FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Název:	Analýza diskusních komentářů na českých zpravodajských serverech
Student:	Martin Vastl
Vedoucí:	Ing. Daniel Vašata, Ph.D.
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	Do konce letního semestru 2019/20

Pokyny pro vypracování

Tématem práce je analýza komentářů v diskusích na zpravodajských serverech pomocí metod strojového učení.

1. Vyberte alespoň jeden zpravodajský server umožňující jednoznačnou identifikaci diskutujících uživatelů a stáhněte dostatečné množství článků spolu s příslušnými diskusními příspěvky.

2. Prozkoumejte současné metody zpracování přirozeného jazyka (NLP) z pohledu jejich využití ke zkoumání relevance diskusních příspěvků vzhledem k obsahu diskutovaného článku.

3. Vyberte alespoň dvě z těchto metod a využijte je k analýze stažených diskusních příspěvků.

4. Zabývejte se zkoumáním relevance komentářů k obsahu článků a případně k jiným obecným tématům, např. politiky. Zaměřte se také na jednotlivé uživatele, tj. zkoumejte, zda je možné je nějakým způsobem klasifikovat a pokuste se detekovat anomální chování.

5. Výsledky analýzy přehledným způsobem prezentujte a komentujte.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Karel Klouda, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 29. ledna 2019

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Analýza diskusních komentářů na českých zpravodajských serverech

Martin Vastl

Katedra aplikované matematiky
Vedoucí práce: Ing. Daniel Vašata, Ph.D.

15. května 2019

Poděkování

Děkuji především vedoucímu mé bakalářské práce Ing. Danielu Vašatovi, Ph.D., za cenné připomínky a věcné rady. Dále bych rád poděkoval mé rodině a všem, kteří mě podporovali.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 15. května 2019 .....

Odkaz na tuto práci

Vastl, Martin. *Analýza diskusních komentářů na českých zpravodajských serverech*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2019.

Abstrakt

Tato práce je zaměřena na možnosti využití metod pro zpracování přirozeného jazyka k analýze komentářů zpravodajského portálu. Hlavním cílem je srovnání modelů BERT, Doc2vec a Doc2vec s předtrénovanými reprezentacemi slov z BERT ke zkoumání relevance komentářů k obsahu článků z portálu. Dalším cílem je aplikace vektorových reprezentací textu k detekci anomálních příspěvků a anomálního chování uživatelů pomocí metody Local outlier factor.

Provedenými experimenty bylo zjištěno, že nejvyšší úspěšnosti ke zkoumání relevance je dosaženo pomocí modelu BERT, a že předtrénované slovní reprezentace nemají pozitivní vliv na zachycení sémantické informace textu oproti metodě Doc2vec. Metoda Local outlier factor, která je použita pro detekci anomálií, je schopna detekovat anomální komentáře i uživatele při využití vektorů z modelu BERT. Na druhou stranu, Doc2vec je v případě detekce anomálií nevhodný a často vrací nesprávné výsledky.

Klíčová slova analýza komentářů zpravodajského portálu, detekce anomálních uživatelů a komentářů, relevance komentářů k článkům, Doc2vec, BERT

Abstract

This thesis is focused on the possibilities of using natural language processing methods to analyze comments on the news portal. The main goal is to compare the ability of BERT, Doc2vec, and Doc2vec with pretrained word vectors from BERT to examine the relevance between the comments and the content of an article from a news portal. Another goal is to use the text vector representations to detect anomalies via the Local outlier factor method.

It was found by experiments, that the best model for text representation is BERT and that the pretrained word vectors have no positive impact on results in comparison of Doc2vec without pretrained vectors. Moreover, the Local outlier factor can detect anomaly comments and users when using vectors from BERT in contrast to Doc2vec text representations which are not good enough for anomaly detection and therefore often returns incorrect results.

Keywords analysis of news portal discussion, detection of anomaly users and comments, comment to article relevancy, Doc2vec, BERT

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Motivací pro výběr tématu analýzy diskuzí na českých zpravodajských portálech byl narůstající počtem informací, které získáváme z diskuzí. Z tohoto důvodu roste poptávka po automatických metodách, které jsou schopny kvantifikovat vlastnosti těchto textů. Mezi základní požadavky patří schopnost nalezení nerelevantních komentářů a detekce anomálního chování uživatelů.

Práce se zabývá získáním dat ze zpravodajského portálu, jejich analýzou za pomocí metod pro zpracování jazyka a prezentací této analýzy. V práci se využívá moderních metod pro převod textu do vektorového prostoru, pro klasifikaci jednotlivých komentářů, ke zkoumání uživatelů a detekci anomálního chování. Výsledkem je analýza a prezentace této analýzy u vybraného českého zpravodajského portálu.

Cíl práce

Cílem teoretické části práce je vytvoření uceleného obrazu o metodách pro zpracování přirozeného jazyka, zejména o metodách reprezentace textu ve vektorovém prostoru a o možnostech aplikace těchto metod na diskuze a články z vybraného zpravodajského portálu. Teoretická část si dále klade za cíl popsat možnosti využití reprezentací textu k detekci anomálií a ke zkoumání relevance komentářů k článkům.

Cílem praktické části je výběr vhodného portálu, který umožňuje jednoznačnou identifikaci uživatelů, pro získání článků a komentářů. Následně využití metod z teoretické části k reprezentaci textů ve vektorovém prostoru a jejich srovnání. Dalším cílem jsou možnosti aplikace metod pro zpracování jazyka ke zkoumání relevance mezi komentáři a články, k detekci anomálního chování uživatelů a k detekci anomálních komentářů. Poslední část si pak bere za cíl prezentaci výsledků z provedené analýzy způsobem, který lze jednoduše interpretovat.

Kapitola 1

Zpracování přirozeného jazyka (NLP)

Zpracování přirozeného jazyka je dle [1] oblast výzkumu, která zkoumá schopnost počítače pochopit a manipulovat přirozený lidský jazyk způsobem, který je k užitku. Cílem výzkumu je získat vědomosti a nástroje k porozumění toho, jak lidé rozumí a používají jazyk. Mezi typické úlohy patří analýza sentimentu textu, klasifikace spamu, převod řeči na text a nebo automatický překlad.

Většinu výše zmíněných problémů lze řešit pomocí metod, které využívají vektorové reprezentace slov případně textu. Z tohoto důvodu jsou metody, které jsou schopny vytvořit kvalitní reprezentace, jsou z tohoto důvodu jednou z důležitou součástí výzkumu v oblasti NLP.

1.1 Reprezentace slov ve vektorovém prostoru

Reprezentace slov ve vektorovém prostoru $\mathbb{R}^d$ je jednou z oblastí, kterou se zpracování přirozeného jazyka zabývá. Cílem této oblasti je zachycení sémantického významu slova do vektoru konstantní délky. Obecně existují dva druhy reprezentace slov ve vektorovém prostoru:

1. vysokodimenzionální řídkým vektorem (one-hot encoding),
2. nízkodimenzionálním hustým vektorem (Word2vec, GloVe, BERT, ...).

Obě tyto metody mají své využití. Reprezentace řídkým vektorem se využívá převážně pro jednoznačnou reprezentaci slova jako vstup do neuronových sítí. Reprezentace hustým vektorem se na druhou stranu používá pro zachycení sémantického významu slova. Tuto reprezentaci lze chápat jako kompresi významu slova do několika složek vektoru.

Tabulka 1.1: Reprezentace slov za pomocí one-hot encoding pro tři slova.

slovo	vektor
muž	$(1, 0, 0)$
chlapec	$(0, 1, 0)$
dům	$(0, 0, 1)$

1.1.1 One-hot encoding

Občas také označováno jako „kód 1 z N“ [2], je definováno takto:

Definice 1 *Nechť slovník je uspořádaná N-tice unikátních slov, pak pro vektor $w_k = (w_{k_1}, ..., w_{k_N})$ reprezentující slovo na k-té pozici a pro jeho j-tou souřadnici vektoru w_k platí:*

$$w_{k_j} := \delta_{jk}, \tag{1.1}$$

kde δ_{jk} je Kroneckerovo delta.

Tabulka 1.1 reprezentuje jednu z možných zakódování slovníku muž, chlapec a dům za pomocí one-hot encoding.

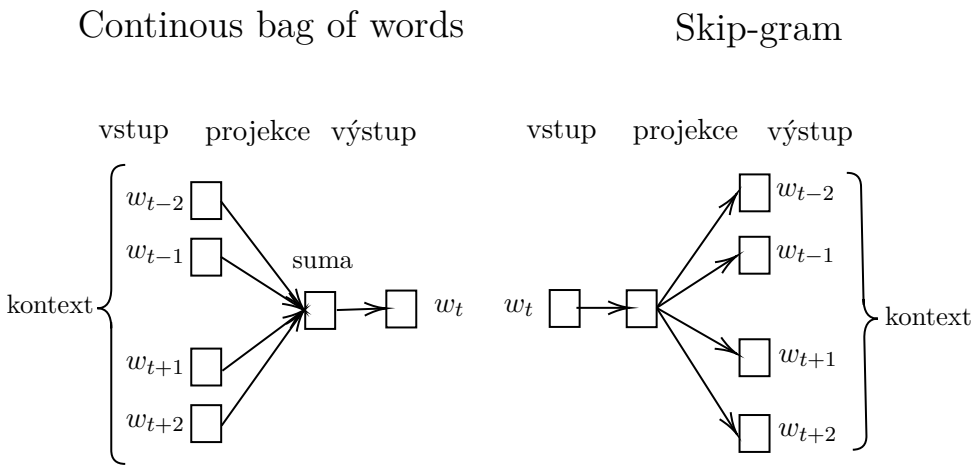
Nevýhodou tohoto přístupu je, že slova, která jsou sémanticky podobná (např. muž a chlapec), nemají podobnou vektorovou reprezentaci. Z tohoto důvodu se one-hot encoding využívá převážně jako vstup do neuronových sítí, ve kterých se následně tvoří vnitřní reprezentace pro jednotlivá slova.

1.1.2 Word2vec

Je soubor metod převodu slova do nízkodimenzionálního prostoru za pomocí neuronové sítě. Hlavní myšlenka Word2vec [3] spočívá v tom, že význam slova se ukrývá v kontextu, ve kterém se slovo objevuje. Tvorba vstupu pro Word2vec využívá okolních slov na pravé a na levé straně středového slova. Proces tvorby kontextu a středového slova je zobrazen v tabulce 1.2.

Tabulka 1.2: Pohyb klouzavého okna přes větu při velikosti okna 1.

Text				Kontext	Středové slovo
Výsledky	dnešní	písemky	jsou	[dnešní]	Výsledky
Výsledky	dnešní	písemky	jsou	[Výsledky, písemky]	dnešní
Výsledky	dnešní	písemky	jsou	[dnešní, jsou]	písemky



Obrázek 1.1: Architektura CBOW a Skip-gram.

První z metod Word2vec se nazývá Continuous bag of words (CBOW), která na základě kontextu předpovídá středové slovo. Vstupem do této metody je uspořádaná n-tice slov, kde n je dvojnásobek velikosti okna a výstupem je předpovězené slovo. Druhou metodou je Skip-gram, který předpovídá kontext na základě středového slova. Vstupem do tohoto modelu je středové slovo a výstupem je předpovězené slovo z kontextu. Schémata architektur obou metod jsou zobrazena na diagramu 1.1.

Z důvodu velké podobnosti obou metod je v následující kapitole popsán pouze CBOW. Principy, které jsou aplikovány na CBOW, lze jednoduše použít i v případě Skip-gram. Důvodem pro výběr CBOW, jako modelu pro popis principu Word2vec, je jeho využití v následujících kapitolách.

CBOW

CBOW je neuronová síť, jejímž vstupem jsou slova z kontextu zakódována pomocí one-hot encoding a výstupem je k hodnot představující pravděpodobnosti, že jednotlivá slova ze slovníku jsou středovým slovem. Triku, kterého se využívá v CBOW, je trénování neuronové sítě na rozdílné úloze, než na které bude později využita. V případě CBOW je to předpověď středového slova na základě kontextu. Velikost kontextu, resp. okna a výsledných vektorů je jedním z hyperparametrů tohoto modelu. Obvyklou hodnotou, která se volí pro velikost okna je 5 a pro dimenzi výsledných vektorů je 300 [3].

Průběh dopředného chodu modelu CBOW je následující. Na vstupu je k slov zakódovaných pomocí one-hot encoding a ty jsou vynásobeny maticí obsahující reprezentace pro jednotlivá slova. Následně jsou výsledné vektory sečteny nebo zprůměrovány. Tento výsledný vektor je předán fully connec-

ted vrstvě (o velikosti počtu slov ve slovníku), která obsahuje softmax jako aktivační funkce. Výpočet softmaxu pro model Word2vec je následující:

$$P(w_t|k) = \frac{e^{\text{score}(w_t,k)}}{\sum_{\text{Pro všechny slova w ze slovníku}} e^{\text{score}(w,k)}}, \tag{1.2}$$

kde k je kontext a score(w, k) je výstup neuronové sítě pro slovo w na základě kontextu k. Výsledkem softmaxu je uspořádaná n-tice s pravděpodobnostmi pro jednotlivá slova. Detail architektury a fungování CBOW je vidět na obrázku 1.2.

Zpětný chod CBOW probíhá tak, že se vypočítá ztrátová funkce a na základě ní se aktualizují váhy neuronové sítě. V případě Word2vec je ztrátová funkce categorical cross entropy mezi výslednou n-ticí a očekávaným slovem zakódovaným pomocí one-hot encoding. Categorical cross entropy je definována jako:

$$H(p, q) = -\sum_x p(x) \cdot \log_2(q(x)), \tag{1.3}$$

kde $p(x)$ je x-tá souřadnice středového slova zakódovaného pomocí one-hot encoding a $q(x)$ je x-tý výstup neuronové sítě. Za pomocí této ztrátové funkce je vypočten gradient a jsou aktualizovány váhy.

Cílem neuronové sítě je tedy maximalizovat pravděpodobnost středového slova na základě kontextu. Po natrénování neuronové sítě obsahuje matice V reprezentace slov ve vektorovém prostoru. Sémantickou podobnost vzniklých reprezentací lze měřit za pomocí kosínovy vzdálenosti. Čím je hodnota kosínové vzdálenosti bližší nule, tím jsou si slova sémanticky podobnější.

$$d_{cos}(A, B) = 1 - \frac{A \cdot B}{||A|| \cdot ||B||}, \tag{1.4}$$

kde $A \in \mathbb{R}^n$ a $B \in \mathbb{R}^n$ jsou vektory reprezentující slova. Někdy se také využívá kosínové podobnosti, která je definována jako

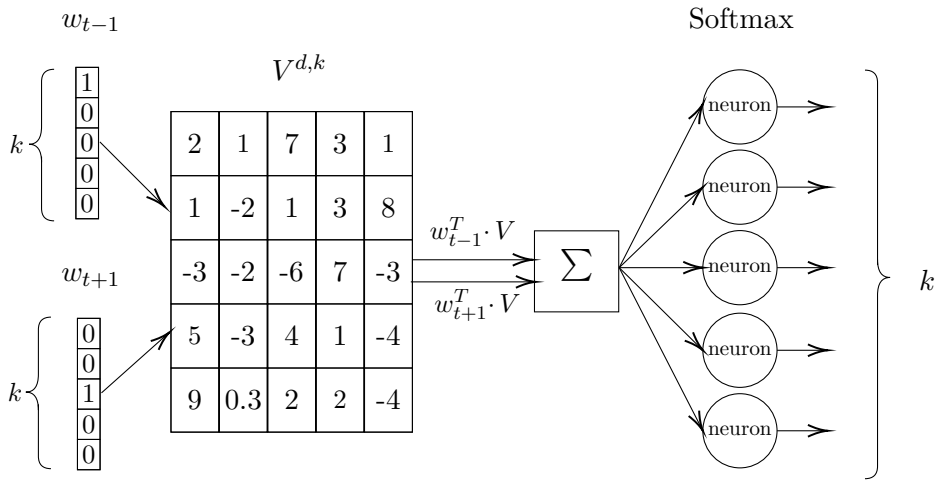
$$K(A, B) = 1 - d_{cos}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{1.5}$$

Srovnání CBOW a Skip-gram

Kvalitu vzniklých vektorových reprezentací produkované těmito modely lze měřit pomocí syntaktického a sémantického testu. Podstatou syntaktického testu je zachytit např. stupňování přídavných jmen.

$$vec(\text{'lepší'}) - vec(\text{'dobrý'}) + vec(\text{'špatný'}) \approx vec(\text{'horší'}) \tag{1.6}$$

Tedy, kdy změna ve slově je syntaktického rázu např. tvorba množného čísla nebo stupňování přídavných jmen. Na druhou stranu, cílem sémantického



Obrázek 1.2: Detail modelu CBOW pro okno o velikosti jedna.

Tabulka 1.3: Porovnání CBOW a Skip-gram. [3]

	CBOW	Skip-gram
Sémantický test	15.5 %	50 %
Syntaktický test	53.1 %	55.9 %
Celkově	36.1 %	53.3 %

testu je zachytit vztahy mezi slovy jako je např. geografická podobnost (země a hlavní město a nebo země a měna).

$$vec(\text{'Německo'}) - vec(\text{'Berlín'}) + vec(\text{'Rakousko'}) \approx vec(\text{'Vídeň'}) \tag{1.7}$$

Srovnání obou těchto metod je zobrazeno v tabulce 1.3, ve které je vidět, že CBOW je signifikantně horší v sémantickém testu. Krom syntaktického a sémantického testu je Skip-gram lepší v zachycení reprezentace i pro méně častá slova. Na druhou stranu trénování CBOW je přibližně třikrát rychlejší [3] než Skip-gram.

1.2 Reprezentace textu ve vektorovém prostoru

Cílem reprezentace textu do vektorového prostoru $\mathbb{R}^d$ je zaznamenání sémantické informace textu do vektoru fixní délky. Vytvořené vektorové reprezentace lze např. využít k analýze sentimentu textu. Stejně jako u slovních reprezentací, lze textové reprezentace porovnávat pomocí kosínovy vzdálenosti.

Výsledné vektory lze interpretovat tak, že každá složka vektoru tvoří určitý koncept a hodnota udává, jak moc je koncept v daném vektoru zastoupen.

1.2.1 Bag-of-words

Jedná se o jednoduchou metodu reprezentace textu do vektorového prostoru. Myšlenka bag-of-words [4] spočívá v následujícím procesu. Nejdříve je nutné všechny dokumenty předzpracovat následujícím způsobem:

1. Rozdělení textu na jednotlivá slova.
2. Odstranění slov, které nepřináší význam dokumentu (spojky, předložky, zájmena, ...).
3. Lemmatizace jednotlivých slov, která převede slova na jejich základní tvary. Např. slovu barvě přiřadí stejný význam jako slovu barva.
4. Stemmatizace, která zajistí převedení slov pouze na kmen slova např. změní slovo jahody na slovo jahod. Z tohoto důvodu budou mít slova jahodový, jahody a jahoda stejný tvar jahod.

Poté, co se takto předzpracují dokumenty, vytvoří se vektor pro každý z dokumentů. Velikost vektoru pro dokument bude počet unikátních slov po stemmatizaci (případně konstantní, pokud se omezí velikost slovníku) a hodnota pro každou složku vektoru bude počet výskytů slova v dokumentu. Tímto postupem vznikne matice, kde řádek reprezentuje dokument a sloupec slovo.

1.2.2 Tf-idf

Z důvodu toho, že slova v modelu Bag-of-words nejsou nikterak vážená, tedy všechny mají stejný sémantický význam pro text, vznikl model Tf-idf [4]. Cílem tohoto modelu je zvýšit váhy slovům, která jsou sémanticky významná pro dokument a snížit váhu slovům, která nejsou. Rozdíl modelu Tf-idf oproti modelu Bag-of-words spočívá tedy pouze v tom, že jednotlivá slova jsou vážená. Dle [4] je výpočet tf-idf následující:

$$\mathrm{tf}_{i,j} = \begin{cases} 1 + \log_{10} f_{ij} & \text{pokud } f_{ij} > 0, \\ 0 & \text{jinak,} \end{cases} \tag{1.8}$$

kde $f_{i,j}$ značí frekvenci slova i v dokumentu j a

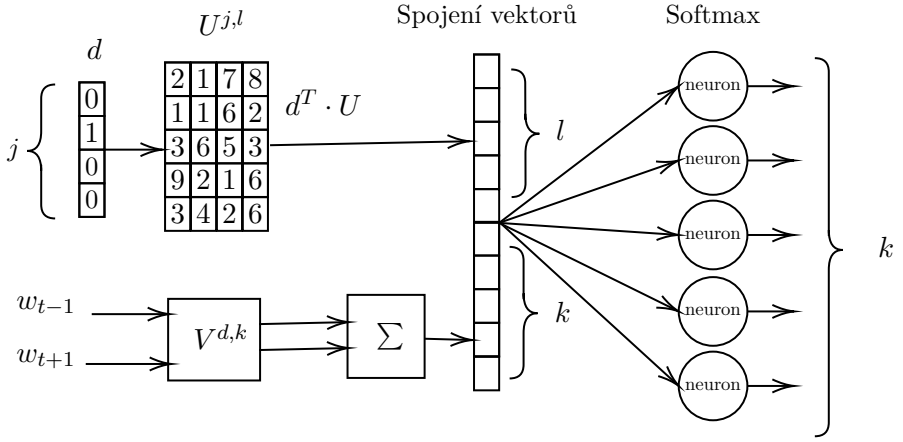
$$\mathrm{idf}_i = \log_{10}\left(\frac{n}{\mathrm{df}_i}\right), \tag{1.9}$$

kde n je celkový počet dokumentů a df_i je počet dokumentů obsahující slovo i, pak pro váhu slova i v j-tém dokumentu platí:

$$w_{i,j} = \mathrm{tf}_{i,j} \cdot \mathrm{idf}_i \tag{1.10}$$

j = počet dokumentů

l = velikost vektoru pro dokument



Obrázek 1.3: DM architektura pro okno o velikosti jedna.

Nevýhodou modelu Bag-of-word a Tf-idf je, že zaznamenávají dokumenty do řídkých vektorů vysoké dimenze, jejich paměťové nároky a to, že neberou v úvahu pořadí a sémantické vlastnosti slov. Některé z těchto problémů lze vyřešit pomocí LSA nebo pomocí invertovaného indexu. I přes tyto neduhy se v praxi ukazuje, že obě metody dávají velice dobré výsledky.

1.2.3 Doc2vec

Doc2vec [5], někdy také označován jako Paragraph vector, je model neuronové sítě, který se snaží vyřešit některé z výše zmíněných problémů. Hlavní myšlenka modelu vychází z Word2vec. Stejně jako Word2vec i Doc2vec obsahuje dva podobné modely. V případě Doc2vec jsou to modely DBOW (distributed bag of words), který vychází z myšlenky Skip-gram a DM (distributed memory), který vychází z CBOW.

