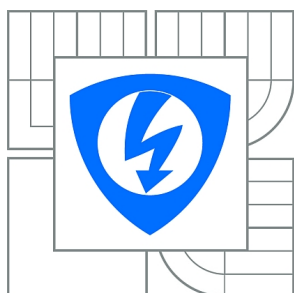


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH
TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

NOVÉ METODY ZPRACOVÁNÍ TEXTU PRO KLASIFIKACI EMOCÍ

NEW METHODS FOR EMOTION RECOGNITION FROM TEXT

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

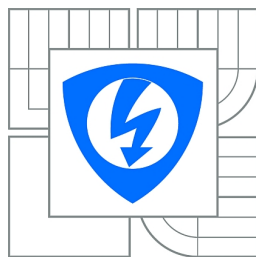
AUTOR PRÁCE
AUTHOR

Bc. JAKUB ONDERKA

VEDOUcí PRÁCE
SUPERVISOR

Ing. JAN MAŠEK

BRNO 2015



**VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ**

**Fakulta elektrotechniky
a komunikačních technologií**

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student: Bc. Jakub Onderka

ID: 125570

Ročník: 2

Akademický rok: 2014/2015

NÁZEV TÉMATU:

Nové metody zpracování textu pro klasifikaci emocí

POKYNY PRO VYPRACOVÁNÍ:

Navrhněte a implementujte v jazyce JAVA nové metody pro zpracování textových dat. Zaměřte se zejména na metody učení bez učitele. S použitím tohoto systému natrénujte model pro klasifikaci emocí z textů a jeho funkčnost ověřte na vytvořených textových databázích.

DOPORUČENÁ LITERATURA:

- [1] Ameeta Agrawal, Aijun An: Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations, IEEE/WIC/ACM, 2012.
- [2] Jiří Materna. 2012. LDA-Frames: an unsupervised approach to generating semantic frames. In Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing. Springer-Verlag, Berlin, Heidelberg.
- [3] Ronen Feldman and James Sanger. 2006. Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, NY, USA.

Termín zadání: 9.2.2015

Termín odevzdání: 26.5.2015

Vedoucí práce: Ing. Jan Mašek

Konzultanti diplomové práce:

doc. Ing. Jiří Mišurec, CSc.

Předseda oborové rady

UPOZORNĚNÍ:

Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

ABSTRAKT

Tato diplomová práce se zabývá možnostmi extrakce emocí z textu, konkrétně strojovými metodami bez učitele. Podrobněji jsou popsány metody sémantického modelování LSA, pLSA a LDA. Byla vytvořena implementace metody LDA v jazyce Java, která byla použita pro emocionální klasifikaci 860 česky psaných dokumentů do šesti odlišných emocí. Maximální přesnost při optimalizaci parametrů modelu byla 24 %.

KLÍČOVÁ SLOVA

emocionální analýza, strojové učení, sémantická analýza, LDA, Java aplikace

ABSTRACT

This master's thesis is about a method for sentimental analysis, especially machine learning methods without teacher. In detail are described method for semantic modeling LSA, pLSA a LDA. It was created a LDA implementation in Java language, which was used to emotional classification of 860 Czech documents to six different emotional categories. Maximal accuracy was 24 % if optimized parameters was used.

KEYWORDS

emotional analysis, machine learning, semantic analysis, LDA, Java application

ONDERKA, Jakub *Nové metody zpracování textu pro klasifikaci emocí*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2015. 49 s. Vedoucí práce byl Ing. Jan Mašek

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Nové metody zpracování textu pro klasifikaci emocí“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno

.....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Ing. Janu Maškovi za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci a také svým rodičům a kamarádům, jmenovitě Bc. Josefu Vlčkovi a Bc. Michalu Jakubíčkoví, za morální podporu.

Brno

.....

(podpis autora)

PODĚKOVÁNÍ

Výzkum popsáný v této diplomové práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno

.....
(podpis autora)

OBSAH

| | |
|---|-----------|
| Úvod | 11 |
| 1 Emoce | 12 |
| 1.1 Rozdělení emocí | 12 |
| 2 Předzpracování přirozeného textu | 14 |
| 2.1 Lexikální analýza | 14 |
| 2.1.1 Tokenizace slov v českém jazyce | 14 |
| 2.1.2 Segmentace vět v českém jazyce | 15 |
| 2.2 Stop slova | 15 |
| 2.3 Doplnění diakritiky | 16 |
| 2.4 Tvarosloví | 17 |
| 2.4.1 Morfologické nástroje | 17 |
| 2.4.2 Práce s tvaroslovím | 18 |
| 3 Metody pro získávání emocí z dokumentu | 20 |
| 3.1 Slovníkové metody | 20 |
| 3.2 Lingvistické metody | 20 |
| 3.3 Strojové učení s učitelem | 21 |
| 3.4 Strojové učení bez učitele | 21 |
| 4 Současné přístupy | 23 |
| 5 Sémantické modelování | 24 |
| 5.1 LSA | 26 |
| 5.2 pLSA | 27 |
| 5.3 LDA | 28 |
| 6 Praktická část | 31 |
| 6.1 Trénovací a testovací data | 31 |
| 6.2 Popis algoritmu | 32 |
| 6.3 Metody pro získání emoce clusteru | 33 |
| 6.3.1 Maximální počet dokumentů | 34 |
| 6.3.2 Procentuální zastoupení | 35 |
| 6.3.3 Vzdálost od clusteru | 35 |
| 6.3.4 Porovnání metod | 36 |
| 6.4 Vliv parametrů na přesnost | 36 |
| 6.4.1 Počet clusterů | 37 |

| | | |
|----------|--|-----------|
| 6.4.2 | Velikost minidávky | 37 |
| 6.4.3 | Přesnost odhadu parametrů | 38 |
| 6.4.4 | Optimalizace parametrů | 39 |
| 6.5 | Sloučení trénovacího a testovacího korpusu | 39 |
| 6.6 | Snížení počtu kategorií | 40 |
| 6.7 | Zhodnocení výsledků | 40 |
| 7 | Závěr | 41 |
| | Literatura | 42 |
| | Seznam symbolů, veličin a zkratk | 45 |
| | Seznam příloh | 46 |
| A | Clustery | 47 |
| B | Ukázkový výstup programu | 48 |
| C | Obsah přiloženého CD | 49 |

SEZNAM OBRÁZKŮ

| | | |
|-----|---|----|
| 5.1 | Model pLSA v PLATE notaci | 27 |
| 5.2 | Model LDA v PLATE notaci | 29 |
| 6.1 | Blokové schema funkce algoritmu | 33 |
| 6.2 | Počet dokumentů náležící clusteru dle emoce | 34 |
| 6.3 | Procentuální zastoupení a vzdálenost dokumentů v clusteru | 36 |
| 6.4 | Přesnost použitých metod | 37 |
| 6.5 | Závislost přesnosti modelu na počtu clusterů | 38 |
| 6.6 | Závislost přesnosti modelu na velikosti minidávky | 38 |
| 6.7 | Závislost přesnosti modelu na přesnosti výpočtu modelu | 39 |

SEZNAM TABULEK

| | | |
|-----|---|----|
| 5.1 | Matice \mathbf{A} | 25 |
| 6.1 | Počet dokumentů v korpusu dle emoce | 32 |
| 6.2 | Emoce clusteru při použití metody Maximální počet dokumentů . . . | 35 |

ÚVOD

Textové informace můžeme rozdělit do dvou kategorií. První, která obsahuje faktické informace, tedy objektivní informace o objektech, událostech a jejich vlastnostech. Druhá jsou subjektivní názory autora textu a obsahují autorovi nálady, přání, předpoklady a pocity.

