Univerzita Pardubice

Fakulta ekonomicko-správní

# Pilier 3: Vplyv jazykovej zložitosti textu na správanie sa stakeholderov na webových stránkach komerčných bánk

Mgr. Ľubomír Benko, Ph.D.

Habilitačná práca

2024

**Prehlásenie**

Prehlasujem, že som túto prácu vypracoval samostatne. Všetky literárne pramene a informácie, ktoré som v práci použil, sú uvedené v zozname bibliografických odkazov.

Bol som oboznámený s tým, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., o autorskom zákone, o právach súvisiacich s autorským zákonom a o zmene niektorých zákonov (autorský zákon), v znení neskorších predpisov, najmä so skutočnosťou, že Univerzita Pardubice má právo na uzatvorenie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona, a s tým, že ak dôjde k použitiu práce mnou alebo bude poskytnutá licencia o využití inému subjektu, je Univerzita Pardubice oprávnená odo mňa požadovať primeraný príspevok na úhradu nákladov, ktoré na vytvorenie diela vynaložila, a to podľa okolností až do ich skutočnej výšky.

Beriem na vedomie, že v súlade s § 47b zákona č. 111/1998 Sb., o vysokých školách a o zmene a doplnení ďalších zákonov (zákon o vysokých školách), v znení neskorších predpisov a smernicou Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávaní, zverejňováni a formální úpravu závěrečných prací, v znení neskorších dodatkov, bude práca zverejnená prostredníctvom Digitálnej knižnice Univerzity Pardubice.

V Nitre dňa 20. 3. 2024

Mgr. Ľubomír Benko, Ph.D.

**Poďakovanie**

Chcel by som sa poďakovať mojim kolegom prof. RNDr. Michalovi Munkovi, PhD. a prof. RNDr. Daši Munkovej, PhD. za odborné rady a konzultácie počas môjho odborného napredovania. Poďakovať by som sa chcel aj spoluautorom odborných článkov, ktorých nápady a odborné znalosti viedli k môjmu napredovaniu v odbore.

Rád by som sa poďakoval svojmu domovskému pracovisku Katedre informatiky, FPVaI UKF v Nitre za vytvorenie priaznivého pracovného prostredia a podmienok na odborný rast. Taktiež by som sa chcel poďakovať Fakulte ekonomicko-správní Univerzity Pardubice za možnosť predložiť habilitačnú prácu.

Obzvlášť sa chcem poďakovať mojej manželke Mgr. Lucii Benkovej, PhD. za podporu a trpezlivosť počas tvorby práce.

**Anotácia**

Predkladaná habilitačná práca sa zameriava na analýzu zložitosti a čitateľnosti dokumentov týkajúcich sa povinne zverejňovaných informácií Pilier 3 na webovom portáli komerčných bánk a na skúmanie ich vplyvu na správanie sa stakeholderov. Výsledky experimentov ukázali, že zverejňovanie informácií Pilier 3 nie je potrebné počas celého roka, ale optimálne iba na začiatku, v prvých týždňoch roka. Záujem stakeholderov sa neobmedzuje len na povinne zverejňované informácie, ale zaujíma ich aj širší kontext informácií o banke. Analýza obsahu webu je realizovaná prostredníctvom metód spracovania prirodzeného jazyka s dôrazom na skúmanie vplyvu zložitosti a čitateľnosti dokumentov na správanie sa stakeholderov. Navrhnutá metodika je založená na kombinácií dát o používaní, štruktúre a obsahu webu, čo viedlo k návrhu ukazovateľov preferencií používateľov na webových portáloch komerčných bánk. Preukázal sa súvis medzi ukazovateľmi preferencií používateľov a metrikami zložitosti alebo čitateľnosti textu. Metriky základných charakteristík čitateľnosti textu ako početnosti tokenov, viet alebo znakov, dosiahla najvyššiu mieru závislosti s preferenciou používateľov. Navyše sa ukázalo, že stakeholderi preferujú rozsiahlejšie dokumenty pred krátkymi dokumentami, z dôvodu ich informačnej bohatosti.

**Kľúčové slová**

Web Usage Mining, Príprava dát, Reference Length, Data Mining, Web Content Mining, Spracovanie prirodzeného jazyka, Čitateľnosť textov, Zložitosť textov, Evalvácia strojového prekladu.

**Title**

Pillar 3: The influence of language complexity on the behaviour of stakeholders on the websites of commercial banks

**Annotation**

The presented habilitation thesis focuses on analyzing the complexity and readability of documents related to the Pillar 3 information disclosures on the web portal of commercial banks and examining their impacts on stakeholder behavior. The results of experiments have shown that Pillar 3 information disclosure is not necessary throughout the year but ideally in the first weeks of the year. The interest of stakeholders is not limited only to mandatory information, but also includes a broader context of bank information. The analysis of web

content is conducted through natural language processing methods with an emphasis on examining the impact of document complexity and readability on stakeholder behavior. The proposed methodology is based on the combination of data on usage, structure, and content of the website and has led to the design of user preference indicators on the web portals of commercial banks. A correlation has been demonstrated between user preference indicators and text complexity or readability metrics. Readability metrics like token, sentence or character frequency have achieved the highest degree of dependence on user preferences. Moreover, it has been shown that stakeholders prefer extensive documents over short documents, due to their information richness.

**Keywords**

Web Usage Mining, Data preparation, Data Mining, Web Content Mining, Natural Language Processing, Text readability, Text complexity, Evaluation of machine translation.

# OBSAH

# ZOZNAM OBRÁZKOV A TABULIEK

# ZOZNAM SKRATIEK

AWL – Academic Word List

BCBS – Basel Committee on Banking Supervision

CBMT – Corpus Based Machine Translation

CEE – Central and Eastern Europe

CTTR – Corrected Type-token ratio

EAWL – Economic Academic Word List

EBMT – Example-Based Machine Translation

GT – Google Translate

MT – Machine Translation

NDW – Number of Different Words

NGSL – New General Service List

NLP – Natural Language Processing

NMT – Neural Machine Translation

PER – Position-independent Error Rate

POS – Part-of-Speech

RBMT – Rule Based Machine Translation

RTTR – Root Type-token ratio

SMT – Statistical Machine Translation

TER – Translation Error Rate

TTR – Type-token ratio

UIH – User interest horizontal

UIV – User interest vertical

WCM – Web Content Mining

WER – Word Error Rate

WUM – Web Usage Mining

# ÚVOD

V každej ekonomike zohrávajú kľúčovú úlohu finančné inštitúcie a zvlášť, ak sú medzinárodne aktívne. Vzhľadom na ich veľký dopad v prípade zlyhania na národné ekonomiky ako aj globálny svet, Bazilejský výbor pre bankový dohľad (BCBS) vytvoril celú radu medzinárodne platných štandardov v oblasti regulácie týchto inštitúcií, známych ako Bazilejské rámce/dohody. Tie sa historicky vyvíjali od roku 1988 a dnes hovoríme už o Bazilej IV, resp. Bazilej V. Na tvorbe týchto dohôd sa významne podieľa Európska komisia, a to prostredníctvom svojej inštitúcie European Banking Authority (EBA). Európska únia sa totiž zaviazala plne implementovať štandardy vyplývajúce z Bazilejských rámcov. EBA podporuje BCBS s cieľom posilniť reguláciu, dohľad a manažovanie rizík bankového sektora. Od roku 2008 sú bazilejské štandardy budované na troch pilieroch: Pilier 1 – minimálne kapitálové požiadavky, Pilier 2 – proces kontroly vykonávaný bankovým dohľadom a Pilier 3 – trhová disciplína. S cieľom poskytnúť účastníkom trhu dostatok informácií a podporovať trhovú disciplínu, stanovil Bazilejský výbor komplexný súbor požiadaviek týkajúcich sa tejto oblasti. Rámec Pilier 3 poskytuje komplexný balík všetkých existujúcich požiadaviek na zverejňovanie nad rámec požiadaviek na kapitálové požiadavky. Pokiaľ nie je uvedené inak, rámec sa vzťahuje na všetky medzinárodne aktívne banky až na najvyššej konsolidovanej úrovni. Podobne ako bazilejské dohody aj Pilier 3 od uvedenia do účinnosti prechádzal viacerými revíziami. Revidované zverejnenia Pilier 3 vychádzajú z piatich hlavných princípov, ktorých základom sú skúsenosti získané z obdobia finančnej krízy v rokoch 2007 – 2009: zrozumiteľnosť, komplexnosť, zmysluplnosť/užitočnosť, konzistentnosť v čase a porovnateľnosť. Frekvencia zverejňovania údajov finančnými inštitúciami sa pohybuje medzi štvrťročnou, polročnou a ročnou frekvenciou, v závislosti od povahy požiadavky.

Komerčné banky v krajinách CEE (Central and Eastern Europe – stredná a východná Európa) majú množstvo špecifík, ktoré nie vždy sú adekvátne zohľadnené v príslušných reguláciách a štandardoch. Spomedzi nich je dôležité spomenúť prevládajúce vlastníctvo veľkými nadnárodnými skupinami, ich právna forma ako subjektov, s akciami, s ktorými sa neobchoduje na kapitálových trhoch (a z toho vyplývajúca aj iná požiadavka na informácie zo strany stakeholderov) alebo biznis modely založené na depozitných klientoch. Ich depozitný klienti – vkladatelia – predstavujú veľmi dôležitú skupinu zainteresovaných strán / stakeholderov. Avšak chýbajú empirické štúdie o správaní a záujmoch depozitných klientov pri využívaní informácií Pilier 3. Regulačné orgány nevedia, do akej miery sú existujúce pravidlá

zverejňovania zmysluplné, hodnotné pre používateľov a pomáhajú predchádzať zlyhaniu trhovej disciplíny, ako tomu bolo počas poslednej finančnej krízy. Pre dosiahnutie cieľov stanovených zo strany regulátorov je kľúčové, aby bol mechanizmus trhovej disciplíny účinný a využívaný v súlade s ich očakávaniami. Ako bolo spomenuté vyššie, v krajinách strednej a východnej Európy je nedostatok štúdií, ktoré by na základe relevantnosti obsahu pre kľúčové zainteresované strany komerčných bánk hodnotili zverejňovanie informácií podľa Pilier. Viaceré výskumy prezentované v tejto práci sú preto zamerané práve na analýzu záujmu o zverejňovanie informácií v rámci Pilier 3 – Trhová disciplína, komerčnými bankami, ktorých akcie nie sú verejne obchodované a kľúčovými stakeholdermi sú depozitní klienti. Význam tejto skupiny podporuje aj fakt, že napríklad na Slovensku takmer polovicu vkladov depozitných klientov tvoria nepoistené vklady a podobný stav možno očakávať aj v iných krajinách strednej a východnej Európy. A práve pre týchto klientov dostupnosť a zrozumiteľnosť informácií o finančnom a rizikovom profile ich finančnej inštitúcie by mali byť jednými z kľúčových faktorov pokiaľ chcú správne manažovať riziká vyplývajúce z umiestnenia ich vkladov do týchto inštitúcií.

## CIEĽ PRÁCE

Hlavný cieľ predkladanej habilitačnej práce vychádza z obsahu a používania webu, cieľom je návrh metodiky zameranej na analýzu zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3 zverejňovanými na stránkach komerčných bánk ako aj na skúmanie ich vplyvu na správanie sa stakeholderov (medzi ktorých môžu patriť klienti, akcionári, regulačné orgány, ale aj široká verejnosť). Navrhovaná metodika kombinuje dáta o používaní, štruktúre a obsahu webu, čím prepája všetky domény webu. Experiment sa realizoval na dátach bankovej inštitúcie, ktoré boli získané za rok 2018 a dokumentami získanými z webového portálu (Príloha M). Naplnenie cieľa je založené na syntéze odbornej práce autora, ktorá je tvorená zjednocujúcim komentárom odkazujúcim na jednotlivé vedecko-výskumné práce zamerané na analýzu správania sa stakeholderov na webovom portáli a na analýzu dát o obsahu webu. Práce boli publikované v impaktovaných vedeckých časopisoch, na zahraničných vedeckých konferenciách a ako kapitoly v monografii. Publikované práce sú súčasťou prílohy habilitačnej práce (Príloha A - M).

Prvým čiastkovým cieľom je **skúmanie správania sa stakeholderov na webovom portáli bankovej inštitúcie pomocou rôznych prístupov**. Za týmto účelom sa realizovala séria experimentov zameraných na analýzu správania sa stakeholderov na webovom portáli.

V prvom rade bolo nutné realizovať predspracovanie dát v procese získavania znalostí z používania webu a zvolenie si vhodnej metodiky na získanie spoľahlivých dát o používaní webu (Príloha C a G). Ďalším krokom bola analýza správania sa stakeholderov v prostredí bankovej sféry (Príloha A, B, D, E a F).

Druhým čiastkovým cieľom je **analýza obsahu webu, vo forme kolekcie dokumentov, pomocou rôznych metód spracovania prirodzeného jazyka**. Viaceré odborné dokumenty a ich lokalizácia pre daný región je v dnešnej dobe vytváraná pomocou strojového prekladu. Z tohto dôvodu sa prvá séria experimentov zamerala na evalváciu strojového prekladu pomocou automatických metrík presnosti prekladu (Príloha J a K). Druhá časť experimentov sa zamerala na skúmanie zložitosti a čitateľnosti týchto dokumentov prostredníctvom rôznych automatických mier (Príloha H, I a L).

# 1 ZÍSKAVANIE ZNALOSTÍ Z WEBU

Pojem získavanie znalostí označuje proces, v ktorom sa pomocou jednej alebo viacerých data mining-ových techník hľadajú vo veľkých dátových zdrojoch vzory, ktoré slúžia na získanie užitočných informácií (Loshin 2013). Proces získavania znalostí si vyžaduje značné množstvo údajov, ktoré musia byť v spoľahlivom stave skôr ako budú podrobené samotnej analýze dát. Pod týmto procesom sa môžu rozumieť úlohy, ktoré zahŕňajú výber dát, predspracovanie dát, transformáciu dát, analýzu dát a interpretáciu výsledkov (Fayyad et al. 1996). Práve pomocou data minig-ových techník sa môžu analyzovať rôzne zdroje dát ako napríklad webové portály, texty alebo databázy a získať z nich rôzne zaujímavé znalosti.

Táto kapitola sa zameriava na dva zdroje dát, ktoré spolu úzko súvisia – dáta o používaní webu a dáta o obsahu webu. Informácie na webe sú najčastejšie uložené v textovej podobe, preto prepojenie oblastí získavania znalostí z webu a znalostí z textu môže priniesť zaujímavé vzory v skúmaní správania sa návštevníkov webových portálov.

## 1.1 ZÍSKAVANIE ZNALOSTÍ Z POUŽÍVANIA WEBU

Webové portály sú zdrojom informácií vyhľadávanými používateľmi. Web Usage Mining (WUM – získavanie znalostí z webu) v sebe zahŕňa porozumenie správania sa používateľov pri navštevovaní webových stránok. Podobnú filozofiu je možné použiť aj pre používateľov informačných systémov, ktorých správanie sa v systéme môže odhaliť prípadné chyby alebo prispieť k vylepšeniu systému. Na zaznamenávanie stôp, či už na webových portáloch alebo v informačných systémoch, slúžia logovacie súbory. Skúmanie logovacích súborov odhalí nielen správanie, ale aj návyky používateľov. Nakoľko sa v logovacích súboroch zaznamenávajú hlavne anonymné údaje, je nutné ich spracovať a pripraviť na analýzu pomocou metódy predspracovania dát. Predspracovanie dát je dôležitou súčasťou WUM, a pre tento účel bolo navrhnutých množstvo techník predspracovania. Cieľom získavania znalostí na základe používania webu je analýza správania sa používateľov pri prechádzaní webu (Srivastava et al. 2000; Romero et al. 2009). Dáta o používaní webu sa zaznamenávajú do logovacieho súboru webového servera, kde je možné z veľkého objemu dát získať informácie pre ich lepšie porozumenie.

Prvou fázou v procese získavania znalostí je porozumenie problematike. Cieľom tejto fázy je pochopiť ciele problému formulovaného z hľadiska modelovania dát. Medzi úlohy získavania znalostí patrí deskripcia a sumarizácia, segmentácia, deskripcia konceptov,

klasifikácia, predikcia a analýza závislostí (Liu 2011). V druhej fáze je cieľom získanie relevantných dát o používaní webu. Zdrojom sú dáta o používaní webu, prípadne informačných systémov a pod. Informačné systémy zväčša evidujú údaje o používaní systému vo vlastnej štruktúre, prevažne vo forme databázy. V prípade webových a proxy serverov sú dáta zaznamenávané v spoločnej štandardnej štruktúre v textovom formáte – v logovacom súbore. Logovací súbor v štandardnej štruktúre – Common Log File (W3C 1995) zaznamenáva informácie o IP adrese, čase a dátume návštevy a pristupovanom objekte. V prípade rozšírenej podoby (Extended Log File – ELF) dokáže zaznamenávať aj údaje o odkazovanom objekte a verzii prehliadača používateľa – User Agent (Liu 2011).

### 1.1.1 PREDSPRACOVANIE DÁT O POUŽÍVANÍ WEBU

Podmienkou dobrej analýzy sú kvalitné dáta. Logovacie súbory sú typické tým, že obsahujú značné množstvo nepodstatných údajov, ktoré by mohli analýzu dát ovplyvniť. V prípade skúmania správania sa používateľov alebo návštevníkov webového portálu je možné pre získanie logovacieho súboru použiť nasledovné metódy (Munk et al. 2010):

- výberové zisťovanie – zisťujú sa odpovede na konkrétne položky dotazníka a návštevník webu si je vedomý predmetu skúmania (Cerna a Poulova 2008),
- Web Usage Mining – analyzuje sa logovací súbor webového servera, ktorý obsahuje informácie o prístupoch na stránky webového portálu bez vedomosti návštevníka, pričom sú jeho údaje do istej miery anonymné (Cooley et al. 1999).

Predpokladom pre prácu s kvalitnými údajmi nie je len ich zber, ale aj príprava pre ďalšie analýzy (príprava dát sa v angličtine označuje pojmami data preprocessing alebo data preparation). Z dôvodu množstva irelevantných údajov v logovacích súboroch, ktoré treba odstrániť, je fáza prípravy dát nielen časovo najnáročnejšia, ale aj veľmi prácna. Losarwar a Joshi (2012) pri analyzovaní fázy prípravy dát vo WUM dospeli k záveru, že v oblasti analýzy webu je táto fáza veľmi dôležitá a vyžaduje si použitie nástrojov, ktoré sa zvyčajne na prípravu dát v iných doménach nepoužívajú. V prípade portálov virtuálneho vzdelávacieho prostredia, autori (Sael et al. 2013) prišli s vlastnou úpravou logovacieho súboru, čím minimalizovali nutnosť prípravy dát a rovno extrahovali všetky potrebné údaje pre analýzu. Napriek tomu, toto riešenie nie je použiteľné v prípade portálov s anonymným prístupom, kde je nutné postupovať klasickým procesom prípravy dát.

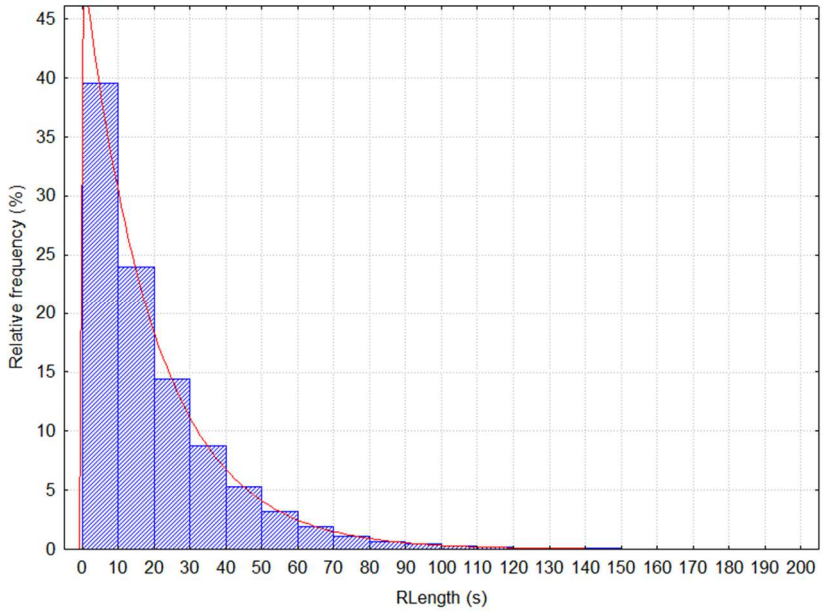
Príprava dát zahŕňa niekoľko krokov. Prvým krokom je čistenie dát od nepotrebných údajov, kde cieľom je odstránenie záznamov, resp. odkazov, ktoré sú irelevantné pre skúmanie správania sa používateľov (Cooley et al. 1999). Medzi takéto odkazy patria hlavne prístupy k obrázkom, flash videám, ikonám kurzora, javascriptom alebo štýlom. Postup identifikácie takýchto záznamov zvyčajne zahŕňa identifikáciu na základe prípony (*.jpg, *.jpeg, *.bmp, *.png, *.gif, *.css, *.js, *.flw, *.swf, *.cur, *.rss, *.ico, *.xml a podobne). Aj pri načítaní len jednej stránky sa všetky tieto požiadavky zapíšu do logovacieho súboru. Okrem požiadavky GET sa do logovacieho súboru zapisujú aj ďalšie požiadavky http protokolu, pričom je potrebné odstrániť aj návratové kódy 4xx/5xx, ktoré identifikujú chybu klienta/servera. Aye (2011) predstavil dva algoritmy pre získavanie dát z databáz. Jeho algoritmus pre čistenie dát v sebe navyše zahŕňal informácie o počte zmazaných údajov a počte unikátnych prístupov na skúmaný webový portál. Srivastava et al. (2015) predstavili algoritmus určený na čistenie logovacieho súboru od nepotrebných dát, v ktorom využívajú daný časový interval a taktiež dokážu vytvoriť sekvenciu záznamov. Nevýhodou predstaveného algoritmu je práca s veľkým objemom dát, kde v prípade čistenia väčších logovacích súborov dochádza k značnému spomaleniu. Spomínaní autori sa nezaoberali čistením logovacích súborov od prístupov robotov vyhľadávacích služieb.