Model DM se od modelu CBOW liší tím, že kromě toho, že má na vstupu zakódovaný kontext pomocí one-hot encoding, má zde i číslo (id dokumentu), které jednoznačně identifikuje dokument. Toto číslo je zakódováno pomocí one-hot encoding a vynásobeno maticí reprezentací jednotlivých dokumentů. Následně je vektor pro dokument spojen nebo sečten s vektorem reprezentujícím kontext. Výsledný vektor je předán softmax klasifikátoru, který vrátí pravděpodobnosti pro jednotlivá slova. Podobnou myšlenku lze použít i pro Skip-gram model, kde je na vstupu středové slovo a id dokumentu a výstupem je slovo z kontextu. Architektura a fungování DM je zobrazeno v diagramu 1.3.

Srovnání modelů DBOW a DM

Při srovnání modelů DBOW a DM vychází model DM lépe [5]. Nejlepších výsledků však lze dosáhnout při zřetězení reprezentací z obou modelů. Pokud je použit pouze jeden z modelů, je doporučeno použít DM, který má téměř podobné výsledky jako jejich spojení.

1.2.4 Bidirectional Encoder Representations from Transformer (BERT)

BERT [6] je model strojového učení vycházející z architektury Transformeru [7]. BERT byl navržen tak, aby se za předtrénovaný model dal napojit jiný model, který je určen na specifickou úlohu. Tímto způsobem lze ušetřit zdroje pro trénování obrovského modelu a soustředit se pouze na specifické vlastnosti úlohy. Autorům modelu se tímto způsobem podařilo dosáhnout v té době nejlepších výsledků v mnoha úlohách z oblasti zpracování přirozeného jazyka jen za pomocí zapojení dopředné sítě. Příkladem můžem být test GLUE [8] nebo question answering [9], ve kterém se modelu podařilo porazit i člověka. S vydáním článku byly zveřejněny i předtrénované modely (mimo jiné i model natrénovaný na 104 největších wikipediích, obsahující i český jazyk), které lze stáhnout a použít.

Předzpracování

Prvním krokem předzpracování vstupu je tokenizace textu za pomocí natrénovaného WordPiece [10] modelu. Trénování Wordpiece modelu probíhá následovně:

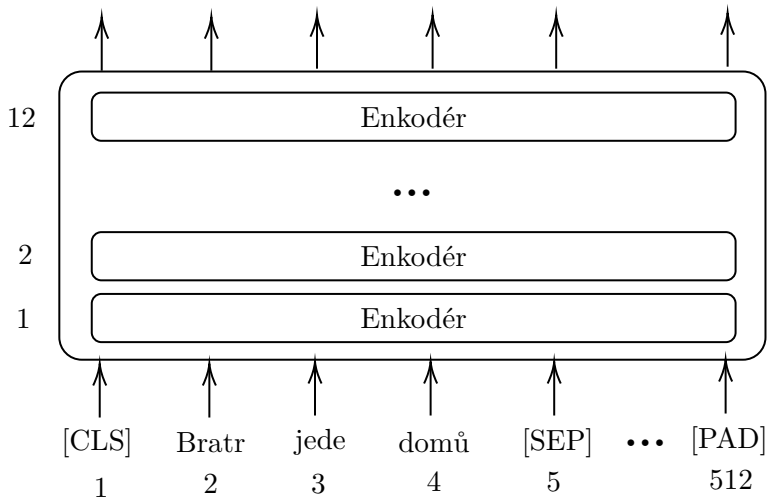
1. Slovník se nainicializuje pomocí základní sady znaků (např. pomocí všech Unicode znaků, které se objevují v českém jazyce).

2. Trénovací dataset je rozdělen na jednotlivé tokeny pomocí slovníku. Tokenizace probíhá hladovým způsobem, tedy text tokenizujeme podle prvního nejdelšího tokenu, který na slovo sedí.

3. Na datasetu rozděleném na tokeny se natrénuje jazykový model, tj. model, který je schopný na základě kontextu předpovědět další slovo.

4. Pro každou dvojici tokenů ze slovníku se vygeneruje nový token tak, že se dvojice spojí. Nakonec se ze všech takovýchto možných spojení dvojic vybere ten token, který když se přidá do slovníku, zvýší nejvíce úspěšnost modelu na trénovacích datech.

5. Proces se vrací zpátky do bodu 2, dokud ve slovníku není dostatek slov, a nebo se úspěšnost modelu nezvýší o minimální hodnotu.



Obrázek 1.4: Architektura BERT. [11]

Výstupem tohoto modelu pro větu *„Nejkrásnější květina je růže.“* po tokenizaci je *['nej', '##kra', '##sne', '##js', '##i', 'k', '##vet', '##ina', 'je', 'ru', '##ze', '.']*. Výhodou využití této tokenizace je, že by se neměla objevit slova, která model nebude znát.

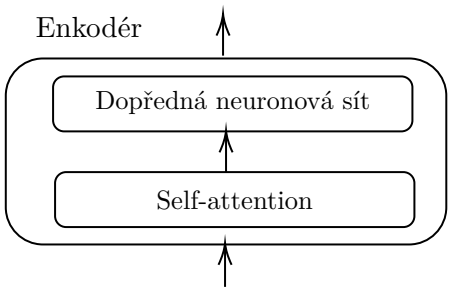
Dalším krokem po tokenizaci je přidání speciálních tokenů. Zleva je připojen token *[CLS]*, který je použit jako výstup modelu a zprava je každá věta doplněna o token *[SEP]*, který odděluje věty. Zbytek vstupu je vyplněn pomocí tokenu *[PAD]*, protože BERT očekává vstup konstantní délky.

Architektura

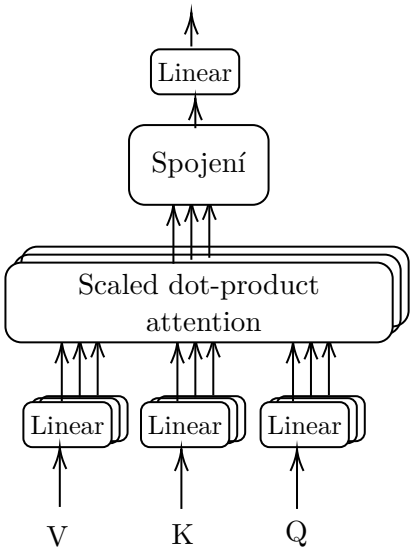
Architektura modelu BERT využívá, v případě menšího modelu, 12 za sebou zapojených enkodérů, které jsou popsány v Transformeru [7]. Architektura celé neuronové sítě je naznačená v diagramu 1.4. Blok enkodéru se skládá ze dvou hlavních částí, ze Self-attention bloku a z dopředné neuronové sítě. Hlavní síla Transformeru vychází převážně ze Self-attention vrstev, jejímž cílem je zachytit vazby mezi jednotlivými slovy. Příkladem je věta *Kočka pije mléko. Ona má totiž hlad.* Od slova *kočka* očekáváme vazbu ke slovu *ona* a *hlad* a toto vrstvy Self-attention zajistí.

Self-attention vrstva se skládá z k vrstev Scaled-dot product attention [7], které jsou lineárně promítány do vektorových prostorů. Tyto projekce jsou naučené při trénování sítě. Výpočet Scale-dot product attention je následující:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{1.11}$$



Obrázek 1.5: Schéma enkodéru. [12]

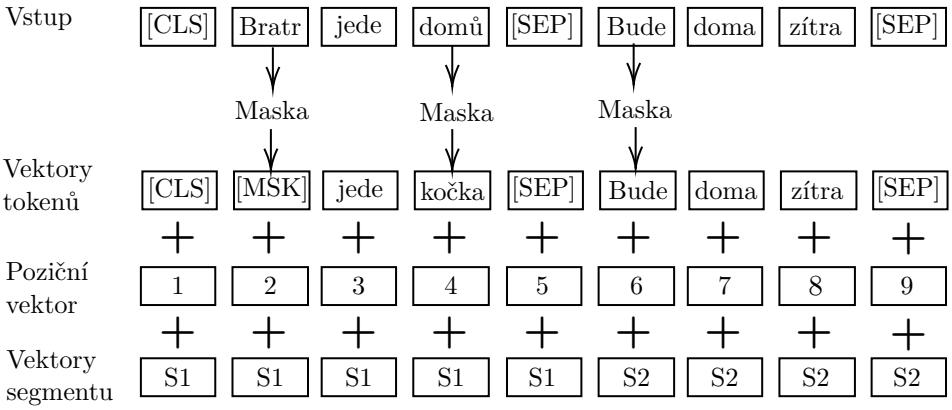


Obrázek 1.6: Diagram multi-head attention. [12]

kde $Q = W_Q E_w$, $K = W_K E_w$, $V = W_V E_w$, d_k je dimenze K a E_w značí vstupní vektor. Stejně jako u výše zmíněných projekčních matic, tak i matice W_Q, W_K a W_V jsou učeny při trénovaní sítě. Důvodem pro normalizaci součinu QK^T hodnotou $\sqrt{d_k}$ je ztrácející se gradient u Softmaxu pro vysoké hodnoty. Tímto způsobem se hodnota uvnitř Softmaxu zmenší a neztrácí se gradient. Po výpočtu Scale-dot product attention jsou výsledky spojeny a vynásobeny naučenou maticí W_O.

Trénování

Model je trénován na dvou různých úlohách. První z nich je předpověď slova na základě kontextu a druhá je určení, zda-li na sebe navazují dvě věty.



Obrázek 1.7: Zpracování vstupu v modelu BERT. [6]

Zpracování vstupu v modelu BERT je vidět na diagramu 1.7. Prvním krokem při trénování je maskování vstupu, které spočívá v tom, že 15 % tokenů ze vstupního textu se ještě zpracuje tímto způsobem. V 80 % případů se token nahradí za speciální token *[MSK]*, který reprezentuje maskované slovo. V 10 % se nahradí slovo za libovolný token a posledních 10 % tokenů se nechá beze změny. U všech takto zpracovaných tokenů se kontroluje předpověď slova. Tato procedura napomáhá tomu, že zbytek modelu BERT neví, u kterého slova bude požadována predikce, a proto si musí držet správnou distribuční funkci pro každý token. Kromě reprezentací pro jednotlivá slova využívá model poziční reprezentace, které jsou trénovány spolu s modelem. Cílem těchto reprezentací je odlišit pozice slov v textu (chceme, aby vektor pro slovo na začátku věty byl jiný než pro stejné slovo na konci). Další součástí je i reprezentace pro segment. Ta modelu pouze říká, zda-li se jedná o první nebo druhou větu a je trénována spolu se zbytkem modelu. Na konci jsou reprezentace slov, reprezentace pozic a reprezentace segmentu sečteny a předány zbytku modelu.

Druhým úkolem, na kterém je model trénován, je, zda-li na sebe dvě věty navazují či ne. Trénování na tomto úkolu je důležité hlavně pro problémy využívající celé věty a ne jen jednotlivá slova. Průběh trénování je jednoduchý, náhodně se vybere věta z textu a je k ní přidána navazující věta s pravděpodobností 50 % a s pravděpodobností 50 % náhodná věta z celého datasetu. Následně se za pomocí klasifikace výstupu pro token *[CLS]* rozhoduje, zda-li se jedná o následující větu či ne. BERT byl na tomto druhu úkolu schopný dosáhnout úspěšnosti okolo 97 % [6] na trénovacím datasetu.

Velikou výhodou tohoto modelu je, že využívá celého kontextu vstupních dat, a proto může využít informaci jak o slovech, která následují, tak i o slovech která již byla. Další důležitou vlastností, která plyne z využití enkodérů z modelu Transformer, je možnost trénování enkodérů paralelně, a tím urych-

lení procesu učení. Nevýhodou modelu je na druhou zvyšující se výpočetní náročnost pro dlouhé sekvence. Náročnost modelu totiž roste kvadraticky s délkou vstupu kvůli využití Self-attention [7]. Dalším problémem, který plyne z využití maskování vstupu, je nutnost velkého počtu epoch pro natrénování modelu.

Kapitola 2

Vybrané metody strojového učení

2.1 Klasifikace

Klasifikace [13] je proces, při kterém počítač specifikuje, do které z k kategorií vstupní data patří. Výstupem klasifikace nemusí být pouze jedno číslo určující kategorii, ale k-tice čísel určující pravděpodobnosti pro jednotlivé kategorie. Příkladem problému, které řeší klasifikace, může být detekce spamu, klasifikace obrázků a nebo rozpoznávaní ručně psaných znaků.

2.1.1 Neuronová síť

V případě klasifikace se obvykle využívá architektury neuronové sítě, která obsahuje vstupní vrstvu, následně alespoň jednu skrytou fully connected vrstvu a nakonec výstupní vrstvu, která obsahuje tolik neuronů, kolik je kategorií. Aktivační funkce poslední vrstvy je v případě klasifikace obsahující více tříd funkce softmax [13]. Výpočet softmaxu je následující:

$$P(w_t) = \frac{e^{\text{score}(w_t)}}{\sum_{\text{Pro všechny kategorie w}} e^{\text{score}(w)}}, \tag{2.1}$$

kde score(w) značí výstupní hodnotu neuronové sítě pro kategorii w. Výstupem je pak k-tice hodnot, kde každá hodnota reprezentuje pravděpodobnost jednotlivé kategorie. Nevýhodou Softmaxu je vysoká náročnost na výpočet při mnoha kategoriích a ztrácející se gradient při trénování pro vysoké hodnoty vnitřní funkce.

2.1.2 Evaluace modelu

Ke zkoumání úspěšnosti modelu je nutné použít metriku, která tuto úspěšnost bude měřit. Mezi základní metriky patří přesnost [13] klasifikace. Nevýhodou

Tabulka 2.1: Vztahy mezi pravdivostí/nepravdivostí.

	Předpověděl pravdu	Předpověděl nepravdu
Správně je pravda	Pravdivě pozitivní (TP)	Falešně negativní (FN)
Správně je nepravda	Falešně pozitivní (FP)	Pravdivě negativní (TN)

Tabulka 2.2: Příklad matice záměn pro tři kategorie.

		Skutečná kategorie		
		auto	kolo	člověk
Předpovězená kategorie	auto	5	10	3
	kolo	2	5	1
	člověk	6	0	12

této metriky je, že v případě, že je dataset nevyvážený např. v datasetu je 90 % dat kategorie 0 a zbytek 1, pak klasifikátor, který vždy předpoví kategorii 0, bude mít úspěšnost 90 %. Výpočet přesnosti je následující:

$$P = \frac{\#\text{ správně klasifikovaných dat}}{\#\text{ všech dat}} \tag{2.2}$$

Jednou z metrik, která výše zmíněný problém řeší, je F-measure. Dle [4] je F-measure definováno jako:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \tag{2.3}$$

kde $P = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$ (precision) a značí procento prvků, které označil jako pravdu a také pravdou byly a $R = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ (recall) značí procento prvků, které jsou na vstupu a byly správně identifikované. Výpočet hodnot TP, FP a FN je vidět v tabulce 2.1. Nejčastěji se využívá F1-measure, která nabývá hodnot od 0 do 1, kde hodnota 1 značí nejlepší možný výsledek. Výpočet F_1-measure je následující:

$$F_1 = \frac{2PR}{P + R} \tag{2.4}$$

Pro zkoumání úspěšnosti modelu na klasifikaci dat, které obsahují více kategorií lze využít matice záměn (confusion matrix) [4]. Matice obsahuje v řádcích skutečnou hodnotu kategorie a ve sloupcích hodnotu kategorie, která byla předpovězena. Příklad této matice je vidět v tabulce 2.2.

Pro nevyvážené datasety lze použít i metriku Matthew's Correlation Coefficient (MCC) [14]. Pro lepší pochopení metriky je její výpočet ilustrován pro binární klasifikaci. Vysvětlení jednotlivých proměnných je v tabulce 2.1.

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{FN} + \mathrm{TN}) \cdot (\mathrm{FP} + \mathrm{TN}) \cdot (\mathrm{TP} + \mathrm{FN})}} \tag{2.5}$$

Výpočet MCC lze zobecnit pro kategorizaci, která obsahuje více tříd než dvě:

$$\mathrm{MCC} = \frac{c \cdot s - \sum_k^K p_k \cdot t_k}{\sqrt{(s^2 - \sum_k^K p_k^2)(s^2 - \sum_k^K t_k^2))}}, \tag{2.6}$$

kde $t_k = \sum_i^K C_{ik}$ je počet, kolikrát se kategorie k měla opravdu objevit, $p_k = \sum_i^K C_{ki}$ je počet kolikrát se kategorie k objevila, $c = \sum_k^K C_{kk}$ je počet správně klasifikovaných dat, C_{ij} je hodnota v i-tém řádku a j-tém sloupci matice záměn (confusion matrix) a $s = \sum_i^K \sum_j^K C_{ij}$ je celkový počet všech dat.

Výstupem je hodnota začínající mezi -1 a 0 (závisející na distribuční funkci kategorií) pro kompletně špatně zvolené předpovědi a 1 pro perfektní předpovědi.

2.2 Redukce dimenzionality

Myšlenka redukce dimenzionality vychází z manifold hypothesis [15], která tvrdí, že vysoce dimenizonální data leží na varietách nižší dimenze. Cílem redukce dimenzionality [16] je mapování dat do prostoru nižší dimenze takovým způsobem, že se zachová co nejvíce relevantních informací. Redukce dimenzionality se používá při vizualizaci (např. do 2D prostoru) a nebo jako způsob zmírnění efektů prokletí dimenzionality. Formálně lze redukci dimenzionality zapsat jako $X \approx f(Y)$, kde $X \in \mathbb{R}^{n,m}$, $Y \in \mathbb{R}^{n,l}$, $l < m$ a f je funkce, která redukuje dimenzi.

2.2.1 Analýza hlavních komponent (PCA)

PCA je metoda, která se používá jak pro analýzu komponent, které mají největší vliv na výsledek, tak i pro redukci dimenze. Myšlenka PCA [17] spočívá v tom, že je potřeba zachytit dimenze dat, které mají největší rozptyl, tedy kde se skrývá nejvíce informace.

Při výpočtu PCA se dle [18] využívá spektrálního rozkladu kovarianční matice $C = \frac{1}{n}AA^T$, kde $A \in \mathbb{R}^{n,m}$ je matice obsahující data. Spektrální rozklad matice C je následující:

$$C = VDV^T, \tag{2.7}$$

kde $V \in \mathbb{R}^{n,n}$ je ortogonální matice obsahující v sloupcích vlastní vektory matice C a $D \in \mathbb{R}^{n,n}$ je diagonální matice obsahující seřazená vlastní čísla

matice C na diagonále. Redukce dimenze dat poté probíhá pomocí projekce $A' = AU_d$, kde U_d je matice obsahující prvních d sloupců matice V.

2.3 Detekce anomálií

Dle [19] je detekce anomálií problémem hledání vzorů v datech, které nevykazují očekávané chování. Mezi obvyklé aplikace těchto metod jsou detekce podvodů v pojišťovnách a bankách, detekce útoků v síťovém provozu a podobně.

2.3.1 Local outlier factor (LOF)

Local outlier factor [20] je metoda pro detekci anomálií, která rozhodne o bodu, zda-li je anomální či ne na základě hustoty okolních bodů v porovnání s k-nejbližšími sousedy, kde k je hyperparametr metody. Princip metody je následující.

K výpočtu LOF je využita *k-distance*(o), která udává vzdálenost bodu mezi bodem o a k-tým nejbližším sousedem. Dalším výpočtem je reachability distance, která se vypočítá jako:

$$\text{reach-dist}(a, b) = \max(\text{k-distance}(b), d(a, b)), \tag{2.8}$$

kde $d(a, b)$ značí vzdálenost mezi body a a b. Následně se vypočítá *Local reachability density*, která udává hustotu bodu.

$$\text{lrd}(a) = \Big(\frac{1}{k}\sum_{n}^{N}\text{reach-dist}(a, n)\Big)^{-1}, \tag{2.9}$$

kde N je množina k-nejbližších sousedů bodu a. Tento výpočet lze interpretovat jako vzdálenost k nejbližšímu clusteru bodů. Finálním výpočtem je hodnota LOF.

$$\text{LOF}(a) = \frac{\text{lrd}(a)}{\frac{1}{k}\sum_{n}^{N}\text{lrd}(n)}, \tag{2.10}$$

kde N je množina k-nejbližších sousedů bodu a. Výpočet lze interpretovat jako poměr mezi hustota bodu a a průměrnou hustotou jeho k-nejbližších sousedů. Pokud $\text{LOF}(a) \approx 1$, pak se nejedná o anomálii, pokud $\text{LOF}(A) \gg 1$, pak se o anomálii jedná.

Ensamble

Z důvodu, že metoda LOF je citlivá na výběr hyperparametru k [20], je využito ensamble modelů. Ensamble LOF využívá myšlenky z [21], ve které se náhodně vybere $\lfloor N/2 \rfloor$ až $N - 1$ příznaků (N značí celkový počet příznaků), které jsou následně použity jako vstup do LOF. Tento proces je několikrát opakován. Výstupem je průměrné LOF skóre přes všechny iterace. Pseudokód LOF je

vidět v algoritmu 1. Anomální data jsou pak klasifikována stejně jako v případě LOF výše. Dle autorů článku [21] tato metoda dává lepší výsledky jak na syntetických, tak i na reálných datech.

Data : Data D určená k detekci anomálií
Iterace : Počet iterací T
Počet příznaků: Počet příznaků n
LOF hodnoty : Výstup LOF hodnoty skore
for $i \leftarrow 0$ **to** T **do**
    pocet_priznaku = `nahodne_cislo`($n/2$, $n-1$);
    data = `vyber_n_priznaku`(D, pocet_priznaku);
    skore = skore + `LOF(data)`;
end
skore = skore/T;

Algoritmus 1: Algortimus ensamble LOF.

Kapitola 3

Realizace

Při realizaci bylo využito programovacího jazyka *Python 3.6* pro implementaci scraperu a k analýze dat, *SQLite* pro ukládání získaných dat a *Jupyter notebook*, který umožňuje kombinaci kódu a textu pro analýzu a prezentaci výsledků.

3.1 Výběr zpravodajského portálu

Výběr zpravodajského portálu byl zúžen na pět zpravodajských portálů, na kterých tráví uživatelé nejvíce času. Důvodem k výběru této metriky byla ochota uživatelů strávit na webu dostatek času a vyjadřovat se k článkům. Statistické informace o jednotlivých portálech [22] lze vidět v tabulce 3.1.

- `novinky.cz` Nejvíce navštěvovaný český portál obsahující i textovou diskuzi, bohužel nelze jednoznačně identifikovat uživatele.
- `idnes.cz` Zpravodajský portál, který lze využít. Obsahuje textovou diskuzi a jednotlivé uživatele lze jednoznačně identifikovat.
- `seznamzpravy.cz` Bohužel neobsahuje textovou diskuzi, a proto ho nelze použít.
- `aktualne.cz` Portál obsahující jak textovou diskuzi, tak i možnost jednoznačně identifikovat uživatele. Výhodou tohoto portálu je možnost stahování komentářů za pomocí API.
- `parlamentnilisty.cz` Zpravodajský portál, který je oproti ostatním portálům relativně málo navštěvován. Na druhou stranu textová diskuze obsahuje dost často vyhraněné názory a jednotlivé uživatelé je možné jednoznačně identifikovat.

