: doi:10.1093/rof/rfaa014

CORREA, Ricardo, Keshav GARUD, Juan-Miguel LONDONO-YARCE a Nathan MISLANG, 2017b. Constructing a Dictionary for Financial Stability [online]. 2017 [cit.

29.10.2025]. Dostupné na: https://www.federalreserve.gov/econres/notes/ifdp-notes/constructing-a-dictionary-for-financial-stability-20170623.htm

COVER, T. a P. HART, 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* [online]. 1967, roč. 13, č. 1, s. 21–27. ISSN 0018-9448, 1557-9654. Dostupné na: doi:10.1109/TIT.1967.1053964

DAVE, Kushal, Steve LAWRENCE a David M. PENNOCK, 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. V: *Proceedings of the 12th international conference on World Wide Web* [online]. New York, NY, USA: Association for Computing Machinery, s. 519–528 [cit. 29.9.2025]. WWW '03. ISBN 978-1-58113-680-7. Dostupné na: doi:10.1145/775152.775226

DU, Kelvin, Frank XING, Rui MAO a Erik CAMBRIA, 2023. FinSenticNet: A Concept-Level Lexicon for Financial Sentiment Analysis. V: *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*: *2023 IEEE Symposium Series on Computational Intelligence (SSCI)* [online]. Mexico City, Mexico: IEEE, s. 109–114 [cit. 29.10.2025]. ISBN 978-1-6654-3065-4. Dostupné na: doi:10.1109/SSCI52147.2023.10371970

HAJEK, Petr a Michal MUNK, 2023. Speech emotion recognition and text sentiment analysis for financial distress prediction. *Neural Computing and Applications* [online]. 2023, roč. 35, č. 29, s. 21463–21477. ISSN 0941-0643, 1433-3058. Dostupné na: doi:10.1007/s00521-023-08470-8

HARTMANN, Jochen, Mark HEITMANN, Christian SIEBERT a Christina SCHAMP, 2023. More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing* [online]. 2023, roč. 40, č. 1, s. 75–87. ISSN 01678116. Dostupné na: doi:10.1016/j.ijresmar.2022.05.005

HUANG, Bo, Xiao YAO, Yinqing LUO a Jing LI, 2023. Improving financial distress prediction using textual sentiment of annual reports. *Annals of Operations Research* [online]. 2023, roč. 330, č. 1–2, s. 457–484. ISSN 0254-5330, 1572-9338. Dostupné na: doi:10.1007/s10479-022-04633-3

ILKOU, Eleni a Maria KOUTRAKI, 2020. Symbolic Vs Sub-symbolic AI Methods: Friends or Enemies? 2020.

JANIESCH, Christian, Patrick ZSCHECH a Kai HEINRICH, 2021. Machine learning and deep learning. *Electronic Markets* [online]. 2021, roč. 31, č. 3, s. 685–695. ISSN 1019-6781, 1422-8890. Dostupné na: doi:10.1007/s12525-021-00475-2

JOSHI, Srivatsa, Vishwas KUMAR, Vishaka VENKATARAMANAN a Kaliprasad C S, 2023. A Review on Neural Networks and its Applications. *Journal of Computer Technology & Applications* [online]. 2023, roč. 14, s. 2023. Dostupné na: doi:10.37591/jocta.v14i2.1062

KANG, Seokho, 2021. k-Nearest Neighbor Learning with Graph Neural Networks. *Mathematics* [online]. 2021, roč. 9, č. 8, s. 830. ISSN 2227-7390. Dostupné na: doi:10.3390/math9080830

KELEBERCOVÁ, Lívia, Michal MUNK, Anna PILKOVÁ a Karol BLIŽNÁK, 2025. Sentiment shifts: sentiment analysis of annual reports during key economic and political periods. *International Journal of Services Technology and Management* [online]. 2025, roč. 30, č. 2/3, s. 166–181. ISSN 1460-6720, 1741-525X. Dostupné na: doi:10.1504/IJSTM.2025.149171

KELEBERCOVÁ, Lívia a Natáliia Časnochova ZOZUK, 2023. Sentiment classification of annual reports in Slovak language using symbolic, subsymbolic and statistic methods. *Procedia Computer Science* [online]. 2023, roč. 225, s. 3508–3516. ISSN 18770509. Dostupné na: doi:10.1016/j.procs.2023.10.346

KHOO, Christopher Sg a Sathik Basha JOHNKHAN, 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science* [online]. 2018, roč. 44, č. 4, s. 491–511. ISSN 0165-5515, 1741-6485. Dostupné na: doi:10.1177/0165551517703514

LOUGHRAN, Tim a Bill MCDONALD, 2011. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. 2011.