Ďalším krokom čistenia dát je odstránenie prístupov robotov vyhľadávacích služieb ako napríklad Google, Yahoo, Bing a pod. Nakoľko roboty pristupujú k webovému portálu sekvenčne, tak nie je vhodné zahrnúť ich aktivitu do skúmania správania sa používateľov. Detekcia robotov prebieha buď na základe ich identifikácie v poli User Agent, alebo na základe IP adresy, ktorú je možné porovnať s databázou robotov (napr. www.robotstxt.org) (Cooley et al. 1999). Vellingiri a Chenthur Pandian (2011) sa sústredili na zlepšenie techník na čistenie dát, hlavne na odstraňovanie prístupov robotov. Okrem už vyššie spomínaných nepotrebných dát a prístupov robotov autori odstránili z logovacieho súboru všetky prístupy, ktoré mali dĺžku prístupu kratšiu ako dve sekundy.

V logovacom súbore sa prioritne zaznamenávajú anonymné údaje o používateľoch, čím vzniká problém s jednoznačnou identifikáciou návštevníka webu. Pri analýze nie je potrebné poznať konkrétnu identitu používateľa, ale rozlišovať medzi jednotlivými používateľmi. Avšak predpoklad, že na identifikáciu používateľa stačí IP adresa, je nesprávny, pretože za jednou IP adresou sa môže nachádzať viacero používateľov. Z toho dôvodu je nutné kombinovať viaceré metódy, ako napríklad využitie poľa Cookie (Pabarskaite a Raudys 2007), prípadne kombinácie IP adresy s poľom User Agent (Srivastava et al. 2000). Viaceré heuristické metódy

využívajú kombináciu IP adresy s poľom User Agent. Ak nastane zmena IP adresy, je zrejmé, že ide o nového používateľa. Ak je IP adresa rovnaká, porovnáva sa pole User Agent, ak nastane zmena, je identifikovaný nový používateľ, v opačnom prípade ide o toho istého používateľa (Srivastava et al. 2000). V prípade, že portál vyžaduje od používateľa registráciu, resp. prihlásenie, je identifikácia používateľov zjednodušená z dôvodu existencie záznamu v logovacom súbore. Používateľ môže navštíviť stránku viackrát, pričom v logovacom súbore sú zaznamenané viacnásobné sedenia (návštevy) pre každého používateľa. Cieľom identifikácie sedení je rozdeliť jednotlivé prístupy každého používateľa do oddelených relácií (Cooley et al. 1999). Sedenie môže byť definované ako postupnosť krokov, ktoré vedú k naplneniu určitej úlohy (Spiliopoulou a Faulstich 1999) alebo ako postupnosť krokov, ktoré vedú k dosiahnutiu určitého cieľa (Ming-Syan Chen et al. 1998). Na identifikáciu sedení sa používajú štruktúrovo-orientované heuristiky, časovo-orientované heuristiky (Liu 2011; Berendt et al. 2003), ako aj kombinácie týchto dvoch prístupov, ktoré predstavujú zaujímavý prístup (Munk a Kapusta 2014).

Metóda Reference Length patrí do skupiny heuristík, ktoré sú kombináciou štruktúrovo a časovo-orientovaných heuristík. Reference Length je založená na predpoklade, že dĺžka času stráveného používateľom na stránke je vo vzťahu s tým, či je stránka klasifikovaná ako obsahová alebo navigačná (Munk a Kapusta 2014; Kapusta et al. 2012a). Na obrázku (Obrázok 1) je znázornený histogram popisujúci rozdelenie premennej Length, ktorá slúži na reprezentáciu času stráveného na stránke webového portálu. Predpokladá sa, že ľavá strana grafu predstavuje navigačné stránky. Tie slúžia návštevníkom hlavne na rýchly prechod k obsahovým stránkam, ktoré sú ich cieľom. Z toho dôvodu pravú stranu tvoria stránky s obsahom, ktorých dĺžka stráveného času má väčší rozptyl.




**Obrázok 1** – Rozdelenie premennej RLength (Munk a Benko 2018)

Na základe predpokladu exponenciálneho rozdelenia premennej je možné vypočítať hraničný čas $C$, ktorý slúži na rozlíšenie navigačných stránok od obsahových. Premenná $RLength$ má exponenciálne rozdelenie

$$f(RLength) = \lambda e^{-\lambda RLength}, \tag{1}$$

$$F(RLength) = 1 - e^{-\lambda RLength}, \tag{2}$$

kde $RLength \geq 0$.

Ak je $p$ relatívna početnosť navigačných stránok, potom sa na odhad hraničného času $C$ využije kvantilová funkcia

$$F^{-1}(p, \lambda) = C = \frac{-\ln(1-p)}{\lambda}, \tag{3}$$

pre $0 \leq p < 1$. Maximálne vierohodný odhad parametra $\lambda$ (priemerná intenzita udalostí) je

$$\hat{\lambda} = \frac{1}{\overline{RLength}}, \tag{4}$$

kde $\overline{RLength}$ je pozorovaný priemer dĺžky návštev.

V okamihu, keď je odhadnutý hraničný čas, sedenie môže byť identifikované porovnaním každého času stráveného na stránke s hraničným časom. Práve hraničný čas rozdelí stránky na navigačné a obsahové podľa dĺžky času stráveného na konkrétnej stránke (Munk a Benko

2018; Munk et al. 2015). Následne, sedenie je sekvencia navštívených stránok s časovou známkou, pre ktorú platí:

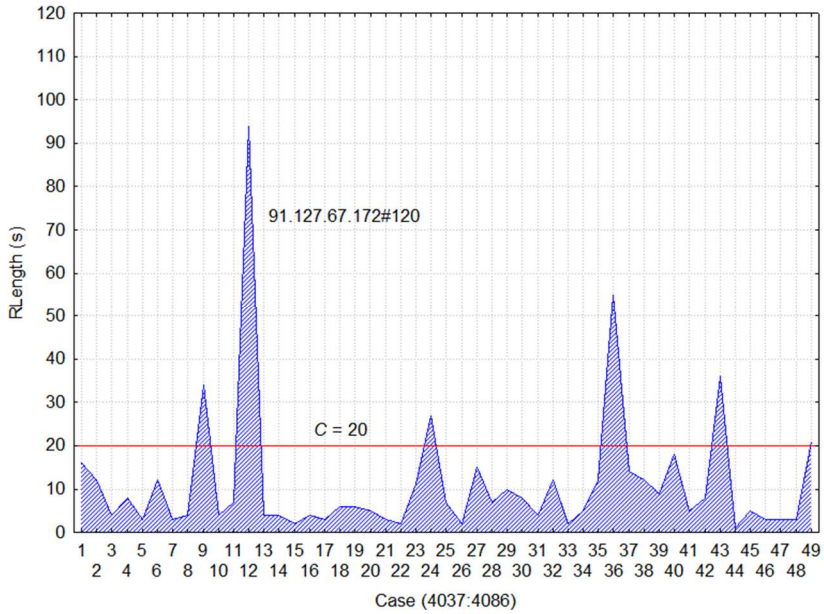
$$\langle USID, \langle URL_1, DTime_1, RLength_1 \rangle, \ldots, \langle URL_k, DTime_k, RLength_k \rangle \rangle, \quad (5)$$

$$RLength_i \leq C, \quad (6)$$

kde $1 \leq i < k$ a pre poslednú stránku sedenia platí:

$$RLength_k > C. \quad (7)$$

Podľa metódy Reference Length je nové sedenie definované od stránky s vlastnosťou (7), pričom prvých $k - 1$ stránok je klasifikovaných ako navigačné stránky a posledná $k$-tá stránka je klasifikovaná ako obsahová.




**Obrázok 2** – Metóda Reference Length (Munk a Benko 2018)

Na obrázku (Obrázok 2) je znázornená sekvencia navštívených stránok z danej IP adresy a agenta, ktorá je usporiadaná podľa času prístupu (os *x*) a času stráveného na stránke (os *y*). Hraničný čas bol 20 sekúnd, kde prvé sedenie je tvorené stránkami 1 až 9, prvých 8 je klasifikovaných ako navigačné stránky a posledná je obsahová stránka. Analogicky sa postupuje pri identifikácií ďalších sedení.

Ak premenná *length* nemá exponenciálne rozdelenie, tak autori (Munk et al. 2013; Kapusta et al. 2013; 2014; Munk et al. 2017a) sa prikláňajú k odhadom na základe kvartilového rozpätia, ktoré nie sú ovplyvnené odľahlými hodnotami, napr. $Q_{III} + 1{,}5Q$, kde $Q_{III}$ je horný kvartil (75. percentil) a $Q$ je kvartilové rozpätie (stredných 50 % hodnôt), t. j. ak je čas na stránke

považovaný za odľahlú hodnotu, začína sa nové sedenie. V opačnom prípade je lepšie použiť na identifikáciu sedení metódu Reference Length.

Dĺžka času stráveného na stránke je daná rozdielom prístupových časov súčasnej stránky a nasledujúcej, pričom sa nedá vypočítať čas poslednej stránky v sekvencii. Metóda Reference Length predpokladá, že každá posledná stránka je obsahová stránka. Môže sa však stať, že z dôvodu neočakávanej udalosti na strane návštevníka (napr. telefonát) je obsahová stránka klasifikovaná ako navigačná stránka. Rovnako je potrebné vziať do úvahy skutočnosť, že pre každého používateľa môže byť každá stránka rôzne klasifikovaná, pre jedného to môže byť navigačná stránka, ale pre druhého obsahová a naopak.

## 1.2 ZÍSKAVANIE ZNALOSTÍ Z OBSAHU WEBU A ZO SPRACOVANIA PRIRODZENÉHO JAZYKA

Návštevníci vyhľadávajú na webových portáloch informácie, ktoré tvoria ich obsah. Web Content Mining (WCM – získavanie znalostí z obsahu webu) extrahuje užitočné informácie alebo znalosti zo štruktúrovaných a neštruktúrovaných dát webu (Liu 2011). Obsah webu môže pozostávať z textu, obrázkov, zvuku, videa alebo štruktúrovaných dát, ako napr. tabuliek. Preto je WCM veľmi prepojený s oblasťou Text Mining (získavanie znalostí z textu). Medzi niektoré z riešených problémov WCM patrí extrahovanie kľúčových slov, zoskupovanie webových dokumentov a klasifikácia webových stránok (Srivastava et al. 2005). Klasifikácia webových stránok je proces priradenia kategórií webovým stránkam na základe predom stanovených kategórií. Z pohľadu návštevníkov webu predstavuje informatívne najdôležitejšiu časť obsahu práve text. V prípade webových portálov komerčných bánk predstavujú textové informácie nielen obsah webu, ale nachádzajú sa aj vo forme dokumentov. Návštevník webu na stránkach strávi určitý čas s cieľom získať informáciu, na tento čas môže mať vplyv viacero faktorov, medzi ktorými môže byť okrem štruktúry webu, aj zložitosť a čitateľnosť hľadaného textu. Na skúmanie textov sa používajú metódy spracovania prirodzeného jazyka (NLP – Natural Language Processing). Jednou z domén NLP je strojový preklad (MT – Machine Translation). Povinne zverejňované informácie na webových portáloch bankovej inštitúcie môžu byť lokalizované do rôznych jazykov. V prípade nedostupnosti lokalizovanej jazykovej verzie povinne zverejňovaných informácií sú s veľkou pravdepodobnosťou tieto informácie preložené pomocou strojového prekladu. Z toho dôvodu bolo nutné sa venovať kvalite strojového prekladu a predstaviť spôsoby evalvácie strojového prekladu. Existujú dva prístupy k procesu strojového prekladu, jeden je založený na pravidlách (Rule Based Machine Translation -

RBMT) a druhý na korpuse (Corpus Based Machine Translation - CBMT). RBMT stavia na lingvistických pravidlách a vyžaduje si širokú škálu gramatických pravidiel. Tento prístup si vyžaduje analýzu a reprezentáciu významu východiskového textu a syntézu (generovanie) jeho ekvivalentu v cieľovom jazyku. Dôležitým kritériom je vytvorenie abstraktnej reprezentácie textu, ktorá je po lexikálnej aj štrukturálnej stránke jednoznačná. Na druhej strane, prístup CBMT nie je založený na pravidlách a gramatike, ale ide o zarovnaný dvojjazyčný korpus. CBMT systém obsahuje v sebe korpusy v strojovo čitateľnej forme, ktoré sú písomného alebo hovoreného charakteru. CBMT si taktiež vyžaduje prekladové znalosti z veľkých dvojjazyčných korpusov. K systémom strojového prekladu založených na korpusoch patrí strojový preklad založený na príkladoch (Example-Based Machime Translation - EBMT) a štatistický strojový preklad (Statistical Machine Translation - SMT).

Štatistický strojový preklad je podobný strojovému prekladu založenému na príkladoch (EBMT). Napriek tomu, že oba prístupy vyžadujú rozsiahle bilingválne korpusy, systémy EBMT „sa učia“ na základe príkladov a systémy SMT na základe štatistiky.

SMT pozostáva z dvoch častí:

- v získaní translačného modelu, tak aj jazykového modelu cieľového textu,
- v dekódovaní východiskovej vety, t. j. v nájdení vhodného prekladu cieľovej vety ($c$), k východiskovej ($v$) s čo najväčšou pravdepodobnosťou.

Translačný model, charakterizujúci adekvátnosť prekladu, poskytuje informáciu o tom, aká je pravdepodobnosť, že reťazec (slovo/fráza) je adekvátnym prekladom iného reťazca, ktorý je natrénovaný na paralelných textoch bilingválneho korpusu. Jazykový model zasa charakterizuje plynulosť prekladu, poskytuje informáciu o tom, aká je pravdepodobnosť, že reťazec (slovo/fráza) je dobre formovaný (štruktúrovaný), pričom je natrénovaný na monolingválnom korpuse cieľového jazyka.

V dnešnej dobe je už bežný neurónový prístup ku strojovému prekladu, ktorý je založený na neurónových sieťach – neurónový strojový preklad (Neural Machine Translation, NMT). NMT dosiahol vynikajúce výkony pri veľkoobjemových prekladoch z angličtiny do francúzštiny (Luong et al. 2015), ako aj z angličtiny do nemčiny (Jean et al. 2015). Spoločnosti Google alebo Systran od konca roku 2016 postupne integrujú NMT do svojich prekladacích nástrojov. NMT chápe vetu ako celok a vytvára asociácie medzi frázami aj v dlhších vetách, t. j. prečíta si všetky východiskové slová až po koniec vety a až potom naraz

začne „vysielať“ jedno cieľové slovo. Výhodou NMT je schopnosť „priamo sa učiť“ (end-to-end), t. j. všetko sa učí ako jednu veľkú úlohu, pri učení neexistujú žiadne extra alebo medzikroky. SMT je založený na frázach, kde rozdeľuje východiskové segmenty (vety) na frázy (Koehn 2010). Počas trénovania SMT vytvára translačný a jazykový model. Počas prekladania vyberie dekóder preklad, ktorý je na základe týchto dvoch modelov najpravdepodobnejší. V princípe SMT dokáže produkovať veľmi dobré výsledky na úrovni fráz v zmysle adekvátneho transferu významu východiskového textu do cieľového jazyka. Avšak častokrát plynulosť prekladu v cieľovom jazyku (gramatika) nespĺňa požadovanú kvalitu. NMT je jedna rozsiahla neurónová sieť, ktorá je natrénovaná formou end-to-end, má malú pamäťovú stopu a schopnosť dobre generalizovať veľmi dlhé sekvencie slov. Dokáže spracovávať východiskové segmenty a transformovať ich na cieľové segmenty, pričom NMT prechádza celými vetami nielen frázami. Nepotrebuje uchovávať rozsiahle translačné (frázové tabuľky) a jazykové modely. Implementovanie NMT dekodéra je jednoduché v porovnaní s ostatnými strojovými prekladmi ako SMT. NMT využíva hlboké strojové učenie (deep learning), ktoré je reprezentované neurónovou sieťou (Bessenyei 2017). Výhodou NMT je, že sa vyhýba mnohým „veľmi krehkým“ prekladateľským návrhom v tradičnom strojovom preklade založenom na frázach. V praxi to znamená, že častokrát zlepšuje plynulosť prekladu na úkor adekvátnosti prekladu. NMT niekedy vyprodukuje vety, ktoré významovo nekorešpondujú s východiskovým textom, čo vyúsťuje do významových posunov a pomerne prekvapujúcich prekladov (hlavne pri natrénovaní na rozsiahlych textových dátach mimo domény).

### 1.2.1 EVALVÁCIA STROJOVÉHO PREKLADU

Napriek tomu, že manuálna evalvácia sa považuje za najspoľahlivejšiu, existujú problémy, s ktorými si nevie poradiť. Papineni et al. (2002) konštatovali, že metódy a metriky manuálnej evalvácie sú príliš pomalé a finančne náročné pre rozvoj systémov strojového prekladu, pre ktorý je dôležitá rýchla spätná väzba o kvalite prekladu. V snahe o zefektívnenie hodnotenia kvality prekladu sa začalo uvažovať o automatických metódach evalvácie strojového prekladu bez intervencie ľudského posudzovateľa. Bolo navrhnutých niekoľko automatických metrík hodnotenia MT kvality za účelom zníženia „času a ľudskej námahy“ počas evalvácie. Metriky automatickej evalvácie poskytujú vysokú efektívnosť a konzistentnosť pri pomerne nízkych nákladoch. Väčšinou sú založené na meraní podobnosti medzi strojovým prekladom, ktorý hodnotí – kandidátom a humánnym (ľudským) prekladom – referenčným. Metriky automatickej evalvácie môžu byť založené na štatistických princípoch (n-gramy alebo

vzdialenosť editovania) alebo na používaní lingvistických štruktúr na morfologickej, syntaktickej alebo sémantickej úrovni.

Medzi najpoužívanejšie metriky patria metriky založené na lexikálnej podobnosti, t. j. určujú podobnosť medzi kandidátom a referenciou na úrovni slov a znakov. Do tejto kategórie patria metriky ako BLEU, METEOR, WER, PER, TER, ROUGE alebo NIST. Lexikálna podobnosť medzi hypotézou (strojovým prekladom) a referenciou (reprezentovanou zväčša ľudským prekladom) sa dá určiť na základe Jaccardovej podobnosti, kosínusovej podobnosti alebo Levenstheinovej podobnosti.

Metrika BLEU (*Bilingual Evaluation Understudy*), ktorú navrhli Papineni a kol. (2002) poskytuje rýchly a lacný spôsob evalvácie modelov MT. Výpočet skóre metriky BLEU (8) je jednoduchý a nezávislý od jazyka. Napriek viacerým nedostatkom tejto metriky je stále vnímaná ako štandardná metrika a nové, navrhnuté metriky sa s ňou porovnávajú. Metrika bola navrhnutá na evalváciu strojového prekladu veľkých korpusov na úrovni n-gramov dĺžky 1-4. Častokrát je dĺžka kandidátskeho a referenčného prekladu rôzna. V takomto prípade sa na nájdenie najlepšieho prekladu používa parameter penalizácie krátkosti prekladu (*brevity penalty,* BP), pomocou ktorého sa zisťuje BP faktor (*multiplicative breavity penalty faktor).* BP faktor (10) nadobúda hodnoty medzi 0 a 1, pričom 1 znamená, že počet slov kandidátskeho a referenčného prekladu je rovnaký. Cieľom metriky je nájsť taký referenčný preklad, ktorý má čo najvyšší BP faktor. Následne sa penalizácia krátkosti prekladu vypočíta pre celý dokument:

$$BLEU(n) = BP\_factor \ * \ exp \sum_{i=1}^{n} Info(w_i) * \log precision_i \ , \qquad (8)$$

kde $Info(w_i)$ je váha *i*–tej premennej $precision_i$ a

$$Info(w_i) = \log_2 \frac{\sum_{r \epsilon R} \sum_{w_{i-1} \epsilon r} w_{i-1}}{\sum_{r \epsilon R} \sum_{w_i \epsilon r} w_i} \ , \qquad (9)$$

$$BP\_factor = min\left(1, \frac{length\_hyp}{length\_ref}\right) . \qquad (10)$$

BP je možné vypočítať pomocou rôznych počtov referenčných slov, čo vplýva aj na výpočet metriky BLEU, z toho dôvodu existujú rôzne verzie metriky BLEU, napríklad IBM BLEU používa vo svojom výpočte priemernú dĺžku referencií. Metrika BLEU patrí medzi tzv. metriky správnosti, čím sa skóre približuje k 1, tým je hypotéza podobnejšia referencii.

Alternatívny prístup ďalších metrík k hodnoteniu kvality strojového prekladu je z pohľadu vzdialenosti editovania (Levensteinova vzdialenosť). Cieľom automatických metrík je určenie miery chybovosti, nezhody, medzi hypotézou a referenciou. Medzi základné metriky chybovosti patria PER (Position – indipendent Error Rate), WER (Word Error Rate) a TER (Translation Error Rate).

Medzi novšie prístupy patria metriky založené na strojovom učení. Metriky evalvácie strojového prekladu založené na strojovom učení dosahujú oveľa lepšie výsledky ako štandardné metriky, najmä v prípade metrík založených na učení s učiteľom. Vyššiu koreláciu s ľudským hodnotením kvality je možné dosiahnuť použitím multilinguálnych a prispôsobiteľných modelov hodnotenia kvality MT. Tieto modely je možné použiť ako metriky na posúdenie kvality prekladu akéhokoľvek konkrétneho MT, čím sa proces evalvácie automatizuje a minimalizuje potrebu anotácie ľudským hodnotiteľom. Rei et al. (2020) navrhli neurónový rámec na trénovanie viacjazyčných modelov hodnotenia strojového prekladu (*Crosslingual Optimized Metric for Evaluation of Translation*, COMET). COMET sa učí pomocou neurónovej siete vyhodnotiť a predpovedať kvalitu MT pre viaceré rôzne jazyky na základe zhody medzi referenciou a hypotézou. Podobne ako pre BLEU, aj pre COMET platí, čím vyššie skóre, tým väčšia podobnosť medzi hypotézou a referenciou.