Tabulka 3.1: Srovnání českých zpravodajských portálů za měsíc březen 2019. [22]

	Uživatelé	Zobrazení	Návštěvy	Čas
novinky.cz	4 556 100	137 161 954	73 327 667	387r 124d
idnes.cz	3 181 523	134 678 363	35 367 929	289r 263d
seznamzpravy.cz	3 965 836	65 922 200	41 892 085	148r 154d
aktualne.cz	2 613 106	55 673 006	17 660 437	108r 339d
parlamentnilisty.cz	765 400	24 534 950	6 537 881	85r 91d

3.2 Získání dat

Důležitou součástí této práce je získání dostatku dat pro analýzu zpravodajského portálu. Mezi výše zmíněnými portály zůstaly pouze tři použitelné portály. Portál `parlamentnilisty.cz` byl z výběru vyřazen z důvodu nižší návštěvnosti. Zbylé dva portály mají textovou diskuzi, umožňují jednoznačnou identifikaci uživatele a obsahují mnoho článků a aktivních uživatelů. Oproti `idnes.cz` nabízí `aktualne.cz` možnost stahování komentářů pomocí API díky využití externí platformy pro práci s diskuzí, která zmenší počet chyb, které se při parsování stránky mohou objevit. Z tohoto důvodu byl vybrán portál `aktualne.cz`.

3.2.1 Webscraping

Při webscrapingu je využito technologie *Python* a knihovny *Beautiful Soup*, která slouží k práci s HTML. Algoritmus pro stažení dat je následující. Nejdříve jsou staženy všechny proxy servery z `www.free-proxy-list.net`. Důvodem použití proxy serveru je možnost blokace IP adresy od zpravodajského portálu, a tím zamezení stahování dat. Kromě použití proxy serveru se také využívá změny *User Agent* v hlavičce HTTP požadavku na náhodný prohlížeč. Mezi každými dvěma úspěšnými HTTP požadavky se dvě sekundy čeká, aby nedocházelo k zatěžování serveru. Při webscrapingu bylo staženo 3 184 článků a 86 193 komentářů. Princip stahování dat je naznačen v algoritmu 2.

3.2.2 Databáze

Pro ukládání a následnou práci s daty se využívá databázového systému SQLite. Tento systém je vhodný pro menší systémy, které nepotřebují složité funkce, které nabízí pokročilejší databázové systémy. Právě v jeho jednoduchosti však tkví síla. Výhoda SQLite spočívá v tom, že je serverless [23], což znamená, že SQLite neběží v separátním procesu jako server, ale program, který chce s databází pracovat, zapisuje data přímo do souboru databáze. Schéma databáze lze vidět na obrázku B.1. Nejdůležitějšími atributy jsou

```
Vstup: Počet stránek ke stažení pocet_stranek
for i ← 0 to pocet_stranek do
    odkazy.pridej(stahni_odkazy(i));
end
foreach odkaz o v odkazy do
    html = stahni_clanek(o);
    uloz_clanek(html);
    diskuze = extrahuj_odkaz_na_diskuzi(html);
    uloz_diskuzi(stahni_diskuzi(diskuze));
end
```

Algoritmus 2: Stahování diskuzí a komentářů z `www.aktualne.cz`.

u článku jeho obsah, perex, titulek a datum a čas zveřejnění. U komentáře to je autor, obsah, datum a čas zveřejnění a počet liků a disliků. Ostatní nepodstatné a osobní údaje byly z databáze smazány.

3.3 Použití metod pro reprezentaci textu

Pro reprezentaci textu vektorem byly využity technologie Doc2vec, BERT a Doc2vec s přetrénovanými reprezentacemi slov. Při trénování bylo využito služby Google collaboratory, která nabízí grafickou kartu a prostředí Jupyter notebook zdarma. Díky této službě je možné trénovat modely mnohem rychleji a bez zátěže na vlastním zařízení.

3.3.1 Doc2vec

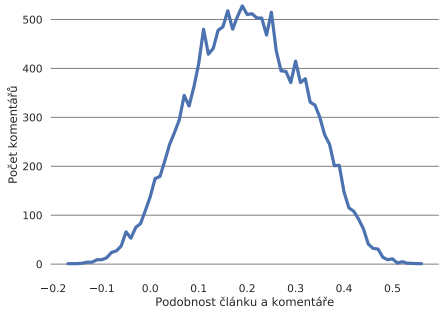
Při implementaci Doc2vec je využito 30 000 nejčastějších slov, méně častá slova jsou pak nahrazena speciálním slovem *[UNK]*. Mimo slova *[UNK]* je do slovníku přidáno slovo *[PAD]*, které slouží jako výplňové slovo na začátku a na konci vstupu. Dále jsou z textu odstraněny všechny netisknutelné znaky navíc (dvě mezery se změní na jednu) a speciální znaky a všechna písmena jsou změněna na malá. Text je na závěr rozdělen na tokeny dle mezer.

Při implementaci Doc2vec DM bylo využito machine learning frameworku Keras. Schéma implementace neuronové sítě 3.2 je identické se schématem 1.3 s rozdílem Flatten vrstvy, která změní rozměry tensoru na 1D vektor. Jedná se pouze o technologickou úpravu, která nemá vliv na výsledek neuronové sítě. Přestože autoři článku [5] doporučují hodnoty hyperparametrů modelu DM, ideální hodnoty jsou specifické pro každou úlohu. Hodnoty obecně doporučovaných a využívaných hyperparametrů jsou vidět v tabulce 3.2. Celý model je trénován na 10 epochách za pomocí Adam optimalizátoru o velikosti batche 256 na všech komentářích, perexech a obsazích článků.

Rozložení podobností článku a komentáře zaokrouhlené na 2 desetinná místa lze vidět na grafu 3.1. Je patrné, že jedná o rozdělení svým tvarem

Tabulka 3.2: Doporučené a využívané hodnoty hyperparametrů.

	Doporučené hodnoty
Velikost okna	10
Velikost vektoru pro slovo	400
Velikost vektoru pro text	400



Obrázek 3.1: Rozdělení hodnot podobnosti zaokrouhlené na dvě desetinná místa pro model Doc2vec.

Tabulka 3.3: Vybraná slova a jim nejbližší slova získaná z Doc2vec.

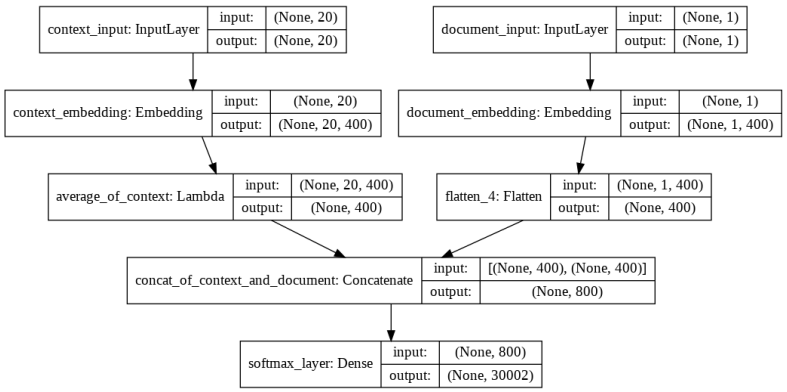
Slovo	Nejbližší slova
ods	spd, kscm, cssd, stan, piráti
usa	rusko, nato, rusku, kldr, ruska
hamáček	pavel, místopředseda, chovanec, milan, předseda
babiš	babiše, babišovi, premiér, bureš, premiéra

dosti podobné normálnímu rozdělení a nabývá hodnot od -0,17 do 0,56 se směrodatnou odchylkou 0,11.

Mimo vzniklých reprezentací pro texty, vzniká při trénovaní Doc2vec i vektorová reprezentace slov. Za pomocí těchto vzniklých reprezentací lze ověřit, zda byla zachycena sémantická informace. Seznam vybraných slov a jim nejpodobnějších je vidět v tabulce 3.3.

3.3.2 BERT

Pro získání vektorových reprezentací bylo využito knihovny Bert-as-service [24], která využívá předtrénovaného modelu BERT-Base, Multilingual Cased, dostupného z `https://github.com/google-research/bert`. Způsob, kterým Bert-as-service získává textové reprezentace, je průměr ze všech reprezentací tokenů z druhé skryté vrstvy od konce. Důvodem k výběru této vrstvy je



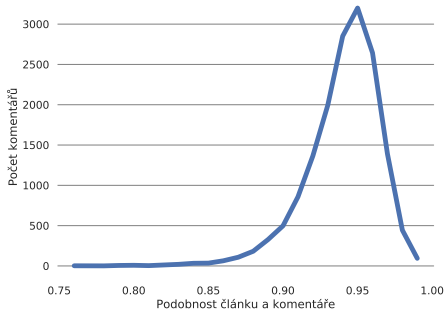
Obrázek 3.2: Schéma modelu Doc2vec vygenerované z Keras.

skutečnost, že dle [24] je poslední vrstva příliš naučená na maskování věty a předpovídání, zda-li dvě věty na sebe navazují, a proto by nedávala vhodné reprezentace. První vrstva je na druhou stranu příliš blízko slovním reprezentacím, a proto lze vybrat libovolnou vrstvu kromě první a poslední.

Z důvodu toho, že BERT pracuje s konstantní délkou (512) vstupu a některé z článků a komentářů jsou delší než tato hodnota, je nezbytné tyto texty rozdělit na bloky o maximální délce 512 znaků. Algoritmus pro získání textové reprezentace pro text delší než 512 tokenů nejdříve rozdělí text na tokeny pomocí WordPiece. Pokud je tokenizovaný text delší než 510 tokenů (vždy bude přidán token *[CLS]* a *[SEP]*), pak je text rozdělen na bloky o 500 tokenech. Pokud je poslední slovo např. *květina*, mohlo by se stát, že bude rozděleno do dvou bloků na *['k', '##vet']* a *['##ina']*). Z tohoto důvodu se volí pouze 500 tokenů, aby v nejhorším případě bylo možné přidat celé slova do rozděleného bloku. Výstupem modelu je vektor o velikosti 768, který reprezentuje celý vstupní text.

Nevýhodou předtrénovaného modelu pro více jazyků je, že obsahuje i tokeny, které se v českém jazyce neobjevují, a tím se zmenšuje možný slovník. Další nevýhodou je, že model je předtrénován na cca 100 jazycích, a proto musí být schopen se všemi jazyky pracovat. Tím se omezuje kvalita reprezentací oproti natrénování modelu pouze na jednom jazyku. Tuto skutečnost lze vyřešit za pomocí natrénování modelu BERT na jednom jazyce. Trénovaní by však bylo nákladné, dle autorů BERT [6] a ceníku `cloud.google.com/tpu/docs/pricing`, by natrénování stálo alespoň 570 dolarů pro menší z modelů.

U modelu BERT oproti model Doc2vec je rozdělení podobnosti užší. Graf rozdělení podobností je vidět na obrázku 3.3, ze kterého je patrné, že má menší rozptyl než model Doc2vec.



Obrázek 3.3: Rozložení podobnosti pro model BERT.

3.3.3 Doc2Vec s předtrénovanou reprezentací slov

Doc2vec s předtrénovanou reprezentací slov vychází jak z modelu BERT, z kterého využívá vektorových reprezentací pro slova, tak i z Doc2vec, kde využívá principu trénování textových reprezentací. Jako vektor reprezentující slovo je vybrán průměrný vektor přes všechny tokeny z předposlední vrstvy. Důvodem k vybrání předposlední vrstvy je, že i samotné slovo může být rozděleno do několika tokenů, a proto je vhodné, aby na ně byla aplikována Attention. Výsledné vektory jsou předány modelu Doc2vec, který je natrénován s doporučenými hodnotami z článku [5], které jsou stejné jako v modelu Doc2vec z tabulky 3.2 s výjimkou velikosti vektoru pro slovo, která je z modelu BERT 768. Stejně jako model Doc2vec i Doc2vec s předtrénovanými reprezentacemi je natrénován na všech komentářích, článcích a jejich perexech.

Výhodou využití vektorů z modelu BERT je množství dat, na kterých byl model natrénován a velikost těchto vektorů. Na druhou stranu se tímto ztrácí variabilita modelu, která je nyní pouze závislá na reprezentaci dokumentů. Nevýhodou oproti modelu BERT je pak nemožnost využití kontextových slovních reprezentací, které BERT nabízí. Kontextová reprezentace je taková, že význam slova a tedy i jeho vektorová reprezentace se mění na základě kontextu. Příkladem může být slovo kohoutek, které může reprezentovat vodovodní kohoutek i malého kohouta. Pro model Word2vec resp. Doc2vec jsou to slova se stejnou reprezentací, pro BERT nikoliv.

3.3.4 Srovnání modelů

Pro srovnání a výběr modelů je využita úloha určení kategorie článku. Tato úloha byla vybrána proto, že texty spadající do stejné kategorie jsou sémanticky podobné, a proto model s největší úspěšností na této úloze nejlépe zaznamená sémantickou informaci textu.

Prvním krokem výběru modelů je získání vektorových reprezentací pro jednotlivé články a úprava dat. Celkově je v datasetu 31 kategorií, kde některé

Tabulka 3.4: Počty článků pro jednotlivé kategorie.

Kategorie	Počet článků
Domácí	1156
Zahraniční	659
Počasí	502
Politika	361
Ekonomika	283
Auto	136
Ostatní	44
Doprava	43

z nich obsahují jen malé množství článků, a proto jsou přesunuty do jiné kategorie. Například kategorie volby, obsahuje pouze pět článků, a proto jsou tyto články přesunuty do kategorie politika. Takovýmto způsobem je zredukován počet kategorií na osm. Výsledek této úpravy je vidět v tabulce 3.4.

Při výběru nejlepšího modelu se postupuje následovně:

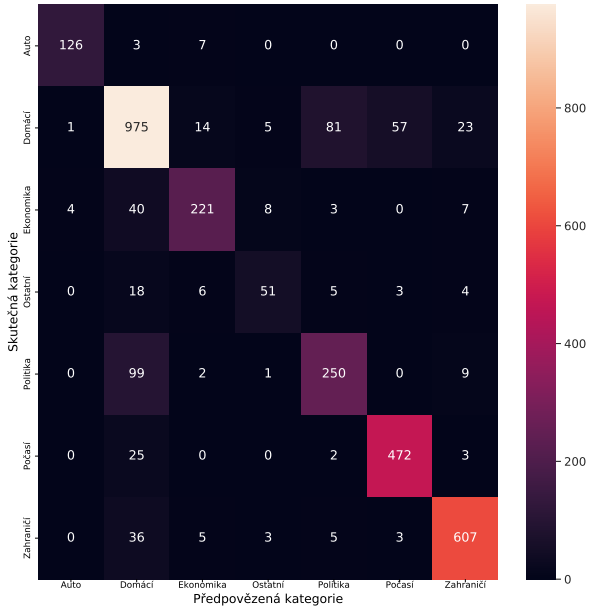
1. Dataset se rozdělí na 3 množiny, trénovací (80 % dat), validační (16 %) a testovací (4 %).

2. Na trénovací množině se natrénují modely k-nebližších sousedů, rozhodovací strom, náhodný les a nakonec neuronová síť.

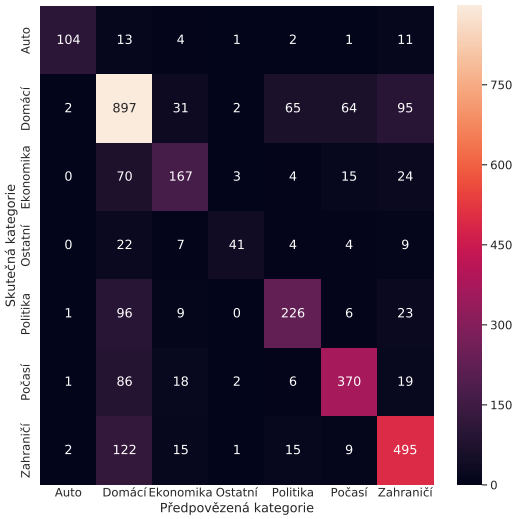
3. Pro každý z modelů se ladí hyperparametry a na validační množině je vybrán model, který má největší Matthews correlation coefficient (důvodem k výběru Matthews correlation coefficient je nevyvážený dataset).

4. Nejlepší model pro reprezentaci textu je ten, který má nejlepší klasifikátor na základě Matthews correlation coefficient na validační množině.

Výsledky pro jednotlivé klasifikátory a modely jsou vidět v tabulce C.1, nejlepším modelem je tedy BERT (a poté Doc2vec. Matice záměn pro BERT pro celý stažený dataset je vidět na obrázku 3.4. Řádky značí skutečnou kategorii, sloupce předpovězenou kategorii klasifikátorem a hodnota v buňce značí počet takto zakategorizovaných článků. V této matici lze vidět, že kategorie Domácí a Politika splývají. Tuto skutečnost lze vysvětlit tím, že existuje pouze malá hranice mezi kategorií Domací a Politika a je pouze na autorovi článku, jakou kategorii článku přiřadí. Stejnou skutečnost lze pozorovat i na matici záměn na obrázku 3.5 pro model Doc2vec, kde je míra špatně zakategorizovaných článku ještě znatelnější.



Obrázek 3.4: Matice záměn pro kategorizaci článků pro model BERT.



Obrázek 3.5: Matice záměn pro kategorizaci článků pro model Doc2vec.

Kapitola 4

Analýza

Dle srovnání z předchozí kapitoly využijeme modelu BERT a Doc2vec pro analýzu článků a komentářů. V obou modelech se podařilo získat nejvyššího Matthews correlation coefficient pro neuronovou síť.

4.1 Základní vlastnosti korpusu

Při webscrapingu bylo staženo 3 184 článků a 86 193 komentářů. Průměrná délka článku je 413 slov a komentáře 26 slov. Projekci článků do 2D prostoru za pomocí PCA lze vidět v příloze na obrázku B.2 (projekce pomocí PCA pro model Doc2vec je na obrázku B.3), ze které je vidět, že modelu BERT ani Doc2vec se nepodařilo dostatečně zaznamenat sémantickou informaci obsaženou v textu. V porovnání s vizualizací [24] na obrázku z B.4, která slouží pro srovnání různých vrstev BERT modelu na titulcích anglicky psaných novin, kde každá barva znamená jinou kategorii, vychází, že model je v případě anglického textu schopen zachytit sémantickou informaci lépe (vektory jsou do prostoru rozloženy dle kategorií). Tuto skutečnost lze vysvětlit tím, že předtrénovaný model je určen pro více jazyků, a proto jeho vnitřní reprezentace nejsou tak dobré, jako pro samostatný anglický jazyk (to stejné platí i pro předtrénovaný slovník). Dalším důvodem může být průměrování vektorových reprezentací pro delší texty, kde není stoprocentně využita Attention na celém textu. Je však možné, že data jsou nelineárně separována ve vyšších dimenzích a projekce pomocí PCA tuto skutečnost není schopna zaznamenat.

4.2 Analýza komentářů

Cílem analýzy komentářů je nalezení nerelevantních komentářů k obsahu článků za pomocí zkoumání vektorových reprezentací pro komentář a pro články. Ke zkoumání podobnosti článku a komentáře se využívá kosínová podobnost,

která říká, jak moc si je článek s komentářem podobný a nabývá hodnot od -1 do 1, kde hodnota 1 značí stejné dokumenty a hodnota -1 úplně odlišné dokumenty.

Při analýze je využito pouze článků, které obsahují nějaké komentáře (např. počasí nebývá tak často diskutované téma) a komentářů, které jsou první úrovně (komentáře druhé a další úrovně nemusí reagovat na obsah článku, a proto nemusí být relevantní) a pro které existuje vektorová reprezentace článku (např. u článků obsahující pouze fotografie). Po takovém zpracování datasetu zůstane 880 článků a 16 139 komentářů.

4.2.1 Klasifikace komentářů

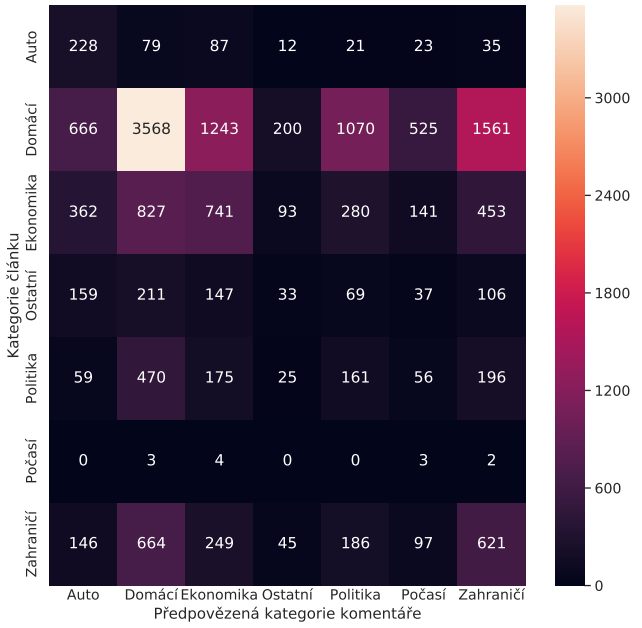
Prvním pohledem, kterým se lze dívat na komentáře je, zda-li kategorie komentáře je stejná jako kategorie článku. Na matici záměn 4.1 pro model BERT je vidět, že kategorie komentáře je málokdy stejná jako kategorie článku. Z matice je dále vidět, že hranice mezi kategorií Domací, Ekonomika a Politika je tenká a klasifikátor dost často určí kategorii jinou. Stejná skutečnost nastává i pro kategorii počasí, u které by se předpokládalo, že její texty jsou tak odlišné, že bude jednoznačně oddělená od ostatních. Při ruční inspekci se však ukazuje, že např. komentáře „Tak se mějte, koblihy.... =)“ nebo „Ze sjezdu jsme vysli mnohem silnejsi a stmelenejsi !“ jsou zakategorizovány jako počasí. Toto stejné platí i pro model Doc2vec, kde dle matice záměn na obrázku B.5 klasifikátor zakategorizuje většinu komentářů u článků z kategorie domácí do kategorie počasí, což s největší pravděpodobností nereflektuje skutečnost.

Počet komentářů, které mají stejnou kategorii jako článek, na který reagují je 5 355 pro model BERT a 2 764 pro Doc2vec z celkového počtu 16 139. Většina komentářů tedy nemá stejnou kategorii. Vysvětlení této skutečnosti může být v tom, že některé kategorie splývají, jak již bylo avizováno výše, klasifikátor špatně vyhodnotil kategorii a nebo je reakce jiné kategorie než článek.

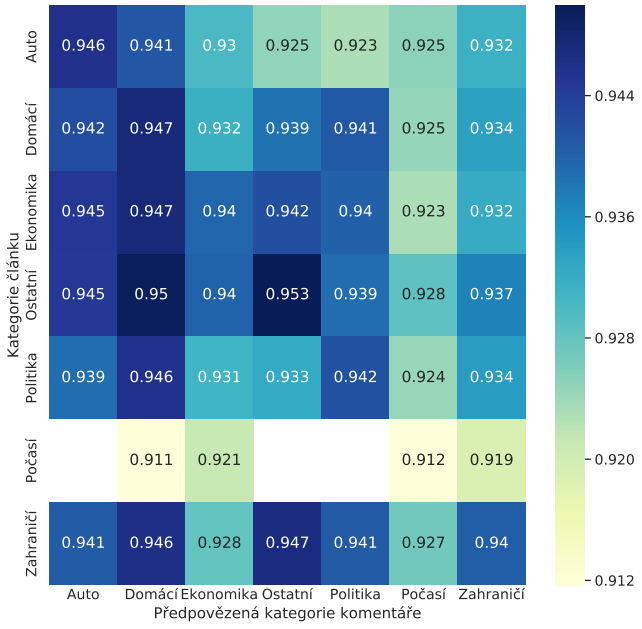
4.2.2 Podobnost článku a komentáře na základě kategorie

Závislost podobnosti komentáře s článkem a kategorií článku a předpovězené kategorie komentáře lze hodnotit za pomocí teplotní matice. Sloupce matice na obrázku 4.2 tvoří předpovězené kategorie, řádky kategorie článků a hodnota je pak průměrná podobnost komentáře při použití modelu BERT. Z matice lze vyčíst, že pokud je komentář kategorie počasí, pak má nízkou podobnost s článkem a také, že pokud jsou článek a komentář stejné kategorie, tak mají obvykle vysokou podobnost (s výjimkou počasí). Tuto skutečnost lze lépe vidět na grafu 4.3 pro model BERT a na grafu 4.4 pro model Doc2vec.