MAHMOOD, Toqeer, Saba NASEEM, Rehan ASHRAF, Muhammad ASIF, Muhammad UMAIR a Mohsin SHAH, 2023. Recognizing factors effecting the use of mobile banking apps through sentiment and thematic analysis on user reviews. *Neural Computing and Applications* [online]. 2023, roč. 35, č. 27, s. 19885–19897. ISSN 0941-0643, 1433-3058. Dostupné na: doi:10.1007/s00521-023-08827-z

MAO, Yanying, Qun LIU a Yu ZHANG, 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences* [online]. 2024, roč. 36, č. 4, s. 102048. ISSN 1319-1578. Dostupné na: doi:10.1016/j.jksuci.2024.102048

MIENYE, Ibomoiye Domor a Nobert JERE, 2024. A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access* [online]. 2024, roč. 12, s. 86716–86727. ISSN 2169-3536. Dostupné na: doi:10.1109/ACCESS.2024.3416838

MIENYE, Ibomoiye Domor a Yanxia SUN, 2022. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access* [online]. 2022, roč. 10, s. 99129–99149. ISSN 2169-3536. Dostupné na: doi:10.1109/ACCESS.2022.3207287

MING, Ruixing, Osama MOHAMAD, Nisreen INNAB a Mohamed HANAFY, 2024. Bagging Vs. Boosting in Ensemble Machine Learning? An Integrated Application to Fraud Risk Analysis in the Insurance Sector. *Applied Artificial Intelligence* [online]. 2024, roč. 38, č. 1, s. 2355024. ISSN 0883-9514, 1087-6545. Dostupné na: doi:10.1080/08839514.2024.2355024

MOHANTY, Aniruddha a Ravindranath C. CHERUKURI, 2023. Sentiment Analysis on Banking Feedback and News Data using Synonyms and Antonyms. *International Journal of Advanced Computer Science and Applications* [online]. 2023, roč. 14, č. 12 [cit. 8.10.2025]. ISSN 21565570, 2158107X. Dostupné na: doi:10.14569/IJACSA.2023.0141294

NASUKAWA, Tetsuya a Jeonghee YI, 2003. Sentiment analysis: capturing favorability using natural language processing. V: *K-CAP03: International Conference on Knowledge Capture 2003*: *Proceedings of the 2nd international conference on Knowledge capture* [online]. Sanibel Island FL USA: ACM, s. 70–77 [cit. 25.10.2025]. ISBN 978-1-58113-583-1. Dostupné na: doi:10.1145/945645.945658

PANDEY, Darpan, Kamal NIWARIA a Bharti CHOURASIA, 2019. Machine Learning Algorithms: A Review. 2019, roč. 06, č. 02.

PATEL, Harsh a Purvi PRAJAPATI, 2018. Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering* [online]. 2018, roč. 6, s. 74–78. Dostupné na: doi:10.26438/ijcse/v6i10.7478

PILKOVÁ, Anna, Slavka ELLEY, Michal MUNK, Lívia KELEBERCOVÁ a Petra BLAŽEKOVÁ, 2025. Unlocking financial insights: decoding sentiment in annual reports to transform bank risk and performance metrics. *International Journal of Services Technology and Management* [online]. 2025, roč. 30, č. 2/3, s. 182–197. ISSN 1460-6720, 1741-525X. Dostupné na: doi:10.1504/IJSTM.2025.149170

SADIA, Azeema, Fariha KHAN a Fatima BASHIR, 2018. An Overview of Lexicon-Based Approach For Sentiment Analysis. 2018.