Tato diplomová práce se zabývá značkováním textu dle subjektivního názoru autora, konkrétněji přímo extrahováním informace o emocích, který tento text vyjadřuje. Jedná se relativně o novou disciplínu, ve které ještě nebyl proveden rozsáhlý výzkum. Před příchodem webových stránek, a obzvláště sociálních sítí obsahujících množství subjektivního textu, by ani nebylo možné využít algoritmy na psaný text. Subjektivní informace před příchodem internetu totiž byly sdělovány převážně ústně bez textové podoby.

A právě s rozvojem sociálních sítí se taktéž stává emocionální analýza důležitá například pro marketingové výzkumy, které dokáží určit, jaké emoce vyjadřují uživatelé těchto sítí na nově uvedený výrobek. Nebo při dotazu na oddělení péče o zákazníky se primárně věnovat těm dotazům, ze kterých je patrné, že jejich autor má negativní zkušenost s daným výrobkem či službou.

Největším přínosem této práce je vytvořená implementace algoritmu metody LDA pro klasifikaci textů v jazyce Java a experimentální ověření přesnosti této metody při určování emoce textu za použití různých parametrů modelu.

V první kapitole je popsáno, jak jsou emoce definovány v psychologické literatuře, do jakých kategorií se dělí a jaké důsledky z toho plynou pro emocionální analýzu. Druhá kapitola popisuje metody, kterými je zpracováván text do podoby, ve které je na něj možné aplikovat algoritmy pro získání emocí z textu. Následující kapitola poté dělí tyto algoritmy podle způsobu fungování. Poté jsou popsány výzkumy, které se taktéž zabývají emocionální analýzou za použití metod strojového učení bez učitele. Další kapitola se podrobněji zabývá metodami strojového učení bez učitele, konkrétně sémantickému modelování, což jsou metody schopné určovat podobnost mezi jednotlivými texty a jejich slovy. Poslední kapitola popisuje vytvořenou aplikaci klasifikující text dle emoce pomocí algoritmu LDA, popsány možné metody klasifikace a uvedeny výsledky měření jednotlivých metod.

1 EMOCE

Emoce procesy, které jsou vytvářeny jak psychicky, tak sociálně. Jejich zážitkovým jádrem jsou city, mají také ale složku fyziologickou, projevující se vegetativními (změna rychlosti srdečního tepu a dýchání) a motorickými reakcemi (mimika a gestikulace) [1].

Co z této definice plyne pro klasifikaci emocí v textu? To, že jsou to sociálně konstruované procesy, znamená, že nejsou dané od narození, ale člověk se je „učí“ během socializace, a tedy různé sociální skupiny mohou používat stejné slova pro vyjádření různých emocí či přiřazovat těmto slovům různou emocionální intenzitu a při kategorizaci textů dle emocí je nutné tedy dbát na kontext, ve kterém jsou tyto texty použity.

Druhou záležitostí je způsob, jakým jsou emoce projevovány. Podle populárně tradovaného psychologického mýtu tvoří slova, která používáme pouze 7 % našich emocí, dalších 38 % vyjádřeno pomocí tónu hlasu a jeho modulace, zbývajících 55 % tvoří výraz tváře a gestikulace. Pokud by byl tento mýtus pravdivý, znamenalo by to, že detekovat emoce pouze z textu by bylo velice nepřesné až nemožné. Tento mýtus ovšem vznikl pouze z jednoho velmi specifického experimentu a nemůže být tedy vztahován na odlišné situace a sám autor se od jeho generalizace distancuje [2].

1.1 Rozdělení emocí

Emoce se dle intenzity a délky trvání dělí do tří kategorií:

- afekt – velmi intenzivní krátkodobá emoce,
- nálada – méně intenzivní dlouhodobější emoce,
- vášně – intenzivní a dlouhodobá emoce.

Dle typu se dělí na základní, které se vyskytují u všech národností a kultur (avšak jejich projevy mohou být odlišné), ale také u vyšších živočichů, a na vyšší (ty jsou evolučně starší). Mezi základní emoce se v psychologické literatuře uvádějí následující typy [1]:

- Úzkost
- Radost a štěstí
- Smutek
- Strach
- Hněv
- Pocit viny
- Stud
- Starost
- Hnus

- Lítost
- Naděje
- Empatie

Tento seznam však není konečný, neboť různí psychologové emoce kategorizují jinak a tedy tyto emoce nemají ustálené definice a záleží na každém člověku, do které kategorie by projevenou emoci zařadil. To ještě ovlivňuje tzv. emoční inteligence (schopnost zvládat své emoce a vcítit se do emocí ostatních), zda bude schopen tento člověk danou emoci vůbec zachytit.

2 PŘEDZPRACOVÁNÍ PŘIROZENÉHO TEXTU

Text můžeme považovat za vektor znaků. Samotné znaky však nenesou sémantický význam sami o sobě, a proto je potřeba tento vektor rozdělit na vyšší jazykové celky – slova a věty. S těmito celky můžeme dále pracovat, jako například doplnit slovům chybějící diakritiku, odstranit sémanticky nevýznamná slova, nebo slova převést na základní tvar u jazyků, které slova ohýbají.

V této kapitole jsou postupně rozebrány tyto postupy v pořadí, jimiž text postupně prochází. Tyto postupy je možné použít na jakýkoliv jazyk oddělující jednotlivá slova, jejich popis je však převážně zaměřen na češtinu.

2.1 Lexikální analýza

Základním problémem při práci s přirozeným jazykem je rozdělení textu na lingvisticky významné jednotky [3].

Tento proces se nazývá segmentace textu a rozděluje se na rozlišení jednotlivých slov a vět. Proces rozdělení věty na slova se nazývá tokenizace (nebo také segmentace slov) a spočívá v nalezení hranice slov, tedy míst, kde dané slovo začíná a kde končí. Takto určené slovo se proto někdy označuje jako token. Segmentace vět má za cíl nalézt jednotky, které se skládají z jednoho nebo více slov a obdobně jako u tokenizace hledá hranice začátku a konce věty.

V praxi nemůžou oba procesy fungovat nezávisle na sobě, jelikož například tečka na konci zkratky *Bc.* neoznačuje ukončení věty, ale náleží k tomuto slovu. Další komplikace nastává v případě, kdy zkratka zakončená tečkou leží na konci věty a tečka tvoří jak hranici slova, tak hranici věty.