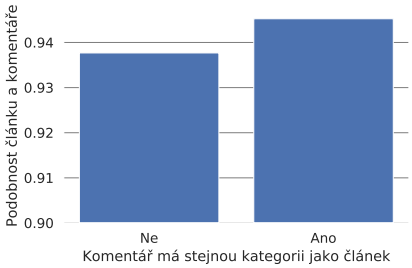
Podobně lze zkoumat teplotní matici na obrázku B.6 pro model Doc2vec, kde se ale trend vyšší podobnosti článku a komentáře, kteří mají stejnou kategorii, ztrácí.



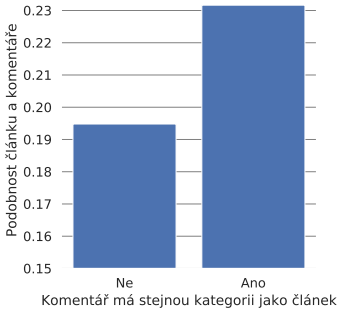
Obrázek 4.1: Matice záměn pro kategorii komentářů a článku pro model BERT.



Obrázek 4.2: Teplotní matice pro kategorie komentářů, kategorie článků a průměrné podobnosti pro model BERT.



Obrázek 4.3: Graf zobrazující průměrnou podobnost komentáře a článku, pokud mají resp. nemají stejnou kategorii pro model BERT.



Obrázek 4.4: Graf zobrazující průměrnou podobnost komentáře a článku, pokud mají resp. nemají stejnou kategorii pro model Doc2vec.

4.2.3 Zkoumání relevance komentáře s článkem

Při zkoumání relevance komentáře s článkem se zaměříme na globální hledisko, tedy komentáře, které mají největší vzdálenost s článkem při využití modelu BERT. Tyto komentáře a titulky článků, na které reagují, jsou vidět v příloze v tabulce C.2. Většinou se jedná o krátké komentáře neobsahující žádnou přidanou hodnotu. Pohled na nejvíce relevantní komentáře je vidět v příloze v tabulce C.3. V této tabulce je vidět, že většina komentářů je věcných a jsou spíše delší. Vysvětlení, proč delší texty jsou podobnější článku než texty kratší spočívá v tom, že v delším textu je lépe využita Attention, která má vliv na kvalitu výsledku.

Stejným způsobem lze analyzovat komentáře pomocí vektorů z metody Doc2vec. Ty nejméně relevantní komentáře jsou vidět v příloze v tabulce C.4 a nejvíce v příloze v tabulce C.5.

Z čistě subjektivního hlediska se zdají být oba modely dobré ke zkoumání relevance. Výhodu modelu BERT spočívá v tom, že dobře detekuje relevantní příspěvky. Nevýhoda modelu spočívá v tom, že délka textu má vliv na kvalitu vytvořených vektorů. V případě nejméně relevantních se zdá být metoda Doc2vec mírně lepší, protože detekuje i delší texty. Na druhou stranu v detekci opravdu relevantních příspěvků se objevuje spoustu komentářů, které relevantní nejsou.

4.2.4 Zkoumání relevance komentáře s článkem u dvou vybraných článků

Mezi 2 nejvíce diskutované články patří *Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková* (303 komentářů) a *ODS vy-*

Tabulka 4.1: Tabulka nejméně relevantních komentářů pro model BERT.

Češi si za drahá data mohou sami ...	*ODS vyloučila Václava Klause ...*
A G R O T E L	Konečně
TO JE ALE OBLUDA!!!!!!!!!!!!!!	Konečně!
ano, bude líp!	Vrazi.
Blbý a blbší	Tak to je nebudu volit.
Ta paní je normální?	V poradku. Naprosto v poradku.

loučila Václava Klause mladšího. Není loajální ke straně, řekl předseda Fiala (276 komentářů). Tabulka 4.1 zobrazuje komentáře s nejmenší kosínovou podobností a tabulka C.7 zobrazuje ty s největší podobností. Stejně jako v případě globálního pohledu i v lokálním pohledu jsou krátké komentáře označeny jako méně relevantní a delší komentáře jako relevantní. Přestože se daří pomocí kosínovo vzdálenosti naleznout relevantní komentáře u článku o cenách dat, v případě článku o Klasovi mladším tomu tak není a občas komentáře sklouzávají mimo téma (např. u předposledního komentáře).

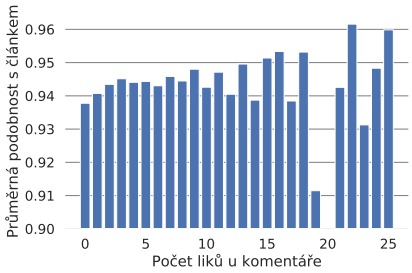
Při využití modelu Doc2vec to dopadá podobně jako při použití modelu BERT. Nejméně relevantní komentáře jsou vidět v příloze v tabulce C.6 a nejvíce relevantní v příloze v tabulce C.8. V tomto případě se model Doc2vec občas detekuje nerelevantní příspěvky jako relevantní a obráceně.

Při srovnání výsledků mezi modelem BERT a Doc2vec se ukazuje, že Doc2vec nebyl tak úspěšný v hledání relevantních a nerelevantních příspěvků oproti modelu BERT. Z tohoto důvodu je model BERT lepší pro zkoumání relevantních komentářů u konkrétního článku.

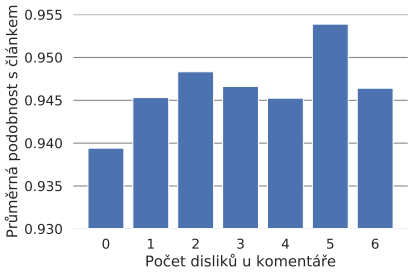
4.2.5 Zkoumání relevance komentáře s článkem v závislosti na kategorii

Kromě zkoumání relevance mezi článkem a komentářem lze zkoumat obecné závislosti mezi podobností komentáře s článkem a atributy komentáře. V následující analýze jsou využity nejdříve vektorové reprezentace z modelu BERT a nakonec jsou srovnány s modelem Doc2vec.

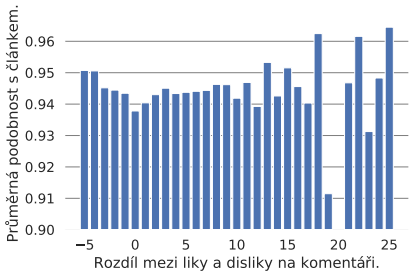
Prvním z hledisek je závislost podobnosti na základě počtu liků. Z grafu 4.5 je vidět, že neexistuje souvislost mezi podobností komentáře a článku a počtem liků na komentáři (konec grafu není reprezentativní, neboť obsahuje málo datových bodů). Co se týká závislosti počtu disliků na průměrné podobnosti článku s komentářem, podobnost paradoxně roste. Tato závislost je zobrazena v grafu 4.6. Stejné chování lze pozorovat na grafu 4.7 a 4.8, které zobrazují vliv rozdílu liků a disliků na podobnosti článku s komentářem. Tato pozorování lze interpretovat tak, že počet disliků resp. liků je špatným ukazatelem rele-



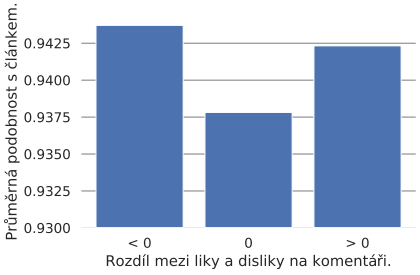
Obrázek 4.5: Závislost mezi průměrnou podobností komentáře a článku a počtem liků na komentáři pro model BERT.



Obrázek 4.6: Závislost mezi průměrnou podobností komentáře a článku a počtem disliků na komentáři pro model BERT.



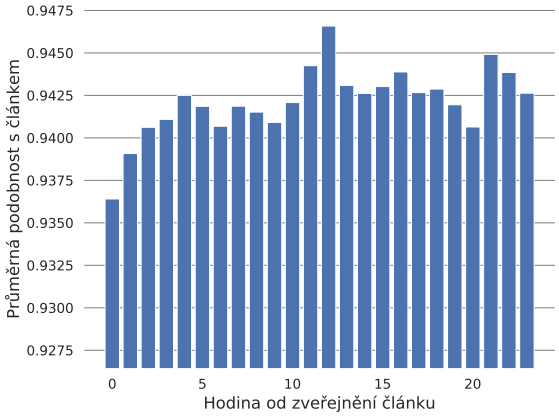
Obrázek 4.7: Závislost mezi průměrnou podobností komentáře a článku a rozdílem počtu liků a disliků na komentáři pro model BERT.



Obrázek 4.8: Závislost mezi průměrnou podobností komentáře a článku a rozdílem počtu liků a disliků na komentáři rozdělená dle nuly pro model BERT.

vantnosti komentáře. Četnosti jednotlivých liků a disliků jsou vidět v příloze v tabulce C.9 a tabulce C.10, z kterých je vidět, že diskutující jsou ochotni příspěvky více likovat než dislikovat.

Jediným rozdílem modelu Doc2vec oproti modelu BERT je v případě závislosti podobnosti na počtu disliků. Jak je vidět na obrázku B.8 podobnost zde zůstává na konstantní hodnotě a nakonec klesá. Otázkou však je, zda-li pozorovaný efekt není způsoben pouze nízkým počtem komentářů, které jsou dislikované. Grafy pro ostatní závislosti z předchozího odstavce, tedy grafy liků a jejich rozdílu jsou zobrazeny v příloze na obrázcích B.7, B.9 a B.10.



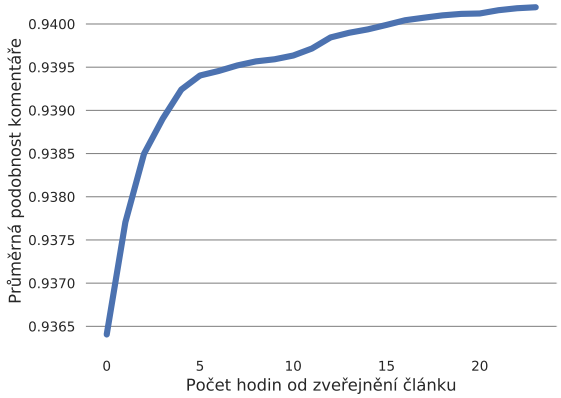
Obrázek 4.9: Graf závislosti podobnosti komentáře a článku a počtem hodin od zveřejnění pro model BERT.

4.2.6 Zkoumání relevance komentáře s článkem v závislosti na čase

Dalším zajímavým pohledem na data je závislost podobnosti komentáře a článku v závislosti na počtu hodin od zveřejnění článku. Jak je vidět z obrázku 4.9, podobnost zveřejněných komentářů v čase pro model BERT roste. Nejpodobnější komentáře jsou obvykle přidány 12 hodin od zveřejnění. Podobně lze takto zanalyzovat závislost podobnosti komentáře a článku na základě doby od zveřejnění s ohledem na jednotlivé kategorie. Některé z kategorií mají rostoucí tendenci v závislosti na čase např. domácí (obrázek B.12) a nebo se drží okolo konstantní hodnoty jako třeba politika (obrázek B.15) nebo ekonomika (obrázek B.13).

V případě Doc2vec podobnost článku závislá na hodině od zveřejnění postupně roste. Graf závislosti podobnosti komentáře s článkem a hodině od zveřejnění je vidět na obrázku B.18. Oproti BERT tedy podobnost komentáře s časem roste. Stejně jako v modelu BERT, některé z kategorií v čase rostou, např. v kategorii domácí na obrázku B.20. Zajímavým rozdílem oproti modelu BERT je, že pro kategorie ekonomika na obrázku B.21 nejdříve podobnost roste až do páté hodiny, pak klesá až do čtrnácté a poté zase roste a nakonec se stabilizuje.

Posledním hlediskem, kterým se lze na data dívat, je průměrná podobnost článku a komentáře v určité hodiny od zveřejnění. Rozdíl oproti předchozímu grafu spočívá v tom, že se zkoumají všechny komentáře, které byly zveřejněny do nějaké hodiny a ne pouze komentáře, které byly zveřejněny v určitou hodinu. Graf 4.10 ukazuje, že průměrná podobnost komentářů v závislosti na čase roste. Interpretace tohoto pozorování je, že komentáře přidané dříve dost často nemusí reagovat na článek, ale reagují pouze na základě klíčových slov



Obrázek 4.10: Průměrná podobnost komentáře k článku po zveřejnění pro model BERT.

(např. Babiš, Kalousek, Huawei, ...) a tedy nereagují přímo k článku. Stejně jako tomu bylo v odstavci výše, lze tuto analýzu udělat pro každou kategorii. Z grafů B.28, B.27, ... lze vidět, že u většiny kategorií článků má podobnost komentáře a článku tendenci růst, výjimkou jsou pak kategorie auto B.26, počasí B.31 a politika B.30. U prvních dvou kategorií lze diskutovat, zdali na výsledek nemá vliv nedostatek dat (jednotlivé kategorie obsahují 485, 12 a 1142 komentářů). U politiky toto však nelze říci, a proto paradoxně komentáře, které jsou přidány pár hodin po zveřejnění, bývají podobnější článkům. Jedno z vysvětlení může být, že komentáře přidané dřív reagují na článek a ty později se od něj oddalují z důvodu započaté diskuze, tedy částečně reagují na výroky ostatní.

Při srovnání modelu BERT s modelem Doc2vec vychází, že se chovají dost podobně. Z grafu závislosti průměrně podobnosti komentářů v čase B.33 lze vidět, že průměrná podobnost v čase roste. To stejné platí pro jednotlivé kategorie s výjimkou kategorie politika. Na grafu B.38 je vidět, že podobnost v čase roste na rozdíl od podobností komentáře a článku v případě modelu BERT. Pro ostatní kategorie, průměrná podobnost komentáře a článku závislá na čase a kategorii roste, jak je vidět na obrázcích B.35, B.36, B.37, B.39 a B.40. Výjimkou je pak kategorie auto B.34, kde nejdříve podobnost roste, a pak oscilujme kolem konstantní hodnoty.

4.2.7 Analýza jednotlivých uživatelů

Při analýze jednotlivých uživatelů jsou bráni v potaz pouze uživatelé, kteří přidali alespoň dvacet příspěvků. Jako nejrelevantnější uživatel je brán uživatel, který má největší průměrnou podobnost s článkem. Obdobně to platí i na druhou stranu tj. uživatel, který přidává nejméně relevantní příspěvky je ten,

který má průměrně nejmenší podobnost s článkem. V tabulce C.11 je vidět výsledek pro model BERT, a že uživatel s největší průměrnou relevanci píše často k tématu a smysluplně, nejsou to pouze výkřiky do tmy. V případě nejméně relevantního přispěvatele je vidět v tabulce C.12, že se jedná pouze o výkřiky do diskuze, které dost často nejsou relevantní a ještě k tomu se takto vyjadřuje opakovaně pod jedním z článků. Stejné pozorování lze učinit i v případě Doc2vec s tím rozdílem, že uživatel s nejmenší průměrnou relevancí nepřispívá pouze krátkými komentáři. Výsledky pro uživatele s nejpodobnějšími příspěvky k článku jsou vidět v tabulce C.13 a s nejméně podobnými v tabulce C.14. Všechny čtyři tabulky jsou zkráceny pouze na pět záznamů.

4.3 Detekce anomálií

Cílem detekce anomálií v komentářích je nalezení takových příspěvků, které se ke zbytku komentářů nehodí. Takovým příkladem může být politické komentáře u článcích o autech nebo počasí. K detekci anomálií je využito algoritmu ensamble Local outlier factor. Důvodem k využití ensamble metody namísto samotného použití algoritmu jsou lepší výsledky pro vysokodimenzionální data.

4.3.1 Anomální uživatelé

Pro zkoumání anomálního chování uživatelů jsou využiti pouze uživatelé, kteří přispěli alespoň 20 příspěvky. Vektor reprezentující uživatele je vypočten tak, že se od vektoru článku odečte vektor komentáře. Výsledný vektor pro uživatele je průměrný vektorů přes všechny jeho příspěvky. Model BERT byl schopen detekovat pouze jednoho uživatele jako anomálního. Jeho příspěvky lze vidět v tabulce C.15. Doc2vec na druhou stranu nalezl dvacet anomálních uživatelů (včetně uživatele z modelu BERT). V tabulce C.16 jsou vidět vybraní uživatelé a jejich příspěvky, které byly nalezeny pomocí algoritmu LOF s využití reprezentací z Doc2vec. Z tabulek je vidět, že algoritmus Local outlier factor byl opravdu schopen nalézt anomální chování uživatele. Ukazuje se, že je lepší využít vektorových reprezentací z modelu Dod2vec, kde je algoritmus schopen nalézt o mnoho více anomálních uživatelů. Zajímavé také je, že v případě uživatelů, kteří vykazují anomální chování dochází k častému komentování u jednoho příspěvku obvykle s nízkou podobností s článkem.

Další možností jak zkoumat anomální uživatele je využít jiného výpočtu k získání vektorové reprezentace. Výpočet výsledného vektoru je následující:

$$(u \cdot w)\frac{w}{||w||^2} - u, \tag{4.1}$$

kde u reprezentuje vektor pro komentář a w vektor pro článek. Výsledný vektor lze interpretovat jako vektor kolmý na komentář a článek. Uživatelé, kteří se

chovají anomálně, jsou pro model BERT zobrazeny v tabulce C.17. V tomto případě se podařilo modelu identifikovat dva anomální uživatele. Stejně jako v předchozím případě se podařilo identifikovat uživatele 1089 a k tomu ještě uživatele 289 (v tabulce je pouze uživatel 289, protože uživatel 1089 je již v tabulce C.15). Pro model Doc2vec je vidět výsledek v tabulce C.18. V tomto případě nalezl algoritmus LOF 31 uživatelů, kteří vykazují anomální chování. Z výsledků je vidět, že některé příspěvky jsou relevantní např. při pohledu na uživatel 1 481 a jeho komentáře se nezdá být nikterak anomální.

4.3.2 Anomální komentáře

Při hledání anomálních komentářů se postupuje stejně jako v předchozí kapitole, tedy za pomocí využití ensamble LOF. Prvním způsobem, jak jsou hledány anomálie, je za pomocí rozdílu vektoru pro článek a komentář. Při využití modelu BERT byl algoritmus schopen detekovat 1075 anomálních příspěvků (cca 6.7 % ze všech příspěvků). Některé vybrané komentáře jsou vidět v příloze v tabulce C.19. Z nalezených anomálií je vidět, že algoritmus nalezl hodně příspěvků, které nejsou anomální a dost často reagují na článek. Příkladem může být komentář *Pokud roste cena novych aut, je zcela nabiledni, ze poroste i cena ojetin, a zvlaste tech mladsich.* u článku *Zánovní ojetiny čeká zdražení. Ceny nebudou klesat ani u lehce jetých dieselů.* Při využití modelu Doc2vec pro reprezentaci textů se algoritmu nepodařilo nalézt žádný anomální komentář.

Při výpočtu vektoru z modelu BERT způsobem v rovnici 4.1 bylo detekováno 162 anomálních příspěvků. Některé z nich jsou zobrazeny v příloze v tabulce C.20. V tomto případě již algoritmus nedetekoval tolik komentářů jako anomálních a také neobsahuje tolik příspěvků, které by nebyly anomální. Subjektivně se zdá, že výpočet pomocí výpočtu 4.1 dává lepší výsledky. Model Doc2vec nalezl 2 012 anomálních příspěvků. Tato hodnota se zdá být příliš velká na to, aby se jednalo o anomální chování. Vybrané příspěvky jsou vidět v příloze v tabulce C.21. Komentáře se nezdají být anomální a dost často bývají i relevantní k tématu.

Dalším způsobem, kde lze aplikovat detekci anomálií, je u konkrétního článku. V tomto případě byl využit nejdiskutovanější článek *Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková.* V tomto případě se podařilo detekovat anomálie pouze v případě využití modelu BERT. Nalezené anomálie pro výpočet pomocí 4.1 je vidět v tabulce C.22 a pro použití pouze vektoru komentáře v tabulce C.23. Z vybraných komentářů je vidět, že se jedná o anomální příspěvky, které nevykazují očekávaný obsah komentáře.

Závěr

Cílem práce bylo vytvořit ucelený pohled na možnosti využití metod pro zpracování přirozeného jazyka k analýze komentářů českého zpravodajského portálu. Při srovnání modelů BERT, Doc2vec a Doc2vec s předtrénovanými slovními reprezentacemi se ukazuje, že BERT a Doc2vec jsou schopny lépe zaznamenat sémantickou informaci textu. V dalším výzkumu by bylo možné zkoumat vliv způsobu získání reprezentací pro text u modelu BERT na výsledky analýzy nebo získání slovních reprezentací pro model Doc2vec s předtrénovanými reprezentacemi.

V analýze se ukázalo, že jak model Doc2vec, tak i model BERT jsou vhodnými nástroji pro získání reprezentací textu a k následnému využití k analýze. Přestože jsou oba modely vhodné k zkoumání relevance příspěvků, ukazuje se, že model BERT je výrazně lepší v nalezení opravdu relevantních příspěvků. Malou nevýhodou je, že při detekci nerelevantních příspěvků vrací nejdříve pouze krátké texty, které sice nerelevantní jsou, ale šlo by je detekovat např. na základě počtu slov v komentáři. Pokud se jedná o model Doc2vec, tak ten je schopný s dobrou úspěšností nalézt relevantní resp. nerelevantní příspěvky. Bohužel dost často jeho výstup obsahuje falešně pozitivní a nebo falešně negativní výsledky. Z tohoto důvodu nelze Doc2vec využít bez ručního ověření výstupu. V případě analýzy jednotlivých uživatelů se ukázaly oba modely srovnatelné a vhodné k hledání uživatelů, které přispívají relevantními resp. nerelevantními příspěvky.

Při detekce anomálií z globálního hlediska vychází model BERT lépe. BERT byl v případě sofistikovanějšího výpočtu schopen nalézt anomální příspěvky na rozdíl od modelu Doc2vec, který detekoval buď žádné nebo téměř osminu všech příspěvků jako anomálních. V případě detekce anomálií pouze u konkrétního článku nebyl model Doc2vec schopen identifikovat anomálie, a proto není pro tuto úlohu vhodný v porovnání s modelem BERT, který byl úspěšný.