Základom rámca je medzijazykový kóder (Crosslingual Encoder) a združovacia vrstva (pooling layer). Ako medzijazykový kóder sa využíva vopred natrénovaný, krížovo-jazykový model (napr. BERT). Vďaka tréningu na údajoch z viacerých jazykov dosahuje dobré výsledky pri klasifikácii dokumentov ako aj pri zovšeobecňovaní pre neznáme jazyky (Rei et al. 2020). Rei et al. (2020) na základe očakávaných výsledkov rozlišuje dve možné architektúry pre metriku COMET:

- model odhadu (*Estimator Model*) – vstupom do modelu sú 3 segmenty: zdrojový text, referencia a hypotéza, ktoré sú nezávisle kódované pomocou predtrénovaného krížového jazykového kódovača so združovacou vrstvou. Vnorenia viet, ktoré sú výstupom združovacej vrstvy sa skombinujú a spoja do jedného vektora, ktorý predchádza do regresora s posuvným riadením (*Feed-Forward regressor*). Celý model je trénovaný na základe minimalizácie strednej kvadratickej chyby (*Mean Squared Error*).
- model hodnotenia prekladu (*Translation Ranking Model*) – vstupom do modelu sú 4 segmenty: zdrojový text, referencia, "lepšia" a "horšia" hypotéza, ktoré

sú kódované rovnako, ako v predchádzajúcom modeli. Model hodnotenia prekladu je trénovaný pomocou straty trojnásobnej marže (*Triplet Margin Loss*), pričom minimalizuje vzdialenosť medzi spomínanými segmentami.

### 1.2.2 ZLOŽITOSŤ A ČITATEĽNOSŤ TEXTU

Zložitosť textu zohráva dôležitú rolu v analýze textu. Pomocou rôznych metrík zložitosti textu dokážeme určiť, či dané texty sú vhodné pre cieľovú skupinu čitateľov alebo nie. Zásadnú rolu zohráva okrem štýlu textu aj jazyk, v ktorom je napísaný. Väčšina automatických metrík zložitosti textu je zameraná prevažne na anglické texty (Fisher et al. 2012). Vzdelávacia iniciatíva pripravujúca amerických študentov pre vzdelávanie a prax definuje vo svojich štandardoch (Common Core State Standards) zložitosť textu ako trojicu navzájom súvisiacich komponentov (Common Core State Standards Initiative 2023):

- Kvalitatívne dimenzie zložitosti textu: V štandardoch sa kvalitatívne dimenzie týkajú aspektov zložitosti textu, ktoré sú najlepšie merateľné alebo merateľné len vnímavým ľudským čitateľom, napr. úrovne významu alebo účelu, štruktúra, jazyková konvenčnosť alebo jasnosť, a požiadavky na znalosti;
- Kvantitatívne dimenzie zložitosti textu: vzťahujú sa na aspekty zložitosti textu, ako je dĺžka či frekvencia slova, dĺžka vety a kohéznosť (súdržnosť) textu, ktoré pre ľudského čitateľa sú ťažké, ak nie nemožné, efektívne vyhodnotiť, najmä v rozsiahlych textoch;
- Čitateľ a úloha: Zatiaľ čo predchádzajúce dva komponenty modelu sa zameriavajú na inherentné vlastnosti zložitosti textu, tretia dimenzia zložitosti textu sa zameriava na premenné špecifické pre konkrétnych čitateľov (ako je motivácia, znalosť a skúsenosť) a pre konkrétne úlohy (ako je účel a zložitosť úlohy). Napríklad, pri určovaní, či je text vhodný pre daného študenta, je potrebné zvážiť aj zadané otázky. Či na základe svojich doterajších vedomostí a skúseností vie odpovedať na danú otázku.

Pri analýze textu je okrem zložitosti dôležitá aj jeho čitateľnosť. Harris a Hodges (1995) definujú čitateľnosť ako jednoduchosť pochopenia textu pomocou štýlu písania, čím sa čitateľnosť textu rozširuje zo schopností čitateľa na analýzu štýlu písania. Naopak Collins a O'Brien (2003) definujú čitateľnosť ako kvalitu a zrozumiteľnosť písaného diela, pričom ide o text, ktorý je zrozumiteľný pre cieľového čitateľa. Čitateľnosť je možné chápať ako rovnováhu medzi schopnosťami čitateľa a samotným textom.

Kvantitatívne miery zložitosti textu sa zameriavajú hlavne na samotné charakteristiky slov a ich výskyt vo vetách a v odsekoch. Gunning (2003) predstavil viac ako sto metrík zložitosti textu, avšak iba zopár z nich sa v praxi používa. Analýza na úrovni slov patrí k prvej úrovni, na ktorú sa zameriava, pretože už samotná dĺžka slova (počet znakov) môže naznačovať do akej miery musí čitateľ slovo dekódovať, pričom jednoslabičné slová sú jednoduchšie na samotné čítanie a porozumenie textu ako viacslabičné. Početnosť slov však nie je možné chápať ako kompletnú metriku, pretože kontext, v ktorom sa slová nachádzajú môže zvyšovať zložitosť textu. Chall a Dale (1995) vytvorili zoznam slov, ktoré pomáhajú určiť zložitosť textu. Čím viac slov zo zoznamu sa v texte nenachádza, tým je daný text zložitejší. Ďalšou úrovňou v prípade kvantitatívnych mier čitateľnosti je dĺžka vety (počet slov) (Kintsch 1974) a s ňou súvisiace charakteristiky.

S ohľadom na plánovaný experiment a lepšiu prehľadnosť, bolo nutné rozdeliť metriky do viacerých kategórií. V prípade prekrytia niektorých kategórií (keď sa metriky nachádzali vo viacerých kategóriách) boli duplicitné metriky počítané pomocou iného nástroja. Zoznam skúmaných metrík a ich rozdelenie do kategórií je inšpirovaný (Lu 2012):

- **Charakteristiky textu** (Text characteristics) [char][1]: je skupina metrík, ktorá je zameraná na základné charakteristiky textu ako sú početnosti, priemer, medián (Gray a Leary 1935). Najčastejšie sa používa počet tokenov, počet viet alebo znakov ako aj počet unikátnych tokenov.
- **Čitateľnosť** (Readability) [read][1]: konvenčné metriky čitateľnosti vznikli hlavne z dôvodu nahradenia zastaralých metrík. Väčšina z nich vychádza z úrovne ročníka, ktorý študenti navštevujú, napríklad metrika Flesch-Kincaid grade level je založená na úrovni ročníkov v Spojených štátoch (Kincaid et al. 1975). Čitateľnosť textu môže byť vnímaná aj ako počet rokov učenia sa, potrebných na pochopenie daného textu: $Flesch - Kincaid = 0{,}39 * \left(\frac{total\ words}{total\ sentences}\right) + 11{,}8 * \left(\frac{total\ syllables}{total\ words}\right) - 15{,}59.$
  Metrika Flesch-Kincaid grade level je zameraná skôr na dĺžku vety, než na dĺžku slova. Ďalšou metrikou používanou v americkom školskom systéme je Gunning Fog Index, ktorá zároveň patrí medzi najpoužívanejšie metriky v súčasnej lingvistike (Spiers et al. 2017). Minimálnou požiadavkou pre výpočet metriky (indexu) je výber časti textu (okna), ktorý obsahuje aspoň sto slov, formálne zapísané: $index = 0{,}4 * \left[\left(\frac{words}{sentences}\right) +\right.$

[1] vektor skúmaných premenných [*x*], kde *x* označuje príslušnú kategóriu čitateľnosti alebo zložitosti textu

$100 * \left(\frac{complex\ words}{words}\right)\Big]$, pričom zložité slová pozostávajú z troch a viacslabičných slov. Texty, ktoré sú určené pre širšie publikum by mali mať index menší než 12. Medzi metriky čitateľnosti patria aj Coleman-Liau index (Coleman a Liau 1975), Automated Reliability Index (Senter a Smith 1967), SMOG (McLaughlin 1969), Flesch reading ease (Flesch 2016).

- **Lexikálna variácia** (Lexical variation) [lex_var][1]: sa vzťahuje na rozsah slovnej zásoby čitateľa v jeho jazykovom prejave (Malvern et al. 2004). Jednou zo základných metrík lexikálnej variácie je počet rôznych slov (number of different words – NDW), ktorá sa používa pri meraní jazykového vývoja dieťaťa (Klee 1992; Miller 1991). Nevýhodou tejto metriky je závislosť od dĺžky jazykovej vzorky, pretože nedokáže porovnať vzorky s rôznou dĺžkou. Možným riešením je skrátenie vzorky na jednotnú dĺžku na základe najkratšej vzorky (Thordardottir a Weismer 2001). Malvern et al. (2004) skracovanie vzoriek vnímajú ako mrhanie užitočnými dátami a preto navrhli dve metódy štandardizácie. V oboch prípadoch sa zo skúmanej vzorky náhodne vyberie súbor čiastkových vzoriek rovnakej dĺžky a následne sa spriemeruje ich NDW, aby sa aproximovala očakávaná hodnota NDW. V prvej metóde sa každá čiastková vzorka skladá zo štandardného počtu slov náhodne vybraných zo skúmanej vzorky. V druhej metóde obsahuje každá čiastková vzorka štandardný počet po sebe nasledujúcich slov zo skúmanej vzorky s náhodným počiatočným bodom. V súvislosti s lexikálnou variáciou počtu rôznych slov skúmal McClure (1991) rôzne pomery vybraných kontextových slovných druhov (počet slovies, podstatných mien, prídavných mien, prísloviek a ich modifikátorov, čo je kombinácia prídavných mien a prísloviek), s rovnakým menovateľom (počet lexikálnych slov).
- **Lexikálna bohatosť** (Lexical richness) [lex_rich][1]: sa vzťahuje na rozsah a rozmanitosť slovnej zásoby v skúmanom texte (McCarthy a Jarvis 2007). Používa sa v kombinácií s lexikálnou variáciou, hustotou a rozmanitosťou (diverzitou), a hovorí o počte rôznych výrazov v texte a rozmanitosti slovnej zásoby. Medzi najpoužívanejšie metriky lexikálnej bohatosti patrí Type-token ratio (TTR) určená vzťahom: $TTR = \frac{T}{N}$, kde $T$ je počet slovných druhov a $N$ celkový počet slov v skúmanom texte (Templin 1957). Nevýhodou tejto metriky je znižovanie pomeru v závislosti od zvyšovania skúmanej vzory (Arnaud 1992). Niektorí autori (Geeraerts et al. 1994; Jarvis 2002) uvádzajú, že lexikálna variácia a diverzita sú podobné vlastnosti, preto boli v prezentovanom experimente habilitačnej práce niektoré metriky uvedené v oboch

kategóriách, a vypočítané pomocou rôznych nástrojov. Ďalšie podobné metriky sú modifikáciou pôvodnej metriky TTR, ako napr. Root TTR (Guiraud 1960), Corrected TTR (Carroll 1964), Bilogarithmic TTR (LogTTR) (Herdan 1964), Uber Index (Dugast 1979) a normalizované TTR (zTTR) (Cvrček a Chlumská 2015).

- **Lexikálna rôznorodosť** (Lexical diversity) [lex_div][1]: je v podstate rozsah a rôznorodosť slovnej zásoby, ktorú v texte používa autor, pričom zohľadňuje kvalitu písania, znalosť slovnej zásoby, všeobecné charakteristiky a socioekonomický status (McCarthy a Jarvis 2007). Autori (Jarvis 2002; Geeraerts et al. 1994; Lu 2012) vnímajú lexikálnu diverzitu ako analógiu s lexikálnou variáciou alebo bohatosťou. V prezentovanom experimente habilitačnej práce boli vybrané metriky, ktoré sú primárne zamerané na lexikálnu diverzitu, t. j. rôznorodosť, aj napriek tomu, že väčšina zo skúmaných metrík je inšpirovaná metrikou TTR. Metrika Measure of textual lexical diversity (MTLD) rozdeľuje text do segmentov a pre každý sa počíta TTR skóre, kde dĺžka textu je premenná, ktorá závisí na hodnote TTR, ktorá hovorí o tom, ako sa rozširujú segmenty. Každý segment sa končí v momente, keď hodnota TTR dosiahne hodnotu 0,72 (McCarthy 2005). Ako ďalšie boli použité metriky Hypergeometric distribution diversity (HD-D) (McCarthy a Jarvis 2007), Herdans' lexical diversity measure (Herdan 1964), Dugasts' lexical diversity measure (Dugast 1979) a Maass' lexical diversity measure (McCarthy a Jarvis 2007).
- **Lexikálna sofistikovanosť** (Lexical sophistication) [lex_sop][1]: nazývaná aj lexikálna zriedkavosť (rareness) meria podiel relatívne neobvyklých alebo abstraktnejších slov v textoch. Linnarud (1986) a Hyltenstam (1988) použili na jej výpočet vzťah: $LS1 = \frac{N_{slex}}{N_{lex}}$, kde $N_{slex}$ je počet sofistikovaných lexikálnych slov a $N_{lex}$ je celkový počet lexikálnych slov v texte. Obaja autori metriku zamerali na študentov angličtiny ako druhého jazyka, Linnarud (1986) definoval sofistikované lexikálne slová ako anglické slová, ktoré sa naučia študenti od 9. ročníka a vyššie v švédskom školskom vzdelávacom systéme. Laufer (1994) vytvoril model Lexical Frequency Profile, ktorý sa zameriava na podiel slovných druhov v texte v kombinácií so zoznamom prvých 1000 najfrekventovanejších slov, druhých 1000 najfrekventovanejších slov a zoznamu univerzitných slov (Xue a Nation 1984). Model ponúka aj metriku lexikálnej sofistikovanosti, pre ktorú platí vzťah: $LS2 = \frac{T_s}{T}$, kde $T_s$ je počet sofistikovaných slovných druhov a $T$ je celkový počet slovných druhov v texte (Wolfe-Quintero et al. 1998). Ďalším z prístupov k lexikálnej sofistikovanosti bola metrika slovesnej

sofistikovanosti (verb sophistication), ktorá sa vypočíta nasledovne: $VS1 = \frac{T_{sverb}}{N_{verb}}$, kde $T_{sverb}$ je počet sofistikovaných slovesných druhov a $N_{verb}$ je celkový počet slovies v texte (Harley a King 1989). Alternatívnym prístupom je upravená slovesná sofistikovanosť (corrected verb sophistication): $CVS1 = \frac{T_{sverb}}{\sqrt{2*N_{verb}}}$, ktorú navrhol Wolfe-Quintero et al. (1998), aby zredukoval efekt veľkosti vzorky. Chaudron a Parker (1990) zvolili analogický prístup k úprave, avšak jej umocnením: $VS2 = \frac{T^2_{sverb}}{N_{verb}}$.

- **Expertné metriky** (Expert metrics) [expert][1]: metrika LIX sa zaraďuje medzi metriky čitateľnosti (Björnsson 1968), ktorá vznikla prioritne pre švédske texty, avšak úspešne bola aplikovaná nezávisle od skúmaného jazyka. Na jej výpočet sa používajú štandardné charakteristiky textu: $LIX = \frac{words}{periods} + \frac{long\ words*100}{words}$, kde *words* je počet slov, *periods* reprezentuje počet interpunkčných znamienok definovaných ako bodka, dvojbodka alebo prvé veľké písmeno a *long words* je počet dlhých slov (viac ako 6 znakov). Výsledné skóre reprezentuje podľa tabuľky (Anderson 1983) stupeň vzdelávania, pričom hodnota väčšia ako 55 indikuje vysoko odborné texty vhodné pre študentov a absolventov vysokých škôl. Anderson (1983) prišiel s optimalizáciou metriky LIX a nazval ju RIX (Rate Index). Definoval ju ako: $RIX = \frac{long\ words}{sentences}$, pričom skóre vyššie než 7,2 reprezentuje rovnako ako pri LIX vysokú zložitosť textu. Anderson (1983) preukázal, že LIX a RIX medzi sebou korelujú takmer dokonale ($r$ = 0,99). O'Hayre (1966) navrhol metriku LINSEAR Write, ktorá je založená na výpočte slabík. Výpočet skóre metriky je nasledovný: $LW = \frac{\frac{easy\ words}{sentences} + 2*hard\ words}{sentences}$, kde *hard words* je počet slov, ktorí obsahujú viac ako dve slabiky a naopak *easy words* sú slová, ktoré pozostávajú z dvoch a menej slabík. Výsledné skóre reprezentuje stupeň školy, pre ktorú je text určený, hodnota 13-16 reprezentuje študenta vysokej školy, zatiaľ čo 17+ reprezentuje absolventa vysokej školy. Všetky tieto metriky sa používajú aj v skúmaní ekonomických textov a preto boli zaradené do spoločnej kategórie. Gunning Fog Index, ktorý sa tiež často používa ako relevantná metrika pre ekonomické texty, bol zaradený do metrík čitateľnosti, na overenie rozdielu medzi týmito metrikami.
- **Slabiky** (Syllable) [syl][1]: ide o podobné metriky, ako v prípade skupiny charakteristík textu, avšak v tomto prípade metriky zachytávajú iba vlastnosti súvisiace s počtom slabík. Nevýhodou týchto metrík je nefunkčnosť, resp. nevyužiteľnosť pre niektoré jazyky.

- **Podiel slovných druhov** (Part of speech ratio) [pos_ratio][1]: tagovanie alebo morfologická anotácia je priradenie lemy a tagu (morfologickej značky) každému tokenu nachádzajúcemu sa v texte. Každý tag pozostáva zo súboru písmen latinskej abecedy, číslic a symbolov. V prípade skúmania čitateľnosti textu stačí identifikovať slovný druh tokenu. Nakoľko je tagovanie časovo náročný proces, v prezentovanom experimente habilitačnej práce bol použitý automatický nástroj Stanza (Qi et al. 2020), ktorý extrahuje podiely pre: číslovky, medzery, podstatné mená, adpozície, determinanty, vlastné podstatné mená, prídavné mená, slovesá, súradnicové spojky, interpunkčné znamienka, príslovky, pomocné výrazy, častice, zámená, podraďovacie spojky, citoslovcia, symboly a iné značky. Nástroj bol vybraný na základe dosiahnutých výsledkov v experimente na anglických textoch, kde dosahoval úspešnosť určenia slovného druhu viac ako 99% (Qi et al. 2018).
- **Ostatné charakteristiky** (Other characteristics) [other]: existuje veľké množstvo metrík čitateľnosti a zložitosti textu, pričom každá používa odlišnú škálu. Z toho dôvodu nebolo možné niektoré metriky zaradiť do vyššie uvedených. Vznikla samostatná kategória obsahujúca rôzne metriky ako napríklad veľkosť súboru v kB (Size of file in kB), ktorá sa pri ekonomických textoch ukazuje byť indikátorom zložitosti textu; Reading time (Demberg a Keller 2008), čo je v experimente podstatná metrika reprezentujúca dĺžku čítania textu v sekundách. Cvrček et al. (2020) vyvinuli nástroj QuitaUp, ktorý je určený na kvantitatívnu stylometrickú analýzu textu. Do nástroja implementovali viaceré už spomínané metriky, ale aj rôzne ďalšie, ktoré boli zaradené do tejto kategórie: h-point, frekvencia hapaxov, entropia, vzdialenosť slovesa, aktivita, deskriptivita, priemerná dĺžka tokenov, tematická koncentrácia, sekundárna tematická koncentrácia (Cvrček et al. 2020). Medzi ostatné metriky boli zaradené aj dve vlastné navrhnuté metriky EAWL a EAWL_unique.

Navrhnutá bola metrika zložitosti textov EAWL, primárne pre aplikovanie na ekonomické texty. Metrika vychádza zo zoznamu slov Economic Academic Word List (EAWL), pozostávajúci z 887 slov, ktoré sa najčastejšie nachádzajú v ekonomických textoch (O'Flynn 2019). EAWL zoznam je podobný Academic Word List (AWL) zoznamu, avšak je vhodnejší pre ekonomické texty, pretože je novší a vznikol ako nadstavba New General Service List (NGSL) zoznamu, ktorý vznikol v roku 2013. Rozdielom medzi nimi je, že EAWL obsahuje menej slovných foriem ako AWL. EAWL obsahuje iba skloňované tvary alebo varianty hláskovania slov, a nie celé skupiny slov, čo znamená, že hoci má viac hesiel ako AWL (887 v porovnaní s 570), celkovo má menej slovných tvarov (1763 v porovnaní s 3112). Skóre

navrhnutej metriky EAWL sa vypočíta ako: $EAWL = \frac{eawl\ words}{words}$, kde $eawl\ words$ je počet všetkých slov textu, ktoré sa nachádzajú v slovníku EAWL a $words$ je celkový počet slov v texte. Ako alternatíva bola navrhnutá optimalizovaná metrika, kde sa skúmali iba jedinečné slová: $EAWL\_unique = \frac{eawl\ words_{unique}}{words_{unique}}$, kde $eawl\ words_{unique}$ je počet unikátnych slov textu, ktoré sa nachádzajú v slovníku EAWL a $words_{unique}$ je celkový počet jedinečných slov v texte.

# 2 VÝSLEDKY VÝSKUMU ANALÝZY SPRÁVANIA SA STAKEHOLDEROV NA PORTÁLI KOMERČNEJ BANKY

Prvým čiastkovým cieľom je **skúmanie správania stakeholderov na webovom portáli bankovej inštitúcie pomocou rôznych prístupov**. Za týmto účelom sa realizovala séria experimentov zameraných na analýzu správania sa stakeholderov na webovom portáli.

Zdrojom dát v experimentoch (Pilková, Munk, Benko et al. 2021a; Munk, Pilkova, Benko et al. 2021c) boli logovacie súbory webových portálov dvoch bankových inštitúcii (logovací súbor prvej bankovej inštitúcie je z obdobia rokov 2009 – 2012, logovací súbor druhej bankovej inštitúcie je z obdobia rokov 2013 – 2018). Oba webové portály mali podobnú štruktúru a pri ich predspracovaní sa postupovalo na základe podobnej metodiky. Podrobný popis logovacieho súboru webového portálu bankovej inštitúcie sa nachádza v článku (Munk, Pilkova, Benko et al. 2021b) (Príloha C). V článku je popísaná aj fáza transformácie dát, v ktorej boli vytvorené nezávislé premenné (prediktory). Premenná *week* bola vytvorená na základe štandardu ISO 8601 a reprezentuje týždne roka. Podobne boli vytvorené premenné *quartal, year* a *year quartal*. Skúmanou premennou bola závislá premenná *category*, ktorá bola vytvorená zlúčením navštívených webových častí do súvisiacich skupín – širších kategórii obsahu webu. Prezentované výskumy sú zamerané na webové časti súvisiace s Pilier 3.