Proces segmentace závisí na použitém korpusu (soubor textů určitého jazyka, někdy specificky zaměřený pouze na určitou oblast – například texty zákonů), tedy na typu zpracovávaných textů, neboť například při zpracování komentářů v diskusích pod zpravodajskými články či statusy a komentáře na sociálních sítích nebývají vždy dodržována gramatická pravidla a celý proces tedy musí být mnohem robustnější než například při zpracování textů novinových článků.

2.1.1 Tokenizace slov v českém jazyce

Pro rozdělení jednotlivých slov v česky psaném textu se používá mezera, odlišení jednotlivých slov by mohlo vypadat jako jednoduchý proces – proházet text znak po znaku a v případě nalezení znaku mezery označit předchozí znaky za znaky tokenu. Jak už ale bylo uvedeno výše, tento přístup vykazuje několik problémů:

- Zkratky zakončené tečkou – v tomto případě by tečka měla náležet k tokenu, aby s ním bylo možné dále pracovat jako se zkratkou.
- Složená slova – některé se nemusí oddělovat spojovníkem a je tedy otázka, zda ze složených slov oddělených spojovníkem vytvořit dva tokeny nebo je naopak spojit v jeden.
- Čísla a data by měly být považováno za jeden token, i když jsou oddělena tečkami či čárkou.
- Emotikony (smajlíky) – skládají se z několika znaků (většinou interpunkčních znamének) a musí zůstat zachovány jako jeden token, aby s nimi bylo možné dále pracovat, neboť právě pro emocionální analýzu jsou podstatné, protože vyjadřují emoci sami o sobě.

2.1.2 Segmentace vět v českém jazyce

České věty bývají zakončeny jedním ze třech interpunkčních znamének: tečka zakončující oznamovací větu, vykřičník ukončující větu rozkazovací nebo zvolací a otazník označující větu tázací. Rozlišit jednotlivé věty ale opět není tak jednoduché:

- Jak již bylo uvedeno, tečka se používá taktéž pro ukončení zkratky, v řadových číslovkách či v češtině nesprávně, ale často jako desetinný oddělovač v číslech nebo pro odlišení tisíců v číslovkách.
- Vykřičník se používá i uprostřed věty obklopen kulatými závorkami dávající důraz na předešlé slovo.
- Samostatnou záležitostí je práce s přímou řečí nacházející se ve větě. Například zacházet s větou *Zuzka volala: “Jakube! Jakube! Stůj!” a utíkala za ním.* jako s jednou větou či ji rozdělit na více vět? Přímá řeč nevyjadřuje postoje autora, a proto by měla být pro potřeby emocionální analýzy označena, aby s ní bylo možné pracovat odlišně (například ji kompletně vypustit). Uvozovky se v textu používají navíc pro označení ironie, tedy přiřazují slovu jiný (opačný) význam, než má dané slovo samo o sobě (například věta *Já miluji Zuzku* a *Já „miluji“ Zuzku* má slovo *miluji* naprosto opačný význam).

2.2 Stop slova

Stop slova jsou taková slova, která nenesou sémantický význam. Většinou se jedná o slova, která se v korpusu vyskytují nejčastěji, ale nemusí to platit pro každé slovo sémanticky nevýznamné (například málo používaná spojka *třebaže*) [4]. V češtině se převážně jedná o slovní druhy:

- spojky (ale, a, i, ...),
- předložky (v, k, u, ...),

- zájmena (já, ty, on, ona, ono, ...),
- často používané slovesa (být, mít).

Odstraněním těchto slov z textu nejenom že dojde ke zjednodušení následujících procesů (sníží se celkový počet zpracovávaných slov), ale také může dojít ke zpřesnění výsledků následného zpracování, neboť například dva texty obsahující velké množství spojek by mohli být považovány za více podobné než v případě, kdy tyto spojky budou odstraněny.

Jelikož se ve většině případů jedná o neohebná slova (předložky, spojky) a není jich velké množství, je možné vytvořit ručně seznam těchto slov a odstraňovat je ještě před tím, než jsou slova převedena lematizátorem (viz níže) na jednotný tvar.

Druhý možný způsob je automaticky projít seznam všech slov v korpusu a odstranit například 10 % nejpoužívanějších slov nebo pomocí metody tf-idf (term frequency-inverse document frequency, popsána níže) nalézt nejméně relevantní slova. U tohoto postupu ale může dojít k odstranění i slov nesoucích sémantický význam, avšak dle použitého algoritmu následného zpracování textů to nemusí znamenat žádnou překážku, neboť i tak by těmito často se vyskytujícími slovy přirazoval nízký význam.

Třetí nejjednodušší způsob je ignorovat všechna jednoznaková a dvouznaková slova.

Pro účely emocionální analýzy může být vhodné zachovat pouze tzv. NAVA slova (z anglického Noun – podstatné jména, Adjective – přídavné jména, Verb – slovesa a Adverb – příslovce), protože právě tato slova jsou nosiče emocí. K tomu však musíme znát slovní druh použitého slova, k čemuž slouží nástroje označované jako Part-of-speech tagging.

2.3 Doplnění diakritiky

Česká gramatika používá diakritická znaménka umístěná nad písmenky. Pokud zpracováváme texty pocházející z diskusí a sociálních sítí, kde uživatelé z historických důvodů nebo kvůli rychlejšímu psaní nepoužívají diakritiku, musíme ji doplnit, neboť poté by slova *stastny* a *šťastný* byly považovány za odlišné. Nejjednodušší řešení je odstranit diakritiku ze všech slov – i když by to vedlo ke spojení různých slov v jedno (například slovo *žebra* by bylo převedeno na slovo *zebra*, které označuje zvíře), pro emocionální analýzu textu by to nemuselo mít podstatný vliv.

Pokud by v následném procesu byl použit lematizátor, pro slova bez diakritiky by nebyl nalezen základní slovní druh, a proto toto řešení není možné použít. Jedním z nejjednodušších řešení je použití slovníku pro kontrolu pravopisu, který obsahuje velkou část existujících slov včetně jejich ohnutých tvarů. V případě, že toto slovo

neobsahuje diakritiku a není nalezeno ve slovníku, je vygenerováno slovo ve všech možných tvarech (například pro slovo *caj* jsou vygenerována slova *čaj*, *cáj* a *čáj*). Pokud je jedno z těchto slov nalezeno ve slovníku, je použito místo slova původního. Tento proces je však velmi náročný pro dlouhá slova bez diakritiky (například pro slovo *nejnarocnější* bude vygenerováno celkem 4607 variant!) a navíc u některých slov může dojít k doplnění diakritiky na jiné slovo, než jaké by se v daném kontextu mělo nacházet (například u slova *presne* nelze bez znalosti kontextu určit, zda se má doplnit na slovo *přesně* nebo *přesné*).