Bibliografie

1. CHOWDHURY, Gobinda G. Natural language processing. *Annual review of information science and technology*. 2003, roč. 37, č. 1, s. 51–89.
2. MÜLLER, A.C.; C, M.A.; GUIDO, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016. ISBN 9781449369903. Dostupné také z: `https://books.google.cz/books?id=1-4lDQAAQBAJ`.
3. MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Efficient Estimation of Word Representations in Vector Space. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. 2013. Dostupné také z: `http://arxiv.org/abs/1301.3781`.
4. JURAFSKY, Daniel; MARTIN, James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000. ISBN 0130950696.
5. LE, Quoc; MIKOLOV, Tomas. Distributed Representations of Sentences and Documents. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. Beijing, China: JMLR.org, 2014, s. II–1188–II–1196. ICML'14. Dostupné také z: `http://dl.acm.org/citation.cfm?id=3044805.3045025`.
6. DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*. 2018, roč. abs/1810.04805. Dostupné z arXiv: `1810.04805`.
7. VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N; KAISER, Ł ukasz; POLOSUKHIN, Illia. Attention is All you Need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN,

S.; GARNETT, R. (ed.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, s. 5998–6008. Dostupné také z: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

8. WANG, Alex; SINGH, Amanpreet; MICHAEL, Julian; HILL, Felix; LEVY, Omer; BOWMAN, Samuel. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018, s. 353–355. Dostupné také z: https://www.aclweb.org/anthology/W18-5446.
9. RAJPURKAR, Pranav; ZHANG, Jian; LOPYREV, Konstantin; LIANG, Percy. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *CoRR*. 2016, roč. abs/1606.05250. Dostupné z arXiv: 1606.05250.
10. SCHUSTER, Mike; NAKAJIMA, Kaisuke. Japanese and Korean Voice Search. In: *International Conference on Acoustics, Speech and Signal Processing*. 2012, s. 5149–5152.
11. ALAMMAR, Jay. *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)* [online]. 2018 [cit. 2019-04-29]. Dostupné z: https://jalammar.github.io/illustrated-bert/.
12. ALAMMAR, Jay. *The Illustrated Transformer* [online]. 2018 [cit. 2019-04-29]. Dostupné z: https://jalammar.github.io/illustrated-transformer/.
13. GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. The MIT Press, 2016. ISBN 0262035618.
14. DEVELOPERS, scikit-learn. *Model evaluation: quantifying the quality of predictions* [online] [cit. 2019-04-20]. Dostupné z: https://scikit-learn.org/stable/modules/model_evaluation.html#matthews-corrcoef.
15. MA, Y.; FU, Y. *Manifold Learning Theory and Applications*. CRC Press, 2011. ISBN 9781466558878. Dostupné také z: https://books.google.cz/books?id=6pr1txA0_yMC.
16. JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.
17. MADSEN, R. E.; HANSEN, L. K.; WINTHER, O. *Singular Value Decomposition and Principal Component Analysis* [online]. 2004 [cit. 2019-04-17]. Dostupné z: http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4000/pdf/imm4000.pdf. Technická zpráva.

18. WANG, J. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer Berlin Heidelberg, 2012. ISBN 9783642274978. Dostupné také z: https://books.google.cz/books?id=0RmZRb2fLpgC.

19. CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly Detection: A Survey. *ACM Comput. Surv.* 2009, roč. 41, č. 3, s. 15:1–15:58. ISSN 0360-0300. Dostupné z DOI: 10.1145/1541880.1541882.

20. BREUNIG, Markus M.; KRIEGEL, Hans-Peter; NG, Raymond T.; SANDER, Jörg. LOF: Identifying Density-based Local Outliers. *SIGMOD Rec.* 2000, roč. 29, č. 2, s. 93–104. ISSN 0163-5808. Dostupné z DOI: 10.1145/335191.335388.

21. LAZAREVIC, Aleksandar; KUMAR, Vipin. Feature Bagging for Outlier Detection. In: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. Chicago, Illinois, USA: ACM, 2005, s. 157–166. KDD '05. ISBN 1-59593-135-X. Dostupné z DOI: 10.1145/1081870.1081891.

22. GEMIUS, NetMonitor SPIR. *Gemius AUDIENCE - Kategorie* [online] [cit. 2019-04-17]. Dostupné z: https://rating.gemius.com/cz/tree/82.

23. *SQLite Is Serverless* [online]. 2018 [cit. 2019-04-17]. Dostupné z: https://www.sqlite.org/serverless.html.

24. XIAO, Han. *bert-as-service* [https://github.com/hanxiao/bert-as-service]. 2018 [cit. 2019-04-29].

Příloha **A**

Seznam použitých zkratek

NLP Natural language processing

HTTP Hyper text transfer protocol

DM Distributed memory

CBOW Continous bag of words

DBOW Distributed bag of words

BERT Bidirectional Encoder Representations from Transformers

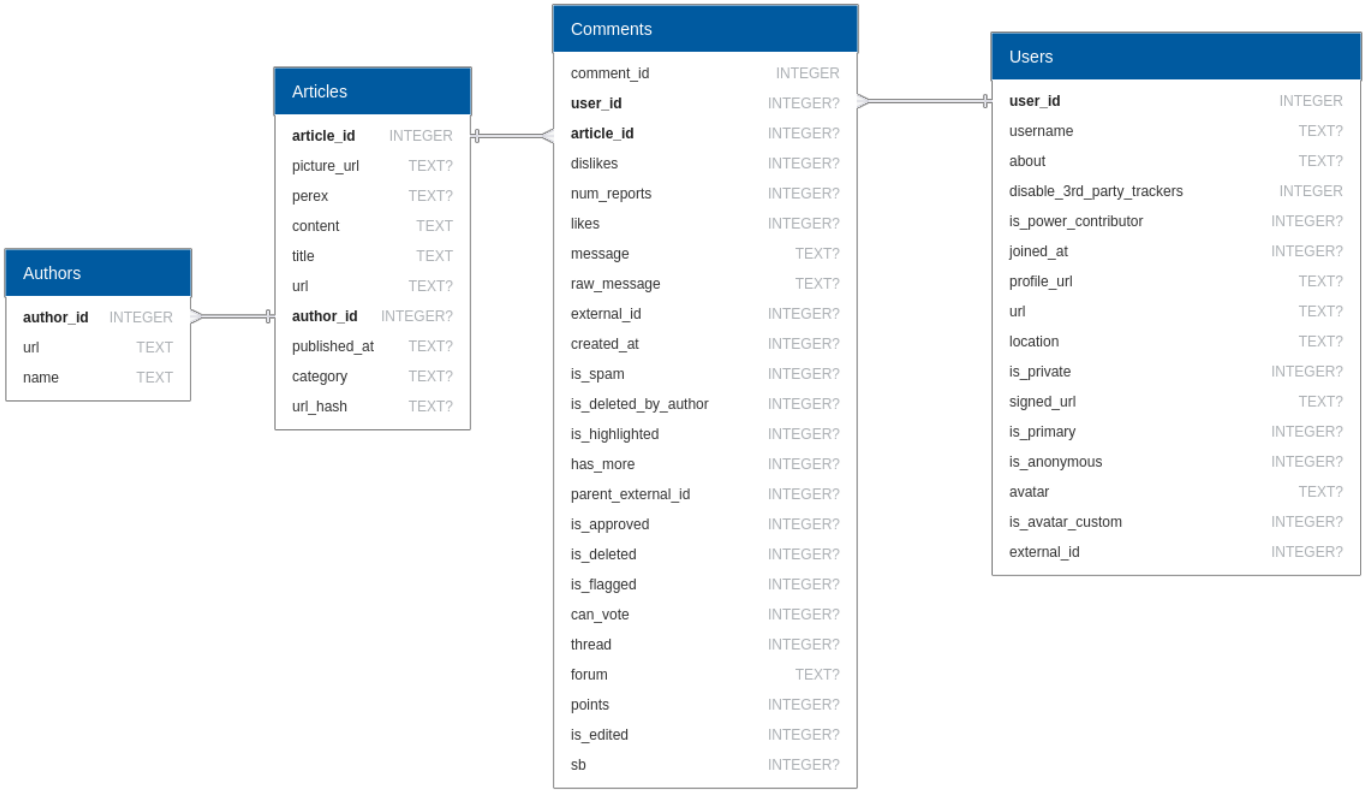
MCC Matthews correlation coefficient

LOF Local outlier factor

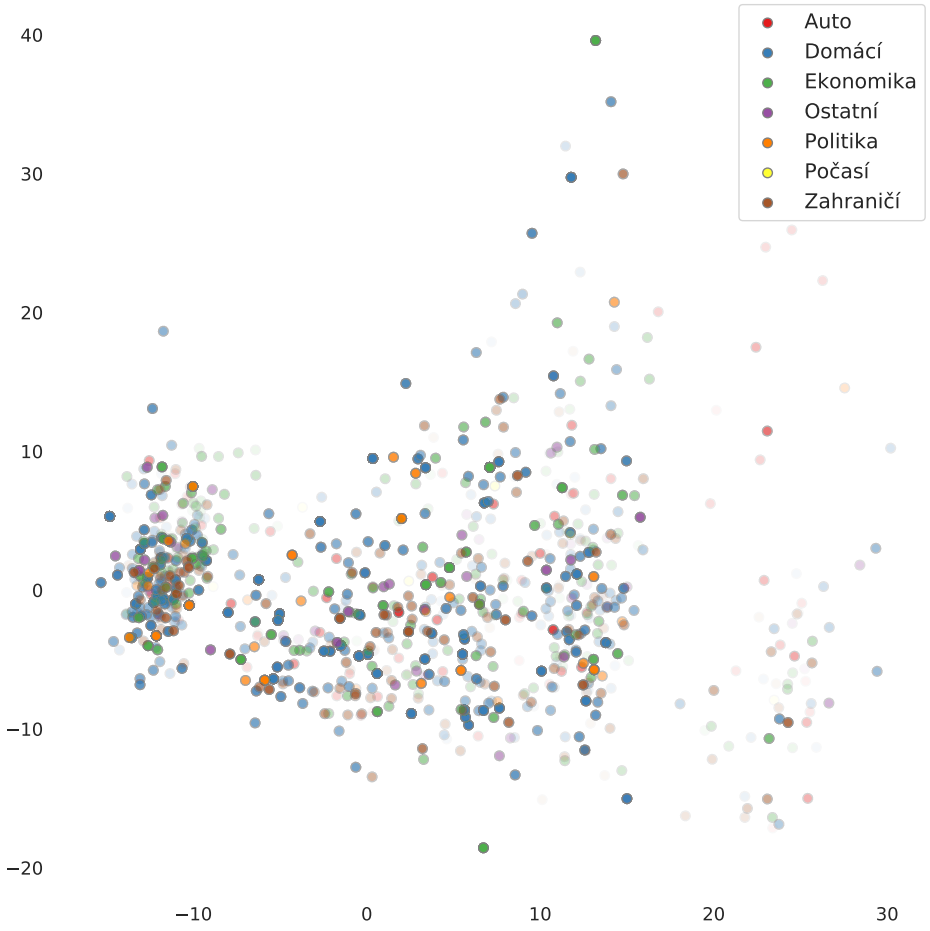
LSA Latent semantic analysis

Příloha B

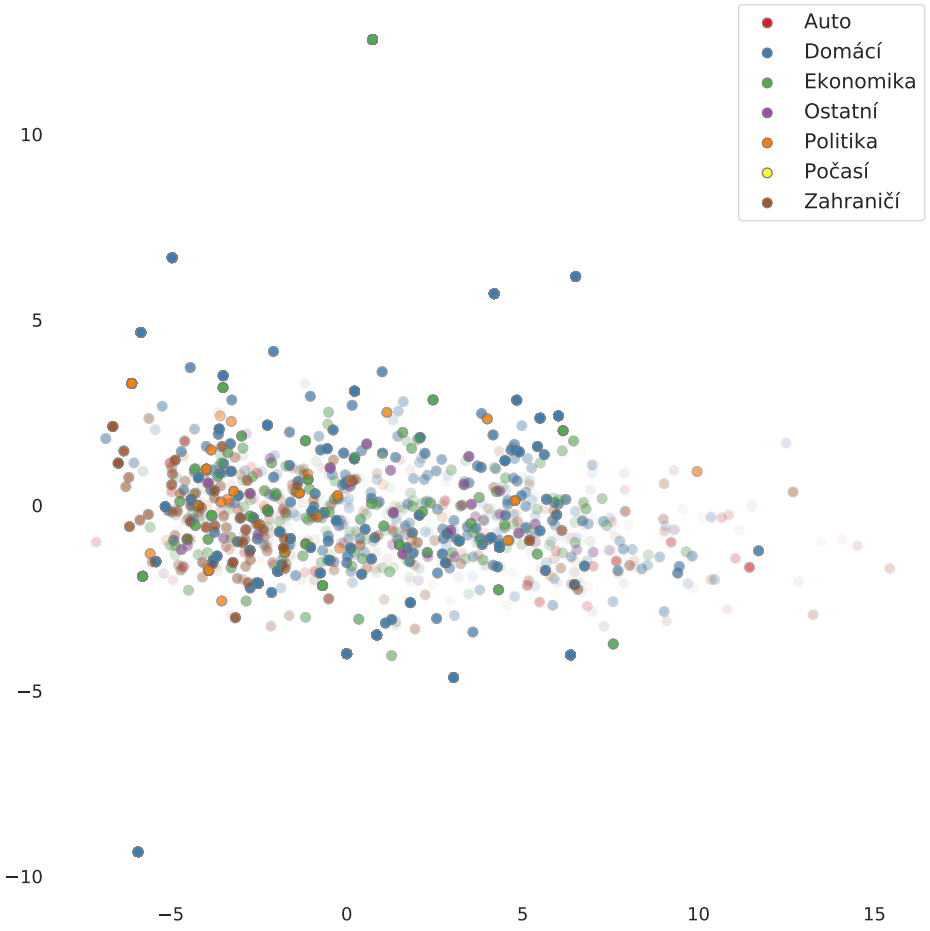
Obrázky



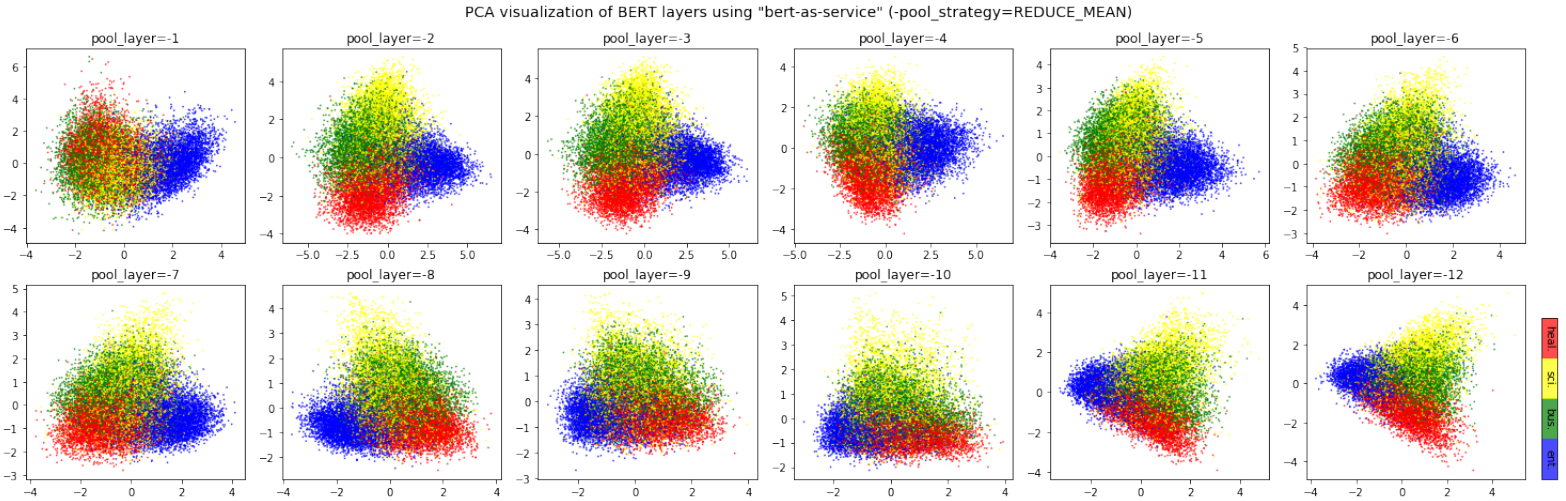
Obrázek B.1: Schéma databáze.



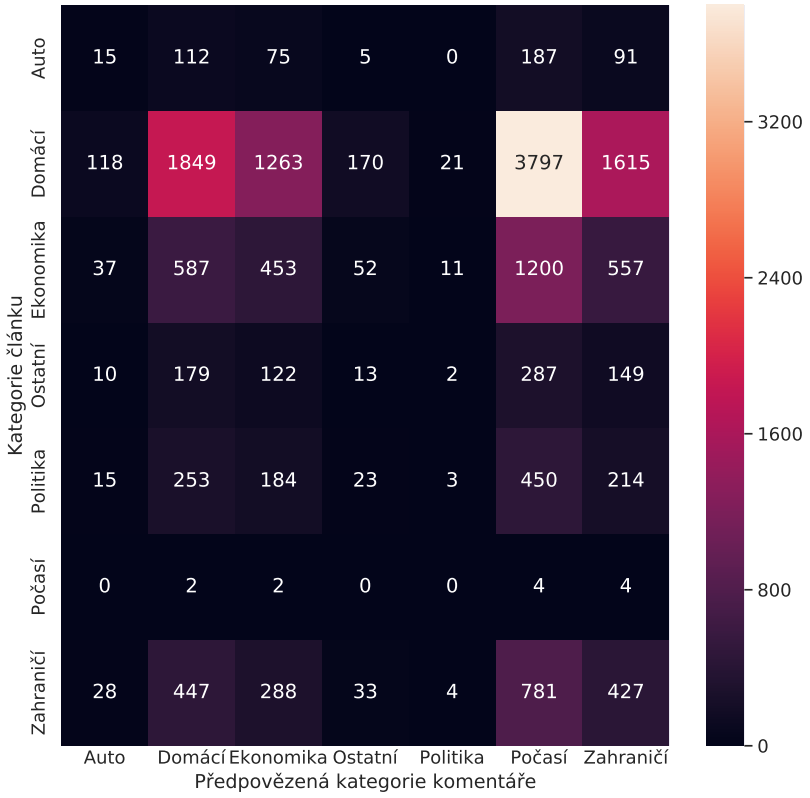
Obrázek B.2: Vizualizaci obsahu článků promítaná do 2D za pomocí PCA pro model BERT.



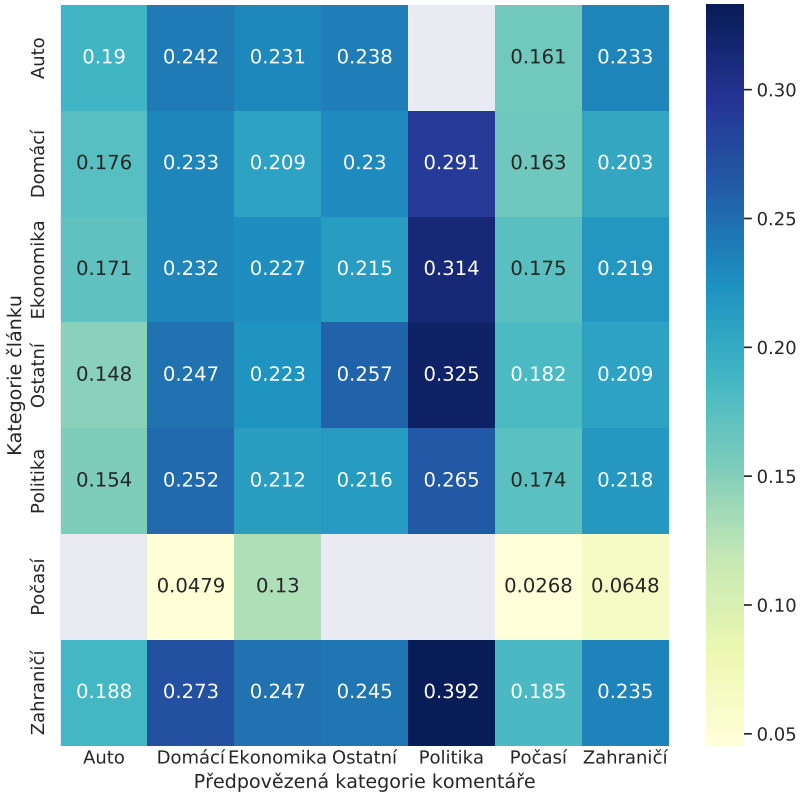
Obrázek B.3: Vizualizaci obsahu článků promítaná za pomocí PCA pro model Doc2vec.



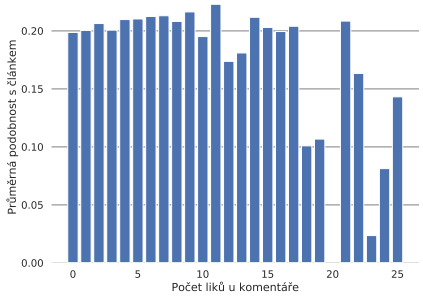
Obrázek B.4: Vizualizace titulků článků z amerických novin na různých vrstvách BERT. Rozdílné barvy značí jiné kategorie. [24]



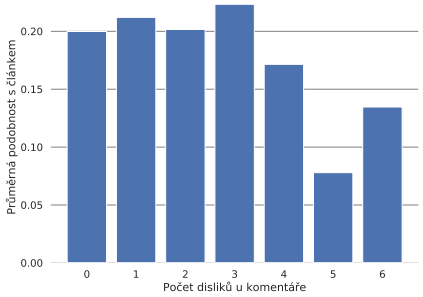
Obrázek B.5: Matice záměn pro kategorizaci komentářů pro modelu Doc2vec.



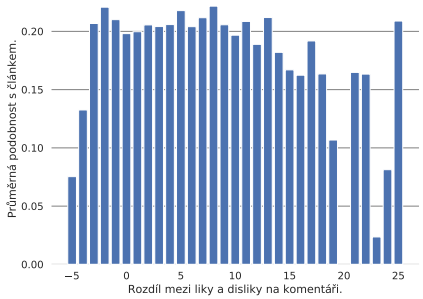
Obrázek B.6: Teplotní matice pro kategorie komentářů, kategorie článků a průměrné podobnosti pro model Doc2vec.



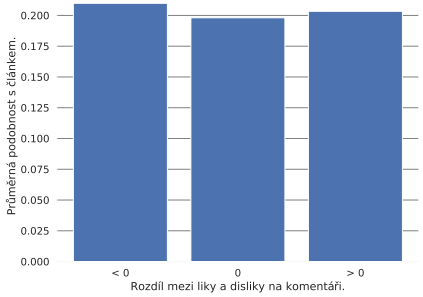
Obrázek B.7: Závislost mezi průměrnou podobností komentáře a článku a počtem liků na komentáři pro model Doc2vec.



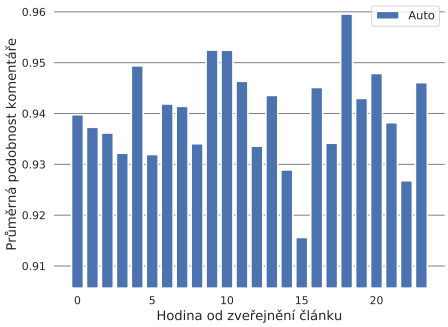
Obrázek B.8: Závislost mezi průměrnou podobností komentáře a článku a počtem disliků na komentáři pro model Doc2vec.