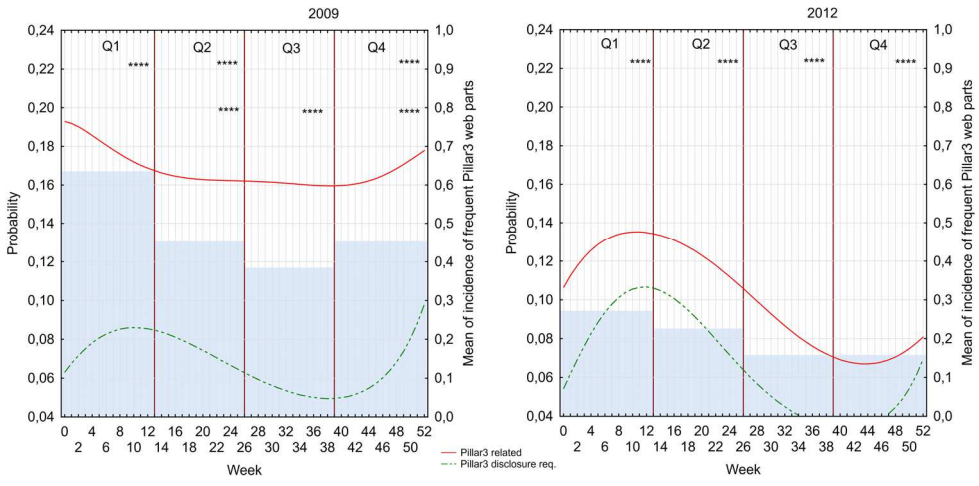
Chyby v predspracovaní údajov z logovacieho súboru môžu zásadne ovplyvniť výsledky analýzy dát a dosiahnuté závery. Z toho dôvodu je nutné napriek dodržiavaniu štandardných postupov prípravy dát vyhodnotiť získané znalosti. V experimente (Svec, Benko et al. 2020) (Príloha G) sa skúmal vplyv prípravy dát na získavanie nových znalostí vo fáze predikcie. Počas fázy vyhodnotenia výsledkov boli objavené prístupy automatizovaných nástrojov na webový portál, ktoré významne ovplyvnili identifikované znalosti, vo forme odhadu pravdepodobností prístupov na webové kategórie. Boli identifikované rozdiely v empirických a teoretických početnostiach prístupov na web a logitoch (Svec, Benko et al. 2020), kde počas 4. hodiny ráno, bola predikcia modelu pre väčšinu kategórií výrazne nadhodnotená, resp. podhodnotená. Na základe evalvácie dát bolo možné identifikovať chybu spôsobenú automatizovaným procesom, ktorý sa pravdepodobne staral o zálohovanie, kontrolu obsahu stránky na prítomnosť vírusov a podobne. Prítomnosť daného procesu významne ovplyvnila získané znalosti, či už na základe typu závislosti alebo vytváraní závislostí na miestach, kde žiadna nebola. Po odstránení týchto prístupov a zopakovaní analýzy dát, boli získané relevantné znalosti, ktoré potvrdila aj opätovná evalvácia dát. Výskum (Svec, Benko et al. 2020) slúžil ako praktická

ukážka ovplyvňovania získaných znalostí v prípade nedôkladného predspracovania dát logovacieho súboru. Podstatnú úlohu v tomto prípade zohrávala aj evalvácia dát, ktorá na všetkých troch skúmaných úrovniach poukazovala na chybu navrhnutého modelu pre konkrétnu kategóriu a čas.

Primárne smerovanie výskumu zameraného na dáta o používaní webu bolo na zhodnotenie správania sa a záujmov stakeholderov na webovom portáli komerčnej banky, ktorých akcie nie sú verejne obchodované. Kľúčovými stakeholdermi sú depozitní klienti, ktorí sa zaujímali o povinné zverejňované informácie v čase turbulentných zmien (Munk, Pilkova, Benko et al. 2021c) (Príloha A) ako aj po nich (Pilková, Munk, Benko et al. 2021a) (Príloha B). Experiment (Munk, Pilkova, Benko et al. 2021c) (Príloha A) bol zameraný na dátovú analýzu počas krízy a po nej, s cieľom identifikovať kľúčové typy informácií, ktoré zaujímajú stakeholderov, ako aj na možnosti optimalizácie politiky zverejňovania daných informácií. V experimente sa pracovalo s logovacím súborom popísaným v článku (Munk, Pilkova, Benko et al. 2021b) a na základe metodiky publikovanej v článku (Munk, Pilkova, Benko et al. 2021a) (Príloha D). V rámci metodiky boli podrobne popísané jednotlivé fázy spracovania údajov získaných z logovacích súborov webového portálu bankovej inštitúcie. Fáza modelovania dát bola zameraná na odhad pravdepodobností prístupov stakeholderov na webové kategórie súvisiace s Pilier 3. Skúmané bolo správanie sa návštevníkov počas obdobia viacerých rokov (2009-2012). Za roky krízy boli zvolené roky 2009 a 2010, pričom roky 2011 a 2012 boli považované za roky, kedy kríza začínala odznievať. Podstatný rozdiel oproti predchádzajúcim výskumom využívajúcim metodiku s multinomiálnym logitovým modelom bol v tom, že v predchádzajúcom prípade sa skúmali hodiny v rámci dňa a ďalšie umelé premenné rozlišujúce skúmané roky. To poskytovalo pohľad na správanie sa stakeholderov počas dní, avšak pre vytvorenie väčšieho obrazu o správaní sa návštevníkov webového portálu bolo potrebné zamerať sa aj na iné časové premenné. V prípade experimentu (Munk, Pilkova, Benko et al. 2021c) boli modelované pravdepodobnosti prístupu na webové kategórie na základe týždňov v rámci roka a umelá premenná identifikujúca obdobie finančnej krízy. Výskum (Pilková, Munk, Benko et al. 2021a) (Príloha B) je pokračovaním predchádzajúceho výskumu, pričom sa skúmalo akým spôsobom sa zmenilo správanie stakeholderov na webovom portáli inej rovnako dôležitej bankovej inštitúcie na Slovensku počas rokov 2016-2018. Predpokladal sa pokračujúci klesajúci trend záujmu o informácie súvisiace s Pilier 3. Napriek tomu, že logovacie súbory pochádzali z dvoch rôznych webových portálov, pomocou transformácie dát bola navrhnutá taxonómia webového portálu, ktorá tvorí

prienik pre webový obsah súvisiaci s informáciami Pilier 3. Vďaka tomu bolo možné sledovať časový trend a sezónnosť v správaní sa stakeholderov vo vzťahu k informáciám Pilier 3.

Cieľom výskumu (Pilková, Munk, Blažeková, Benko 2021b) (Príloha E) bolo zhodnotiť záujem stakeholderov o dve skupiny zverejňovaných informácií: *Pillar3 disclosure requirements* a *Pillar 3 related* počas obdobia rokov 2009-2012. Rovnako bolo cieľom zhodnotiť robustnosť overením výsledkov pomocou dvoch prístupov založených na rôznych časových premenných: týždeň a kvartál počas rokov 2009-2012. Prvý prístup bol založený na metodike popísanej v článku (Munk, Pilkova, Benko et al. 2021a). Druhý prístup sa zaoberal získavaním vzorov správania sa návštevníkov webu počas kvartálov. Výsledky boli spracované pomocou asociačnej analýzy s cieľom extrahovať frekventované položkové množiny s minimálnou podporou 1%. Predpokladom boli podobné výsledky, pričom týždenná analýza poskytuje podrobnejší prehľad správania sa návštevníkov webu v porovnaní s analýzou kvartálov. Grafy (Obrázok 3) vizualizujú pravdepodobnosti prístupov na skúmané webové kategórie súvisiace s trhovou disciplínou počas rokov 2009-2012. Rok 2009 (Obrázok 3) bol označený ako rok finančnej krízy, pričom najväčší záujem o webové kategórie bol na prelome rokov. Počas roka záujem o tieto kategórie klesá a koncom roka opäť začína stúpať. Hviezdičky nachádzajúce sa v grafoch označujú homogénne skupiny výskytu frekventovaných položkových množín. V roku 2009 ich najviac bolo identifikovaných v prvom kvartáli a najmenej v treťom. V roku 2012 (Obrázok 3) dochádza k výraznému prepadu záujmu o dané kategórie v priebehu celého roka. Napriek tomu je najväčší záujem hlavne v prvom kvartáli na začiatku roka.




**Obrázok 3** – Vizualizácia pravdepodobností kategórií súvisiacich s trhovou disciplínou počas rokov 2009 a 2012 (Pilková et al. 2021b)

Podrobná analýza, pravdepodobností prístupov stakeholderov počas týždňov na webový portál s povinne zverejňovanými informáciami komerčnej banky, ukázala, že výsledky korešpondujú s výsledkami analýzy kvartálov (Munk et al. 2017b). Najväčší záujem stakeholderov o informácie súvisiace s Pilier 3 bol počas prvého kvartálu, hlavne v období okolo 10. týždňa. 10. týždeň patril k obdobiu s najväčším záujmom o dané webové kategórie. Na základe výsledkov bolo možné konštatovať, že frekvencia povinného štvrťročného zverejňovania výsledkov nie je pre trhovú disciplínu nutná. Predpoklad, že aj po rokoch krízy (2012-2015) bude záujem o informácie Pilier 3 naďalej klesať sa potvrdil. Rovnako sa potvrdilo, že správanie sa stakeholderov vo vzťahu k zverejňovaným informáciám Pilier 3 už viac nepreukazuje žiadny trend alebo sezónnosť. Zvýšený záujem o dané informácie bol iba v čase krízy a jej následným odznievaním (2009-2011). Nasledujúce roky (2012-2018) boli charakteristické nízkym záujmom o tieto informácie, čo potvrdili dáta dvoch webových portálov bankových inštitúcií. Predchádzajúce revízie Pilier 3 regulátormi nepriniesli významný vplyv na zvyšovanie záujmu o tieto informácie.

Výskum v kapitole (Blažeková, Benko et al. 2021) (Príloha F) bol zameraný na skúmanie času stráveného na skúmaných webových stránkach v kontexte obsahu webových kategórií súvisiacich so zverejňovanými informáciami Pilier 3. Výsledky ukázali, že nadpriemerný čas bol stakeholdermi strávený na webovej stránke výročných správ (*annual reports*). Najviac času strávili návštevníci na stránkach poskytujúcich všeobecné informácie o banke a súvisiace informácie Pilier 3. Druhou najmenej navštevovanou kategóriou bola *Pillar3 Q-terly Information*, ktorá súvisí s povinne zverejňovanými informáciami. Podrobnejšia analýza tejto kategórie ukázala, že obsahuje niektoré webové časti, ktoré majú vysoký čas strávený návštevníkmi webu. To môže indikovať buď dôležitý obsah, ktorý návštevníci webu hľadajú alebo naopak, príliš veľa informácií na stránkach. Dosiahnuté výsledky potvrdzujú predchádzajúce výsledky (Munk et al. 2017b; Benko et al. 2020), ktoré ukázali, že návštevníci webu sa nezaujímajú o samotné informácie súvisiace s informáciami Pilier 3, ale skôr spolu s výročnými správami alebo informáciami o banke.

Na základe dosiahnutých výsledkov prezentovaných experimentov boli identifikované zaujímavé oblasti, ktoré môžu zvýšiť záujem stakeholderov o informácie Pilier 3 zhrnuté do nasledujúcich odporúčaní (Pilková, Munk, Blažeková, Benko 2021b):

- Zlepšiť štandardizáciu, teda harmonizáciu zverejňovania informácií vnútroštátnymi orgánmi požiadaviek a požiadaviek na zverejňovanie informácií na úrovni EÚ (Pilier 3 a národné požiadavky).
- Zvýšiť porovnateľnosť zverejňovaných informácií vytvorením jednej spoločnej šablóny (vizuálne predpísaných tabuliek), ktorú by v ideálnom prípade vytvorili regulačné orgány, s cieľom zaviesť jednotnosť.
- Znížiť frekvenciu zverejňovania informácií Pilier 3 vzhľadom na nízky záujem zainteresovaných strán o štvrťročné zverejňovanie informácií.
- Odlíšiť ročné zverejňovanie (formou zvýšenia objemu informácií) v porovnaní so štvrťročným, v prípade, ak sa štvrťročné zverejňovanie informácií použije na zníženie objemu zverejňovaných informácií.
- Zahrnúť informačné oblasti (povinne alebo dobrovoľne), o ktoré sa zainteresované strany zaujímajú (obchodné správanie inštitúcie, stratégia, reputácia, štruktúra, vlastníctvo, poslanie, hodnoty) a ovplyvňujú rizikovú pozíciu inštitúcie.
- Zabezpečiť dodržiavanie povinných požadovaných informácií zo strany regulačných orgánov - najmä obmedziť vynechávanie požadovaných informácií inštitúciami bez uvedenia dôvodu.
- Uložiť pravidlá pre umiestnenie zverejnených dokumentov, ktoré by mali byť na identifikovateľnom mieste časti webovej stránky.
- Povinnosť používať anglický jazyk ako spoločný jazyk pre zverejňovanie informácií.

# 3 VÝSLEDKY VÝSKUMU ANALÝZY DÁT Z OBSAHU WEBU

Druhým čiastkovým cieľom je **analýza obsahu webu vo forme dokumentov pomocou rôznych metód spracovania prirodzeného jazyka**. Analýza dokumentov pochádzajúcich z webu má podstatný vplyv na ďalší náš výskum, ktorý sa zameral na jazykovú zložitosť a čitateľnosť textu. Predpokladom je, že zložitosť textu, konkrétne jeho obsah, ktorý je zverejňovaný na stránkach komerčných bánk v súvislosti s Pilier 3 ovplyvňuje správanie sa stakeholderov. Na základe doterajších výsledkov našich štúdií a stanovených odporúčaní o používaní anglického jazyka ako univerzálneho jazyka pre zverejňovanie informácií, sme sa rozhodli skúmať iba anglické texty obsahujúce informácie Pilier 3. Výsledky tohto experimentu sú prezentované v predkladanej habilitačnej práci. Nakoľko pracujeme s dátami z roku 2018, tak nie ku všetkým slovenským dokumentom existovali oficiálne preklady do angličtiny. Z toho dôvodu bolo nutné niektoré dokumenty strojovo preložiť do angličtiny. Čo vyvolalo otázku kvality strojového prekladu a sekundárny výskum zameraný na hodnotenie kvality strojového prekladu. Výsledky prezentovaných výskumov vznikali v rámci projektov zameraných na evalváciu strojového prekladu dokumentov rôzneho štýlu extrahovaných z rôznych webových portálov.

Hodnotením kvality strojového prekladu sa zaoberali výskumy Benko et al. (2022) (Príloha J) a Benko et al. (2024a) (Príloha K). Benko et al. (2022) prepojili manuálnu evalváciu strojového prekladu s automatickými metrikami, kde pomocou analýzy chýb sa hľadali asociácie medzi kategóriami chýb a automatickými metrikami evalvácie strojového prekladu založenými na lexikálnej podobnosti. Výsledky štúdií indikujú, že nie všetky automatické metriky založené na n-gramoch (lexikálnej podobnosti) alebo vzdialenosti editovania by mali byť implementované do modelu hodnotenia kvality MT strojových prekladov z angličtiny do flektívnej slovenčiny. Pri určovaní kvality strojového prekladu vzhľadom na syntakticko-sémantickú korelatívnosť (plynulosť a adekvátnosť) stačí brať do úvahy metriky BLEU-4, NIST a CharacTER, pričom výsledky by mohli byť aplikovateľné aj pre iné flektívne jazyky. Následne bol prostredníctvom analýzy rezíduí porovnaný štatistický strojový preklad s neurónovým strojovým prekladom. Skúmalo sa, či zmena paradigmy vplýva na kvalitu strojového prekladu (Benko et al. 2024a). Porovnané boli dva prístupy k strojovému prekladu (SMT a NMT) použitím dvoch odlišných systémov (Google translátora a MT nástroja Európskej komisie pre preklad), pričom na evalváciu MT boli použité automatické metriky chybovosti. Predmetom skúmania boli publicistické dokumenty extrahované z webu pre

jazykový pár angličtina-slovenčina a nemčina-slovenčina. Výsledky analýz preukázali, že neurónové MT dosahovali štatisticky významne vyššiu kvalitu ako štatistické MT bez ohľadu na to, ktorý nástroj na preklad bol použitý. Neurónové MT generované nástrojom Google Translátorom (GT) dosahovali štatisticky významne najnižšiu chybovosť. Na druhej strane štatistické MT generované nástrojom mt@ec dosahovali štatisticky významne najvyššiu chybovosť. Predpoklad o vyššej kvalite neurónového MT v porovnaní so štatistickým MT sa potvrdil bez ohľadu na jazykový pár ako aj nástroj MT.

Výskum v článku Benko et al. (2024b) (Príloha L) bol zameraný na ďalšiu úlohu spracovania prirodzeného jazyka, konkrétne na porovnanie automatických nástrojov pre morfologickú anotáciu slovenského jazyka (POS taggers). Cieľom výskumu bolo navrhnúť metodiku na porovnávanie taggerov pre flektívne jazyky a nízko-zdrojové jazyky, a zároveň nájsť najefektívnejší automatický nástroj pre morfologickú anotáciu, tzn. ktorý dosahuje najlepší výkon na základe presnosti. Výber efektívneho (v zmysle presnosti) nástroja má veľký vplyv na metriky zložitosti textu, ktoré sú definované na základe tokenizácie ako napríklad pomer slovných druhov v texte (podiel podstatných miest, prídavných mien, slovies atď.). Článok sa zameriava na porovnanie najznámejších nástrojov TreeTagger (Schmid et al. 2007), RNNTagger (Schmid 2019), MorphoDita (online verzia a desktopová aplikácia) (Straka a Straková 2014), UDPipe2 (Straka a Straková 2017) a Stanza (Qi et al. 2020). Na tento účel bol zvolený pomerne jednoduchý a krátky ručne anotovaný subkorpus Slovenského korpusu závislostí (SDC) (Gajdošová a Šimková 2016). Nevýhodou použitého datasetu sú hlavne krátke vety (Benko a Benková 2022). Texty boli rozdelené na umelecké a náučné texty. Výsledky výskumu (Benko et al. 2024b) ukázali využiteľnosť POS taggerov v prípade málo zdrojového slovenského jazyka. Štyri zo šiestich skúmaných nástrojov dosiahli vysoký výkon vzhľadom na presnosť určenia, pričom RNNTagger sa ukázal najpresnejší pre oba typy textov. Prínos dosiahnutých výsledkov z hľadiska cieľa prezentovanej habilitačnej práce spočíva vo výbere a následnom aplikovaní pri metrikách zložitosti textu, ktoré sú založené na identifikácií slovných druhov v textoch.

Kvalita prekladu zohráva významnú úlohu v porozumení obsahu textu a extrahovaní dôležitých informácií, ktoré sú poskytované čitateľom. Práve na kvalitu prekladu sa zamerali výskumy v článkoch Munkova, Munk, Benko et al. (2021b) (Príloha H) a Benko et al. (2023) (Príloha I). Munkova, Munk, Benko et al. (2021b) skúmali vplyv kvality strojového prekladu na jazykovú zložitosť na úrovni slova a vetnej štruktúry. Cieľom štúdie bolo nájsť a overiť nový prístup k hodnoteniu kvality strojového prekladu na základe čitateľnosti a zložitosti textu.

Navrhovaná metodika bola založená na evalvácii frekventovaných tagsetoch strojového a post-editovaného strojového prekladu ako aj na základe frekventovaných POS tagsetoch a sumarizačných pravidiel. Prínos navrhovanej metodiky spočíva v identifikácií systematických a nie náhodných chýb. Munkova, Munk, Benko et al. (2021b) taktiež preukázali, že pre technické texty, MT systémy produkujú preklad s prijateľnou úrovňou kvality. Bol navrhnutý originálny a doteraz nepoužívaný unikátny prístup využívajúci miery zložitosti textu. Na dosiahnuté výsledky nadviazal výskum v článku Benko et al. (2023), v ktorom sa autori snažili prepojiť lexikálnu rôznorodosť vyjadrenú metrikami zložitosti textu s typmi (kategóriami) chýb strojového prekladu. Výsledky štúdie ukázali, že chyby vznikajúce v neurónovom MT súvisia s lingvistickými vlastnosťami založenými na početnostiach. Zaujímavým zistením štúdie je, že nie všetky metriky lexikálnej rôznorodosti súvisia s frekvenciou každého typu chýb. Štatisticky významnú závislosť s frekvenciou chýb v oblasti syntakticko-sémantickej korelatívnosti, súvetnej syntaxi a lexikálnej sémantike, vykázali iba metriky RTTR a CTTR. Limitáciou výskumu bolo obmedzenie sa iba na publicistické texty, ktoré sú pomerne ľahko čitateľné. Ukázalo sa, že čitateľnosť textu nezávisí od frekvencie chýb, čo umožnilo použiť strojový preklad v skúmaní ekonomických textov, konkrétne v oblasti bankovníctva.

Dosiahnuté výsledky výskumu analýzy dát z obsahu webu indikujú, že v prípade absentujúcich jazykových lokalizácií dokumentov, je možné vychádzať z ich strojových prekladov.

# 4 VPLYV JAZYKOVEJ ZLOŽITOSTI NA SPRÁVANIE SA STAKEHOLDEROV NA WEBOVÝCH STRÁNKACH KOMERČNÝCH BÁNK

Primárny cieľ prezentovanej habilitačnej práce vychádza z obsahu a používania webu. Cieľom práce je návrh metodiky zameranej na analýzu zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3 zverejňovanými na stránkach komerčných bánk a skúmanie ich vplyvu na správanie sa stakeholderov. Súčasťou tohto procesu je vytvorenie ukazovateľov preferencií používateľov na webovom portáli bankovej inštitúcie a overenie vzťahu preferencií používateľov k zložitosti zverejňovaných textov. Experiment vychádza z dát bankovej inštitúcie získaných za rok 2018 a z dokumentov získaných z webového portálu (Benko et al. 2024c) (Príloha M). Motiváciou prezentovaného výskumu bolo nadviazať na dosiahnuté výsledky predchádzajúcich výskumov (Pilková et al. 2021a; Munk et al. 2021c; 2015; 2017b; Benko et al. 2020; Pilková et al. 2021b; Blažeková et al. 2021) a podrobne preskúmať obsah povinne zverejňovaných informácií v kontexte zložitosti a čitateľnosti textu.