Druhou možností je vytvořit konečný automat vycházející z korpusu a najít slovo s diakritikou podle četnosti použití v použitém korpusu [5]. Tuto metodu používá například systém CzAccent vyvinutý na Masarykově univerzitě. Výhodou tohoto postupu je rychlejší nalezení kýženého slova a větší pravděpodobnost nalezení správného slova, neboť doplňuje nejčastěji používanou variantu slova. Stále však zůstává problém s doplňováním dle kontextu.

Třetí variantou je systém vycházející opět z korpusu, ale pracující s N-gramy, tedy například z dvojic po sobě jdoucích slov. Při doplňování diakritiky by bylo možné vycházet z těchto dvojic a pokud by bylo v korpusu uvedena dvojice slov *jíme žebra*, při doplňování diakritiky do dvojice slov *jíme zebra* by byl doplněn správný tvar slova *žebra*. Nepodařilo se mi ovšem najít, zda by se tento systém pro doplňování diakritiky pro češtinu používal.

2.4 Tvarosloví

Gramatika českého jazyka a ostatních slovanských jazyků pracuje s ohýbáním slov (v menší míře se vyskytuje i v angličtině a němčině), takže se v přirozeném textu nacházejí slovní tvary než přímo jednotlivá slova. Pokud na tento jev náležitě nezareagujeme, povede k odlišení slov se stejným významem pouze jinak ohnutá v důsledku kontextu věty. Například ve větách *dobrý muž* a *dobrá žena* vyjadřuje slovo *dobrý* stejnou emoci, ovšem prostým porovnáním znaků slova by nedošlo k jejich propojení a při následném zpracování by byly považována za dvě odlišná slova. Naším cílem tedy je dosáhnout toho, aby různě ohnutá slova byla považována za stejná [6].

2.4.1 Morfologické nástroje

V lingvistických disciplínách se pro popis slova používá tzv. morfů, ze kterých se jednotlivá slova skládají a nesou lexikální nebo gramatický význam.

Lexikální morfy nazýváme také jako kmeny slov. Tento morf zůstává u slova po odtržení všech gramatických morfů. Zjednodušeně řečeno se jedná o tu část slova,

která zůstává stejná ve všech tvarech daného slova. Problémem je, že v češtině existují slova, která mají několik různých kmenů (například slovo *dům* má dva kmeny: *dům* a *dom-*). Lexikální morfy se ještě dělí na submorfy (odvozovací předpony, kořeny a odvozované přípony), která se ale používají spíše pro studium historického vývoje slov a pro počítačové zpracování přirozeného jazyka nemá velký význam, v českém jazyce by je bylo ovšem možné použít pro hledání odvozených slovních druhů z jiného slovního druhu (například přídavného jména *knižní* od podstatného jména *kniha*).

Gramatické morfy jsou nejmenší prvky slova, které nesou určitý gramatický význam. Dělí se na:

- koncovky,
- gramatické předpony (například nej- vytvářející tvar třetího stupně u přídavných jmen a příslovčí),
- infixy (prvky vkládané dovnitř kmene) a interfixy (posloupnosti hlásek prokládané mezi hlásky kmene) – v českém jazyce tyto morfy neexistují.

Jazyky, které odlišují různé tvary slov (jako čeština) můžeme rozdělit do dvou topologických tříd:

- **třída flexivních jazyků**, ve kterých gramatické morfy kombinují více gramatických významů. Mezi tuto třídu jazyků patří i čeština, neboť například pádové koncovky v sobě neoddělitelně obsahují jak informaci o pádu i čísle. Pro počítačové zpracování jazyka je navíc důležité, že různé morfy mají stejný význam v závislosti na třídě slov (v češtině například koncovky vyjadřující 2. pád jednotného čísla podle různých skloňovacích vzorů).
- u **aglutinačních jazyků** naopak typicky jednotlivé gramatické morfy nesou jen jeden elementární gramatický význam. Mezi tyto jazyky patří například maďarština nebo finština.

Důležitou schopností jazyků je možnost skládat slova, která mají více plnohodnotných lexikálních morfů, neboli kořenů. Tento systém skládání slov má nejvyvinutější němčina, v češtině je typický pro některá přídavná jména (například slovo *rychloobrátkové*), v některých složitějších případech se v písmu tyto kmeny oddělují spojovníkem (například *vědecko-technický*).

2.4.2 Práce s tvaroslovím

Pro řešení problému s tvaroslovím můžeme zvolit dva základní nástroje – derivátor slovních tvarů a lemmatizátor.

Derivátor je algoritmus, který k danému slovu vygeneruje všechny možné gramatické tvary případně i odvozeniny. Například pro slovo *počítač* by tento algoritmus vygeneroval tvary: *počítač*, *počítače*, *počítači*, *počítači*, *počítačem*. V češtině se

pro skloňování podstatných jmen používá 14 různých vzorů, pro přídavné jména jen vzory dva. Pokud bychom tedy dokázali zjistit, do které třídy slovo patří, vygenerovat všechny tvary přidáním správných koncovek podle příslušného vzoru se nejeví jako velký problém. Avšak slova s latinským původem, pomnožná podstatná jména a vlastní jména cizího původu se těmito vzory neřídí ve všech případech. Dalším problémem komplikujícím skloňování je to, že u některých slov dochází při skloňování ke změně kmene (např. kořen slova *dům* se v druhém pádě mění na *dom-*).

Lematizátor je v podstatě inverzní proces k derivátoru a tedy pro slovo v jakémkoliv podobě vrací základní (slovníkový) tvar, holý kmen (v anglické literatuře se pro tento nástroj častěji používá označení **stemmer**) nebo jen kořen slova. Jaký výstup lematizátoru je nejvhodnější záleží na způsobu použití a zpracovávanému jazyku. Při použití kmene slov v češtině je nejeví jako vhodné, neboť některé slova s různým významem mají stejné kmeny (slova *past* a *pasta* by tak byla považována za stejná) a některé slova nemají, jak již bylo uvedeno výše, pouze jeden kmen a nastává tedy otázka, který kmen by měl lematizátor pro tyto slova vracet, a proto se spíše používá převod na základní tvar.

Dalšími možnými způsoby je porovnávání podobnosti slov se slovníkem základních tvarů. Ke zjištění zvukové podobnosti slov se používá metoda SOUNDEX, která je však vytvořena pro anglický jazyk a pro složitější češtinu by mnoho slov s odlišným významem bylo považováno za stejná. Další možností je převést dané slovo na trigramy (posloupnost tří po sobě jdoucích hlásek) a porovnání s míry podobnosti s trigramy slov nacházejících se v korpusu a přiřazovat k danému slovo další v případě vysoké míry shody (obdobně jako derivátor). Pokusy s těmito metodami v anglickém jazyce ale nevedly k lepším výsledkům.

Pro detekci emocí v přirozeném textu je vhodné z uvedených nástrojů využít lematizátoru, neboť derivátor vstup derivátoru musí být slovo v základním tvaru, čehož není možné docílit v přirozených textech a navíc vede ke vzniku mnoha dalších slov, které by bylo nutné dále zpracovávat a měnit jejich váhu v závislosti na tom, kolik slov derivátor vygeneroval.