SAGI, Omer a Lior ROKACH, 2018. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery* [online]. 2018, roč. 8, č. 4, s. e1249. ISSN 1942-4787, 1942-4795. Dostupné na: doi:10.1002/widm.1249

SUYAL, Manish a Parul GOYAL, 2022. A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning. *International Journal of Engineering Trends and Technology* [online]. 2022, roč. 70, č. 7, s. 43–48. ISSN 22315381. Dostupné na: doi:10.14445/22315381/IJETT-V70I7P205

TAUNK, Kashvi, Sanjukta DE, Srishti VERMA a Aleena SWETAPADMA, 2019. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. V: *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*: *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* [online]. Madurai, India: IEEE, s. 1255–1260 [cit. 16.10.2025]. ISBN 978-1-5386-8113-8. Dostupné na: doi:10.1109/ICCS45141.2019.9065747

VERMA, Binita a Ramjeevan Singh THAKUR, 2018. Sentiment Analysis Using Lexicon and Machine Learning-Based Approaches: A Survey. V: Basant TIWARI, Vivek TIWARI, Kinkar Chandra DAS, Durgesh Kumar MISHRA a Jagdish C. BANSAL, ed. *Proceedings of International Conference on Recent Advancement on Computer and Communication* [online]. Singapore: Springer Singapore, Lecture Notes in Networks and Systems, s. 441–447 [cit. 12.10.2025]. ISBN 978-981-10-8197-2. Dostupné na: doi:10.1007/978-981-10-8198-9_46

XU, Qianwen Ariel, Victor CHANG a Chrisina JAYNE, 2022. A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal* [online]. 2022, roč. 3, s. 100073. ISSN 2772-6622. Dostupné na: doi:10.1016/j.dajour.2022.100073

YUSUF, Ali Sahabi, Ab Razak CHE HUSSIN a Abdelsalam H. BUSALIM, 2018. Influence of e-WOM engagement on consumer purchase intention in social commerce. *Journal of Services Marketing* [online]. 2018, roč. 32, č. 4, s. 493–504. ISSN 0887-6045. Dostupné na: doi:10.1108/JSM-01-2017-0031

ZAIDI, Abdelhamid a Asamh Saleh M. AL LUHAYB, 2023. Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression. *Mathematical Problems in Engineering* [online]. 2023, roč. 2023, č. 1, s. 5525675. ISSN 1024-123X, 1563-5147. Dostupné na: doi:10.1155/2023/5525675

ZHOU, Zhi-Hua, 2021. *Machine learning*. Prel. Shaowu LIU. Singapore: Springer. ISBN 978-981-15-1966-6.

ZHOU, Zhi-Hua, 2025. *Ensemble Methods: Foundations and Algorithms* [online]. 2. vyd. Boca Raton: Chapman and Hall/CRC [cit. 26.10.2025]. ISBN 978-1-003-58777-4. Dostupné na: doi:10.1201/9781003587774

ZOU, Xiaonan, Yong HU, Zhewen TIAN a Kaiyuan SHEN, 2019. Logistic Regression Model Optimization and Case Analysis. V: *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*: *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)* [online]. Dalian, China: IEEE, s. 135–139 [cit. 4.11.2025]. ISBN 978-1-7281-3299-0. Dostupné na: doi:10.1109/ICCSNT47585.2019.8962457