Zložitosť a čitateľnosť odborných textov je vyhodnocovaná pomocou rôznych automatických mier navrhnutých viacerými autormi. Gunning (2003) predstavuje viac ako sto metrík zložitosti textu, avšak iba niekoľko z nich sa používa. Väčšina z nich sa používa na zisťovanie základných charakteristík textu ako dĺžka vety, počet slovných druhov a pod. (Sadeek Quaderi a Varathan 2024; Awan et al. 2021). Texty, ktoré sa analyzujú sú primárne určené pre učenie sa anglického jazyka ako druhého jazyka, čo smeruje k skúmaniu skôr edukačných ako odborných textov (Maqsood et al. 2022). Ehara (2022) sa zameral na skúmanie čitateľnosti úvodných textov informatiky. Vo svojom výskume porovnával BERT klasifikáciu s konvenčnými metrikami čitateľnosti ako sú Flesch-Kincaid Grade Level (Kincaid et al. 1975), ARI (Senter a Smith 1967), Coleman-Liau Index (Coleman a Liau 1975), Flesch Reading Ease (Flesch 2016), Gunning Fog Index (Gunning 2003), LIX (Björnsson 1968), SMOG Index (McLaughlin 1969), RIX (Anderson 1983) a Dale-Chall Index (Chall a Dale 1995). Ehara (2022) analyzoval prioritne texty extrahované z GitHub-u (návody k softvéru) a abstrakty vedeckých článkov zverejnených v rámci ACL Anthology a PubMed. Navrhol metriku čitateľnosti založenú na slovníku, ktorú porovnal s konvenčnými metrikami. Výsledky výskumu ukázali vyššiu koreláciu než konvenčné metriky, konkrétne, že vedecké texty sú nečitateľné pre stredne pokročilých študentov a naopak návody k softvéru sú študentov väčšinou čitateľné. Podobný výskum Ehara (2021) realizoval s ekonomickými (spravodajskými) textami, avšak zameral

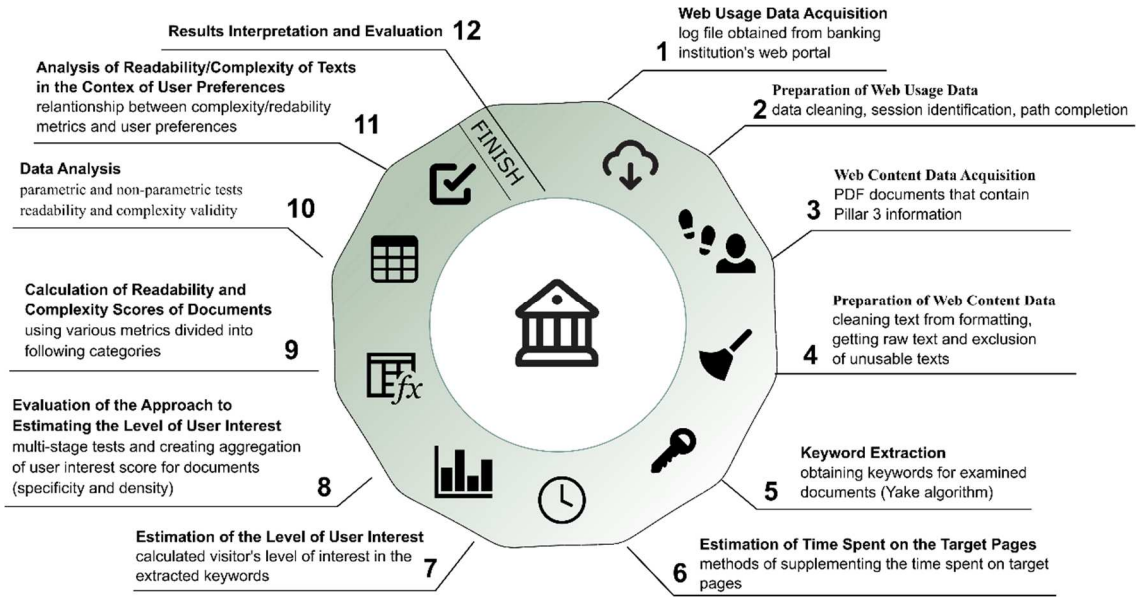
sa iba na porovnanie prístupu na báze BERT a slovníku. Výsledky ukázali, že väčšina textov bola čitateľná pre stredne pokročilých študentov, pričom 2,4% z nich neboli zrozumiteľné pre študentov.

Guay et al. (2015) sa zamerali na skúmanie zložitosti finančných výkazov (*financial statement*), v prípade čitateľnosti použili metriku ReadIndex, ktorá pozostávala zo skóre metrík: Flesch-Kincaid Grade Level, LIX, RIX, Gunning Fog Index, ARI a SMOG. Výsledky analýzy ukázali, že medzi všetkými šiestimi metrikami čitateľnosti je vysoká miera korelácie ako aj s ReadIndex-om. Prínos ich výskumu spočíva v tom, že napriek tomu, že zložité finančné výkazy negatívne ovplyvňujú informačné prostredie, tak niektoré firmy sa pokúšajú tieto vplyvy zmierniť dobrovoľným zverejňovaním ďalších informácií. Moreno a Casasola (2016) sa zamerali na analýzu čitateľnosti výročných správ v španielčine prostredníctvom upravenej metriky Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index, LIX a RIX. Zistili, že výročné správy v španielčine vykazujú známky ťažšej čitateľnosti, čím potvrdili zistenia štúdií, ktoré boli zamerané na anglické výročné správy. Toerien a du Toit (2024) sa zamerali na zverejňované správy ohľadom rizika v Južnej Afrike. Analyzovali sa výročné správy (*annual reports*) zozbierané za roky 2005-2021, počas obdobia, kedy boli v krajine zavádzané viaceré štandardy týkajúce sa zverejňovania informácií vychádzajúce z praxe v EÚ. Skúmali texty pomocou metrík čitateľnosti Flesch-Kincaid a Gunning Fog Index ako aj metrík, ktoré sa zameriavajú na počet slov, dĺžku vety a podobne. Výsledky naznačujú (Toerien a du Toit 2024), že zverejňované správy majú nízku čitateľnosť. Zavedenie štandardov, ktoré majú zvyšovať čitateľnosť a znižovať zložitosť zverejňovaných správ, sa ukázali ako neefektívne, pričom je potrebné prehodnotiť formu informácií, aby zaujali širšiu verejnosť. Za limitáciu výskumu považujú nielen problémy s určením zložitých slov, vetnej štruktúry a jej dĺžky, ale aj nedostatok kontextu. Toerien a du Toit (2024) odporúčajú rozšíriť metriky o ďalšie, ktoré by mohli lepšie vysvetľovať zložitosť zverejňovaných informácií.

Na základe preštudovanej literatúry bolo smerovanie prezentovaného výskumu zamerané na skúmanie zložitosti textu súvisiaceho so zverejňovanými informáciami Pilier 3 na stránkach komerčných bánk a jeho vplyvu na správanie sa stakeholderov.

## 4.1 METODIKA VÝSKUMU

Metodika výskumu (Obrázok 4) bola inšpirovaná viacerými výskumami (Munk et al. 2021c; 2021a; Munkova et al. 2021a; Pilková et al. 2021a; Yao et al. 2017) a podrobnejšie je popísaná v nasledujúcich podkapitolách a v článku (Benko et al. 2024c) (Príloha M).




**Obrázok 4** – Metodika výskumu zameraného na jazykovú zložitosť a čitateľnosť textu (Benko et al. 2024c)

### 4.1.1 ZÍSKAVANIE A PRÍPRAVA DÁT O POUŽÍVANÍ WEBU

Výskum vychádza z dvoch dátových zdrojov. Prvým zdrojom je logovací súbor získaný z webového portálu bankovej inštitúcie a druhým zdrojom sú dokumenty extrahované z webového portálu na základe logovacieho súboru. Logovací súbor obsahoval prístupy na webový portál počas celého roku 2018 a prešiel fázou prípravy dát, ktorá pozostávala z čistenia dát, identifikácie používateľov/sedení a dopĺňania ciest (Munk et al. 2021a). Počas fázy predspracovania dát boli vytvorené ďalšie potrebné premenné ako *Category* a *Subcategory*, ktoré slúžili na prepojenie webových častí portálu, na ktoré pristupovali návštevníci. Taxonómia webového portálu, na základe ktorej boli webové časti rozdelené, sa nachádza v Tabuľke 1. Prioritou boli informácie súvisiace s informáciami Pilier 3, preto bol logovací súbor zredukovaný na sedenia, ktoré obsahovali aspoň jeden prístup na jednu zo skúmaných kategórií. Takto upravený logovací súbor pozostával z 265 216 záznamov.

**Tabuľka 1** – Taxonómia webového portálu komerčnej banky

| Kategória (Category) | Podkategória (Subcategory) |
|---|---|
| /Pilar3 disclosure requirements/ | /financial_statement/ |
| /Pilar3 disclosure requirements/ | /information_about_bank/ |
| /Pillar3 related/ | /annual_reports/ |
| /Pillar3 related/ | /financial_reports/ |
| /Pillar3 related/ | /covered_bonds/ |
| /Pillar3 related/ | /information_for_investors_except_shareholders/ |
| /Pillar3 related/ | /information_for_shareholders/ |

### 4.1.2 ZÍSKAVANIE DÁT Z OBSAHU WEBU

Pre skúmanie vplyvu zložitosti textov na návštevnosť webového portálu, bolo nutné extrahovať texty, ku ktorým stakeholderi pristupovali. Skúmaný webový portál bankovej inštitúcie zverejňuje všetky informácie týkajúce sa Pilier 3 vo formáte PDF dokumentov. Z logovacieho súboru boli extrahované priame odkazy na dané dokumenty a pomocou webového crawlera bola získaná väčšina dokumentov. Nakoľko išlo o logovací súbor z roku 2018, nie všetky dokumenty boli prístupné, avšak podarilo sa extrahovať viac ako 90 % navštívených dokumentov. Viac ako polovica dokumentov bola v anglickom jazyku spolu s ich oficiálnym prekladom do slovenčiny. Dokumenty, ktoré obsahovali iba obrázky, tabuľky alebo grafy a duplikované jazykové verzie boli odstránené z kolekcie dokumentov (celkovo bolo odstránených 178 dokumentov). Niektoré slovenské dokumenty nemali sprístupnenú anglickú verziu, preto bol ich anglický preklad vytvorený pomocou MT systému Google Translátora (preložených bolo 97 dokumentov). Na analýzu textu bolo použitých 226 dokumentov rozdelených do jednotlivých kategórií podľa taxonómie (Tabuľka 1). Z týchto dokumentov bol pomocou PDF OCR nástroja extrahovaný čistý text, ktorý sa následne použil na analýzu zložitosti textu a hľadanie záujmu používateľov.

### 4.1.3 EXTRAKCIA KĽÚČOVÝCH SLOV

Pre potreby analýzy správania sa návštevníkov webu bolo potrebné z jednotlivých dokumentov extrahovať kľúčové slová. Bol použitý algoritmus Yake (Campos et al. 2020), ktorý dosahuje lepšie výsledky než štandardné techniky na identifikáciu kľúčových slov, ako napríklad TF-IDF, TextRank, KP-Miner alebo Rake (Campos et al. 2020). TF-IDF je častokrát používaná štatistická miera, ktorá určuje význam kľúčového slova vzhľadom na význam v jednom dokumente naprieč všetkými dokumentami celého korpusu. Experimenty preukázali, že v prípade odborných textov, nástroje ako Yake, KEA alebo KP-Miner dosahujú lepšie

výsledky než TF-IDF (Sarwar et al. 2021; Sarwar a Noor 2021). Nakoľko je prezentovaný experiment habilitačnej práce zameraný na dokumenty, ktoré sú odborného charakteru z oblasti bankovníctva, bol na extrakciu kľúčových slov zvolený nástroj Yake, ktorý bol implementovaný v jazyku Python. Algoritmus pozostával z nasledujúcich krokov (Campos et al. 2020):

- Predspracovanie textu a identifikácia kandidátnych pojmov – predbežné spracovanie dokumentu do strojovo čitateľného formátu s cieľom identifikovať potenciálne kandidátne pojmy, vďaka čomu sa zlepšuje účinnosť algoritmu,
- Extrakcia charakteristík – vstupom je zoznam jednotlivých pojmov reprezentovaných súborom štatistických znakov,
- Odhad skóre slov – spája vlastnosti do jedného skóre, ktoré odráža dôležitosť pojmov,
- Generovanie n-gramov a odhad skóre kandidátnych kľúčových slov – vygeneruje kandidátne kľúčové slová (prostredníctvom n-gramovej konštrukčnej metodológie) a priradí im skóre na základe ich dôležitosti,
- Deduplikácia a ranking – porovnáva podobné kľúčové slová pomocou miery podobnosti vzdialenosti deduplikácie. Zoznam konečných kľúčových slov je následne zoradený podľa ich skóre relevantnosti.

Pre každý skúmaný dokument bolo extrahovaných 100 kľúčových slov. Na základe dosiahnutého skóre boli vybrané kľúčové slová, ktoré spĺňali zvolenú hranicu pre každú kategóriu dokumentov. Hranica bola zvolená na základe priemerného skóre získaného pre dokumenty z danej kategórie. Kľúčové slová, ktoré mali vyššie skóre ako priemer, boli z ďalšej analýzy vylúčené (nižšie skóre znamená zaujímavejšie kľúčové slovo). V tabuľke 2 sa nachádza početnosť extrahovaných kľúčových slov pre jednotlivé kategórie ako aj s počtom skúmaných dokumentov. Táto tabuľka obsahuje súčet všetkých kľúčových slov za dokumenty vrátane duplicitných kľúčových slov, ktoré boli rovnaké pre niektoré dokumenty z tej istej kategórie.

**Tabuľka 2** – Počet dokumentov a kľúčových slov extrahovaných pre skúmané kategórie a podkategórie webového portálu

| Kategória | Podkategória | Počet dokumentov | Počet kľúčových slov |
|---|---|---|---|
| /Pilar3 disclosure requirements/ | /financial_statement/ | 54 | 1367 |
| /Pilar3 disclosure requirements/ | /information_about_bank/ | 45 | 1554 |
| /Pillar3 related/ | /annual_reports/ | 17 | 1034 |
| /Pillar3 related/ | /financial_reports/ | 18 | 1134 |
| /Pillar3 related/ | /covered_bonds/ | 85 | 5939 |
| /Pillar3 related/ | /information_for_investors_ except_shareholders/ | 2 | 120 |
| /Pillar3 related/ | /information_for_shareholders/ | 5 | 312 |

### 4.1.4 ODHAD ČASU STRÁVENÉHO NA CIEĽOVÝCH STRÁNKACH

Dôležitým parametrom pri skúmaní čitateľnosti textu je odhad času tzv. cieľovej (obsahovej) stránky. Vo fáze identifikácie sedení pomocou metódy Reference Length (Kapusta et al. 2012b; Munk et al. 2015) sa stránky webového portálu v sedení rozdelia na navigačné a obsahové (cieľové) stránky. Cieľovými stránkami v prezentovanom experimente sú skúmané dokumenty, avšak pre ne nebolo možné určiť presný čas, ktorý na nich návštevníci strávili, a preto boli navrhnuté rôzne spôsoby odhadu tohto času. Nakoľko časové okno bolo vo fáze prípravy dát stanovené na 3 600 sekúnd (60 minút), jeden zo spôsobov vychádzal práve z daného časového okna. Ďalším spôsobom bolo použiť metriku čitateľnosti textu prostredníctvom času čítania (*reading_time*) jednotlivých dokumentov, doplnením priemerného času čítania dokumentu danej podkategórie alebo doplnením konkrétneho času čítania pre každý dokument. Navrhli sa štyri spôsoby doplnenia času stráveného na obsahových stránkach a vytvorili nové premenné v dátovom súbore:

- štandardný prístup bez doplnenia času (*length*),
- doplnenie 3 601 sekúnd pre všetky obsahové stránky v sedeniach (*length3601*),
- doplnenie priemerného času čítania pre dokumenty zo skúmaných podkategórií (*lengthRT_cat*),
- doplnenie konkrétneho času čítania pre jednotlivé dokumenty (*lengthRT_doc*).

Vzhľadom k tomu, že dĺžka času stráveného na stránke vstupovala do celého postupu odhadu úrovne záujmu používateľov, bol celý postup vykonaný pre všetky štyri prístupy doplnenia času obsahových stránok.

### 4.1.5 ODHAD ÚROVNE ZÁUJMU POUŽÍVATEĽOV

Úroveň záujmu návštevníka o extrahované kľúčové slová bola vypočítaná z času stráveného na stránke (Yao et al. 2017):

$$time_u(c_j, k_i) \begin{cases} \frac{length_u(c_j)}{m}, & \text{ak } k_i \text{ je v } c_j \\ 0, & \text{ak } k_i \text{ nie je v } c_j \end{cases},$$

kde $length_u(c_j)$ označuje dĺžku času, ktorú návštevník $u$ strávi na webovej časti kategórie $c_j$, ktorá obsahuje extrahované kľúčové slová $\{K_1, K_2, \dots, K_i\}$, pre celkový počet kľúčových slov $m$ vo všetkých kategóriách.

Celkový čas $sum_u$, ktorý návštevník stránky strávi nad určitým kľúčovým slovom $K_i$ v $KT_u$, je vypočítaný nasledovne:

$$sum_u(c_j, k_i) = \begin{cases} \sum_{i=j}^{f} time_u(c_j, k_i), & \text{ak } c_j \text{ je v } KT_u \\ 0, & \text{ak } c_j \text{ nie je v } KT_u \end{cases}.$$

Cieľom daného postupu je predpovedať prístup k určitému kľúčovému slovu, ktorý je založený na informáciách o navigácií návštevníkov. Ak má používateľ slovo, ktoré je pre neho zaujímavé či podstatné, opakovane navštevuje niektoré stránky, na ktorých s vysokou pravdepodobnosťou strávi viac času ako na iných stránkach. Na predpovedanie prístupu návštevníka webu k určitému kľúčovému slovu bol použitý model UISM (User interest structure model) (Yao et al. 2017), ktorý kombinuje dáta o obsahu, štruktúre a používaní webu a prepája všetky domény webu. UISM model je definovaný ako:

- súbor stavov: $Q = \{q_1, q_2, \dots, q_n\}$, počiatočný stav sa začína v $q_1$, každé $q_i$ reprezentuje webovú kategóriu,
- súbor kľúčových slov: $K = \{k_1, k_2, \dots, k_n\}$, $K$ obsahuje všetky kľúčové slová zo všetkých webových kategórií Q,
- pravdepodobnosť prechodu stavu $P_1(q \rightarrow q')$ medzi dvomi kategóriami, ktorá je definovaná nasledovne:

$$P_1(q \rightarrow q') = \frac{count\ (q \rightarrow q')}{count\ (q)},$$

kde $(q \rightarrow q')$ označuje cestu používateľa, ktorý najprv navštívil kategóriu $q$ a následne kategóriu $q'$. Yao et al. (2017) predstavili v rámci UISM dva rôzne prístupy k výpočtu $P_1$ na základe štruktúry webového portálu. Rozlišovali tzv. vertikálnu a horizontálnu štruktúru. V prípade vertikálnej štruktúry $count\ (q \rightarrow q')$ reprezentuje počet sedení, v ktorých

používateľ navštívil kategóriu $q'$ hneď po kategórií $q$. Teda obe kategórie sa musia nielen nachádzať v tom istom sedení, ale musia byť navštívené priamo po sebe. Na druhej strane, v prípade prístupu horizontálnej štruktúry, $count\ (q \rightarrow q')$ reprezentuje počet sedení, v ktorých sa nachádzajú obe kategórie $q$, $q'$, avšak nie sú obmedzené tým, že musia po sebe priamo nasledovať.

Vo všetkých stavoch $q$ existuje pravdepodobnosť rozdelenia $P_2\ (k_i|q)$ pre každé kľúčové slovo $k_i$ z $K$:

$$P_2\ (k_i|q) = \frac{\sum_{u=1}^{N} sum_u(q, k_i)}{\sum_{u=1}^{N}(\sum_{m=1}^{M} sum_u(q, k_i))},$$

ktorá sa označuje ako pravdepodobnosť symbolu pozorovania skrytého Markovovho modelu. Počíta sa na základe celkového času stráveného používateľom na danom kľúčovom slove na danej webovej kategórie. Predstavuje pravdepodobnosť záujmu o dané kľúčové slovo.

Pre sedenie $S^l$ ($l$ reprezentuje dĺžku sedenia) a záujem používateľa, je možné vyjadriť úroveň záujmu používateľa o dané kľúčové slovo $R\ (k|S_l)$, ktorého výpočet je nasledovný:

$$R\ (k|S^l) = P_1(q_{start} \rightarrow q_1) \times P_2(k|q_1) \times P_1(q_1 \rightarrow q_2) \times P_2(k|q_2) \times \ldots$$
$$\times P_1(q_{l-1} \rightarrow q_l) \times P_2(k|q_l)\ .$$

Ak je $R(k|S^l)$ väčšie alebo sa rovná $C$ – čo je hraničná hodnota spoľahlivosti, potom $R(k|S^l)$ je zaujímavé sedenie, pretože používatelia s rovnakým záujmom môžu pristupovať ku kategóriám sedenia. Hraničná hodnota spoľahlivosti bola nastavená podľa Yao et al. (2017) v intervale $10^{-3}$ až $10^{-7}$. V prípade klasického skrytého Markovovho modelu sa hraničná hodnota pohybuje v intervale $10^{-5}$ až $10^{-10}$.