U lematizátoru je ale třeba zvážit jeho výstup, jelikož lematizátor vracející lexikální morf (kmen) slova, jak už bylo uvedeno, by v českém jazyce vedl k problematickým výsledkům a speciálně pro použití v emoční analýze by například u třetího stupně přídavných jmen odtrhl gramatickou předponu *nej-*, která nese jinou emocionální informaci, než u slova bez této předpony. Lematizátor by tedy měl vracet základní tvar slova.

3 METODY PRO ZÍSKÁVÁNÍ EMOCÍ Z DOKUMENTU

Mezi základní metody pro získávání emocí z textu dokumentu patří slovníkové metody pracující s vytvořeným slovníkem slov s přiřazenou emocí, lingvistické metody vycházející z pravidel gramatiky jazyka a strojové učení, které po natrénování modelu na korpusu je schopno automaticky určovat typ emoce nacházející se v textu [7].

Tyto metody se dělí na strojové učení s učitelem, kdy je model natrénován pomocí dokumentů s již určenou emocí (tedy správnou „odpovědí“) a model posléze pro nový neoznačený dokument vrací dle podobnosti s ostatními dokumenty nejpravděpodobnější emoci. Při použití strojových metod učení bez učitele naopak není potřeba mít data označena, ale tyto metody automaticky zpracují text, ve kterém naleznou podobnosti. Pro získání typu emoce v dokumentu je poté potřeba porovnat slova označující danou emoci s modelem.

3.1 Slovníkové metody

Při používání těchto metod se většinou využívá ručně vytvořených slovníků slov spolu s informací, jakou emoci tyto slova nesou. Vytvoření tohoto slovníku je zdlouhavá ruční práce, počet emocionálních kategorií je u těchto metod omezen vytvořeným slovníkem a nedokáže označit větu, které obsahuje emoce ovšem bez přímého uvedení slova tuto emoci nesoucí (například ve větě: *Malý Petr dostal na Vánoce hodně dárků*). Dalším problémem je, že tyto metody pracují pouze s jednotlivými slovy bez ohledu na kontext, ve kterém jsou použity, což může ovlivnit typ vyjádřené emoce. Tyto slovníky existují například pro anglický jazyk, pro češtinu se mi nepodařilo žádný veřejně dostupný slovník nalézt.

3.2 Lingvistické metody

Další z možných metod je porozumění lingvistických pravidel, kde se určení emoce využívá principu založeného na základě několika předpřipravených typech frází obvykle vyjadřující subjektivní význam. U těchto metod dochází k vytažení frází obsahující přídavná jména nebo příslovce, neboť významy ukázaly, že právě tyto slovní druhy jsou dobrými ukazateli toho, že daná fráze je subjektivní. Přesto však bez uvedení kontextu není možné určit orientaci emoce. Proto algoritmus vyjme z fráze dvě po sobě jdoucí slova, kde jedno z nich je přídavné jméno nebo příslovce a druhé

je slovo určující kontext. Tyto algoritmy pracují s tabulkou jaké slovní druhy, které se nacházejí za sebou, budou použity.

Například algoritmus [11] založený na této metodě vybírá mimo jiné slova podle vzoru, kdy první slovo musí být přídavné jméno a následující podstatné jméno v jednotném nebo množném čísle, což znamená, že z věty *This camera produces beautiful pictures* vybere slova *beautiful pictures*. V následujícím kroku dojde ke spočtení pravděpodobnosti pomocí metody PMI (pointwise mutual information), že se uvedené slovo nachází v korpusu blíž slovům *excellent* nebo *poor*, čímž je zjištěna orientace (kladná nebo záporná) tohoto slova.

Nevýhodou těchto metod je pracné vytvoření tabulky slovních druhů, která je navíc závislá na použitém jazyku (zatímco ve zmíněném algoritmu bylo pro angličtinu použito jen pět pravidel, pro složitější češtinu by jich muselo být více a byly by složitější) [7] .

3.3 Strojové učení s učitelem

Strojové metody pro určení emoce většinou pracují s učitelem. Tento přístup však vyžaduje velké množství textu, které musí člověk ručně zařadit k příslušné emoci, kterou tento text vyjadřuje. Tento korpus se posléze použije k natrénování modelu (jako nejúčinnější se ukázala metoda SVM [12]). Přestože dosahují dobrých výsledků, alespoň v češtině neexistuje žádná veřejná databáze, která by obsahovala text spolu s určenou emocí. Navíc databáze z jedné domény textů nebude dobře fungovat při použití na doménu jinou (například označená databáze pocházející z komentářů pod webovými zprávami bude odlišná od databáze vycházející z hodnocení produktů na e-shopu).

Existují taktéž metody spojující slovníkové metody a strojového učení s učitelem. Slovníkové metoda se použije pro natrénování modelu a metoda poté dokáže určit, na rozdíl od přímé slovníkové, emoce i u textu, ve kterém se přímo nevyskytuje slovo uvedené ve slovníku.

3.4 Strojové učení bez učitele

Tyto metody pracují s automatickou kategorizací textu dle použitých slov a jejich frekvence. Mezi ně patří metody sémantického modelování: latentní sémantická analýza (LSA), pravděpodobnostní LSA (pLSA) a latentní Dirichletova alokace (LDA). Metoda LDA bude použita pro emocionální analýzu v této práci a proto bude popsána blíže spolu s ostatními metodami, ze kterých principiálně vychází. Princip

těchto metod spočívá v tom, že slova nacházející se v dokumentech často spolu budou mít podobný význam bez ohledu na pozici, ve které se v textu nachází (tzv. bag-of-words). Používají se ale i jiné metody, například založené na Pointwise mutual information [7].

4 SOUČASNÉ PŘÍSTUPY

Klasifikace emocí pomocí metod strojového učení bez učitele byla použita již v předchozích výzkumech. V této kapitole jsou popsány ty, jsou relevantní k této diplomové práci.

Článek Chenghua Lina a Yulana He [8] využívá pro určení emoce jimi vytvořenou metodu Joint Sentiment/Topic (JST) Model, která rozšiřuje vrstvu LDA o emocionální klasifikátor. Jedná se tedy o metodu strojového učení bez učitele, pro zvýšení účinnosti je možné použít seznam emocionálně zabarvených slov spolu s příslušnou kategorií. Tato metoda byla testována pro kategorizaci 2 000 hodnocení filmů do třech kategorií: pozitivní, neutrální a negativní. Bez využití seznamu emocionálně zabarvených slov byla přesnost v určení pozitivní nebo negativní emoce přibližně 60 %, při použití různých seznamů slov byla přesnost až 84 %. Autoři tuto metodu porovnávali s metodami založenými na SVM, která dosahovaly přesnosti až 90 %.