ZOZNAM PRÍLOH

PRÍLOHA A - TABUĽKA POPISNÝCH ŠTATISTÍK PRE MODELY STROJOVÉHO UČENIA

Príloha A

Model	N	Mean	Median	Min	Max	Lower Q	Upper Q	Range	IQR	CV
LogReg_TFIDF_acc	10	0,878	0,879	0,843	0,901	0,874	0,889	0,058	0,014	1,889
LogReg_TFIDF_pre	10	0,841	0,846	0,776	0,882	0,821	0,865	0,105	0,044	3,755
LogReg_TFIDF_rec	10	0,763	0,756	0,721	0,815	0,738	0,791	0,094	0,053	4,351
LogReg_TFIDF_f1	10	0,796	0,797	0,746	0,844	0,771	0,817	0,097	0,046	3,884
LogReg_BOW_acc	10	0,860	0,857	0,845	0,882	0,851	0,867	0,037	0,017	1,386
LogReg_BOW_pre	10	0,779	0,783	0,746	0,817	0,761	0,794	0,071	0,032	2,852
LogReg_BOW_rec	10	0,804	0,806	0,766	0,836	0,792	0,814	0,070	0,022	2,462
LogReg_BOW_f1	10	0,790	0,786	0,770	0,826	0,772	0,801	0,056	0,028	2,238
SVM_TFIDF_acc	10	0,857	0,857	0,839	0,872	0,848	0,867	0,033	0,019	1,447
SVM_TFIDF_pre	10	0,786	0,787	0,747	0,830	0,783	0,800	0,084	0,016	3,132
SVM_TFIDF_rec	10	0,763	0,762	0,718	0,811	0,736	0,782	0,093	0,045	4,034
SVM_TFIDF_f1	10	0,773	0,776	0,736	0,804	0,749	0,792	0,068	0,043	3,138
SVM_BOW_acc	10	0,864	0,862	0,848	0,882	0,855	0,872	0,034	0,017	1,296
SVM_BOW_pre	10	0,800	0,799	0,773	0,832	0,783	0,817	0,059	0,034	2,716
SVM_BOW_rec	10	0,768	0,762	0,735	0,815	0,749	0,780	0,080	0,031	3,400
SVM_BOW_f1	10	0,782	0,778	0,756	0,818	0,769	0,790	0,062	0,021	2,714
SGD_TFIDF_acc	10	0,862	0,860	0,843	0,880	0,860	0,872	0,036	0,012	1,333
SGD_TFIDF_pre	10	0,794	0,797	0,763	0,817	0,779	0,813	0,054	0,034	2,466
SGD_TFIDF_rec	10	0,769	0,768	0,726	0,825	0,744	0,788	0,098	0,045	3,870
SGD_TFIDF_f1	10	0,780	0,779	0,742	0,820	0,771	0,792	0,077	0,021	2,830
SGD_BOW_accuracy	10	0,865	0,864	0,843	0,884	0,858	0,875	0,041	0,017	1,554
SGD_BOW_precision	10	0,799	0,803	0,767	0,838	0,780	0,811	0,071	0,031	2,988
SGD_BOW_recall	10	0,786	0,783	0,769	0,827	0,770	0,791	0,058	0,020	2,267
SGD_BOW_f1	10	0,792	0,788	0,768	0,827	0,783	0,806	0,059	0,023	2,372
Ridge_TFIDF_acc	10	0,867	0,865	0,853	0,886	0,862	0,870	0,033	0,008	0,985
Ridge_TFIDF_pre	10	0,832	0,825	0,806	0,882	0,815	0,847	0,077	0,031	2,825
Ridge_TFIDF_rec	10	0,741	0,737	0,723	0,775	0,730	0,749	0,052	0,019	2,205
Ridge_TFIDF_f1	10	0,778	0,776	0,762	0,806	0,771	0,783	0,044	0,012	1,713
Ridge_BOW_acc	10	0,869	0,870	0,855	0,886	0,863	0,874	0,031	0,012	1,103
Ridge_BOW_pre	10	0,829	0,828	0,779	0,860	0,820	0,851	0,081	0,031	2,819
Ridge_BOW_rec	10	0,750	0,754	0,717	0,779	0,736	0,766	0,063	0,030	2,802
Ridge_BOW_f1	10	0,783	0,783	0,746	0,808	0,780	0,795	0,062	0,015	2,206
PA_TFIDF_acc	10	0,848	0,847	0,829	0,872	0,841	0,850	0,043	0,010	1,685