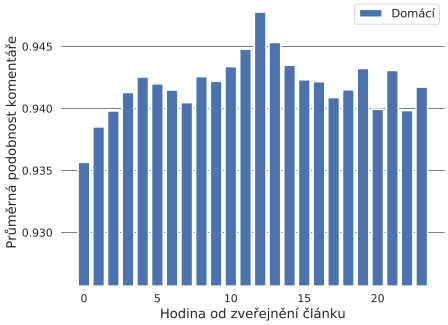
Obrázek B.9: Závislost mezi průměrnou podobností komentáře a článku a rozdílem počtu liků a disliků na komentáři pro model Doc2vec.



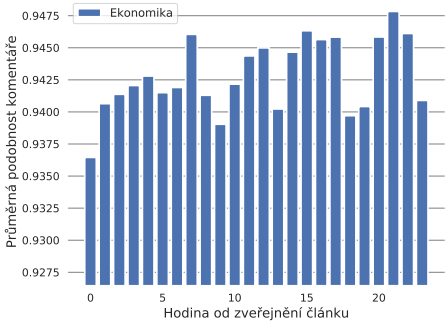
Obrázek B.10: Závislost mezi průměrnou podobností komentáře a článku a rozdílem počtu liků a disliků na komentáři rozdělená dle nuly pro model Doc2vec.



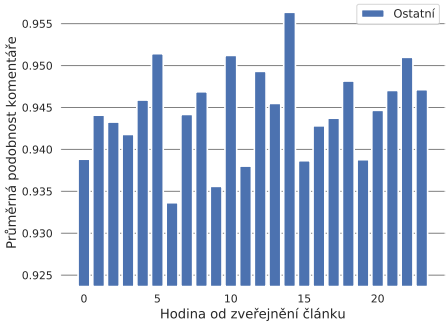
Obrázek B.11: Graf závislosti podobnosti komentáře s článkem pro kategorii auto na čase zveřejnění pro model BERT.



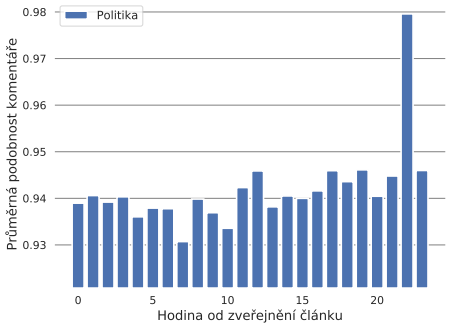
Obrázek B.12: Graf závislosti podobnosti komentáře s článkem pro kategorii domácí na čase zveřejnění pro model BERT.



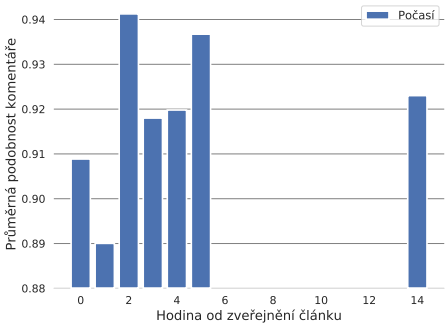
Obrázek B.13: Graf závislosti podobnosti komentáře s článkem pro kategorii ekonomika na čase zveřejnění pro model BERT.



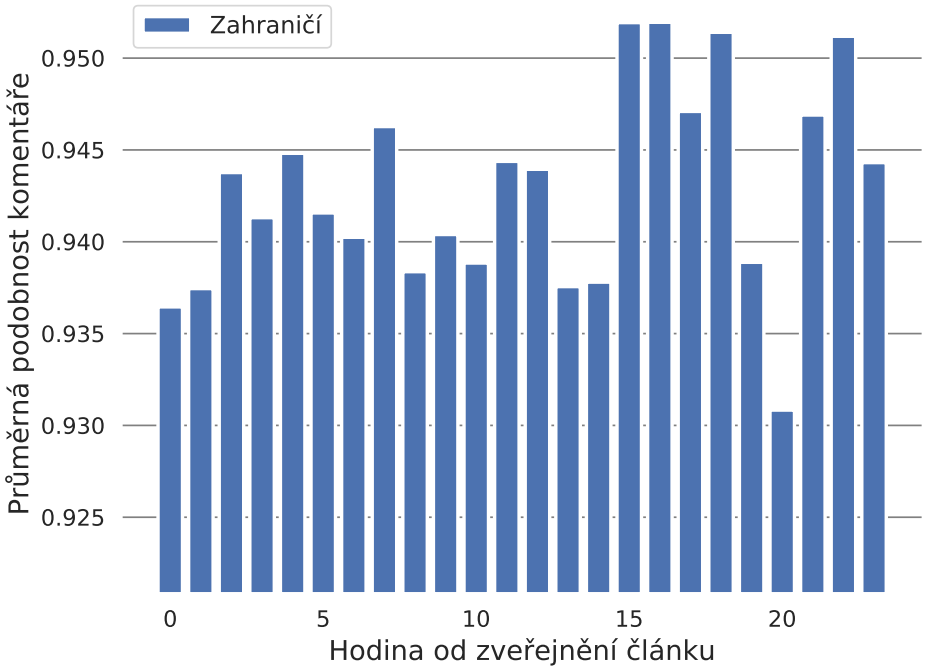
Obrázek B.14: Graf závislosti podobnosti komentáře s článkem pro kategorii ostatní na čase zveřejnění pro model BERT.



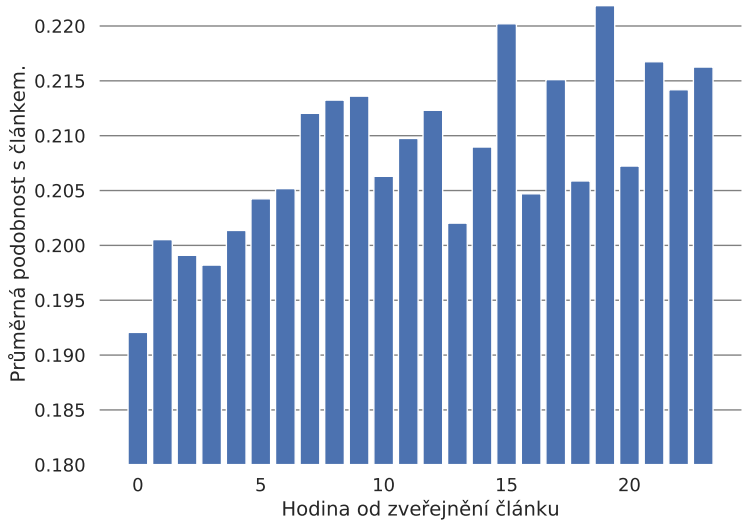
Obrázek B.15: Graf závislosti podobnosti komentáře s článkem pro kategorii politika na čase zveřejnění pro model BERT.



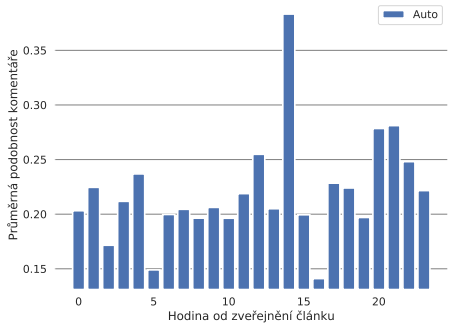
Obrázek B.16: Graf závislosti podobnosti komentáře s článkem pro kategorii počasí na čase zveřejnění pro model BERT.



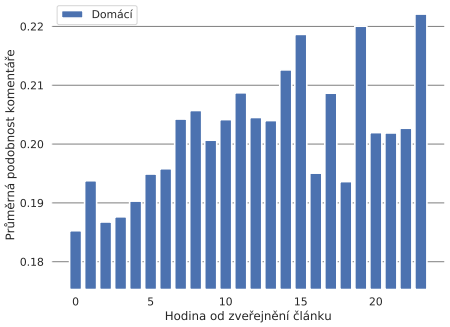
Obrázek B.17: Graf závislosti podobnosti komentáře s článkem pro kategorii zahraničí na čase zveřejnění pro model BERT.



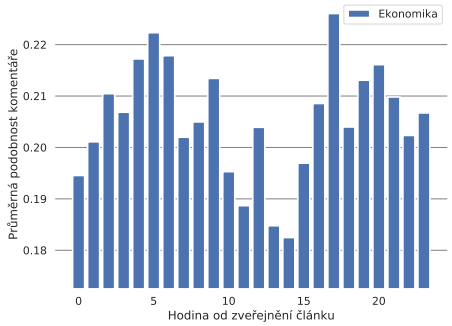
Obrázek B.18: Graf závislosti podobnosti komentáře a článku a počtem hodin od zveřejnění pro model Doc2vec.



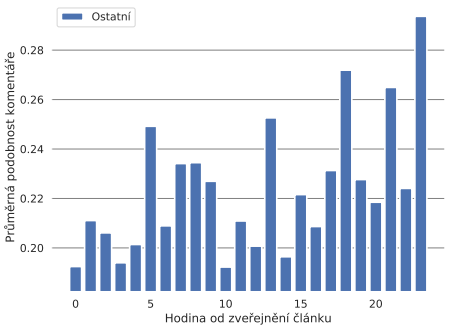
Obrázek B.19: Graf závislosti podobnosti komentáře s článkem pro kategorii auto na čase zveřejnění pro model Doc2vec.



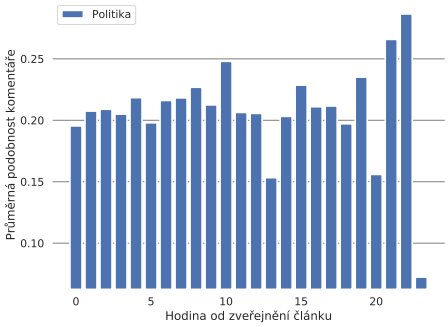
Obrázek B.20: Graf závislosti podobnosti komentáře s článkem pro kategorii domácí na čase zveřejnění pro model Doc2vec.



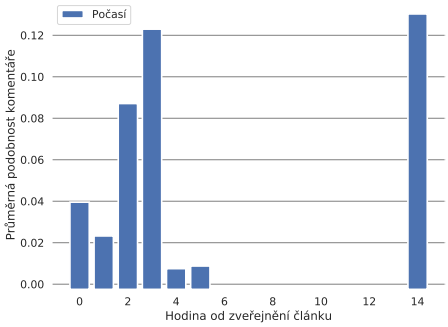
Obrázek B.21: Graf závislosti podobnosti komentáře s článkem pro kategorii ekonomika na čase zveřejnění pro model Doc2vec.



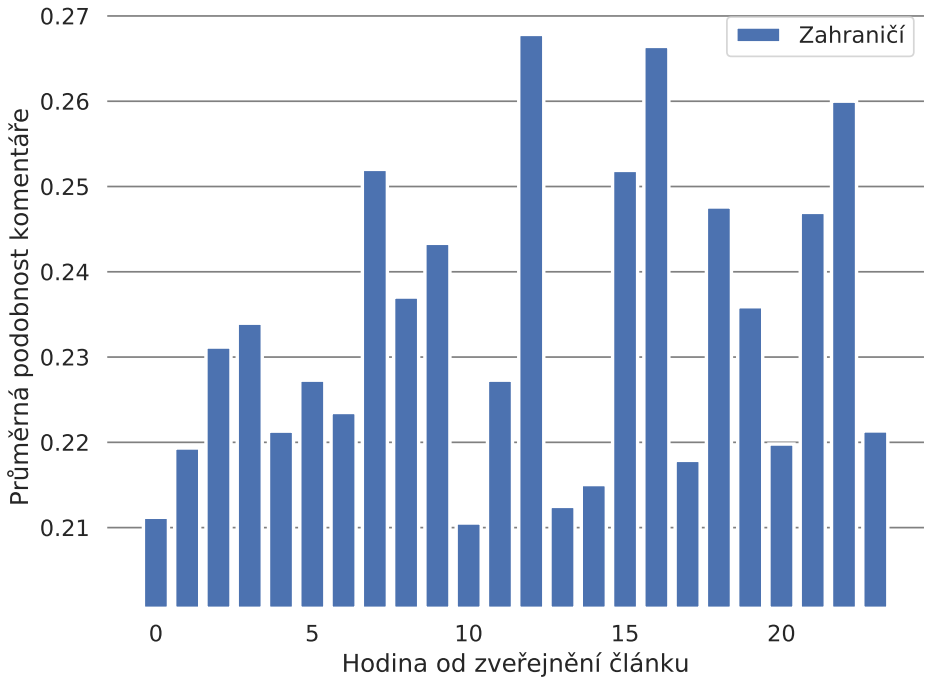
Obrázek B.22: Graf závislosti podobnosti komentáře s článkem pro kategorii ostatní na čase zveřejnění pro model Doc2vec.



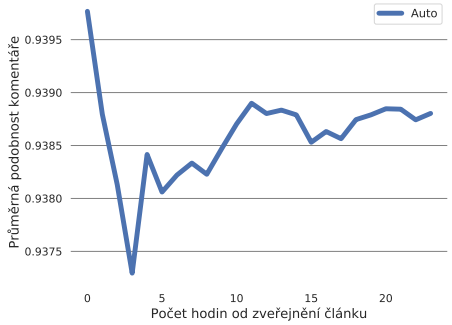
Obrázek B.23: Graf závislosti podobnosti komentáře s článkem pro kategorii politika na čase zveřejnění pro model Doc2vec.



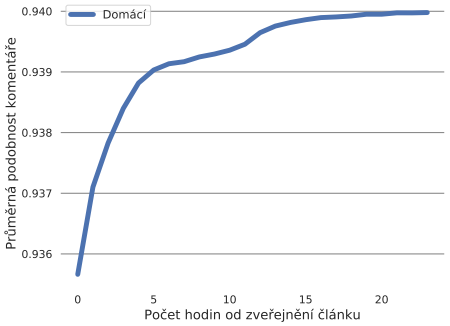
Obrázek B.24: Graf závislosti podobnosti komentáře s článkem pro kategorii počasí na čase zveřejnění pro model Doc2vec.



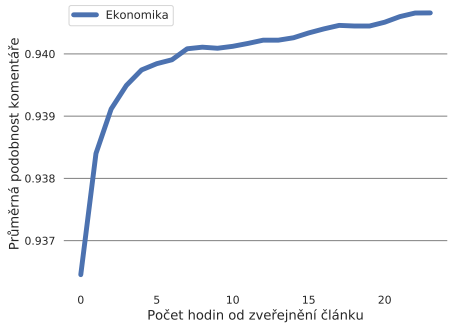
Obrázek B.25: Graf závislosti podobnosti komentáře s článkem pro kategorii zahraničí na čase zveřejnění pro model Doc2vec.



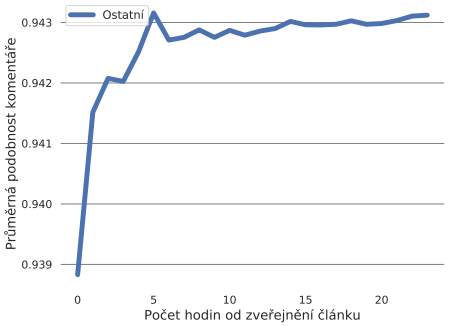
Obrázek B.26: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii auto v závislosti na čase pro model BERT.



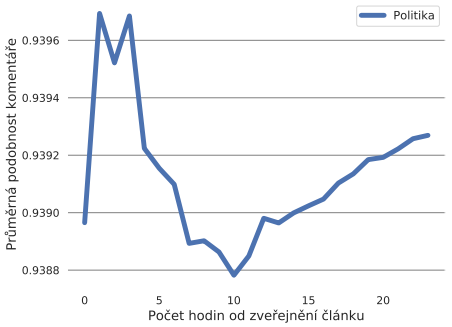
Obrázek B.27: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii domácí v závislosti na čase pro model BERT.



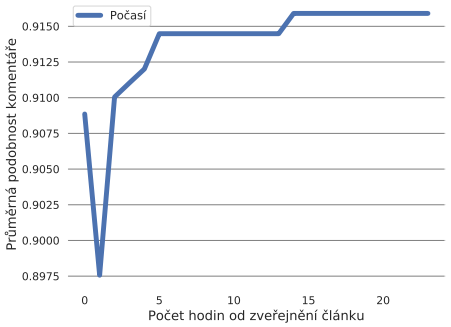
Obrázek B.28: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii ekonomika v závislosti na čase pro model BERT.



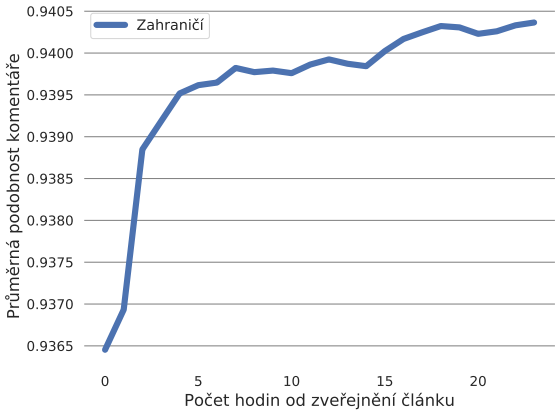
Obrázek B.29: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii ostatní v závislosti na čase pro model BERT.



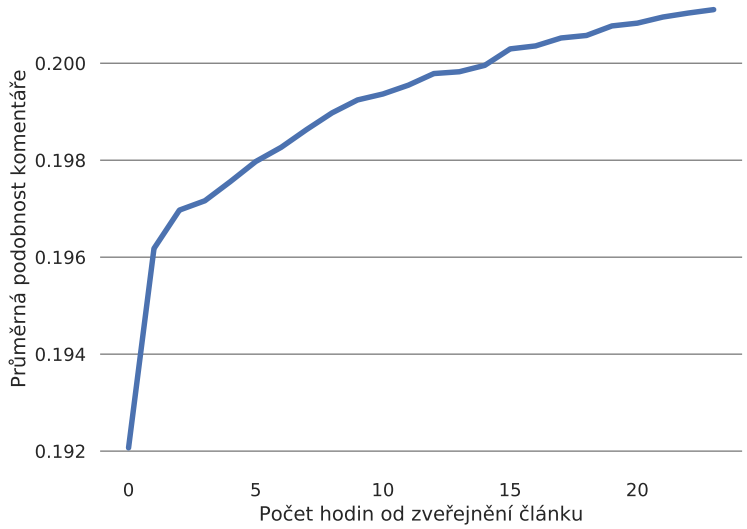
Obrázek B.30: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii politika v závislosti na čase pro model BERT.



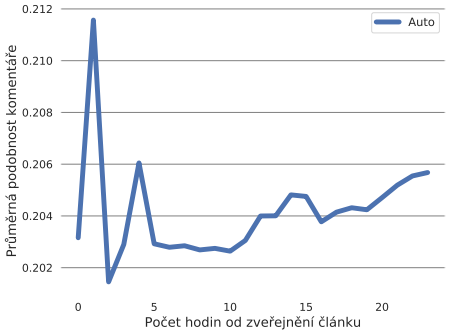
Obrázek B.31: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii počasí v závislosti na čase pro model BERT.



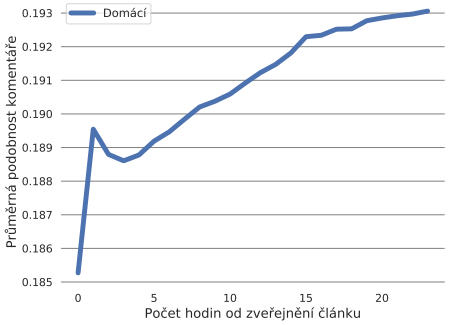
Obrázek B.32: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii zahraničí v závislosti na čase pro model BERT.



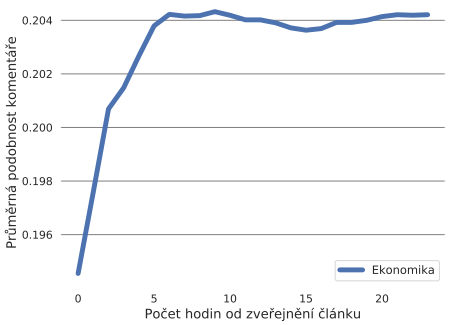
Obrázek B.33: Průměrná podobnost komentáře a článku v závislosti od zveřejnění pro model Doc2vec.



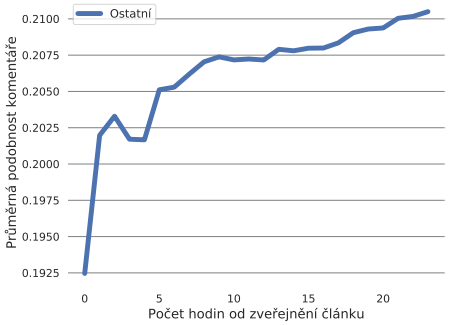
Obrázek B.34: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii auto v závislosti na čase pro model Doc2vec.



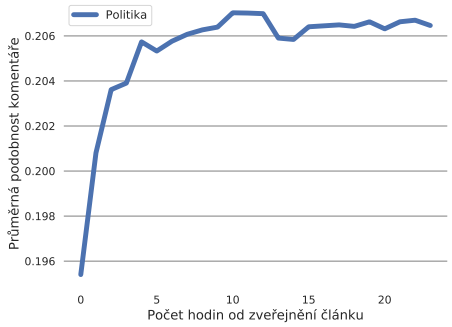
Obrázek B.35: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii domácí v závislosti na čase pro model Doc2vec.



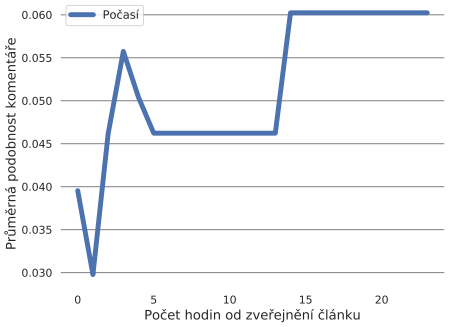
Obrázek B.36: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii ekonomika v závislosti na čase pro model Doc2vec.



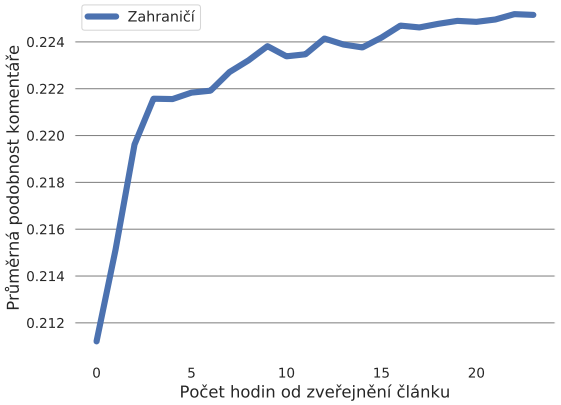
Obrázek B.37: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii ostatní v závislosti na čase pro model Doc2vec.



Obrázek B.38: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii politika v závislosti na čase pro model Doc2vec.



Obrázek B.39: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii počasí v závislosti na čase pro model Doc2vec.