Práce [9] porovnává tři metody pro detekci emocí: dvě využívající učení s učitelem (Naïve Bayes a SVM) a metodu strojového učení bez učitele označovanou jako SO-PMI-IR. Tato metoda spočívá v extrakci předem určených slovních spojení z dokumentu dle slovních druhů u kterých je následně určeno, s jakou pravděpodobností se nachází v blízkosti slov „excellent” a nebo „poor” v dokumentech korpusu. Emoce celého dokumentu je následně zjištěna rozdílem pozitivních a negativních pravděpodobností nebo za využití metody sémantického modelování LSA (tato metoda je poté označována jako SO-LSA). Pro určení emoce slovního spojení je zapotřebí velkého korpusu, autoři ve své práci využili vyhledávač, který prohledával webové stránky. Tyto metody byly použity pro detekci pozitivní nebo negativní emoce na 2 000 dokumentech, přesnost Naïve Bayes byla 83 %, SVM 79 % a přesnost SO-PMI-IR 84 %.

Hybridní metoda HDP-LDA (Hierarchical Dirichlet process-Latent Dirichlet Allocation) byla použita v článku [10] a má vyřešit problém samostatné metody LDA, která nedokáže odlišit faktické a emocionální slova a věty přiřazuje jen jeden aspekt. Pomocí HDP je určen počet aspektů, které budou použity pro model LDA. Jedná se o podobnou metodu jako JST, neboť taktéž rozšiřuje funkci LDA, ale měla by poskytovat vyšší přesnost, jelikož umožňuje odlišit faktické a emocionální slova. Tato hypotéza se potvrdila i v experimentálních výsledcích, kdy byla metoda použita pro určení pozitivní, negativní nebo neutrální emoce v recenzích restaurací a byl využit seznam emocionálně zabarvených slov – pro použité korpusy HDP-LDA dosahovala vždy lepších výsledků než JST.

5 SÉMANTICKÉ MODELOVÁNÍ

Vstupem těchto metod je dokument, který byl předzpracován metodami uvedenými v kapitole 2. Předzpracování přirozeného textu, tedy základní tvary slov po lematizaci. Pro každý dokument d nacházející se v korpusu D je poté vytvořen vektor, který má délku jako seznam všech unikátních slov použitých v daném korpusu. Pokud se v tomto dokumentu dané slovo vyskytuje, je na pozici vektoru tohoto slova uvedena hodnota 1, v opačném případě 0. Z toho vyplývá, že tento vektor bude velmi řídký, neboť daný dokument bude obsahovat pouze několik slov nacházejících se v celém vektoru. Tyto vektory se následně skládají do matice \mathbf{A} , kde každý řádek označuje dokument a každý sloupec slovo z korpusu [13].

Mějme tři dokumenty d_n

$$d_1 = \text{Modré peří na zem padá padá padá neroztáva} \quad (5.1)$$

$$d_2 = \text{Jako kámen k zemi padal jako krásný modrý kámen} \quad (5.2)$$

$$d_3 = \text{Padá tam kde roste tráva modré modré peří sněží} \quad (5.3)$$

po provedení předzpracování vzniknou tři vektory d'_n

$$d'_1 = (\text{modré, peří, zem, padat, neroztávat}) \quad (5.4)$$

$$d'_2 = (\text{kámen, zem, padat, krásný, modré}) \quad (5.5)$$

$$d'_3 = (\text{padat, růst, tráva, modré, peří, sněžit}) \quad (5.6)$$

seznam slov nacházejících se ve všech transformovaných dokumentech seřazených dle abecedy pak označíme jako

$$V = (\text{kámen, krásný, modré, neroztávat, padat, peří, růst, sněžit, tráva, zem}) \quad (5.7)$$

výsledná reprezentace pak získá podobu tabulky 5.1.

Místo prostého přiřazení hodnoty 1 při výskytu slova nebo 0 je vhodnější využít již dříve zmíněnou metodu tf-idf. Ta určuje, jaký je výskyt daného slova v dokumentu relevantní vůči slovům v celém korpusu. Tato hodnota přímo úměrně narůstá s tím, jak často je použito v daném dokumentu, avšak je posunuto o to, jak toto slovo často použito v celém korpusu – takže jsou zvýhodněna slova specifická pro daný dokument a potlačena slova, která se vyskytují často ve většině textů.

Výpočet této hodnoty je následující [14]: Jako term frequency $\text{tf}(t, d)$ je považován počet výskytů slova t v dokumentu d , tedy $\text{tf}(t, d) = f(t, d)$. Existují ale i odlišné

Tab. 5.1: Matice **A**

| slovo | d'_1 | d'_2 | d'_3 |
|------------|--------|--------|--------|
| kámen | 0 | 1 | 0 |
| krásný | 0 | 1 | 0 |
| modré | 1 | 1 | 1 |
| neroztávat | 1 | 0 | 0 |
| padat | 1 | 1 | 1 |
| peří | 1 | 0 | 1 |
| růst | 0 | 0 | 1 |
| sněžit | 0 | 0 | 1 |
| tráva | 0 | 0 | 1 |
| zem | 1 | 1 | 0 |

metody pro výpočet tf, jako například rozšířená frekvence, která bere ohled na délku dokumentu. V tom případě

$$\text{tf}(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{\max\{f(w, d) : w \in d\}} \quad (5.8)$$

Inverse document frequency určuje, kolik informace dané slovo nese, tedy jak často (tedy spíše jak málo) se vyskytuje ve všech dokumentech. Spočítá se jako

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (5.9)$$

kde:

- N značí délku korpusu
- $|\{d \in D : t \in d\}|$ je počet dokumentů d v korpusu D , ve kterých se vyskytuje slovo t

Výsledná hodnota vznikne vynásobením obou hodnot:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \quad (5.10)$$

U takto vytvořené matice můžeme spočítat podobnost dvou dokumentů (řádků) použitím kosinovy podobnosti pro tyto vektory výpočtem kosinu úhlu, které tyto vektory svírají. Jelikož tyto vektory vždy svírají úhel od 0 do 90 stupňů (hodnota po tf-idf nemůže být záporná), je výsledná hodnota od 0 do 1, kde 1 značí stejný dokument [13].

$$\text{podobnost}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (5.11)$$

Nevýhodou tohoto přístupu je vysoký počet dimenzí (sloupců matice), které jsou z velké části prázdné. Prostor matice roste exponenciálně a rozdíly ve vzdálenostech klesají. Pro zmenšení velikosti můžeme vynechat ty dimenze, které jsou nenulové pouze v jednom nebo v několika málo dokumentech.

Výše uvedený postup je společný pro všechny tři následující metody sémantického modelování. Nyní bude popsáno, jak snížit počet dimenzí vytvořením témat, které tvoří skupina slov s podobným významem, kdy každému slovu je přiřazena určující míra příslušnosti slova k danému tématu. Důvod aplikace tohoto kroku je:

- Další snížení prostoru matice, neboť práce s velkou maticí je výpočetně a paměťově náročná.
- Snížení šumu, tedy odstranění slov, které se v dokumentech vyskytují jen sporadicky, čímž vznikne matice vhodnější pro další použití.
- Dojde k nahrazení slov, které se opravdu v dokumentu vyskytují, skupinou slov se stejným významem či slov relevantních pro daný dokument.