Model	N	Mean	Median	Min	Max	Lower Q	Upper Q	Range	IQR	CV
PA_TFIDF_pre	10	0,773	0,770	0,740	0,820	0,752	0,790	0,080	0,038	3,511
PA_TFIDF_rec	10	0,749	0,743	0,717	0,805	0,727	0,758	0,087	0,031	3,974
PA_TFIDF_f1	10	0,759	0,754	0,729	0,806	0,740	0,763	0,077	0,023	3,503
PA_BOW_acc	10	0,867	0,867	0,853	0,894	0,855	0,872	0,041	0,017	1,443
PA_BOW_pre	10	0,845	0,844	0,814	0,885	0,825	0,861	0,072	0,035	2,759
PA_BOW_rec	10	0,719	0,706	0,691	0,777	0,694	0,753	0,087	0,059	4,420
PA_BOW_f1	10	0,767	0,753	0,741	0,822	0,749	0,783	0,081	0,035	3,433
MNB_TFIDF_acc	10	0,803	0,801	0,776	0,843	0,797	0,809	0,067	0,012	2,259
MNB_TFIDF_pre	10	0,701	0,698	0,658	0,757	0,684	0,718	0,099	0,034	4,006
MNB_TFIDF_rec	10	0,680	0,674	0,649	0,763	0,660	0,688	0,114	0,028	4,699
MNB_TFIDF_f1	10	0,689	0,685	0,653	0,760	0,678	0,688	0,106	0,010	3,942
MNB_BOW_acc	10	0,802	0,801	0,780	0,829	0,783	0,814	0,048	0,031	2,282
MNB_BOW_pre	10	0,708	0,694	0,668	0,770	0,678	0,747	0,102	0,069	5,412
MNB_BOW_rec	10	0,659	0,657	0,633	0,707	0,642	0,675	0,074	0,033	3,459
MNB_BOW_f1	10	0,680	0,669	0,647	0,729	0,658	0,704	0,081	0,046	4,111
CNB_TFIDF_acc	10	0,771	0,768	0,745	0,805	0,763	0,778	0,060	0,014	2,324
CNB_TFIDF_pre	10	0,652	0,646	0,619	0,702	0,638	0,666	0,083	0,028	3,607
CNB_TFIDF_rec	10	0,710	0,716	0,660	0,769	0,679	0,726	0,109	0,047	4,487
CNB_TFIDF_f1	10	0,675	0,670	0,642	0,730	0,658	0,690	0,088	0,032	3,863
CNB_BOW_acc	10	0,781	0,778	0,749	0,807	0,773	0,790	0,058	0,017	2,075
CNB_BOW_pre	10	0,665	0,661	0,625	0,708	0,651	0,675	0,083	0,024	3,560
CNB_BOW_rec	10	0,693	0,686	0,667	0,756	0,671	0,697	0,089	0,026	3,975
CNB_BOW_f1	10	0,677	0,672	0,643	0,729	0,666	0,677	0,086	0,011	3,594
RF_TFIDF_SVD_acc	10	0,827	0,824	0,817	0,841	0,821	0,838	0,024	0,017	1,100
RF_TFIDF_SVD_pre	10	0,826	0,828	0,780	0,868	0,816	0,841	0,088	0,026	2,903
RF_TFIDF_SVD_rec	10	0,611	0,611	0,572	0,641	0,594	0,630	0,069	0,035	3,530
RF_TFIDF_SVD_f1	10	0,673	0,669	0,642	0,704	0,661	0,692	0,063	0,031	2,978
DT_TFIDF_SVD_acc	10	0,740	0,743	0,706	0,783	0,725	0,754	0,077	0,029	3,199
DT_TFIDF_SVD_pre	10	0,597	0,601	0,543	0,663	0,574	0,621	0,120	0,047	5,996
DT_TFIDF_SVD_rec	10	0,593	0,587	0,539	0,676	0,579	0,606	0,137	0,027	6,334
DT_TFIDF_SVD_f1	10	0,594	0,597	0,541	0,669	0,578	0,603	0,128	0,025	5,972
KNN_TFIDF_SVD_acc	10	0,819	0,812	0,802	0,843	0,809	0,833	0,041	0,024	1,760
KNN_TFIDF_SVD_pre	10	0,835	0,834	0,731	0,908	0,807	0,866	0,176	0,058	6,282
KNN_TFIDF_SVD_rec	10	0,580	0,576	0,528	0,632	0,568	0,597	0,104	0,028	4,748