Obrázek B.40: Graf závislosti průměrné podobnosti komentáře s článkem pro kategorii zahraničí v závislosti na čase pro model Doc2vec.

Příloha C

Tabulky

Tabulka C.1: Porovnání modelů, klasifikátorů a jejich úspěšnost.

Model	Klasifikátor	Hyperparametry	Hodnota MCC
BERT	k-nejbližších sousedů	počet sousedů: 6 vzdálenost: kosínova	0,52
	rozhodovací strom	maximální hloubka: 6	0,38
	náhodný les	maximální hloubka: 9 počet stromů: 24	0,53
	neuronová síť	velikost skryté vrstvy: 50	*0,62*
Doc2vec	k-nejbližších sousedů	počet sousedů: 26 vzdálenost: euklidova	0,27
	rozhodovací strom	maximální hloubka: 8	0,08
	náhodný les	maximální hloubka: 6 počet stromů: 17	0,11
	neuronová síť	velikost skryté vrstvy: 36	*0,34*
Doc2vec s BERT reprezentacemi	k-nejbližších sousedů	počet sousedů: 29 vzdálenost: kosínova	0.24
	rozhodovací strom	maximální hloubka: 12	0,06
	náhodný les	maximální hloubka: 7 počet stromů: 23	0,11
	neuronová síť	velikost skryté vrstvy: 106	*0,23*

Tabulka C.2: Deset komentářů s nejmenší kosínovo podobností s článkem pro model BERT.

Komentář	Titulek článku
tragédi	Dráhy plánují omezit wi-fi ve vlacích. Lidé stahují moc dat, operátoři zrušili dohodu

Dog-eat-dog?	SPD zrušilo regionální klub. Šéfoval mu Volný, který chce kandidovat proti Okamurovi
To je..🤢🤢🤢😭😭😭	ANO by ve volbách získalo přes 30 procent, druzí jsou piráti, ukázal průzkum
Make Britain great again. Remain!	Ze zákulisí summitu o brexitu: Mayová byla chabá, její vystoupení ”zatraceně špatné”
:-)	Policista v civilu se snažil vybírat pokuty. Prezidium už zahájilo kázeňské řízení
God save the Queen!	Britové zamítli divoký brexit. Mayová zkusí napotřetí prosadit dohodu s EU
:-)	Ve zlínské nemocnici se 29 lidí nakazilo salmonelózou. Zdrojem nákazy mohlo být kuře
💇‍♂️	Babiš odletěl do Spojených států. Na schůzku s Trumpem si beru svoje grafy, řekl
Mauzoleum...	Zeman chce stavět rodinný dům v Lánech, koupil si pozemek kousek od zámku
:-)_)	Skláním se před vaší moudrostí, loučila se hejtmanka Karlovarského kraje se Zemanem

Tabulka C.3: Pět komentářů s největší kosínovo podobností s článkem pro model BERT.

Komentář	Titulek článku
"... starý způsob přemýšlení, že jenom díky střední škole s maturitou a univerzitě si zajistíte dobrý život, už neplatí, ... "Neplatí proto, že maturitu dostane prakticky každý a u vysoké je to podobné. Dříve měli maturitu jen ti velmi dobří a vysokou ti nejlepší, a po škole zastávali odpovídající místa, taková, která by lidé s menšími schopnostmi nezvládli. Vzdělání bylo atributem vysoké kvalifikace a schopností. Jenomže sociální inženýři zaměnili a dodnes zaměňují příčinu a důsledek. Vypozorovali, že vzdělaní lidé se mají obecně lépe než nevzdělaní, z toho vyvodili, že klíčem je vzdělání a tudíž když budou mít všichni vzdělání, budou se mít všichni dobře. To je ovšem evidentní nesmysl, ze vzdělání se se stala víceméně formální věc, ale poměr schopných a neschopných zůstal stejný. Mimochodem, něco podobného tady bylo před více než sto lety, tehdy se za klíč k úspěchu považovala gramotnost. Jistě, kdo uměl číst a psát, byl před negramotnými zvýhodněn. Jakmile se naučili číst a psát všichni, bylo po výhodách. Co je úplně šílené, jsou názory některých humanitně vzdělaných teoretiků (Eduin apod.), že vzdělání má být co nejširší, protože dnes člověk přece nedělá jednu práci celý život, ale vystřídá deset i více zaměstnání. Jistě, v MacDonaldu, u pásu nebo v neziskovce žádná specializace není nutná. Pokud ale někdo dosáhnout v jakékoliv oblasti úpěchu, musí o ní vědět podstatně více než ostatní.	Finské děti jsou ve škole méně a excelují. Třída už tady není kostel, říká učitelka

Tak jsem si přečetl materiály dostupné na stránkách FF UK včetně všech příloh. Zahraniční posuzovatelé Kovářovi v posudcích celkem ostře nakládají, ale mám pocit, že s výjimkou jediného všichni pracovali jenom s textem ”obžaloby”, nikoliv dotyčnými prameny (nebo je jen tak povšechně projeli). Kovář i Pánek vypracovali poměrně rozsáhlou obhajobu, kde vyvracejí konkrétní nařčení (přestože u Pánka je to místy dost přitažené za vlasy a snaha pomoct Kovářovi z toho čouhá jak sláma z bot). Etická komise se s tím bohužel vypořádala dosti stručně - místo toho, aby Kovářovi bezezbytku vyvracela jeho body, tak spíš zdůrazňuje skutečnost, že mluví o doktorandech jako o lhářích... (přestože se nějakému rozboru věnují, to zase ne že ne). Celkově mi z toho vychází vlastně to, co předtím - opisoval, ale ne natolik, aby ho to mělo stát kariéru... Respektive opisoval - přeložil kus původního textu a pak do něj nasázel své poznámky a postřehy, rozšířil ho, místy třeba protáhl nějakou citaci (možná právě proto, aby ho v budoucnu nebylo možné napadnout z plagiátorství). Podle mého věděl, že jedná neeticky, ale zároveň si celkem poctivě projel to, na co jeho klíčový autor odkazoval a trošku si s tím pohrál, takže jistá přidaná hodnota je nezpochybnitelná. Není to tedy v pravém slova smyslu překlad, jak tvrdili doktorandi a i jeden zahraniční posuzovatel, ovšem práci si jednoznačně ulehčoval.	Odhalit Kováře žádalo od studentů odvahu, ale metály rozdávat nebudu, říká děkan

Starší články odjinud to podávají malinko jinak: Když policejní zásahová jednotka vnikla na Hanušův pozemek, údajně si myslel, že jde o přepadení, proto vytáhl legálně drženou pistoli, aby bránil sebe i rodinu. V následné přestřelce skončilo v jeho těle 14 střel. To celkem chápu - podle mě, když člověku vpadnou do baráku cizí lidi, normální reakce je bránit se. Sice by toho měl nechat v momentě, kdy pozná policajty, ale netuším, jak rychle se to semlelo a jak přesně to tam vypadalo. Po zatčení starší z Hanušů dokonce plánoval vraždu státní zástupkyně, která dozorovala vyšetřování případu. Měl požádat své dva spoluvězně ve vazební cele, aby ho zkontaktovali s člověkem z podsvětí, který by byl schopen žalobkyni zabít. Slíbil jim za to peníze. Oba recidivisté se ale zalekli a o Hanušově vražedném plánu informovali dopisem přímo dotyčnou žalobkyni. No, takže to zřejmě není tak hodný chlapec, jak by se jen z tohohle článku snad mohlo zdát. Fehim Hanuša se v červnu 1998 dostal do konfliktu s mužem, který na něj zavolal policisty. Hanuša se synem ujížděli z Plzně. U obce Letkov zastavili a začali střílet na policejní hlídku, která je pronásledovala. Jednoho policistu zabili ranou do srdce, druhého zranili, sami byli také postřeleni. To taky nevypadá, že by policajti chtěli nějak bezdůvodně obtěžovat zrovna jeho. Prostě se choval tak, až na něj zavolali benga. Takže ano, pořád je šílený, že policajti k němu vlítnou domů a prostřílí ho, pořád je děs, jak dlouho to soudy řeší, ale nějak mi ho už není moc líto.	Policie se spletla a rozstřílela mu břicho. Po 23 letech vysoudil odškodné 150 tisíc

Zdejší diskuze svědčí o typu čtenářů tohoto čehosi, napodobujícího internetové noviny. Nováková se pokusila o zkratku na nevhodném místě a to se jí vymstilo. :) Já bych na jejím místě hájil raději svobodu rozhodování ohledně využívání internetu a do výsledné pomoci občanům bych připustil pouze zásadní vliv státu při vyhlášení nové aukce kmitočtů a důraz na zamezení kartelovým dohodám operátorů. STÁTU TOTIŽ NIC JINÉHO NEPŘÍSLUŠÍ!!! O to více mi vadí, že nikomu nevadí blemtání zástupce (údajně pravicové) ODS o snížení ceny dat státem! Vždyť je to úplně stejná NEHORÁZNOST, jako žvatlání paní ministrině. Kdyby kdysi tzv. pravicová ODS nedělala levicovou politiku v tomto oboru, mohl jsem mít IT firmu jako velký ISP. Jenomže právě POPULISTICKÁ GESTA ODS s wifi zadarmo na náklady státu nebo obce v určitých oblastech mi vzali jistotu, že politici nezmění pravidla dříve, než zaplatím úvěry! Ještě dnes bych zde viděl příležitost k velkému podnikání, ale už nemám tolik let života na vybudování firmy. Tím spíše, že politici a jejich lobisté jsou schopni vše zmařit vydáním nového zákona. Nechte na lidech, zda chtějí taková data kupovat a nedělejte ze sebe spasitele pro ostatní, když to soudě podle jejich chování evidentně nechtějí.	Zopakovala to osmkrát. ČT vyvrátila, že slova Novákové o datech vytrhla z kontextu

Už jsem se zde vyjadřoval. Jsem podnikatel, s kolegyní máme menší firmu. Máme 12 zaměstnanců, z toho 9 VŠ, minimální fluktuace. 3 ženy. Za těch 20 let se jich protočilo 11. Všichni zaměstnanci dělají stejnou práci. Nebo přesněji jsou na stejné pozici. Máme takovou zkušenost, že chlap je výkonnější a průbojnější. Pouze 1 žena z celkových 11 se dokázala chlapům vyrovnat. A možná proto je sama, bez manžela. Ty ostatní by to asi zvládly taky, ale to by musely chtít. Nechtějí. A to moc dobře ví, že bychom je ohodnotili odpovídajícím způsobem. Raději mají pohodu. Starají se o děti, nedělají přesčasy nebo ani nemají full time. Proč taky, rodinu živí manžel. Mě jako podnikatele zajímá jenom výkon a podle toho platím. Nezajímá mě praxe, pohlaví, národnost... Jen odvedená práce. Proto jim až na tu 1 uvedenou výjimku platím méně a sedí to s údaji v článku. Nejsem blázen ani charita. Po těchto zkušenostech vím, že pro firmu je nejlepší mít zaměstnané muže, které doplní 2-3 ženy. A nejlépe typ Douchová, jinak muže včetně mě rozptylují :)))).... Ne, tohle byl jen vtip:)) Moje kolegyně - společnice se mnou v tomto souhlasí. Je to 20 let zkušeností z reálného světa.	On čtyřicet tisíc, ona dvaatřicet. Rozdíl v odměňování je alarmující, říká analytička

Tabulka C.4: Deset komentářů s nejmenší kosínovo podobností s článkem pro model Doc2vec.

Komentář	Titulek článku
A fašismem dnešní doby je globalismus vzešlý z marxismu.	"Každá doba má svůj fašismus."Demokracie je zranitelná, varovala Albrightová
Nejvíce znečištěná výfuky je Praha a české město s nejdelší délkou života je Praha.	Dýcháme strašné karcinogenní koktejly. Rodiny se musí vzdát více aut, říká Brabec
Je to šmejd a kreje ho Zeman.	Mynář v Osvětimanech načerno vybudoval rybník. Z něj bez souhlasu zasněžuje sjezdovku

Však je to politik.	Kdysi mírnil strach z uprchlíků, teď otáčí. Porovnejte si, jak Šefčovič mění názory
Ale zase uznejte, naprosto infarktová neděle.	Andrej Babiš obhájil funkci předsedy ANO. Hlasovalo proti němu 13 delegátů
50%	Decimace á la římské legie. ČSSD a odbory odmítají snížit počet státních zaměstnanců
Clinton to měl na střídačku, jednou Monica, jednou Madeleine - brrrrr!	"Každá doba má svůj fašismus."Demokracie je zranitelná, varovala Albrightová
hříbek ukazuje kouličky.... ovšem na nesprávném místě.....	Hřib odmítl půjčit svazu bojovníků za svobodu Brožíkův sál kvůli metálu pro Ondráčka
Od krvavé Madly dostal glejto? Styď se, chlape!! Hned to zahoď!	Pomáhal Bushovi i Nixonovi. Příběh Čechoameričana Maleka, který radil prezidentům
Jaký pussydent, takový kancléř.	Mynář v Osvětimanech načerno vybudoval rybník. Z něj bez souhlasu zasněžuje sjezdovku

Tabulka C.5: Pět komentářů s největší kosínovo podobností s článkem pro model Doc2vec.

Komentář	Titulek článku
Češi zešíleli z šílených polských krav. V hovězím jsme soběstačný, takže bych dal zákaz dovozu jakéhokoliv hovězího.	Češi zešíleli, nenecháme se vydírat. Můžeme kontrolovat pivo, hrozí polský ministr
Chudaci turisti....	Taxíky nebudou muset mít taxametr, šoféři zkoušky z místopisu, schválila vláda

Zdravým selským rozumem lze snadno dospět ke zjištění, že pro zřízení Centrálního mozku pražské dopravy v době elektronizace a digitalizace, je budování monstrózního stavebního komplexu zhola zbytečné. Daňovým poplatníkům nezbývá než doufat, že k takovému zjištění dospěje i "Nejlepší stavař mezi lékaři"a "Nejlepší lékař mezi stavaři", t. j. současný pražský primátor a jeho vážený poradní sbor.	Centrální mozek pražské dopravy ještě roky stát nebude. Praha vypoví smlouvu stavařům
Míň chtít se učit, míň dobrých známek mít. jako bezdomáči dají nejmenší uhlíkovou stopu, letadly nikdy neletíce.	Míň chtít, míň mít, pochválili ministři demonstrující děti. Poté vyhlásili boj suchu
Faltynek to rekl jasne, v Cesku nevladne vlada, tam vladne hnuti ..... komunisti byli take hnuti. Komunisticke hnuti !	Faltýnek: Sobotka tlačil kvůli mýtu na Ťoka, zasáhl jsem jako místopředseda hnutí

Tabulka C.6: Tabulka zobrazující nejméně relevantní komentáře pro vybrané články pro model Doc2vec.

Češi si za drahá data mohou sami . . .	*ODS vyloučila Václava Klause . . .*
Loupežníci.	Přeskupují se síly před volbami.
Já je převezu a opráším morseovku. :)	v SPD se uvolnilo místo
To je kosočtverec.	Timto se v ODS otevrela vratka pro zbrusu novy oblicej.
A už je z jejího žvástu virál...	Kam se nyní vrtnou Klausi juniore, rudí a nebo hnědí? a nebo nějaký jejich hybrid, třeba proputinovští?
TO JE ALE OBLUDA!!!!!!!!!!!!!	pak zhasni

Tabulka C.7: Pět komentářů s největší kosínovo podobností s vybranými články pro model BERT.

Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková	*ODS vyloučila Václava Klause mladšího. Není loajální ke straně, řekl předseda Fiala*
ANO je strana rovných příležitostí. I naprostý De Ment může být ministrem, pokud má dostatečně rád vůdce! Zaplať pánbůh za EU, protože bez jejich omezení "svobody a nezávislosti"by to v ANO-česku vypadla tak, že kilo brambor bude za kilo, a ministr zemědělství by nám vysvětlil že je to proto, že jich kupujeme málo. Jinak to byla EU, kdo nařídil ceny za Roaming - A ONO TO ŠLO! Byla to EU kdo nařídil snížit výkon vysavačů - A MÁME TU KONEČNĚ VYSAVAČE MÍSTO TOPENÍ - ŠLO TO! Bohužel, my v ČR si zvolíme STBáka s Putinovým poskokem.	Komicke je jak se tu celozivotni odpurci ODS snazi vzajemne presvedcit, ze bez Klause Jr. teda ODS volit zase nebudou. Mozna by dost pomohlo, kdyby se s nazory Klause skutecne seznamili. Protoze pak by i Kalousek vypadal jako socialista nejhrubsiho zrna.

Ta paní je asi blázen? Ona si myslí, že když začneme víc utrácet a víc platit za data operátorům, tak ti nám je pak zlevní?Muhehe. Tak větší pitomost jsem v životě neslyšel. Operátor se chová tržně a chce maximalizovat zisk. To že mu někdo zvedá tržby a dává víc peněz není v oligopolní ekonomice žádný podnět pro zlevňování a konkurenční boj. Možná v případě dokonalé konkurence (teorie) by to tak skutečně fungovalo. Ale při konstantním rovnoměrném rozložení trhu mezi tři velké hráče, kdy nastal status qou...tedy nulový produktový souboj o zákazníka, souboj probíhá pouze na marketingovém (brandovém) poli, tak že všichni tři hráči maximalizovali své marže a případná cenová válka by jim nic nepřinesla, maximálně by snížila marže všem hráčům a trh by se ustálil na obdobných tržních podílech ale s daleko nižšími cenami.... prostě oligopol nemá potřebu reagovat tak jak předpokládá ministrině.... Což mě vede k jedné otázce? Co tam tahle nedostudovaná (vycházím z toho že ve vyšších ročnících na vše se to učí, dokonalá konkurence jako taková se bere v prváku a hnedka v druháku se konstatuje, že nic takového jako dokonalá konkurence nikdy nemůže existovat) hlupačka proboha dělá? Nechtěl by jít příště dělat ministra instalatér?	Kritika, nebo nedej buh vlastni nazor se v tzv.demokraticke strane, ktera je zrozena z prekabatenych komunistu resi hezky po soudruzsku, vyhozenim. Soudruh Husak pomylene spolustraniky odmenoval stejnym zpusobem. Zvyky jsou zelezna kosile.

Rada paní ministryně je pro operátory nad zlato. Češi plaťte ještě dražší data než nyní a pak budete všichni ták móc štastní, že nám i sluníčkoví lidé budou tiše závidět. Jenže operátora vůbec nezajímá, co člověk potřebuje a co ne. Prostě mu vše vrzne do základního tarifu, i když danou službu vůbec nepotřebuje a včetně wifi příjmu, který v dané lokalitě není dostupný.	No jo, ve sněmovně a předsedou výboru bude do konce volebního období, zda se berou penízky i za vyšetřování OKD nevím. Zda si troufne založit něco nového nevím, zda by o něho stáli jinde také ne. Nejlépe by učinil, kdyby se vrátil učit matematiku.
Ještě by tu byla jedna hypotéza, jestli oni operátoři z nás nedolují peníze jak na Klondajku, aby mohli křížově dotovat tarify v zemích, kde je opravdový konkurenční boj. Je to to samé jako u drogerie DM, kde je zde u nás jejich totožné zboží někdy i o polovinu dražší (nikdy ale levnější) než v jejich domovině Deutschlandu, protože tam zuří v tomto odvětví pravý konkurenční boj a firmy se ho snaží ustát přeléváním peněz z jedné země do druhé, jenom aby přežili a nemuseli místo na trhu opustit. No a ministryně jim v tomto dělá jenom zástěrku, protože tento náš systém kope za korporace a ne za lidi. Takže líp asi opravdu nebude, i kdyby jsem platili jak mourovatí.	Den plpec, Sorosova dcerka Čaputová drtivě vede a juniora vyhodili z ODS...Nicméně chleba se bude lámat 24. a 25.5. Souboj SPD versus ”demokratický”blok+Piráti+ČSSD slibuje drama...A šance Tomia Okamury uspět jsou rázem větší, neb reprezentuje jedinou euroskeptickou stranu na českém politickém jevišti.
Tady ta paní mi byla podezřelá už v době, kdy se zavádělo EET. To měla kopat za podnikatele, ale na místo toho lezla do ři/ti Babišovi. Téhle paní nelze nic věřit, ostatně jako v dnešní době skoro každému druhému člověku, poněvadž lidi už úplně zblbli z peněz, sociálních sítí a chytrých mobilů.	Zvonik od Matky Bozi by byl lidem ukradeny kdyby nenosil prijmeni Klaus. Je totiz vniman predevsim jako Klaus. A jelikoz je jen spatnym stinem sveho otce resil zajem o svou osobu extremni retorikou. ODS se vyloucenim nemocneho Klause zbavila pytle s piskem.

Tabulka C.8: Pět komentářů s největší kosínovo podobností s vybranými články pro model Doc2vec.

Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková	*ODS vyloučila Václava Klause mladšího. Není loajální ke straně, řekl předseda Fiala*
Obhajoba monopolů 3 operátorů ministrem za ANO v přímém přenosu !	Klaus junior = chteny MUCEDNÍK demokraticke strany ODS. Jaka uzasna tecka jeho politicke kariery. A ted to muze konecne po jeho, nacionalistcky rozjet ...!
Zahajuji poradu. Náš produkt se bestseller. Výroba nestíhá, nejsou lidi, materiál. Je třeba nutně přijmout potřebná opatření !! .....Navrhuji proto ZLEVNIT PRODUKT !!	Nazory pana Klause ml. se mi (na rozdil od jeho tatika) vetsinou zamlouvaji. Skoda pro ODSS.
Tady ta paní mi byla podezřelá už v době, kdy se zavádělo EET. To měla kopat za podnikatele, ale na místo toho lezla do ři/ti Babišovi. Téhle paní nelze nic věřit, ostatně jako v dnešní době skoro každému druhému člověku, poněvadž lidi už úplně zblbli z peněz, sociálních sítí a chytrých mobilů.	Když ODS odešel i velikán naší konzervativní pravice Pospíšil, zbyla dneškem v této straně jen jediná rozumná persona Kubera....
Zajímavé je, když dám operátorovi výpověď obratem mi dá tarif 499kč měsíčně, 10Gb dat, neomezené volání a SMS. Tento tarif ovšem oficiálně neexistuje. Ministryni ihned odvolat, buď je blbkka, nebo ji platí nějaký operátor.	Kandidáti na prezidenta nesmí býti členy politcké strany. Prvního kandidáta už teď známe.
Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková Tak tohle jsou ministři za ANO? No potěš protěž!	Správně Fialo ! Jak tě to učili Klausovi - hlavně RAZANTNĚ !

Tabulka C.9: Počet komentářů pro jednotlivé liky a disliky.

Počet	Like	Dislikes
0	8 154	14 172
1	3 552	1 609
2	1 789	275
3	929	60
4	586	17
5	340	4
6	228	2
7	174	0
8	112	0
9	86	0
10	52	0
11	35	0
12	24	0
13	26	0
14	17	0
15	11	0
16-25	24	0

Tabulka C.10: Počet komentářů pro rozdíl liků a disliků.

Počet	Like – Dislikes
-5	4
-4	7
-3	22
-2	91
-1	700
0	7875
1	3322
2	1668
3	848
4	533
5	333
6	230
7	150
8	106
9	71
10	45
11	39
12	26
13	16
14	8
15-25	21

Tabulka C.11: Pět nejrelevantnějších komentářů od uživatele s největší průměrnou relevancí pro model BERT.