5.1 LSA

Základem latentní sémantické analýzy (LSA, někdy též nazývána jako LSI – latentní sémantické indexování) je tzv. singular value decomposition (SVD). Matice \mathbf{A} je aproximována maticí s nízkou hodnotí \mathbf{A}' .

Toho je dosaženo postupnou volbou os původního prostoru tak, aby byl zachován co největší rozptyl dat. Jde tedy o techniku nejmenších čtverců, která minimalizuje součet druhých mocnin změn. Nejdříve je třeba rozhodnout, kolik témat bude vytvořeno. Čím větší počet jich bude, tím jemnější rozdělení by se mělo vytvořit, avšak nebude možnost zachytit rozdíly u příbuzných témat [15].

Hlavní výhodou SVD pro použití v oblasti zpracování přirozeného jazyka je, že ignoruje rozptyl pod zadanou hranici pro silné snížení množství dat, ale zároveň zachovává hlavní závislosti. Vychází z principu lineární algebry, který určuje, že matice \mathbf{A} může být rozdělena do součinu tří matic – ortogonální matice \mathbf{U} , diagonální matice \mathbf{S} a transpozice ortogonální matice \mathbf{V} [16]:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \quad (5.12)$$

kde:

- $\mathbf{U}^T\mathbf{U} = \mathbf{I}$,

- $\mathbf{V}^T \mathbf{V} = \mathbf{I}$,
- sloupce matice \mathbf{U} jsou ortogonální vlastní vektory $\mathbf{A}\mathbf{A}^T$,
- sloupce matice \mathbf{V} jsou ortogonální vlastní vektory $\mathbf{A}^T \mathbf{A}$ a
- \mathbf{S} je diagonální matice obsahující odmocniny vlastních čísel z matic \mathbf{U} nebo \mathbf{V} v sestupném pořadí.

Po tomto kroku řádky matice \mathbf{V} reprezentují jednotlivé dokumenty a podobnost jednotlivých dokumentů je možné určit porovnáním řádků v matici \mathbf{VS} . Obdobně jsou slova reprezentována řádkovými vektory v matici \mathbf{U} a podobnost jednotlivých slov může být zjištěna porovnáním řádků v matici \mathbf{US} [16].

Pokud odstraníme triplet (pár vektorů z matice \mathbf{U} a \mathbf{V}^T s příslušnou hodnotu z matice \mathbf{S}) s nejmenší singulární hodnotou, získáme nejlepší aproximaci matice s nižší hodnotí \mathbf{A}' , kde

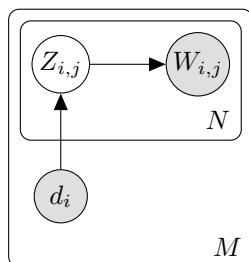
$$\Delta = \|\mathbf{A} - \mathbf{A}'\|_2 \quad (5.13)$$

je nejmenší [17].

V tomto procesu se pokračuje, dokud není získán požadovaný počet témat (pro typické velké korpusy se jedná přibližně o odstranění 300 tripletů).

5.2 pLSA

pLSA (pravděpodobnostní LSA, také známé pod označením Aspekt Model) je podobná metoda, která místo SVD používá statistické nástroje. Výhoda oproti LSA je ve stabilnějším matematickém základu a v dosahování lepších výsledků.



Obr. 5.1: Model pLSA v PLATE notaci

Předpokládejme, že v korpusu chceme identifikovat témata $z \in \{z_1, z_2, \dots, z_K\}$. Taktéž předpokládejme, že s každým tématem z je spojeno rozdělení pravděpodobnosti přes všechna slova $t \in \{t_1, t_2, \dots, t_M\}$, která se v korpusu vyskytují. Hledáme tedy funkci, která pro každé téma z přiřazuje každému slovu ze slovníku jeho pravděpodobnost, tedy $P(t|z)$. Dále obdobně předpokládejme, že pro každým dokumentem $d \in \{d_1, d_2, \dots, d_N\}$ z korpusu existuje funkce $P(z|d)$, která udává pravděpodobnost

výskytu tématu z v dokumentu. Spojením těchto dvou funkcí se poté pro určení pravděpodobnosti výskytu slova t v dokumentu d použije funkce [18]:

$$P(t|d) = \sum_{k=1}^K P(t|z_k, d)P(z_k|d) = \sum_{k=1}^K P(t|z_k)P(z_k|d). \quad (5.14)$$

Pravděpodobnost $P(t|z, d)$ můžeme zjednodušit na $P(t|z)$, neboť je důležité rozdělení pravděpodobnosti na globální úrovni, nezávislé na dokumentu.

Použitý model předpokládá, že slovo dokumentu je generováno právě z jednoho tématu a jsou na sobě navzájem v rámci tohoto dokumentu nezávislá. Z toho plyne pravděpodobnost dokumentu:

$$P(d) = \prod_{m=1}^M P(t_m|d)^{c(t,d)} \quad (5.15)$$

$$\log(P(d)) = \sum_{m=1}^M c(t_m, d) \log(P(t_m|d)) \quad (5.16)$$

kde $c(t, d)$ značí počet výskytů slova t v dokumentu d . Celý výraz můžeme zjednodušit na logaritmus pravděpodobnosti celého korpusu D :

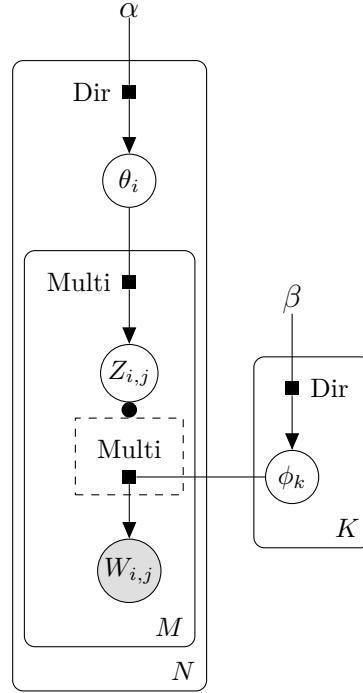
$$\log(P(D)) = \sum_{n=1}^N \sum_{m=1}^M c(t_m, d_n) \log(P(t_m|d_n)). \quad (5.17)$$

Stále však neznáme, jakým způsobem zjistit $P(t|z)$ a $P(z|d)$. Tyto pravděpodobnosti určíme pomocí metody založené na maximalizaci věrohodnosti dat. Pokud tyto neznámé označíme za X , pak budeme hledat takové parametry X , které maximalizují pravděpodobnost $P(D|X)$. K tomu se používá například algoritmus Expectation-Maximization. Poté získáme pravděpodobnost $P(z|d)$ pro libovolný dokument a téma. Pro každý dokument tedy bude existovat vektor témat, jehož složky budou určovat pravděpodobnost výskytu tématu v dokumentu.