Model	N	Mean	Median	Min	Max	Lower Q	Upper Q	Range	IQR	CV
KNN_TFIDF_SVD_f1	10	0,643	0,638	0,590	0,706	0,627	0,668	0,116	0,041	5,079
MLP_TFIDF_SVD_acc	10	0,845	0,845	0,824	0,875	0,836	0,848	0,051	0,012	1,568
MLP_TFIDF_SVD_pre	10	0,801	0,799	0,762	0,852	0,774	0,822	0,090	0,048	4,073
MLP_TFIDF_SVD_rec	10	0,690	0,683	0,659	0,748	0,672	0,698	0,089	0,026	4,024
MLP_TFIDF_SVD_f1	10	0,732	0,727	0,708	0,791	0,713	0,741	0,083	0,028	3,374
XGB_TFIDF_SVD_acc	10	0,842	0,838	0,826	0,880	0,836	0,846	0,053	0,010	1,771
XGB_TFIDF_SVD_pre	10	0,834	0,823	0,790	0,890	0,810	0,863	0,100	0,054	3,871
XGB_TFIDF_SVD_rec	10	0,647	0,636	0,624	0,723	0,635	0,655	0,099	0,020	4,484
XGB_TFIDF_SVD_f1	10	0,708	0,697	0,683	0,784	0,689	0,722	0,101	0,032	4,242
LGBM_TFIDF_SVD_acc	10	0,847	0,845	0,831	0,889	0,841	0,846	0,058	0,005	1,835
LGBM_TFIDF_SVD_pre	10	0,876	0,879	0,836	0,925	0,858	0,893	0,089	0,035	3,291
LGBM_TFIDF_SVD_rec	10	0,642	0,632	0,607	0,740	0,624	0,646	0,133	0,023	5,757
LGBM_TFIDF_SVD_f1	10	0,710	0,702	0,675	0,807	0,693	0,711	0,132	0,017	5,067

*IQR = Medzikvartilové rozpätie; CV = Variačný koeficient

PRÍLOHA B - TABUĽKA TESTOV NORMALITY PRE MODELY STROJOVÉHO UČENIA

Príloha B

	N	**max *D***	**K-S *p***
LogReg_TFIDF_accuracy	10	0,211	$p > 0,20$
LogReg_TFIDF_precision	10	0,154	$p > 0,20$
LogReg_TFIDF_recall	10	0,151	$p > 0,20$
LogReg_TFIDF_f1	10	0,115	$p > 0,20$
LogReg_BOW_accuracy	10	0,154	$p > 0,20$
LogReg_BOW_precision	10	0,184	$p > 0,20$
LogReg_BOW_recall	10	0,140	$p > 0,20$
LogReg_BOW_f1	10	0,153	$p > 0,20$
SVM_TFIDF_accuracy	10	0,196	$p > 0,20$
SVM_TFIDF_precision	10	0,250	$p > 0,20$
SVM_TFIDF_recall	10	0,103	$p > 0,20$
SVM_TFIDF_f1	10	0,135	$p > 0,20$
SVM_BOW_accuracy	10	0,144	$p > 0,20$
SVM_BOW_precision	10	0,175	$p > 0,20$
SVM_BOW_recall	10	0,185	$p > 0,20$
SVM_BOW_f1	10	0,167	$p > 0,20$
SGD_TFIDF_accuracy	10	0,220	$p > 0,20$
SGD_TFIDF_precision	10	0,136	$p > 0,20$
SGD_TFIDF_recall	10	0,106	$p > 0,20$
SGD_TFIDF_f1	10	0,143	$p > 0,20$
SGD_BOW_accuracy	10	0,118	$p > 0,20$
SGD_BOW_precision	10	0,115	$p > 0,20$
SGD_BOW_recall	10	0,196	$p > 0,20$
SGD_BOW_f1	10	0,179	$p > 0,20$
Ridge_TFIDF_accuracy	10	0,259	$p > 0,20$
Ridge_TFIDF_precision	10	0,245	$p > 0,20$
Ridge_TFIDF_recall	10	0,200	$p > 0,20$
Ridge_TFIDF_f1	10	0,201	$p > 0,20$
Ridge_BOW_accuracy	10	0,115	$p > 0,20$
Ridge_BOW_precision	10	0,175	$p > 0,20$
Ridge_BOW_recall	10	0,133	$p > 0,20$
Ridge_BOW_f1	10	0,231	$p > 0,20$
PA_TFIDF_accuracy	10	0,230	$p > 0,20$
PA_TFIDF_precision	10	0,130	$p > 0,20$
PA_TFIDF_recall	10	0,218	$p > 0,20$
PA_TFIDF_f1	10	0,241	$p > 0,20$
PA_BOW_accuracy	10	0,232	$p > 0,20$
PA_BOW_precision	10	0,140	$p > 0,20$
PA_BOW_recall	10	0,264	$p > 0,20$
PA_BOW_f1	10	0,287	$p > 0,20$
MNB_TFIDF_accuracy	10	0,215	$p > 0,20$
MNB_TFIDF_precision	10	0,191	$p > 0,20$
MNB_TFIDF_recall	10	0,274	$p > 0,20$