### 4.1.6 ZHODNOTENIE PRÍSTUPU K ODHADU ÚROVNI ZÁUJMU POUŽÍVATEĽOV

Na základe vyššie spomenutého postupu bola vytvorená dátová matica, ktorá obsahovala horizontálnu/vertikálnu úroveň záujmu používateľov (UIH a UIV) pre jednotlivé extrahované kľúčové slová zo skúmaných dokumentov. Cieľom výskumu je prepojiť zložitosť textov so záujmom o navštevované webové kategórie. Z tohto dôvodu bolo nutné identifikovať vhodnú hraničnú hodnotu spoľahlivosti, prístup k úrovni záujmu používateľa (horizontálny alebo vertikálny) a prístup k odhadu času stráveného na cieľovej stránke.

Výsledky popisnej štatistiky ukázali, že dopĺňanie odhadu času stráveného na obsahových stránkach nemá vplyv na záujem používateľov. Z toho dôvodu bol zvolený prístup s doplnením

času na základe času čítania, tzn., ak používateľ hľadal daný dokument, tak pravdepodobne na dokumente strávil určitý čas, ktorý môže reprezentovať čas čítania daného dokumentu. Výsledky tiež ukázali, že vertikálny prístup k odhadu záujmu používateľov prináša najmenej užitočných informácií, pretože skúmané dokumenty sa v sedeniach nachádzajú v sekvencií za sebou zriedkavo. Z toho dôvodu bol do ďalšej analýzy vybraný iba horizontálny prístup, kedy sa dokumenty nachádzajú v sedeniach nezávisle od poradia návštevy.

Keďže navrhnutý postup odhadu záujmu používateľov autorov Yao et al. (2017) bol založený na kľúčových slovách a ich záujmu na základe hraničnej hodnoty spoľahlivosti, nebolo možné vyhodnotiť záujem používateľov o konkrétne dokumenty. Z toho dôvodu boli navrhnuté agregácie, ktoré mali reprezentovať záujem používateľov o jednotlivé dokumenty odhliadnuc od hraničnej hodnoty spoľahlivosti. Na základe výsledkov bol zvolený horizontálny prístup, ktorý vykazoval lepšiu rozlišovaciu schopnosť ako vertikálny prístup. Analogický postup odhadu úrovne záujmu používateľov je možné aplikovať nielen na dokumenty, ale aj na webové podkategórie alebo iné skúmané webové časti. Každé kľúčové slovo charakterizuje daný dokument iným spôsobom, a preto navrhnuté agregácie sa snažia zohľadniť váhu pre daný dokument. Agregácie boli navrhnuté na základe nasledovných vlastností s vytvorenými váhami:

- Záujem používateľa na základe špecificity kľúčových slov v dokumente: $UIH_{s_i} = \sum_{k=1}^{K_i} specificity_k * UIH_k$, kde $UIH$ je odhad úrovne záujmu používateľa o dané kľúčové slovo $k$ a $specificity_k = \ln \frac{n_i}{N}, i = 1, \ldots, N$, kde $N$ je celkový počet skúmaných dokumentov a $n$ je počet dokumentov, ktoré obsahujú dané kľúčové slovo $k$.
- Záujem používateľa na základe hustoty kľúčových slov v dokumente: $UIH_{d_i} = \sum_{k=1}^{K_i} density_k * UIH_k$, kde $UIH$ je odhad úrovne záujmu používateľa o dané kľúčové slovo $k$ a $density_k = \frac{n_i}{N}, i = 1, \ldots, N$, kde $N$ je celkový počet skúmaných dokumentov a $n$ je počet dokumentov, ktoré obsahujú dané kľúčové slovo $k$.

Napríklad kľúčové slovo „client“ sa nachádza v 83 dokumentoch z 226 skúmaných dokumentov. Prirodzeným logaritmom podielu týchto čísiel sa získa hodnota váhy -1,00169. Odhadnutá hodnota záujmu používateľov UIH pre toto kľúčové slovo je 0,002064. Vynásobením týchto dvoch hodnôt sa dosiahne hodnota -0,0020675, ktorá určuje špecificitu pre skúmané kľúčové slovo. Hodnota váhy v absolútnych číslach je bližšie k nule, ak je kľúčové

slovo všeobecnejšie. Analogicky sa postupuje aj v prípade hustoty, čo v tom prípade je hodnota 0,000758. Hodnota váhy je vyššia, čím kľúčové slovo je všeobecnejšie, tzn. nachádza sa vo viacerých dokumentoch. Na základe absolútnych čísiel špecifickosti je možné vidieť, že výsledky sú podobné, ale skóre sú dva opačné extrémy. Týmto spôsobom sa získali váhy pre jednotlivé kľúčové slová v dokumentoch, ktoré sa následne agregovali pre každý dokument.

Ako ďalšie ukazovatele preferencií používateľov sa vypočítali podpora (*support*), entropia (Shannon 1948) a počet sedení, v ktorých je dokument cieľová stránka. Vychádzalo sa z identifikovaných sedení pomocou metódy Reference Length (Kapusta et al. 2012b; Munk et al. 2015) a podpory daných dokumentov v identifikovaných sedeniach. Podpora vyjadruje záujem o navštevované dokumenty. Slúži ako referencia v prezentovanom experimente, v ktorom je snahou identifikovať ďalšie ukazovatele, ktoré by mali korelovať s validným kritériom. Jedným z takýchto ukazovateľov je entropia sedení (*entropy*) s dôrazom na zloženie jednotlivých sedení. Neusporiadanosť je charakterizovaná rôznorodosťou návštevností používateľa rôznych kategórií webového portálu počas sedenia. Pričom pri výpočte entropie sedenia sa vychádzalo z entropie definovanej (Shannon 1948): $entropy_s = -\sum_{x \in X} p(x) \log_n p(x)$, kde $n$ je počet stránok v sedení $s$ a $p(x)$ je pravdepodobnosť výskytu stránky $x$ v sedení. Ak sa entropia rovná 1, sedenia obsahovali webové stránky z rôznych kategórií. Ak sa entropia rovná 0, potom všetky stránky v sedení pochádzali z jednej kategórie, pričom používateľ hľadal cielene informáciu z danej kategórie. Entropia bola určená pre každé sedenie a pre každý dokument pričom výslednou entropiou bola priemerná hodnota pre každý dokument.

Druhým ukazovateľom bola premenná *target*, ktorá reprezentovala počet sedení, v ktorých bol dokument cieľovou (obsahovou) stránkou. Cieľová stránka je stránka, ktorej čas strávený používateľom na stránke je väčší než hraničný čas (Kapusta et al. 2012b), čo indikuje záujem používateľa o obsah a cieľ jeho hľadania. V prípade prezentovaného experimentu, cieľom návštevníkov bol obsah skúmaných dokumentov.

### 4.1.7 VÝPOČET SKÓRE ČITATEĽNOSTI A ZLOŽITOSTI DOKUMENTOV

V prezentovanom experimente habilitačnej práce sa aplikovalo niekoľko metrík zložitosti a čitateľnosti s motiváciou identifikovať tie, ktoré budú najlepšie charakterizovať zložitosť a čitateľnosť dokumentov z povinne zverejňovaných informácií súvisiacich s Pilier 3. Metriky boli rozdelené do viacerých kategórií pre lepšiu prehľadnosť. V prípade niektorých kategórii dochádza k prekrytiu (spôsobené prekrytím kategórií), preto boli niektoré metriky použité

duplicitne, avšak v takom prípade boli počítané pomocou iného nástroja. Skúmané metriky boli implementované pomocou jazyka Python alebo boli použité externé nástroje (Cvrček et al. 2020; Lu 2012; 2011; 2010; Lu a Ai 2015). Podrobnejší popis jednotlivých kategórií bol uvedený v kapitole 1.2.2, pre zvýšenie prehľadnosti metodiky sa na tomto mieste uvádza iba zoznam skupín metrík čitateľnosti a zložitosti textu: Charakteristiky textu [char][2], Čitateľnosť [read][2], Lexikálna variácia [lex_var][2], Lexikálna bohatosť [lex_rich][2], Lexikálna rôznorodosť [lex_div][2], Lexikálna sofistikovanosť [lex_sop][2], Expertné metriky [expert][2], Slabiky [syl][2], Podiel slovných druhov [pos_ratio][2], Ostatné charakteristiky [other][2].

Po realizovaní všetkých spomenutých krokoch bol výsledkom dátový súbor, ktorý obsahoval skúmané dokumenty a k nim odhad úrovne záujmu o dokumenty, časové charakteristiky, počet sedení, v ktorých je dokument cieľovou stránkou, entropiu sedení a charakteristiky na základe zložitosti a čitateľnosti textu.

## 4.2 VÝSLEDKY EXPERIMENTU

Analýza závislosti medzi úrovňou záujmu používateľov a čitateľnosťou/zložitosťou textu pozostávala z viacerých krokov (Benko et al. 2024c) (Príloha M). V prvom kroku bolo nutné vyhodnotiť, ktorá z agregácii úrovne záujmu používateľov najlepšie vystihuje skúmané dokumenty v kombinácií s časovými charakteristikami a ostatnými používateľsky zameranými charakteristikami (počet sedení, v ktorých je dokument cieľovou stránkou a entropiou sedení) a podporou.

*Bol stanovený predpoklad, že ukazovatele preferencií používateľov (počet sedení, v ktorých je dokument cieľová stránka, entropia sedení a úroveň záujmu používateľa zohľadňujúci špecificitu a hustotu kľúčového slova v dokumente) budú relevantné, v zmysle rozlišovacej schopnosti a miery vysvetlenia návštevnosti.*

V ďalšom kroku sa porovnávala agregácia s premennou podpory (support), ktorá vyjadruje záujem o navštívené dokumenty a slúži ako validné kritérium. Výsledky agregácií (Benko et al. 2024c) ukázali, že počet sedení, v ktorých je dokument cieľová stránka; entropia sedení a úroveň záujmu používateľa zohľadňujúci hustotu kľúčového slova v dokumente, sú relevantné. Nové ukazovatele navrhnuté v experimente, reprezentujúce preferenciu používateľa (počet sedení, v ktorých je dokument cieľová stránka a entropia sedení) dosiahli rovnakú rozlišovaciu schopnosť, pričom v oboch prípadoch bola dosiahnutá veľmi veľká

[2] vektor skúmaných premenných [*x*], kde *x* označuje príslušnú kategóriu čitateľnosti alebo zložitosti textu

štatisticky významná korelácia. Oba ukazovatele sa ukázali ako zaujímavé aj z hľadiska miery vysvetlenia návštevnosti. Počet sedení, v ktorých je dokument cieľová stránka vysvetľuje 96 % variability podpory a entropia sedení vysvetľuje 79 % variability podpory.

Zaujímavý výsledok priniesli neparametrické odhady v prípade počtu relácií, v ktorých je dokument cieľovou stránkou a v prípade entropie sedení, kde podkategória výročných správ (*annual reports*) dosiahla najvyšší priemer poradí (mean rank) v prípade počtu sedení, v ktorých je dokument cieľovou stránkou a najnižší priemer poradí v prípade entropie sedení (na základe absolútnych výsledkov špecifickosti možno vidieť, že výsledky sú podobné, ale skóre je opačné voči počtu sedení). V prípade ostatných skúmaných ukazovateľov, nedosahovala podkategória výročných správ také vysoké hodnoty.

*Bol stanovený predpoklad, že najvýkonnejší (v zmysle rozlišovacej schopnosti a miery vysvetlenia návštevnosti) ukazovateľ preferencií používateľov z hľadiska času stráveného na stránke bude čas strávený na stránke zohľadňujúci čas čítania dokumentu.*

Globálna nulová hypotéza sa zamieta na hladine významnosti 0,001 (lengthRT_doc_mean: $F(6, 219) = 3{,}538$, $p < 0{,}001$; $H(6, N = 226) = 122{,}387$, $p < 0{,}001$), ktorá tvrdí, že neexistuje žiadny štatisticky významný rozdiel v preferenciách používateľa vyjadrený časom stráveným na stránke pri zohľadnení času čítania dokumentu medzi skúmanými podkategóriami obsahu. To znamená, že medzi skúmanými podkategóriami obsahu boli identifikované rozdiely v časových hodnotách. V prípade času stráveného na stránke pri zohľadnení času čítania dokumentu bola v prípade podkategórie *annual reports* dosiahnutá štatisticky významne najväčšia preferencia používateľov ($p < 0{,}05$). V prípade ďalších časových premenných sa globálna nulová hypotéza nezamieta ($p > 0{,}05$).

### 4.2.1 VALIDITA ZLOŽITOSTI A ČITATEĽNOSTI TEXTU

Nasledujúcim krokom bolo zameranie sa na validitu metrík čitateľnosti a zložitosti. Kvôli prehľadnosti boli skúmané metriky rozdelené do desiatich kategórií podľa spoločných vlastností. Vo výskumoch zameraných na ekonomické texty sa používajú rôzne metriky čitateľnosti, pričom časť autorov preferuje metriku Gunning Fog Index a druhá skupina autorov preferuje metriku LIX alebo RIX (Ebaid 2023; Guay et al. 2015; Moreno a Casasola 2016). Z toho dôvodu bola v experimente vytvorená kategória expertných metrík LIX, RIX a LinsearWrite. Metrika Gunning Fog Index bola zaradená medzi metriky čitateľnosti, s ktorými sa často kombinuje. Bola vypočítaná korelácia medzi vektormi jednotlivých skupín

metrík a skupiny expertných metrík, ktorých vektor premenných [expert] predstavuje validné kritérium. Výsledky poukazujú na fakt, že všetky kategórie sú štatisticky významné, avšak najvyššiu mieru závislosti s expertnými metrikami dosahujú metriky čitateľnosti (skupina obsahuje aj metriku Gunning Fog Index) a metriky základných charakteristík textu. Metriky čitateľnosti a expertné metriky dokážu spoločne odhaliť podobné znaky zložitosti a čitateľnosti finančných textov. Z toho dôvodu metrika Gunning Fog Index nebola zaradená medzi expertné metriky, čo nemalo vplyv na výsledky experimentu. Za validné kritérium tak mohli byť zvolené akékoľvek metriky čitateľnosti.

Posledná skupina metrík (ostatné) obsahovala metriky zložitosti textu rôznorodého charakteru, preto k nej nebolo možné pristupovať ako k vektoru konzistentných metrík. Porovnanie bolo realizované pomocou viacnásobnej analýzy v kombinácií premenná vs. vektor expertných metrík. Vo všetkých prípadoch bol viacnásobný korelačný koeficient medzi jednotlivými premennými a vektorom expertných metrík štatisticky významný na hladine významnosti 0,001, okrem poslednej (*other_Size_in_kB*), ktorá bola na hladine významnosti 0,01.

Výsledkom validity metrík čitateľnosti a zložitosti textu je, že všetky skupiny skúmaných metrík sú použiteľné. Navrhnutá metrika *other_eawl* založená na slovníku ekonomických slov dosiahla štatisticky významnú strednú mieru závislosti. Od nej odvodená metrika *other_eawl_unique*, ktorá brala do úvahy iba jedinečné slová dosiahla dokonca štatisticky významnú veľkú mieru závislosti voči vektoru expertných metrík. Ukázalo sa, že tieto metriky majú potenciál charakterizovať čitateľnosť ekonomických textov.

### 4.2.2 ANALÝZA ČITATEĽNOSTI/ZLOŽITOSTI TEXTOV V KONTEXTE PREFERENCIÍ POUŽÍVATEĽA

Hlavným cieľom výskumu bolo zistiť, aký je vzťah medzi metrikami zložitosti/čitateľnosti a preferenciami používateľa (počet sedení, v ktorých je dokument cieľová stránka, entropia sedení, čas strávený na stránke zohľadňujúci čas čítania dokumentu a úroveň záujmu používateľa na základe hustoty kľúčových slov v dokumente). Analýza závislostí bola vykonaná pre každú skupinu metrík zvlášť, v kombinácií s ukazovateľmi preferencií používateľa. Viacnásobná analýza sa aplikovala pre skupinu metrík a následne bola vypočítaná jednorozmerná analýza pre jednotlivé metriky danej skupiny v kombinácií s ukazovateľmi preferencií používateľa.

Podrobné výsledky analýzy pre všetky skupiny metrík sa nachádzajú v článku (Benko et al. 2024c) (Príloha M). Na základe dosiahnutých výsledkov boli identifikované skupiny metrík, ktoré dosiahli najvýznamnejší vzťah so skúmanými ukazovateľmi preferencií používateľa.

**Úroveň záujmu používateľa vs. čitateľnosť/zložitosť textu**

Metriky skupiny [pos_ratio] dosiahli najvyššiu mieru závislosti s úrovňou záujmu používateľa na základe hustoty kľúčových slov v dokumente (*Multiple R* = 0,848; *Multiple R2* = 0,720; *Adjusted R2* = 0,697), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Zaujímavé výsledky dosiahli v tejto skupine metriky popisujúce podiel vlastných mien a slovies. Zo skupiny ostatných metrík [other] sa ukázalo, že viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,01 pre metriku *eawl_unique* s úrovňou záujmu používateľa na základe hustoty kľúčových slov v dokumente. Výsledky ukázali, že úroveň záujmu používateľa na základe hustoty kľúčových slov v dokumente súvisí hlavne s podielom čísloviek (*numerals*), podielom vlastných mien (*proper noun*), podielom slovies (*verbs*) a podielom unikátnych ekonomických slov (*eawl_unique*). Dá sa predpokladať, že vyšší počet vlastných mien ($r$ = 0,4) a slovies ($r$ = 0,2) môže znamenať vyššiu úroveň záujmu používateľa. Naopak nižší počet čísloviek ($r$ = -0,2) v texte naznačuje vyššiu úroveň záujmu používateľa. Rovnako, navrhnutá metrika podielu unikátnych ekonomických slov dokáže zachytiť úroveň záujmu používateľa. Dôvodom môže byť fakt, že generované kľúčové slová sú prevažne vlastné mená, ktoré sa nachádzajú aj v zozname ekonomických slov. Z hľadiska interpretácie metrík zložitosti, je vhodné používať metriku *eawl_unique* z viacerých dôvodov:

- nevyžaduje si použitie nástroja tretej strany, pomocou ktorého je nutné vykonať časovo náročnú morfologickú anotáciu na identifikáciu počtu slovných druhov;
- zoznam ekonomických slov sa môže rozširovať, a tým neustále zlepšovať presnosť metriky;
- ekonomické slová sú zrozumiteľné pre odborníkov na finančníctvo, na druhej strane môžu byť menej zrozumiteľné pre bežných stakeholderov. Ukázalo sa, že vyšší podiel týchto slov značí vyšší záujem používateľov ($r$ = 0,2), čo môže naznačovať, že o dané dokumenty majú záujem hlavne odborníci na finančníctvo.

Napríklad v prípade podkategórie *information-for-shareholders-not-investors* dokument „slovakiavub_presentation_for_investors_062018“ mal nadpriemernú úroveň záujmu 0,03029

a *eawl_unique* hodnotu 0,08, t. j. 8% unikátnych ekonomických slov v dokumente. Na druhej strane dokument podkategórie *annual-reports* s názvom „vubannualreport14“ mal podpriemernú úroveň záujmu 0,00001 a *eawl_unique* hodnotu 0,02, čo znamená, že nižšie percento ekonomických slov môže znižovať záujem o daný dokument medzi expertami na danú oblasť.

**Čas strávený na stránke vs. čitateľnosť/zložitosť textu**

Metriky skupiny [char] a [lex_rich] dosiahli najvyššiu mieru závislosti s časom stráveným na stránke zohľadňujúcim čas čítania dokumentu ([char]: *Multiple R* = 0,566; *Multiple R2* = 0,321; *Adjusted R2* = 0,269; [lex_rich]: *Multiple R* = 0,558; *Multiple R2* = 0,311; *Adjusted R2* = 0,258), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Výsledky ukázali, že čas strávený na stránke zohľadňujúci čas čítania dokumentu značne súvisí s početnosťou slovných jednotiek textu, ako je počet znakov ($r = 0,3$), tokenov ($r = 0,2$) a jedinečných tokenov ($r = 0,2$). Potvrdili to aj výsledky skupiny lexikálnej bohatosti, kde obe metriky, zachytávajúce mieru rôznych slovných druhov v texte, sa ukázali ako zaujímavé ($r = -0,2$). V prípade metrík zo skupiny [other] je pozitívne, že metrika času čítania *other_RT* ($r = 0,3$) súvisí s časom stráveným na stránke zohľadňujúcim čas čítania dokumentu. Výsledky tiež ukázali, že vyšší čas strávený na stránke súvisí s väčším rozsahom (väčšou dĺžkou) textov, čo potvrdzujú aj metriky súvisiace s časom čítania a veľkosťou daného dokumentu ($r = 0,3$). Zaujímavým zistením bolo, že nižšia miera rôznych slovných druhov ($r = -0,2$) v textoch zvyšuje čas strávený na stránkach.

Napríklad v prípade podkategórie *annual-reports* mal dokument „ar_2017_en_final _web“ nadpriemernú úroveň priemerného času stráveného na stránkach 188 sekúnd a hodnoty metrík pre daný dokument dosahovali nadpriemerné skóre: počet tokenov = 106317, počet viet = 3378, čas čítania = 7484 sekúnd.

**Počet sedení, v ktorých je dokument cieľová stránka vs. čitateľnosť/zložitosť textu**

Metriky skupiny [char] dosiahli najvyššiu mieru závislosti s počtom cieľových stránok (*Multiple R* = 0,628; *Multiple R2* = 0,394; *Adjusted R2* = 0,348), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Výsledky ukázali, že počet sedení, v ktorých je dokument cieľová stránka, podobne ako čas strávený na stránke, súvisí hlavne s početnosťou slovných jednotiek textu, ako počet znakov ($r = 0,4$), tokenov ($r = 0,4$) a jedinečných tokenov ($r = 0,4$). Väčšia dĺžka textu reprezentovaná dĺžkou vety ($r = 0,3$) a vyšší počet znakov, tokenov a aj jedinečných tokenov v dokumente má vplyv

na väčší počet sedení, v ktorých je dokument cieľová stránka. Tieto výsledky potvrdzujú aj metriky zo skupiny [other], či už veľkosť súboru ($r = 0,4$) alebo čas čítania ($r = 0,4$). Podobne aj metrika h-point (v prípade väčších textov ($r = 0,3$)) a entropia textu ($r = 0,2$) potvrdzujú dosiahnuté výsledky: čím väčšia diverzita slovníka je v dokumente identifikovaná, tým je väčší počet sedení, v ktorých je dokument cieľová stránka.