Komentář	Titulek článku
pokr. Co ovšem ten novodobý socialismus znamená v praxi: napojení na Rusko, hájit bezvýhradně politiku Kremlu, ať prosazují cokoli. Vše, jak vládce Kremlu rozhodne, to je dobré. Potom při řízení státu to znamená totální chaos, naprostou neschopnost. Nefungující ekonomika. A taky nasazení armády a policie proti lidem, zákazy cestovat do zahraničí, mrtvé při demonstracích opozice. O socialismu 21. století mluvil už prezident Cháves, současný prezident Venezuely, nebo spíš usurpátor, který ovládá silové složky, ten to dovedl k dokonalosti. Napojení na Rusko, taky humanitární pomoc z Ruska - a rozstřílení té z Ameriky. Uzavřené hranice. Země, která je nesmírně bohatá na ropu a má necelých 30 mil. obyvatel, tak místo bohatství všech, např. neexistence daní pro fyzické osoby, jak mají některé arabské státy Zálivu, tak naprostá chudoba. Lidi nemají co jíst, děti silně podvyživené, umírají. Kdo jídlo má, tak jí jednou denně, lidé výrazně zhubli. Chybí i jiné základní věci nutné k životu. Ovšem armáda a policie na hranicích, zákazy vycestovat nebo utéci do sousedních zemí. Zkrátka stačí se podívat na reálný socialismu. Který "funguje"i bez ruských vojáků na území země, ale je tu jen politické napojení.	Ideální uspořádání je novodobý socialismus, říká kandidát KSČ do europarlamentu
Slovákům to můžeme jen přát, nebo tiše závidět. Každopádně nebudou mít prezidenta toho typu, co máme my, nebo by měli, kdyby zvolili ze svých třeba Fica, nebo Kotlebu, či z dřívějších Mečiara. I kdyby to Čaputová nevyhrála, tak to nebude taková katastrofa jak u nás Zeman nebo dříve Klaus. Kde je ta doba, kdy jsme měli prezidenta Havla. Tehdy to člověk bral jako samozřejmost a až nyní dokáže ocenit, co jsme v něm měli. K některým jeho krokům jsem byl kritický, ale byl to pan prezident. Nebyla to loutka, které tahají za provázky z Kremlu.	Čaputová je pro druhé kolo v dobré pozici. Šefčovič podcenil odpor vůči Ficovi

U té malířky nevím, je to docela fyzicky náročné. Kdo si sám doma maloval, myslím štětkou a ne válečkem, tak ví, že to je spíš pro chlapa. Určitě jsou fyzicky zdatné ženy, co to zvládnou. Ale jinak asi spíš jen něco lehčího, nevím, jestli ta podlahařina je fyzické méně náročná. I když i ta kuchařka, ta má taky fyzickou námahu. Krom různého zvedání hlavně celý den stojí. Což se týká i holiček, kadeřnic. Jinak jsem nadšený, když vidím, že tu jsou učni, co je to baví. Hlavně aby tam chodili hoši. Určitě lepší dobrý řemeslník, než průměrný či podprůměrný maturant. Z hlediska uživení se je to řemeslo asi lepší. I když ne všude dostanou dost peněz. Ale dnes není problém si tu maturitu dodělat a mít ji v záloze, až to fyzicky nebudou zvládat. Nevím, jestli to žena vydrží dělat celý život. Taky je pravda, že učební obor je přece jen méně náročný, pokud jde o studijní předpoklady. Ale je v pořádku, když tam budou chodit ti méně nadaní ale zato manuálně šikovní. Nebo i ti nadaní ale současně velmi šikovní. Dnes tu byl rozhovor s učitelkou, která rok učila na učilišti, takže představu o kvalitě školství mám. Ovšem tam to bývalo i za socialismu, nyní ovšem je lecdky mizerná úroveň i na běžných sš. Zas mi nepálí, že tolik nebude uměl češtinu nebo jazyky, hlavně když budou umět řemeslo a budou mít nutné technické znalosti. U řemeslníka mě hrubky v pravopise tak nepálí jak u inženýra nebo lékaře.	Z kuchařky podlahářkou. Dívky se vrhají na mužská řemesla, jsou pečlivé a lépe mluví
První kolo ještě není definitivní vítězství, u nás proti Zemanovi jasně demokratičtí kandidáti (sem nepočítám Topolánka, ten jim spíše jen ubíral hlasy) měli dohromady jasnou většinu. Zdálo se, že Zeman nemá moc kde brát. Jenže druhé kolo dopadlo tak, že to sice těsně, ale vyhrál. U nás byla potíž, že ti demokratičtí kandidáti, tedy Hilšer, Drahoš a Horáček, se nedokázali sjednotit. Pokud by včas dva z nich odstoupili a podpořili třetího, mohlo to dopadnout jinak. Zdá se, že paní Čaputová má podporu velmi výraznou a vítězství by jí nemělo nic vzít.	Čaputová je pro druhé kolo v dobré pozici. Šefčovič podcenil odpor vůči Ficovi

Existuje i jiná věc, někdo šéfuje určité skupině akademických pracovníků, např. oddělení na katedře, nebo nějaký týp pro grant ap. Řadoví pracovníci a doktorandi odvedou mravenčí práci a ten vedoucí týmu to jaksi shrne do své publikace. Jelikož ti lidé to přímo nepublikují, nebo publikují něco jiného, ale za tým je publikace přece na vedoucím, tak nelze říci, že to od nich opsal. Pouze využil práce cizích lidí, kterou sepsal, zaznamenal věci, co mu lidé říkali na různých poradách týmu ap. Toto bohužel nikdo nepostihne. Jen dodám, že vím, o čem píši.	Odhalit Kováře žádalo od studentů odvahu, ale metály rozdávat nebudu, říká děkan

Tabulka C.12: Pět nejméně relevantních komentářů od uživatele s nejmenší průměrnou relevancí pro model BERT.

Komentář	Titulek článku
Dycky MOST!	ODS dotahuje piráty, sociální demokraté klesli na úroveň komunistů, ukazuje průzkum
Dycky ANO!	ODS dotahuje piráty, sociální demokraté klesli na úroveň komunistů, ukazuje průzkum
Dycky Most!	Fajnšmekři ať klidně cupují hříbky zombie kudlou, vyvrací Brusel zprávu o zákazu nožů
Opouštím Bakalovu propagandu a stávám se tak slušným člověkem, nashledanou!	ODS dotahuje piráty, sociální demokraté klesli na úroveň komunistů, ukazuje průzkum
ANO si udržuje nedostižný náskok!	ODS dotahuje piráty, sociální demokraté klesli na úroveň komunistů, ukazuje průzkum

Tabulka C.13: Pět nejrelevantnějších komentářů od uživatele s největší průměrnou relevancí pro model Doc2vec.

Komentář	Titulek článku
V roce 2017 se dovezlo zruba 37 tisic tun hoveziho masa, nejvetsim dodavatelem bylo Polsko, nechci problematiku nijak zasadne bagatelizovat, ale pokud se obcas nekde objevi par set kilogramu zavadneho masa, tak je to relativne male mnozstvi. Kdybych byl paranoidni, tak bych se mozna mohl ptat, zda pan ministr, sam vyznamny podnikatel v potravinarstvi a osoba velmi aktivni v zemedelskych a potravinarskych sdruzenich, neni tak trochu ve stretu zajmu, stejne jako vetsina vladnouci strany v cele s panem premierem.	Polské maso za argentinské nikdo nevydával, přiznal Toman. Kontroly budou pokračovat
Ja myslim, ze to s ceskym statem pujde podobne jako bez nej, jsme malinky trh co do poctu prodanych aut (necela 2 % EU prodeju), navic se tu prodavaji spise levnejsi automobily. Aby lide a firmy misto lacinych aut nakoupily drahe elektromobily, tak by dotace musela byt pomerne extremni, relativne casty evropsky bonus je do 5000 EUR, kdyz uvazime rozdil cen elektromobilu a levnych aut, tak je to vicemene zanedbatelna castka. Cestu danovych ulev take nevidim prilis realne vzhledem k tomu, ze drazsi auta tu kupuji zejmena platci DPH, tak odpusteni DPH prilis nepomuze. Silnicni dan je jiz elektromobilum odpoustena. Nevim, zda by pomohlo odpusteni dane z elektriny, ale vzhledem k jeji vysi 28,30 Kc/MWh to neocekavam. A v neposledni rade stat nema penize nazbyt i v dobe konjunktury a rekordnich vyberu dani hospodari na dluh a nedostava se mu penez na infrastrukturni projekty, ktere narozdil od elektromobilu poslouzi cele spolecnosti.	Škoda Auto vzkazuje české vládě: Elektromobilitu musíte podpořit, bez vás to nepůjde

Muj nazor je, ze Rozhlas by mel byt privatizovan a koncesionarske poplatky zruseny. Na druhou stranu musim prijmout smutny fakt, ze existuje a je to organizace s rozpoctem cca 2,5 miliardy. To, ze reditel podobne velke organizace bude jezdit autem za 1,3 milionu, neni snad nic tak hrozneho. Jestli ma nekdo pocit, ze 400 tisic navic* za auto porizovane zhruba na 5 let zachrani zamestnanost nebo vysi platu v podobne organizaci, tak s nim musim celkem zasadne nesouhlasit. Vemme to treba ve srovnani s naklady na plat pana reditele, ty jsou zhruba 3,8 milionu korun rocne, tedy cca 19 milionu korun za 5 let, je opravdu tak hrozne podobnemu cloveku koupit auto za 1,6 milionu s DPH? * je otazka, zda je to navic, Skoda Superb a Volvo S90 jsou naprosto neporovnatelna auta, obzvlast pokud maji plnit reprezentativni funkci a vozit vazene hosty.	Rozhlas koupil řediteli luxusní vůz za 1,3 milionu. Pohrdá zaměstnanci, míní odbory
Pan reditel je zrejme vtipalek, ale bohuzel si neuvedomuje, kdy je vhodne vtipkovat. Ted ma z ostudy kabat a mozna ho za to nejaky aktivni politik vyhodi, aby ziskal levny bodik u volicu, fakt je ten, ze to muze byt docela skoda, treba nemocnici vede dobre a treba i dobre vychazi se svymi zamestnanci.	Kdo jiný může jíst rohlík po pacientech, když je nesní? Šéf nemocnice sestry pobouřil

V zakone je, ze prezident jmenuje, neni tam podminka, ze musi s vyberem souhlasit nebo ze muze jmenovani odmitnout, take neni uvedena lhuta, takze se tomu da teoreticky vyhnout odkladem na neurcito, na druhou stranu mi to prijde ponekud hloupe, osobne bych podobne formulace bez uvedeni lhuty vykladal tak, ze se dany ukon provede bez zbytecnych odkladu. Nicmene ja nejsem pravnik a prezidentovi pravnici to vidi jinak. Z meho pohledu nejlepsi reseni problemu s jmenovanim nechtene osoby zvolil sveho casu prezident Klaus, kdyz v pripade jmenovani Iva Mathe rektorem nahradil obvykle osobni jmenovani dopisem. Tim nechci Klause nejak adorovat, v jinych pripadech se choval podobne jako Zeman, napriklad ve zname kauze nejmenovani mladych soudcu. Kazdopadne reseni ala Klaus-Mathe u Zemana nepripada v uvahu, protoze profesorum bezne osobne jmenovani nepredava. Je to celkove pomerne nestastna situace, ktera lze resit trpelivosti nebo zmenou zakona, takova zmena by vsak dle meho nazoru byla pomerne nestastnym signalem, ze jsme ztratili iluze o osobnosti prezidenta, nejenom Zemana, ale hlavne jeho nasledovniku.	Karlova univerzita podala žaloby kvůli Zemanovu rozhodnutí nejmenovat profesory

Tabulka C.14: Pět nejméně relevantních komentářů od uživatele s nejmenší průměrnou relevancí pro model Doc2vec.

Komentář	Titulek článku
hříbek ukazuje kouličky.... ovšem na nesprávném místě.....	Hřib odmítl půjčit svazu bojovníků za svobodu Brožíkův sál kvůli metálu pro Ondráčka
milí trolíci koukám že se zmůžete akorát na žvásty a ne na oponování a argumenty....ano pro vás neschopné vovce je každé vedení dobré i z eu	Brexit bez dohody by vyšel Česko až na jedno procento HDP, říká Petříček
pusťte nás na ně!	Rebelové ve stylu Jamese Deana i partneři ČSSD. Piráti mají plán, jak porazit Babiše

babiši babiši s petříčky další volby nevyhraješ	Komunisté budou ladit taktiku proti Petříčkovi. Možná výrazně přitvrdíme, zaznívá
prostě kalousci fialové gazdíci farští a další nýmandi se z projevu pana Babiše poserou.....a to znamená že to Babiš dělá dobře.....jen do té americké řiti by neměl tolik lézt to mu zlomí vaz	”Zneužívání výhod, drahá služební auta, prostě papalášství,”kritizuje ANO Babiš

Tabulka C.15: Tabulka zobrazující anomální uživatele při využití rozdílu mezi článkem a příspěvkem pro model BERT.

Uživatel	Komentář	Titulek
1089	Dycky ANO!	ODS dotahuje piráty, sociální demokraté klesli na úroveň komunistů, ukazuje průzkum
1089	A počkejte na ty preference, až Babiš sundá sebestředného blba Trumpa!	ODS dotahuje piráty, sociální demokraté klesli na úroveň komunistů, ukazuje průzkum
1089	Německo má přijmout teroristy z IS.....	SPD se chystá obnovit severomoravský klub. Zrušení byla manažerská chyba, tvrdí Volný
1089	Vždy, když vzpomenu na Kalouskovo působení ve vládě si řeknu, že se máme nejlépe v porevoluční historii a to díky tomu, že Kalouskové byli lidem z politiky odejiti!	Důchodová komise se poprvé sešla. Řešit bude nejprve nižší penze žen, pak příjmy

Tabulka C.16: Tabulka zobrazující anomální uživatele při využití rozdílu mezi článkem a příspěvkem pro model Doc2vec.

Uživatel	Komentář	Titulek
121	Nevim o co se ten Laska vlastne snazi. Zemanovi uz volebni preference zvysit nemuze, jedine snad, ze by chtel nejak vyplnit diru po chybejicich komicich...	Senátor Láska má kostru žaloby na Zemana. Vyčítá mu tlak na soudy a dalších 56 věcí
121	V diskusnim foru Aktualne existuji zjevne dva tabory, tabor pokojnych diskuteru proti taboru demagogickych bojovniku proti samym principum demokracice konkretne proti zcela demokraticky zvolenym politikum .	Policie stále nevyslechla Babiše juniora. Mluví s ním pouze přes advokáta
121	Pro utechu zdejsim prislusnikum Elity : Tech 32% volicu ANO tvori naprosta spodina spolecnosti, duchodci, lide na podporach, postizeni exekucemi, lide sotva se zakladnim vzdelanim, obyvatele Chanova a samozrejme bezdomovci. Elito ulevilo se ti ?	Volby by nyní vyhrálo Babišovo ANO. Piráti a ODS podle průzkumu zaostávají
1764	Tak ahoj proletáři,pěkně tu diskutujte, stejně vás všichni poitici potřebují jen na jedno.....	Rozchod s Klausem mladším nebude dramatický, lidé kvůli němu ODS nevolí, říká Benda
1764	Milí voliči,nebuďte hloupí a volte zase hromadně ODS, abychom vám poté zase mohli utáhnou opasky, jste rozežraní a něco by z vás zase káplo....	Rozchod s Klausem mladším nebude dramatický, lidé kvůli němu ODS nevolí, říká Benda
157	Rychle sdílejte než to smažou!!	Fanoušci SPD podpořili na internetu pomoc syrským sirotkům. Hnutí anketu hned smazalo
157	to je nuda.. furt to samy.	Andrej Babiš obhájil funkci předsedy ANO. Hlasovalo proti němu 13 delegátů

157	Na volebních preferencích to ale stejně nic nezmění. Andrej může mít kauzu dvacátou, třicátou, a pořád bude mít velkou podporu. Bureš má své hloupé ovečky sesbírané chytře napříč celým politickým spektrem, které tvoří pevné vůdcovo voličské jádro.	Zde jsou zásadní důkazy úřadu v Černošicích. Babiš může dál ovládat svá média

Tabulka C.17: Tabulka zobrazující anomální uživatele při využití kolmého vektoru na článek a příspěvek pro model BERT.

Uživatel	Komentář	Titulek
289	V Moskvě vyhlásí státní smutek.	ODS vyloučila Václava Klause mladšího. Není loajální ke straně, řekl předseda Fiala
289	Zeman mu beztak zařídí nějaké místo velvyslance	ODS vyloučila Václava Klause mladšího. Není loajální ke straně, řekl předseda Fiala
289	A co říká Ústřední výbor NSDAP ?	Petříček pracuje v rozporu s národními zájmy, tvrdí KSČM. Jeho odvolání však nežádá

Tabulka C.18: Tabulka zobrazující anomální uživatele při využití kolmého vektoru na článek a příspěvek pro model Doc2vec.

Uživatel	Komentář	Titulek
282	...jděte někam ! Podle Kremlu má tady nejlepší mozek Eman !	Obyčejná čórka, zlobí se piráti na Babiše kvůli vykradení tématu mobilních operátorů
282	.V římse toho krbu není zabudován odposlech ? Špatné...špatné -velké lajdáctví !	Menší pocta než Nečasovi. Trump neubytoval Babiše v nejexkluzivnějším hotelu světa

282	NO a co na to s. Filip -Burešův poňoukatel ? Nebude se cítit dotčen ?	Menší pocta než Nečasovi. Trump neubytoval Babiše v nejexkluzivnějším hotelu světa
1764	Pseudopravice už neloví kapříky,ale zase rozdává udice....	ODS vyloučila Václava Klause mladšího. Není loajální ke straně, řekl předseda Fiala
1764	Tak stačí jen dvě nové sazby DPH a hnedle jsme evropskej tygr, prdůchům dá pan Fiala udice,aby měli co žrát...	ODS chce dvě sazby DPH a snížení odvodů pro zaměstnance se zkráceným úvazkem
1764	A co paní Němcová, už se o ní zase pokouší infarkt?	Poslanci ODS vyzvali Klause ml. k odchodu z klubu, ten odejít odmítá
123	neřekneme a neřekneme jako důůkaz vám musí stačit že to řekli naši drazí pruhovaní bratři co nás bombardovali....	V čem jste hrozbou, neupřesníme. Většina podkladů je tajná, píše NÚKIB firmě Huawei
123	moji milí eurohujeři vaše vyhlídky vám nezávidím.....	Brexit bez dohody by vyšel Česko až na jedno procento HDP, říká Petříček
123	tak už po třetí se zdejších odborníků ptám koho babiš udal abych ho mohl potrestat hlubokým opovržením ale zatím jsem se nedověděl ani jedno jméno koho udal.....tak pro prosím netajte ale veřejně oznamte....nebo se odpovědi nedočkám?	Rebelové ve stylu Jamese Deana i partneři ČSSD. Piráti mají plán, jak porazit Babiše

Tabulka C.19: Tabulka zobrazující anomální komentáře při využití rozdílu mezi článkem a příspěvkem pro model BERT.

Komentář	Titulek
Pokud roste cena novych aut, je zcela nabiledni, ze poroste i cena ojetin, a zvlaste tech mladsich.	Zánovní ojetiny čeká zdražení. Ceny nebudou klesat ani u lehce jetých dieselů
A G R O T E L !	Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková

na to jak jsou všechna auta předražená, by toto zařízení mělo být ze zákona již v autě , aby je mohl bez problémů používat každý	Nebezpeční řidiči dodávek: Za volantem telefonují 35 minut denně, 50 % bez handsfree
Protoze zasilkove domy pouzivaji stale vetsi baliky na 50 % vycpane ochranych materialem, chce posta na teto ”novince”vydelat a prechazi pri vypoctu cen od vahy na objem.	Pošta mění ceny při posílání balíků. Nově bude záležet na velikosti zásilky

Tabulka C.20: Tabulka zobrazující anomální komentáře při využití kolmého vektoru mezi článkem a příspěvkem pro model BERT.

Komentář	Titulek
Faltýnek jako mesiáš	Faltýnek: Sobotka tlačil kvůli mýtu na Ťoka, zasáhl jsem jako místopředseda hnutí
Mestsky management level Agrofert.	Torzo nové radnice na Palmovce chce Praha 8 prodat. Dostavba by stála příliš, tvrdí
Ano, chteji konec svobodneho internetu jak ho zname. Pokud tento paskvil projde budou nasledovat dalsi a dalsi restrikce. Zavedeme absolutismus EU kde svoboda budou jen prazdna slova.	Polské noviny bojují za ochranu autorství na webech. Vyšly s prázdnou titulní stranou
Tibéééééééét	Sedm stovek úřadů v Česku vyvěsí tibetskou vlajku. Připomíná 87 tisíc obětí povstání

Tabulka C.21: Tabulka zobrazující anomální komentáře při využití kolmého vektoru mezi článkem a příspěvkem pro model Doc2vec.

Komentář	Titulek
Fuuuuj, zase FAKE NEWS, s Čínou přeci není žádný byznys, Zeman jenom kecal. (Mimochodem anketní otázka: kolik % BMW a Mercedesu vlastní Arabové?)	Výrobce miniaut Smart bude z 50 % čínský. Polovičním vlastníkem se má stát Geely
To se skočím mrknout na výplatnici jestli nekecají.	Lidé pracující pro automobilky vydělají o třetinu víc než jinde. Průměr je 41 tisíc
Je smutné, že celní unii jako naprosté integrační minimum omezující mnoho ekonomických škod napáchaných Brexitem musí Toryům připomínat levičák Corbyn a hrabat na tom politické body. Mayová pokračuje v tragické cestě svého předchůdce Camerona.	Podpoříme váš brexit, navrhl Mayové šéf labouristů Corbyn. Má ale pět podmínek
Babiš s chotí v Bílém domě Česku určitě neudělali ostudu.	Americká cla jako největší problém. Babiš jednal v Bílém domě s prezidentem USA
Tak bydlet v centru má výhody i nevýhody... Třeba na Vysočině je spousta vesnic, kde je ticho nejenom v kostele....	Nová pravidla omezí noční život na pražské náplavce. Nalákat chce i druhá strana řeky

Tabulka C.22: Tabulka zobrazující anomální komentáře pro článek *Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková* při využití kolmého vektoru na článek a komentář u modelu BERT.

A G R O T E L !
Tak to je tedy ministerský materiál.

Tabulka C.23: Tabulka zobrazující anomální komentáře pro článek *Češi si za drahá data mohou sami, nemají používat wi-fi, míní ministryně Nováková* při využití vektoru komentáře u modelu BERT.

A G R O T E L !
TO JE ALE OBLUDA!!!!!!!!!!!!!!
Blbý a blbší
Bože do čehos to Duši dal
Tak to je tedy ministerský materiál.
mládek reloaded?
Neměl by ten článek být označený jako placená reklama?

Příloha D

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `analysis` ..... zdrojové kódy pro analýzu
  - `scraper` ..... zdrojové kódy pro stahování dat
- `docs` ..... text a přílohy práce