	N	**max *D***	**K-S *p***
MNB_TFIDF_f1	10	0,364	$p < 0,15$
MNB_BOW_accuracy	10	0,173	$p > 0,20$
MNB_BOW_precision	10	0,226	$p > 0,20$
MNB_BOW_recall	10	0,183	$p > 0,20$
MNB_BOW_f1	10	0,250	$p > 0,20$
CNB_TFIDF_accuracy	10	0,153	$p > 0,20$
CNB_TFIDF_precision	10	0,199	$p > 0,20$
CNB_TFIDF_recall	10	0,158	$p > 0,20$
CNB_TFIDF_f1	10	0,153	$p > 0,20$
CNB_BOW_accuracy	10	0,172	$p > 0,20$
CNB_BOW_precision	10	0,205	$p > 0,20$
CNB_BOW_recall	10	0,234	$p > 0,20$
CNB_BOW_f1	10	0,304	$p > 0,20$
RF_TFIDF_SVD_accuracy	10	0,334	$p < 0,20$
RF_TFIDF_SVD_precision	10	0,204	$p > 0,20$
RF_TFIDF_SVD_recall	10	0,108	$p > 0,20$
RF_TFIDF_SVD_f1	10	0,152	$p > 0,20$
DT_TFIDF_SVD_accuracy	10	0,139	$p > 0,20$
DT_TFIDF_SVD_precision	10	0,137	$p > 0,20$
DT_TFIDF_SVD_recall	10	0,177	$p > 0,20$
DT_TFIDF_SVD_f1	10	0,205	$p > 0,20$
KNN_TFIDF_SVD_accuracy	10	0,247	$p > 0,20$
KNN_TFIDF_SVD_precision	10	0,154	$p > 0,20$
KNN_TFIDF_SVD_recall	10	0,205	$p > 0,20$
KNN_TFIDF_SVD_f1	10	0,134	$p > 0,20$
MLP_TFIDF_SVD_accuracy	10	0,235	$p > 0,20$
MLP_TFIDF_SVD_precision	10	0,145	$p > 0,20$
MLP_TFIDF_SVD_recall	10	0,193	$p > 0,20$
MLP_TFIDF_SVD_f1	10	0,260	$p > 0,20$
XGB_TFIDF_SVD_accuracy	10	0,231	$p > 0,20$
XGB_TFIDF_SVD_precision	10	0,179	$p > 0,20$
XGB_TFIDF_SVD_recall	10	0,315	$p > 0,20$
XGB_TFIDF_SVD_f1	10	0,331	$p < 0,20$
LGBM_TFIDF_SVD_accuracy	10	0,385	$p < 0,10$
LGBM_TFIDF_SVD_precision	10	0,121	$p > 0,20$
LGBM_TFIDF_SVD_recall	10	0,304	$p > 0,20$
LGBM_TFIDF_SVD_f1	10	0,331	$p < 0,20$

PRÍLOHA C - TABUĽKA POPISNÝCH ŠTATISTÍK PRE JAZYKOVÉ MODELY

Príloha C

Model	N	Mean	Median	Min	Max	Lower Q	Upper Q	Range	IQR	CV
slovakbert_BPE_acc	10	0,820	0,811	0,790	0,872	0,800	0,838	0,082	0,038	3,188
slovakbert_BPE_pre	10	0,753	0,743	0,701	0,884	0,717	0,774	0,183	0,057	7,185
slovakbert_BPE_rec	10	0,679	0,677	0,620	0,742	0,642	0,712	0,122	0,070	6,377
slovakbert_BPE_f1	10	0,706	0,695	0,656	0,774	0,663	0,737	0,118	0,073	6,165
xlmroberta_unigram_acc	10	0,745	0,737	0,727	0,790	0,728	0,758	0,063	0,031	2,777
xlmroberta_unigram_pre	10	0,434	0,429	0,242	0,680	0,243	0,644	0,438	0,402	42,780
xlmroberta_unigram_rec	10	0,440	0,386	0,333	0,632	0,333	0,567	0,298	0,234	27,971
xlmroberta_unigram_f1	10	0,417	0,376	0,281	0,640	0,281	0,585	0,360	0,305	36,041
skblm_BPE_acc	10	0,802	0,802	0,780	0,821	0,797	0,810	0,041	0,013	1,576
skblm_BPE_pre	10	0,758	0,753	0,704	0,821	0,731	0,786	0,117	0,056	4,922
skblm_BPE_rec	10	0,578	0,590	0,520	0,606	0,555	0,598	0,087	0,044	4,872
skblm_BPE_f1	10	0,628	0,641	0,577	0,662	0,613	0,644	0,085	0,031	4,262