Napríklad v prípade podkategórie *information-about-bank* mal dokument „pillar-iii_15_12_en“ nadpriemernú úroveň počtu sedení (35), v ktorých je dokument cieľová stránka. Hodnoty metrík pre daný dokument dosahovali nadpriemerné skóre: počet tokenov = 34367, počet viet = 3952, čas čítania = 2526 sekúnd, h-point = 54 a entropia textu = 9,59.

**Entropia sedení vs. čitateľnosť/zložitosť textu**

Metriky skupín [char] a [pos_ratio] dosiahli najvyššiu mieru závislosti s entropiou sedení ([char]: *Multiple R* = 0,646; *Multiple R2* = 0,418; *Adjusted R2* = 0,373; [pos_ratio]: *Multiple R* = 0,630; *Multiple R2* = 0,397; *Adjusted R2* = 0,348), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Výsledky ukázali, že entropia sedení taktiež súvisí s početnosťou znakov ($r = -0,5$), tokenov ($r = -0,4$) a jedinečných tokenov ($r = -0,4$) v dokumente. Dokazujú to metriky zo skupiny [pos_ratio], kde vyšší podiel vlastných mien ($r = -0,4$) v dokumente znižuje entropiu sedení, t. j. návštevník hľadá v sedeniach súvisiace informácie. Výsledky tiež ukázali, že nižšia entropia sedení zodpovedá väčšej dĺžke textov. Dokazujú to metriky času čítania ($r = -0,5$) a veľkosti súboru ($r = -0,2$), ako aj väčšia entropia dokumentu ($r = -0,3$), ktorá reprezentuje mieru rôznych slovných druhov v dokumente.

Napríklad v prípade podkategórie *financial-reports* mal dokument „polrocna-financna-sprava-za-rok-2012“ nadpriemernú úroveň entropie sedení = 0,99990 a hodnoty metrík pre daný dokument dosahovali podpriemerné skóre: počet tokenov = 30354, počet viet = 396, čas čítania = 1188 sekúnd, h-point = 44 a entropia textu = 8,64.

# ZÁVER

Habilitačná práca bola zameraná na prepojenie zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3 zverejňovanými na stránkach komerčných bánk a správanie sa stakeholderov na webových stránkach. Realizované experimenty boli smerované na skúmanie správania sa stakeholderov na webovom portály bankovej inštitúcie. Ukázalo sa, že príprava dát má vplyv na kvalitu analýzy dát a získavanie znalostí z webu (Munk, Pilkova, Benko et al. 2021b; Svec, Benko et al. 2020), čo viedlo k vytvoreniu metodiky, ktorá bola základom pre všetky vykonané experimenty (Munk, Pilkova, Benko et al. 2021a). Výsledky viacerých experimentov potvrdili (Pilková, Munk, Benko et al. 2021a; Munk, Pilkova, Benko et al. 2021c; Pilková, Munk, Blažeková, Benko 2021b), že zverejňovanie povinných informácií Pilier 3, nie je potrebné v priebehu celého roka. Preukázali to výsledky na báze rôznych časových premenných, či už týždňov alebo kvartálov. Kombináciou rôznych metód realizovaných v experimentoch bolo dosiahnuté zlepšenie výsledkov získaných z údajov z logovacieho súboru webového portálu, čo prispelo k lepšiemu pochopeniu správania sa stakeholderov v prípade informácií súvisiacich s Pilier 3. Analýza zameraná na skúmanie času stráveného na webových stránkach (Blažeková, Benko et al. 2021) ukázala, že záujem iba o povinne zverejňované informácie nie je až taký vysoký. Stakeholderov zaujíma širší kontext informácií o banke a výročných správach. Na základe dosiahnutých výsledkov boli stanovené odporúčania, ktoré môžu zvýšiť záujem stakeholderov o informácie Pilier 3.

Skúmanie obsahu webu vo forme dokumentov a práca s nimi má podstatný vplyv na ďalší výskum, ktorý sa zameral na jazykovú zložitosť a čitateľnosť textov. Na základe stanovených odporúčaní o používaní anglického jazyka ako jednotného jazyka pre zverejňovanie informácií, bolo predefinované, že v prezentovanom experimente habilitačnej práce sa budú skúmať iba anglické texty obsahujúce informácie Pilier 3. Z toho dôvodu bolo ďalšie smerovanie zamerané na skúmanie kvality strojového prekladu, keďže v súčasnosti veľké korporácie používajú na lokalizáciu svojich produktov pre daný región strojový a nie humánny preklad. Výsledky experimentu (Benko et al. 2022) indikujú, že nie všetky automatické metriky založené na lexikálnej podobnosti (n-gramoch alebo vzdialenosti editácie) by mali byť implementované do modelu určovania kvality MT, či už ekonomických alebo iných typov textov prekladaných z anglického jazyka do flektívnej slovenčiny. V ďalších experimentoch sa pracovalo prevažne s metrikami, ktoré boli vhodné pre jazykový pár slovenčina-angličtina. Ďalším prínosom pri analýze chybovosti strojových prekladov bolo prepojenie analýzy rezíduí na identifikáciu

konkrétnych segmentov alebo textov, v ktorých systémy strojového prekladu dosahujú vyššiu chybovosť (Benko et al. 2024a). Významným prínosom je schopnosť identifikovať segmenty a lingvisticky charakterizovať segmenty, v ktorých strojový preklad vykazoval chybovosť v zmysle adekvátnosti a plynulosti do slovenčiny. Výsledky výskumu (Benko et al. 2024b) ukázali, že používanie POS taggerov by mohlo byť v prípade slovenského jazyka prínosné. Štyri nástroje zo šiestich skúmaných taggerov dosiahli vysoký výkon, v zmysle presnosti, pri lingvistickom anotovaní do 15-pozičného tagsetu. Použitie RNNTagger (najefektívnejšieho nástroja) by malo byť preferované pre generovanie morfologických značiek pre slovenský jazyk. Dosiahnuté výsledky v oblasti spracovania textu poukázali na skutočnosť, že v prípade absentujúcich jazykových mutácií (lokalizácií), je možné vychádzať z výstupov systémov strojového prekladu a použiť ich strojové preklady. Zložitosť textov sa ukázala ako jeden z atribútov, ktorý môže tiež poukázať na chybovosť strojového prekladu (Benko et al. 2023). Experiment (Munkova, Munk, Benko et al. 2021b) bol zameraný na vplyv jazykovej zložitosti na úrovni slov a vetnej štruktúry, pričom za hlavný prínos je možné považovať navrhnutú metodiku, ktorá zohľadňuje miery jazykovej zložitosti, slovných druhov, frekventovaných tagsetov, asociačných pravidiel a ich sumarizácie.

Cieľom prezentovaného experimentu (Benko et al. 2024c) bolo navrhnúť metodiku zameranú na analýzu zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3, ktorý sa podarilo splniť. Hlavným prínosom habilitačnej práce do odboru je navrhnutie metodiky prepájajúcej zdroje dát o používaní, obsahu a štruktúre webu. Metodika viedla k návrhu ukazovateľov preferencií používateľov na webových portáloch komerčných bánk a umožnila skúmať, či zložitosť a čitateľnosť povinne zverejňovaných informácií má vplyv na vytvorenie preferencie používateľov. Počas realizovania experimentu boli navrhnuté dve metriky zložitosti textu *eawl* a *eawl_unique*, ktoré sú vhodné pre ekonomické texty. Za hlavný prínos experimentu je možné považovať ukazovatele preferencií používateľov z hľadiska návštevnosti: entropia sedení, počet sedení, v ktorých dokument je cieľová stránka a úrovne záujmu používateľa na základe hustoty kľúčových slov v dokumente, a z hľadiska času: čas strávený na stránke zohľadňujúci čas čítania dokumentu. Ukazovatele orientované na návštevnosť boli porovnané s podporou, ktorá vystupovala ako referencia. Preukázal sa súvis medzi ukazovateľmi preferencií používateľov a metrikami zložitosti alebo čitateľnosti. Skupina metrík základných charakteristík, ako sú rôzne početnosti tokenov, viet, znakov dosiahla najvyššiu mieru závislosti s preferenciou používateľov na základe entropie sedení; počtu sedení, v ktorých je dokument cieľová stránka a času stráveného na stránke zohľadňujúceho čas čítania dokumentu.

Najvýznamnejší prínos habilitačnej práce spočíva v prepojení oblasti spracovania prirodzeného jazyka a používateľských preferencií, ktoré reprezentujú doménu webu.

Dĺžka textu hrá podstatnú úlohu v jeho zložitosti a čitateľnosti. Rozsiahlejšie dokumenty obsahujú väčšie množstvo informácií, a preto sú podľa dosiahnutých výsledkov pre stakeholderov zaujímavejšie a preferujú ich pred krátkymi dokumentami. Z hľadiska ukazovateľa úrovne záujmu používateľa na základe hustoty kľúčových slov v dokumente sa preukázala ako zaujímavá navrhnutá metrika *eawl_unique*, ktorá bola navrhnutá za účelom charakterizovať zložitosť ekonomických textov.

Zaujímavý výsledok priniesli neparametrické odhady v prípade počtu sedení, v ktorých je dokument cieľovou stránkou a v prípade entropie sedení, kde podkategória výročných správ dosiahla najvyšší priemer poradí v prípade počtu sedení, v ktorých je dokument cieľovou stránkou a najnižší priemer poradí v prípade entropie sedení. Potvrdzujú to aj výsledky výskumu správania stakeholderov na portáli skúmanej bankovej inštitúcie (Blažeková, Benko et al. 2021; Pilková, Munk, Blažeková, Benko 2021b; Pilková, Munk, Benko et al. 2021a; Munk, Pilkova, Benko et al. 2021c).

Z hľadiska čitateľnosti a zložitosti textu sa ukázalo, že všetky kategórie metrík čitateľnosti/zložitosti sú štatisticky významné, pričom najvyššiu mieru závislosti s expertnými metrikami dosahujú metriky čitateľnosti a metriky základných charakteristík textu. Metriky čitateľnosti a expertné metriky dokážu spoločne popísať podobné znaky zložitosti a čitateľnosti finančných textov. Viac ako polovica skúmaných dokumentov (151 z 226) dosiahla skóre Gunning Fog Index, ktoré reprezentuje úroveň absolventov vysokej školy (*index* > 17), čo naznačuje vysoko odborné dokumenty. Podobné výsledky boli dosiahnuté aj pre metriku LIX (*index* > 56), ktorá viac ako polovicu dokumentov (157 z 226) identifikovala ako odborné texty. Na druhej strane v kombinácií metrík čitateľnosti/zložitosti s ukazovateľmi preferencií používateľov, v prípade metriky Gunning Fog Index sú viacnásobné koeficienty korelácie štatisticky významné na hladine významnosti 0,01 len v spojení s entropiou sedení. V prípade ostatných ukazovateľov sú viacnásobné koeficienty korelácie štatisticky nevýznamné. V experimente bolo implementovaných 110 metrík zložitosti/čitateľnosti zaradených do 10 skupín. Výsledky analýzy identifikovali iba jednu metriku (*pos_ratio_PROPN*), pre ktorú boli koeficienty korelácie štatisticky významné v prípade všetkých ukazovateľov preferencií používateľov. Znamená to, že z hľadiska preferencií používateľov, podiel vlastných mien v dokumentoch zvyšuje záujem o tieto dokumenty zo strany stakeholderov.

Dosiahnuté výsledky potvrdzujú výsledky predchádzajúcich štúdií (Pilková et al. 2021a; Munk et al. 2021c). Prezentovaný výskum nadviazal na predchádzajúce zistenia veľmi nízkeho záujmu o informácie Pilier 3 zo strany investorov komerčných bánk pôsobiacich v strednej a východnej Európe. Z doterajších výsledkov štúdií možno vyvodiť, že skupina klientov, ktorí sa o tieto informácie zaujímajú v tomto type komerčných bánk, sú tí, ktorí majú v banke nepoistené vklady (právnické osoby, fyzické osoby s vkladmi nad 100 tis. EUR). Ako však ukazujú výsledky prezentovaného výskumu, títo klienti majú väčší záujem o menej náročné a čitateľnejšie texty, ako sú výročné správy. Keďže informácie Pilier 3 a ďalšie dokumenty Pilier sú zložitejšie a ťažšie čitateľné, majú o ne menší záujem. To vedie k dôležitému záveru pre regulátorov: zvýšenie záujmu o tieto informácie v tomto type bánk si vyžaduje nájsť spôsoby ich prezentácie, aby boli menej zložité a čitateľnejšie.

Limitácií hlavného výskumu je niekoľko. Prvou je počet extrahovaných dokumentov, ktorých bolo 226, z toho v niektorých skúmaných podkategóriách bolo možné extrahovať iba niekoľko dokumentov. Skúmané dáta o používaní webu a obsahu webu pochádzali z roku 2018, ktorý je považovaný za najstabilnejšie obdobie, keďže nešlo o turbulentné obdobie. Nevýhoda skúmania starších dát spočíva v nedostupnosti všetkých dát a v zmene štruktúry webového portálu. Získavanie obsahu webu je možné cez archívne záznamy webu, ktoré ukladajú webovú stopu z daného obdobia. Avšak ani to nemusí garantovať použiteľnosť extrahovaných dokumentov.

Prínos habilitačnej práce nie je iba do oblasti bankovníctva, ale aj do vzdelávania. V rámci predmetov magisterského štúdia Objavovanie znalostí a Hĺbkováanalýza dát vyučovanými autorom práce, študenti pracujú s rôznymi zdrojmi dát. Študenti sa učia ako funguje proces získavania znalostí práve pomocou domény webu, pretože táto oblasť ponúka najlepšie zdroje dát – štruktúrované aj neštruktúrované. Úlohou študentov je pochopiť procesy prípravy dát a ich následné spracovanie pre potreby analýzy dát. Pri riešení semestrálnych projektov postupujú na základe metodiky (Munk, Pilkova, Benko et al. 2021a) a skúmajú okrem dát o používaní webu, aj dáta o obsahu webu, kde v textoch získaných z webových portálov hľadajú znalosti a skúmajú zložitosť textu.

Ďalšie smerovanie výskumu sa zameria na porovnanie ukazovateľov preferencií používateľov a čitateľnosti dokumentov naprieč obdobím viacerých rokov s ohľadom na turbulentné obdobia (globálna finančná kríza, pandémia a pod.) a revíziu zverejňovaných informácií.

# ZOZNAM POUŽITEJ LITERATÚRY


ANDERSON, Jonathan, 1983. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*. **26**(6), 490–496.

ARNAUD, Pierre J.L., 1992. Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In: Pierre J.L. ARNAUD a H. BÉJOINT, ed. *Vocabulary and Applied Linguistics*. London, UK: Palgrave Macmillan.

AWAN, Malik Daler Ali, Nadeem Iqbal KAJLA, Amnah FIRDOUS, Mujtaba HUSNAIN a Malik Muhammad Saad MISSEN, 2021. Event classification from the Urdu language text on social media. *PeerJ Computer Science* [online]. **7**, e775. ISSN 2376-5992. doi:10.7717/peerj-cs.775

AYE, Theint Theint, 2011. Web log cleaning for mining of web usage patterns. *2011 3rd International Conference on Computer Research and Development* [online]. **2**, 490–494. doi:10.1109/ICCRD.2011.5764181

BENKO, Ľubomír a Lucia BENKOVÁ, 2022. Comparison of Novel Approach to Part-Of-Speech Tagging of Slovak Language. In: *DIVAI 2022 – The 14th international scientific conference on Distance Learning in Applied Informatics*. Štúrovo, Slovakia: Wolters Kluwer, s. 327–333.

BENKO, Ľubomír, Lucia BENKOVA, Dasa MUNKOVA, Michal MUNK a Danylo SHULZENKO, 2022. Error Classification Using Automatic Measures Based on n-grams and Edit Distance. In: *Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2022* [online]. Springer, Cham, s. 345–356. doi:10.1007/978-3-031-20319-0_26

BENKO, Ľubomír, Petra BLAŽEKOVÁ, Michal MUNK a Anna PILKOVÁ, 2020. Time Spent on Web Page as an Indicator of Interest. In: *DIVAI 2020 The 13 th international scientific conference on Distance Learning in Applied Informatics*. s. 489–497. ISBN 9788075988416.

BENKO, Ľubomír, Dasa MUNKOVÁ a Michal MUNK, 2023. Relationship Between Linguistic Complexity and MT Errors in the Context of Inflectional Languages. In: *Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2023* [online]. Springer, Cham, s. 546–557. doi:10.1007/978-3-031-42430-4_45

BENKO, Ľubomír, Dasa MUNKOVA, Michal MUNK, Lucia BENKOVA a Petr HAJEK, 2024a. The use of residual analysis to improve the error rate accuracy of machine translation. *Scientific Reports (v recenznom konaní)*.

BENKO, Ľubomír, Dasa MUNKOVA, Mária PAPPOVÁ a Michal MUNK, 2024b. Comparison of various approaches to tagging for the inflectional Slovak language. *PeerJ Computer Science (v recenznom konaní)*.

BENKO, Ľubomir, Anna PILKOVA, Michal MUNK a Slavka ELEY, 2024c. Pillar 3: The impact of language complexity on the preferences of commercial bank website users. *Expert Systems with Applications (v recenznom konaní)*.

BERENDT, Bettina, Bamshad MOBASHER, Miki NAKAGAWA a Myra SPILIOPOULOU, 2003. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. *Lecture Notes in Computer Science* [online]. **2703**, 159–179. doi:10.1007/978-3-540-39663-5_10

BESSENYEI, Gabor, 2017. Neural Machine Translation: The Rising Star. *Memsource* [online] [vid. 2022-10-04]. Dostupné z: https://www.memsource.com/blog/2017/09/19/neural-machine-translation-the-rising-star/?utm_source=mailchimp&utm_medium=email&utm_content=blog_article

BJÖRNSSON, Carl Hugo, 1968. *Lasbarhet*. Stockholm, Sweden: Bokforlaget Liber.

BLAŽEKOVÁ, Petra, Ľubomír BENKO, Anna PILKOVÁ a Michal MUNK, 2021. Is Pillar 3 a Good Tool for Stakeholders in CEE Commercial Banks? In: *Studies in Systems, Decision and Control* [online]. Springer, s. 421–440. doi:10.1007/978-3-030-76632-0_15

CAMPOS, Ricardo, Vítor MANGARAVITE, Arian PASQUALI, Alípio JORGE, Célia NUNES a Adam JATOWT, 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* [online]. **509**, 257–289. ISSN 00200255. doi:10.1016/j.ins.2019.09.013

CARROLL, John Bissell, 1964. *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

CERNA, Miloslava a Petra POULOVA, 2008. VISIT RATE OF INTERNET PORTALS AND UTILIZATION OF THEIR TOOLS AND SERVICES. *E & M EKONOMIE A MANAGEMENT*. **11**(4), 132–143. ISSN 1212-3609.

COLEMAN, Meri a Ta Lin LIAU, 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*. **60**, 283–284.

COLLINS, John William a Nancy P. O'BRIEN, 2003. *The Greenwood Dictionary of Education*. Greenwood Press.

COMMON CORE STATE STANDARDS INITIATIVE, 2023. *English Language Arts Standards* [online]. [vid. 2024-02-02]. Dostupné z: https://corestandards.org/wp-content/uploads/2023/09/ELA_Standards1.pdf

COOLEY, R, B MOBASHER, J SRIVASTAVA a OTHERS, 1999. Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*. **1**(1), 5–32.

CVRČEK, Václav, Radek ČECH a Miroslav KUBÁT, 2020. QuitaUp – nástroj pro kvantitativní stylometrickou analýzu. *Czech National Corpus and University of Ostrava* [online] [vid. 2023-07-21]. Dostupné z: https://korpus.cz/quitaup/

CVRČEK, Václav a Lucie CHLUMSKÁ, 2015. Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguistics* [online]. **39**(3), 309–325. ISSN 0304-3487. doi:10.1007/s11185-015-9151-8

DEMBERG, Vera a Frank KELLER, 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* [online]. **109**(2), 193–210. ISSN 00100277. doi:10.1016/j.cognition.2008.07.008

DUGAST, Daniel, 1979. *Vocabulaire et stylistique: Théâtre et dialogue*. Geneva, Switzerland: Slatkine-Champion.

EBAID, Ibrahim El-Sayed, 2023. IFRS adoption and the readability of corporate annual reports: evidence from an emerging market. *Future Business Journal* [online]. **9**(1), 80. ISSN 2314-7210. doi:10.1186/s43093-023-00244-x

EHARA, Yo, 2021. To What Extent Can English-as-a-Second Language Learners Read Economic News Texts? In: *Proceedings of the Third Workshop on Economics and Natural Language Processing* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 62–68. doi:10.18653/v1/2021.econlp-1.9

EHARA, Yo, 2022. Neural Language Model-based Readability Assessment of Computer Science Introductory Texts for English-as-a-Second Language Learners. In: *Proceedings of the*

*44th Annual Conference of the Cognitive Science Society*. Toronto, Canada: eScholarship, s. 1698–1704.

FAYYAD, Usama M., Gregory PIATETSKY-SHAPIRO a Padhraic SMYTH, 1996. From Data Mining to Knowledge Discovery in Databases [online]. **17**(3), 37–54. ISSN 0738-4602. doi:10.1609/AIMAG.V17I3.1230

FISHER, Douglas, Nancy FREY a Diane LAPP, 2012. *Text Complexity: Raising Rigor in Reading*. International Reading Association.

FLESCH, Rudolf, 2016. How to Write Plain English. *University of Cantenbury* [online] [vid. 2024-01-21]. Dostupné z: https://web.archive.org/web/20160712094308/http:/www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

GAJDOŠOVÁ, Katarína a Mária ŠIMKOVÁ, 2016. Slovak Dependency Treebank. *http://hdl.handle.net/11234/1-1822*.

GEERAERTS, Dirk, Stefan GRONDELAERS a Peter BAKEMA, 1994. *The Structure of Lexical Variation* [online]. DE GRUYTER MOUTON. ISBN 978-3-11-014387-4. doi:10.1515/9783110873061

GRAY, William Scott a Bernice Elizabeth LEARY, 1935. *What Makes a Book Readable: With Special Reference to Adults of Limited Reading Ability*. Chicago, IL: University of Chicago Press.