*IQR = Medzikvartilové rozpätie; CV = Variačný koeficient

PRÍLOHA D - TABUĽKA VIACSTUPŇOVÉHO TESTOVANIA NAJVÝKONNEJŠÍCH MODELOV

Príloha D

Správnosť	**Mean**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
skblm_BPE_accuracy	0,802	****						
slovakbert_BPE_accuracy	0,820		****					
XGB_TFIDF_SVD_accuracy	0,842			****				
MLP_TFIDF_SVD_accuracy	0,845			****				
LGBM_TFIDF_SVD_accuracy	0,847			****	****			
SVM_TFIDF_accuracy	0,857				****	****		
LogReg_BOW_accuracy	0,860					****	****	
SGD_TFIDF_accuracy	0,862					****	****	
SVM_BOW_accuracy	0,864					****	****	
SGD_BOW_accuracy	0,865					****	****	
PA_BOW_accuracy	0,867					****	****	
Ridge_TFIDF_accuracy	0,867					****	****	
Ridge_BOW_accuracy	0,869						****	****
LogReg_TFIDF_accuracy	0,878							****

Presnosť	**Mean**	**1**	**2**	**3**	**4**	**5**
slovakbert_BPE_precision	0,753	****				
skblm_BPE_precision	0,758	****	****			
LogReg_BOW_precision	0,779		****	****		
SVM_TFIDF_precision	0,786			****		
SGD_TFIDF_precision	0,794			****		
SGD_BOW_precision	0,799			****		
SVM_BOW_precision	0,800			****		
MLP_TFIDF_SVD_precision	0,801			****		
Ridge_BOW_precision	0,829				****	
Ridge_TFIDF_precision	0,832				****	
XGB_TFIDF_SVD_precision	0,834				****	
LogReg_TFIDF_precision	0,841				****	
PA_BOW_precision	0,845				****	
LGBM_TFIDF_SVD_precision	0,876					****

Citlivosť	**Mean**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
skblm_BPE_recall	0,578	****							
LGBM_TFIDF_SVD_recall	0,642		****						
XGB_TFIDF_SVD_recall	0,647		****						
slovakbert_BPE_recall	0,679			****					
MLP_TFIDF_SVD_recall	0,690			****					
PA_BOW_recall	0,719				****				
Ridge_TFIDF_recall	0,741					****			
Ridge_BOW_recall	0,750					****	****		
SVM_TFIDF_recall	0,763						****		
LogReg_TFIDF_recall	0,763						****		
SVM_BOW_recall	0,768						****	****	
SGD_TFIDF_recall	0,769						****	****	
SGD_BOW_recall	0,786							****	****
LogReg_BOW_recall	0,804								****

F1	**Mean**	**1**	**2**	**3**	**4**	**5**	**6**
skblm_BPE_f1	0,628	****					
slovakbert_BPE_f1	0,706		****				
XGB_TFIDF_SVD_f1	0,708		****				
LGBM_TFIDF_SVD_f1	0,710		****				
MLP_TFIDF_SVD_f1	0,732			****			
PA_BOW_f1	0,767				****		
SVM_TFIDF_f1	0,773				****	****	
Ridge_TFIDF_f1	0,778				****	****	****
SGD_TFIDF_f1	0,780				****	****	****
SVM_BOW_f1	0,782				****	****	****
Ridge_BOW_f1	0,783				****	****	****
LogReg_BOW_f1	0,790					****	****
SGD_BOW_f1	0,792					****	****
LogReg_TFIDF_f1	0,796						****

Poznámka: **** - $p > 0.05$