GUAY, Wayne R., Delphine SAMUELS a Daniel J. TAYLOR, 2015. Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure. *SSRN Electronic Journal* [online]. ISSN 1556-5068. doi:10.2139/ssrn.2564350

GUIRAUD, Pierre, 1960. *Probl`emes et m´ethodes de la statistique linguistique*. Dordrecht, Netherlands: Springer.

GUNNING, Thomas G., 2003. The Role of Readability in Today's Classrooms. *Topics in Language Disorders*. **23**(3), 175–189.

HARLEY, Brigit a Mary Lou KING, 1989. Verb Lexis in the Written Composition of Young L2 Learners. *Studies in Second Language Acquisition*. **11**(4), 415–439.

HARRIS, Theodore L. a Richard E. HODGES, 1995. *The Literacy Dictionary: The Vocabulary of Reading and Writing*. International Reading Association. ISBN 9780872071384.

HERDAN, Gustav, 1964. *Quantitative linguistics*. London, UK: Butterworths.

HYLTENSTAM, Kenneth, 1988. Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development* [online]. **9**(1–2), 67–84. ISSN 0143-4632. doi:10.1080/01434632.1988.9994320

CHALL, Jeanne Sternlicht a Edgar DALE, 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

CHAUDRON, Craig a Kate PARKER, 1990. Discourse Markedness and Structural Markedness: The Acquisition of English Noun Phrases. *Studies in Second Language Acquisition*. **12**, 43–64.

JARVIS, Scott, 2002. Short texts, best-tting curves and new measures of lexical diversity. *Language Testing*. **19**(1), 57–84.

JEAN, Sébastien, Kyunghyun CHO, Roland MEMISEVIC a Yoshua BENGIO, 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* [online]. Beijing, China: Association for Computational Linguistics, s. 1–10. doi:10.3115/v1/P15-1001

KAPUSTA, Jozef, Michal MUNK a Martin DRLÍK, 2012a. Cut-off time calculation for user session identification by reference length. In: *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings*.

KAPUSTA, Jozef, Michal MUNK a Martin DRLÍK, 2012b. Cut-off time calculation for user session identification by reference length. In: *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings* [online]. 2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012. ISBN 9781467317405. doi:10.1109/ICAICT.2012.6398500

KAPUSTA, Jozef, Michal MUNK, Peter SVEC a Anna PILKOVA, 2014. Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal. *Procedia Computer Science*. **29**, 1779–1790.

KAPUSTA, Jozef, Anna PILKOVA, Michal MUNK a Peter SVEC, 2013. Data pre-processing for web log mining: Case study of commercial bank website usage analysis. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. **61**(4), 973–979.

KINCAID, Peter J., Robert P. FISHBURNE JR., Richard L. ROGERS a Brad S. CHISSOM, 1975. *Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel*.

KINTSCH, Walter, 1974. *The Representation of Meaning in Memory* [online]. Lawrence Erlbaum. ISBN 9781317744894. Dostupné z: doi:10.4324/9781315794563

KLEE, Thomas, 1992. Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders*. **12**, 28–41.

KOEHN, Philipp, 2010. *Statistical Machine Translation*. Cambridge University Press. ISBN 0521874157, 9780521874151.

LAUFER, Batia, 1994. The Lexical Profile of Second Language Writing: Does It Change Over Time? *RELC Journal* [online]. **25**(2), 21–33. ISSN 0033-6882. doi:10.1177/003368829402500202

LINNARUD, Moira, 1986. *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund, Sweden: CWK Gleerup.

LIU, Bing, 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* [online]. ISBN 978-3-642-19459-7. doi:10.1007/978-3-642-19460-3

LOSARWAR, Vijayashiri a Madhuri JOSHI, 2012. Data Preprocessing in Web Usage Mining. In: *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore*. s. 1–5.

LOSHIN, David, 2013. Knowledge Discovery and Data Mining for Predictive Analytics. In: *Business Intelligence* [online]. Elsevier, s. 271–286. doi:10.1016/B978-0-12-385889-4.00017-X

LU, Xiaofei, 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*. **15**(4), 474–496.

LU, Xiaofei, 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers's language development. *TESOL Quaterly*. **45**(1), 36–62.

LU, Xiaofei, 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal* [online]. **96**(2), 190–208. ISSN 00267902. doi:10.1111/j.1540-4781.2011.01232.x

LU, Xiaofei a Haiyang AI, 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*. **29**, 16–27.

LUONG, Minh-Thang, Ilya SUTSKEVER, Quoc V. LE, Oriol VINYALS a Wojciech ZAREMBA, 2015. Addressing the Rare Word Problem in Neural Machine Translation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* [online]. Beijing, China: Association for Computational Linguistics, s. 11–19. doi:10.3115/v1/P15-1002

MALVERN, David, Brian RICHARDS, Ngoni CHIPERE a Pilar DURÁN, 2004. *Lexical Diversity and Language Development* [online]. London: Palgrave Macmillan UK. ISBN 978-1-4039-0232-0. doi:10.1057/9780230511804

MAQSOOD, Shazia, Abdul SHAHID, Muhammad TANVIR AFZAL, Muhammad ROMAN, Zahid KHAN, Zubair NAWAZ a Muhammad Haris AZIZ, 2022. Assessing English language sentences readability using machine learning models. *PeerJ Computer Science* [online]. **7**, e818. ISSN 2376-5992. doi:10.7717/peerj-cs.818

MCCARTHY, Philip M., 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Memphis, TN. PhD. Thesis. The University of Memphis.

MCCARTHY, Philip M. a Scott JARVIS, 2007. vocd: A theoretical and empirical evaluation. *Language Testing* [online]. **24**(4), 459–488. ISSN 0265-5322. doi:10.1177/0265532207080767

MCCLURE, Erica, 1991. A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning*. **2**, 141–154.

MCLAUGHLIN, Harry G., 1969. SMOG Grading - a New Readability Formula. *Journal of Reading*. **12**(8), 639–646.

MILLER, Jon F, 1991. Quantifying productive language disorders. In: Jon F. MILLER, ed. *Research in child language disorders: A decade of progress*. Austin, TX: Pro-Ed, s. 211–220.

MING-SYAN CHEN, Ming-Syan, Jong Soo JONG SOO PARK a P.S. YU, 1998. Efficient data mining for path traversal patterns. *IEEE Transactions on Knowledge and Data Engineering* [online]. **10**(2), 209–221. ISSN 10414347. doi:10.1109/69.683753

MORENO, Alonso a Araceli CASASOLA, 2016. A Readability Evolution of Narratives in Annual Reports. *Journal of Business and Technical Communication* [online]. **30**(2), 202–235. ISSN 1050-6519. doi:10.1177/1050651915620233

MUNK, Michal a Lubomir BENKO, 2018. Using Entropy in Web Usage Data Preprocessing. *Entropy* [online]. **20**(1), 67. doi:10.3390/e20010067

MUNK, Michal, Ľubomír BENKO, Mikuláš GANGUR a Milan TURČÁNI, 2015. Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. *E+M Ekonomie a Management* [online]. **18**(3), 144–159. doi:10.15240/tul/001/2015-3-013

MUNK, Michal, Martin DRLIK, Lubomir BENKO a Jaroslav REICHEL, 2017a. Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. *IEEE Access* [online]. **5**, 8989–9004. ISSN 21693536. doi:10.1109/ACCESS.2017.2706302

MUNK, Michal a Jozef KAPUSTA, 2014. *Web Usage Mining: Príprava a modelovanie dát*. Nitra: Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-558-0692-1.

MUNK, Michal, Jozef KAPUSTA, Peter ŠVEC a Milan TURČÁNI, 2010. Data Advance Preparation Factors Affecting Results of Sequence Rule Analysis in Web Log Mining. *E+M Ekonomie a Management*. **13**(4), 143–160.

MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021a. Methodology of stakeholders' behaviour modelling based on time. *MethodsX* [online]. **8**, 101570. ISSN 22150161. doi:10.1016/j.mex.2021.101570

MUNK, Michal, Anna PILKOVA, Ľubomír BENKO, Petra BLAZEKOVA a Peter SVEC, 2021b. Pillar 3–Pre-processed web server log file dataset of the banking institution. *Data in Brief* [online]. **39**, 107672. ISSN 23523409. doi:10.1016/j.dib.2021.107672

MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021c. Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times. *Expert Systems with Applications* [online]. **185**, 115503. ISSN 09574174. doi:10.1016/j.eswa.2021.115503

MUNK, Michal, Anna PILKOVA, Lubomir BENKO a Petra BLAŽEKOVÁ, 2017b. Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times. *Journal of Business Economics and Management* [online]. **18**(5), 954–973. doi:10.3846/16111699.2017.1360388

MUNK, Michal, Anna PILKOVA, Jozef KAPUSTA, Peter SVEC a Martin DRLIK, 2013. Pillar 3 and Modelling of Stakeholders' Behaviour at the Commercial Bank Website during the Recent Financial Crisis. *Procedia Computer Science* [online]. **18**, 1747–1756. ISSN 18770509. doi:10.1016/j.procs.2013.05.343

MUNKOVA, Dasa, Michal MUNK, Ľubomír BENKO a Petr HAJEK, 2021a. The role of automated evaluation techniques in online professional translator training. *PeerJ Computer Science* [online]. **7**, e706. ISSN 2376-5992. doi:10.7717/peerj-cs.706

MUNKOVA, Dasa, Michal MUNK, Ľubomír BENKO a Jiri STASTNY, 2021b. MT Evaluation in the Context of Language Complexity. *Complexity* [online]. **2021**, 1–15. ISSN 1099-0526. doi:10.1155/2021/2806108

O'FLYNN, James Adam, 2019. An Economics Academic Word List (EAWL): Using online resources to develop a subject-specific word list and associated teaching-learning materials. *Journal of Academic Language and Learning*. **13**(1).

O'HAYRE, John, 1966. *Gobbledygook has gotta go*. U.S. Government Printing Office.

PABARSKAITE, Zidrina a Aistis RAUDYS, 2007. A process of knowledge discovery from web log data: Systematization and critical review. *Journal of Intelligent Information Systems* [online]. **28**(1), 79–104. ISSN 09259902. doi:10.1007/s10844-006-0004-1

PAPINENI, Kishore, Salim ROUKOS, Todd WARD a WeiJing ZHU, 2002. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. s. 311–318.

PILKOVÁ, Anna, Michal MUNK, Ľubomír BENKO, Petra BLAŽEKOVÁ a Jozef KAPUSTA, 2021a. Pillar 3: Does banking regulation support stakeholders' interest in banks financial and risk profile? *PLOS ONE* [online]. **16**(10), e0258449. ISSN 1932-6203. doi:10.1371/journal.pone.0258449

PILKOVÁ, Anna, Michal MUNK, Petra BLAŽEKOVÁ a Ľubomír BENKO, 2021b. Web usage analysis: Pillar 3 information assessment in turbulent times. In: Mohammad Z. ABEDIN, Kabir HASSAN, Petr HAJEK a Mohammed M. UDDIN, ed. *The Essentials of Machine Learning in Finance and Accounting* [online]. Routledge, s. 24. ISBN 9780367480813. doi:10.4324/9781003037903

QI, Peng, Timothy DOZAT, Yuhao ZHANG a Christopher D. MANNING, 2018. Universal Dependency Parsing from Scratch. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 160–170. doi:10.18653/v1/K18-2016

QI, Peng, Yuhao ZHANG, Yuhui ZHANG, Jason BOLTON a Christopher D. MANNING, 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 101–108. doi:10.18653/v1/2020.acl-demos.14

REI, Ricardo, Craig STEWART, Ana C FARINHA a Alon LAVIE, 2020. COMET: A Neural Framework for MT Evaluation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 2685–2702. doi:10.18653/v1/2020.emnlp-main.213

ROMERO, Cristóbal, Sebastián VENTURA, Amelia ZAFRA a Paul de BRA, 2009. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers and Education*. **53**(3), 828–840.

SADEEK QUADERI, Shah Jafor a Kasturi Dewi VARATHAN, 2024. Identification of significant features and machine learning technique in predicting helpful reviews. *PeerJ Computer Science* [online]. **10**, e1745. ISSN 2376-5992. doi:10.7717/peerj-cs.1745

SAEL, N, A MARZAK a H BEHJA, 2013. Web Usage Mining data preprocessing and multi level analysis on Moodle. In: *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on* [online]. s. 1–7. ISSN 2161-5322. doi:10.1109/AICCSA.2013.6616427

SARWAR, Talha Bin a Noorhuzaimi Mohd NOOR, 2021. An Experimental Comparison of Unsupervised Keyphrase Extraction Techniques for Extracting Significant Information from Scientific Research Articles. In: *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)* [online]. IEEE, s. 130–135. ISBN 978-1-6654-1407-4. doi:10.1109/ICSECS52883.2021.00031

SARWAR, Talha Bin, Noorhuzaimi Mohd NOOR, M. Saef Ullah MIAH, Mamunur RASHID, Fahmid Al FARID a Mohd Nizam HUSEN, 2021. Recommending Research Articles: A Multi-Level Chronological Learning-Based Approach Using Unsupervised Keyphrase Extraction and Lexical Similarity Calculation. *IEEE Access* [online]. **9**, 160797–160811. ISSN 2169-3536. doi:10.1109/ACCESS.2021.3131470

SENTER, RJ a EA SMITH, 1967. *Automated Readability Index*.

SHANNON, C. E., 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* [online]. **27**(3), 379–423. ISSN 00058580. doi:10.1002/j.1538-7305.1948.tb01338.x

SCHMID, Helmut, 2019. Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage* [online]. New York, NY, USA: ACM, s. 133–137. ISBN 9781450371940. doi:10.1145/3322905.3322915

SCHMID, Helmut, Marco BARONI, Eros ZANCHETTA a Achim STEIN, 2007. The Enriched TreeTagger System. In: *Proceedings of the EVALITA 2007 workshop*.

SPIERS, Harry, Nikul AMIN, Raj LAKHANI, Andrew J. MARTIN a Parag M. PATEL, 2017. Assessing Readability and Reliability of Online Patient Information Regarding Vestibular Schwannoma. *Otology & Neurotology* [online]. **38**(10), e470–e475. ISSN 1531-7129. doi:10.1097/MAO.0000000000001565

SPILIOPOULOU, Myra a Lukas C. FAULSTICH, 1999. WUM: A Tool for Web Utilization Analysis. In: *The World Wide Web and Databases* [online]. Springer Berlin Heidelberg, s. 184–203. doi:10.1007/10704656_12

SRIVASTAVA, Jaideep, Robert COOLEY, Mukund DESHPANDE a Pang-ning TAN, 2000. Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data. *Text* [online]. **1**(2), 12–23. ISSN 19310145. doi:10.1145/846183.846188

SRIVASTAVA, Jaideep, Prasanna DESIKAN a Vipin KUMAR, 2005. Web Mining - Concepts, Applications, and Research Directions. In: *Foundations and Advances in Data Mining*. Springer, Berlin, Heidelberg, s. 275–307.

SRIVASTAVA, Mitali, Rakhi GARG a P. K. MISHRA, 2015. Analysis of Data Extraction and Data Cleaning in Web Usage Mining. In: *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) - ICARCSET '15* [online]. New York, New York, USA: ACM Press, s. 1–6. ISBN 9781450334419. doi:10.1145/2743065.2743078

STRAKA, Milan a Jana STRAKOVÁ, 2014. MorphoDiTa: Morphological Dictionary and Tagger. *http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0*.

STRAKA, Milan a Jana STRAKOVÁ, 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 88–99. doi:10.18653/v1/K17-3009

SVEC, Peter, Lubomir BENKO, Miroslav KADLECIK, Jan KRATOCHVIL a Michal MUNK, 2020. Web Usage Mining: Data Pre-processing Impact on Found Knowledge in Predictive Modelling. *Procedia Computer Science* [online]. **171**, 168–178. ISSN 18770509. doi:10.1016/j.procs.2020.04.018

TEMPLIN, Mildred, 1957. *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

THORDARDOTTIR, Elin T. a Susan Ellis WEISMER, 2001. High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders* [online]. **36**(2), 221–244. ISSN 1368-2822. doi:10.1080/13682820118239

TOERIEN, Franz Eduard a Elda DU TOIT, 2024. Fighting through the Flesch and Fog: the readability of risk disclosures. *Accounting Research Journal* [online]. **37**(1), 39–56. ISSN 1030-9616. doi:10.1108/ARJ-03-2023-0094

VELLINGIRI, J. a S. CHENTHUR PANDIAN, 2011. A novel technique for web log mining with better data cleaning and transaction identification. *Journal of Computer Science* [online]. **7**(5), 683–689. ISSN 15493636. doi:10.3844/jcssp.2011.683.689

W3C, 1995. *Configuration File of W3C httpd* [online] [vid. 2022-01-23]. Dostupné z: https://www.w3.org/Daemon/User/Config/Logging.html

WOLFE-QUINTERO, Kate, Shunji INAGAKI a Hae-Young KIM, 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. Honolulu, US: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

XUE, Guo-yi a Paul NATION, 1984. A university word list. *Language Learning and Communication*. **3**, 215–229.

YAO, Zheng, X. WANG a J. LUAN, 2017. Using Hidden Markov Model to Predict the Web Users' Linkage. *Journal of Residuals Science & Technology* [online]. **14**(3), 554–565. doi:10.14355/jrst.2017.1403.053

# PRÍLOHY: ZOZNAM POUŽITÝCH PUBLIKOVANÝCH PRÁC

Príloha A: MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021. Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times. *Expert Systems with Applications*. **185**, 115503. doi:10.1016/j.eswa.2021.115503 (**Web of Science, 2021IF: 8.665, Q1**) [WoS: 2, Scopus: 1]

Príloha B: PILKOVÁ, Anna, Michal MUNK, Ľubomír BENKO, Petra BLAŽEKOVÁ a Jozef KAPUSTA, 2021. Pillar 3: Does banking regulation support stakeholders' interest in banks financial and risk profile? *PLOS ONE*. **16**(10), e0258449. doi:10.1371/journal.pone.0258449 (**Web of Science, 2021IF: 3.752, Q2**) [WoS: 0, Scopus: 0]

Príloha C: MUNK, Michal, Anna PILKOVA, Ľubomír BENKO, Petra BLAZEKOVA a Peter SVEC, 2021. Pillar 3–Pre-processed web server log file dataset of the banking institution. *Data in Brief*. **39**, 107672. doi:10.1016/j.dib.2021.107672 (**Web of Science; Scopus**) [WoS: 1, Scopus: 0]

Príloha D: MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021. Methodology of stakeholders' behaviour modelling based on time. *MethodsX*. **8**, 101570. doi:10.1016/j.mex.2021.101570 (**Web of Science; Scopus**) [WoS: 0, Scopus: 1]

Príloha E: PILKOVÁ, Anna, Michal MUNK, Petra BLAŽEKOVÁ a Ľubomír BENKO, 2021. Web usage analysis: Pillar 3 information assessment in turbulent times. In: Mohammad Z. ABEDIN, Kabir HASSAN, Petr HAJEK a Mohammed M. UDDIN, ed. *The Essentials of Machine Learning in Finance and Accounting*. Routledge, s. 24. doi:10.4324/9781003037903 (**Scopus**) [Scopus: 1]

Príloha F: BLAŽEKOVÁ, Petra, Ľubomír BENKO, Anna PILKOVÁ a Michal MUNK, 2021. Is Pillar 3 a Good Tool for Stakeholders in CEE Commercial Banks? In: *Studies in Systems, Decision and Control*. Springer, s. 421–440. doi:10.1007/978-3-030-76632-0_15 (**Scopus**) [Scopus: 0]

Príloha G: SVEC, Peter, Lubomir BENKO, Miroslav KADLECIK, Jan KRATOCHVIL a Michal MUNK, 2020. Web Usage Mining: Data Pre-processing Impact on Found

Knowledge in Predictive Modelling. *Procedia Computer Science*. **171**, 168–178. doi:10.1016/j.procs.2020.04.018 (**Scopus**) [Scopus: 10]

Príloha H: MUNKOVA, Dasa, Michal MUNK, ĽUBOMÍR BENKO a Jiri STASTNY, 2021. MT Evaluation in the Context of Language Complexity. *Complexity*. **2021**, 1–15. doi:10.1155/2021/2806108 (**Web of Science, 2021IF: 2.121, Q2**) [WoS: 2, Scopus: 0]

Príloha I: BENKO, Ľubomír, Dasa MUNKOVÁ a Michal MUNK, 2023. Relationship Between Linguistic Complexity and MT Errors in the Context of Inflectional Languages. In: *Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2023*. Springer, Cham, s. 546–557. doi:10.1007/978-3-031-42430-4_45 (**Scopus**) [Scopus: 0]

Príloha J: BENKO, Ľubomír, Lucia BENKOVA, Dasa MUNKOVA, Michal MUNK a Danylo SHULZENKO, 2022. Error Classification Using Automatic Measures Based on n-grams and Edit Distance. In: *Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2022*. Springer, Cham, s. 345–356. doi:10.1007/978-3-031-20319-0_26 (**Web Of Science**, **Scopus**) [WoS:0, Scopus: 0]

Príloha K: BENKO, Ľubomír, Dasa MUNKOVA, Michal MUNK, Lucia BENKOVA a Petr HAJEK, 2024. The use of residual analysis to improve the error rate accuracy of machine translation. *Scientific Reports* (v recenznom konaní od 2023, 3. kolo) (**Web of Science, 2022IF: 4.6, Q2**)

Príloha L: BENKO, Ľubomír, Dasa MUNKOVA, Mária PAPPOVÁ a Michal MUNK, 2024. Comparison of various approaches to tagging for the inflectional Slovak language. *PeerJ Computer Science* (v recenznom konaní od 2023, 2. kolo) (**Web of Science, 2022IF: 3.8, Q2**)

Príloha M: BENKO, Ľubomír, Anna PILKOVA, Michal MUNK a Slavka ELEY, 2024. Pillar 3: The impact of language complexity on the preferences of commercial bank website users. *Expert Systems with Applications* (v recenznom konaní od 2024, 1. kolo) (**Web of Science, 2022IF: 8.5, Q1**)