Poté můžeme určit podobnost dvou dokumentů jako u LSA. Nevýhodou této metody je nemožnost predikce vektoru témat, které nebyly v trénovací množině a náchylnost k přeučování. Oba tyto problémy řeší metoda LDA a proto se pLDA v praxi nepoužívá [19].

5.3 LDA

Latentní Dirichletova alokace (LDA) je metoda, která vychází z pLSA, avšak na rozdíl od ní předpokládá, že témata mají Dirichletovskou apriorní pravděpodobnost rozdělení [19].



Obr. 5.2: Model LDA v PLATE notaci

Tato metoda vychází z binomického rozdělení pravděpodobnosti, které je definováno jako:

$$f(k|n, p) = \binom{n}{k} p^k (1 - p)^{n-k}. \quad (5.18)$$

V případě, že neznáme parametr p , ale existuje jen představa, je možné jej modelovat pomocí dalšího rozdělení pravděpodobnosti. K tomu se používá tzv. beta rozdělení, jehož definice je následující:

$$f(p|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1 - p)^{\beta-1}. \quad (5.19)$$

a jehož podoba se zásadně mění dle parametrů α a β . Binomické a beta rozdělení je definováno pouze pro dvourozměrný prostor, a protože LDA pracuje ve vícerozměrném prostoru, je třeba použít vícerozměrné varianty těchto funkcí. Vícerozměrná varianta binomického rozdělení se nazývá multinomické rozdělení (v matematickém zápisu označovaná jako *Mul*) a vícerozměrná varianta beta rozdělení je Dirichletovo rozdělení (*Dir*), od kterého získala tato metoda také svůj název.

Obdobně jako u pLSA při snaze získat určitý počet témat vyjádříme pravděpodobnost generování korpusu, a poté budeme hledat takové hodnoty parametrů, které tuto pravděpodobnost maximalizují. Na rozdíl od metody pLSA se však používají aproximativní metody jako například variační metody nebo Gibbsovo vzorkování.

Výhodou LDA oproti pLSA je fixní počet parametrů, tedy není náchylná k přeučování a dosahuje lepších výsledků. Hlavní výhodou je však možnost odhadovat témata i pro nové dokumenty, které nebyly použity k trénování modelu.

6 PRAKTICKÁ ČÁST

Pro ověření možnosti klasifikovat texty dle emocí metodou bez učitele byla zvolena metoda LDA, která z analýzy možných metod vzešla jako dosahující nejlepších výsledků. Taktéž umožňuje generovat témata i pro dokumenty, které nebyly použity pro natrénování modelu, což je potřeba pro ověření účinnosti algoritmu.

Implementace, která je součástí přílohy C, byla provedena dle zadání v programovacím jazyce Java a vycházela z implementace popsané v článku Hoffman, Blei, Bach: Online Learning for Latent Dirichlet Allocation [20]. Její výhodou je, že parametry modelu upravuje přímo po zpracování jednoho dokumentu, nebo po zpracování minidávky (minibatches, omezují šum) a ne po dávkách, jako jiné metody, které provádějí několikanásobný průchod všemi dokumenty dokud nedojde ke konvergenci modelu. Tato výhoda se projeví v menší časové náročnosti trénování modelu a také tím, že není nutné udržovat zpracovávané dokumenty v paměti a po zpracování je ho možné z paměti odstranit. Z téhož článku vychází také implementace LDA použitá v knihovně Gensim [21] pro jazyk Python, vůči které byla implementace vytvořená v této práci testována na správnou funkčnost.

V knihovně jsou využity mimo standardní knihovny jazyka Java také upravená třída `GammaDistribution` z matematické knihovny Math projektu Apache Commons, která slouží k výpočtu náhodných čísel dle rozdělení gama a je distribuována pod open-source licencí Apache a metoda `psi` počítající funkci digama $\psi(x)$ z matematické knihovny Cephes napsanou v jazyce C a distribuovanou pod licencí BSD. Ostatní třídy a metody byly vytvořeny v rámci této diplomové práce.

6.1 Trénovací a testovací data

Pro otestování funkce algoritmu byly použity dokumenty v českém jazyce rozdělené na dva korpusy:

- D_{train} – dokumenty určené pro natrénování modulu.
- D_{test} – dokumenty sloužící k porovnání dokumentů s natrénovaným modelem.

Každý dokument se skládá z již předzpracovaných slov a je zařazen pod jednu z následujících emocí:

- Afraid (znepokojení) – negativní
- Anger (naštvaní) – negativní
- Neutral (neutrální)
- Sadness (smutek) – negativní
- Satisfaction (spokojenost) – pozitivní
- Surprise (překvapení) – pozitivní

Tab. 6.1: Počet dokumentů v korpusu dle emoce

| Korpus | Afraid | Anger | Neutral | Sadness | Satisfaction | Surprise | Celkem |
|-------------|--------|-------|---------|---------|--------------|----------|--------|
| D_{train} | 89 | 110 | 22 | 129 | 149 | 22 | 521 |
| D_{test} | 42 | 48 | 21 | 87 | 120 | 21 | 339 |

Počet dokumentů dle korpusu a emoce je uveden v tabulce 6.1. Jedná se již o předzpracované dokumenty, u kterých bylo provedeno odstranění stop slov a slova byla převedena na základní tvar. Soubory obsahují přímo jednotlivá slova oddělená mezerami, tudíž není třeba pro rozdělení dokumentů na jednotlivá slova (tokeny) používat složité tokenizátory.

Trénovací dokumenty obsahují celkem 3 455 unikátních slov, testovací pak 2 894, z čehož však 1 509 není obsaženo v trénovací množině dokumentů, což znamená, že pro ověření funkce algoritmu je využito 1 385 slov.

6.2 Popis algoritmu

Program načítá dokument z korpusu D_{train} , rozdělí jej na jednotlivá slova za pomoci třídy `java.util.StringTokenizer`. Tím vznikne obdobná matice, jako je uvedená v tabulce 5.1. Jelikož by však byla většina buněk vyplněna nulami (jeden dokument z trénovací množiny obsahuje jen přibližně 15 unikátních slov z 3 455), je v implementaci použita datová struktura tzv. řídkého pole (spojový seznam) pro snížení paměťové náročnosti.

Na tuto matici je poté aplikována metoda pro tf-idf, tak jak je popsána v kapitole 5. Následně je tato matice použita pro natrénování modelu LDA.

Vstupními parametry metody jsou:

- matice skládající se z dokumentů a slov,
- počet clusterů, do kterých mají být jednotlivé dokumenty zařazeny,
- přesnost parametrů modelu (vyšší přesnost znamená zpomalení trénování),
- počet dokumentů v tzv. minidávce (minibatch), v rámci níž neproběhne přepočítání parametrů celého modelu (vyšší počet zvýší rychlost výpočtu a navíc může snížit „šum“ modelu).

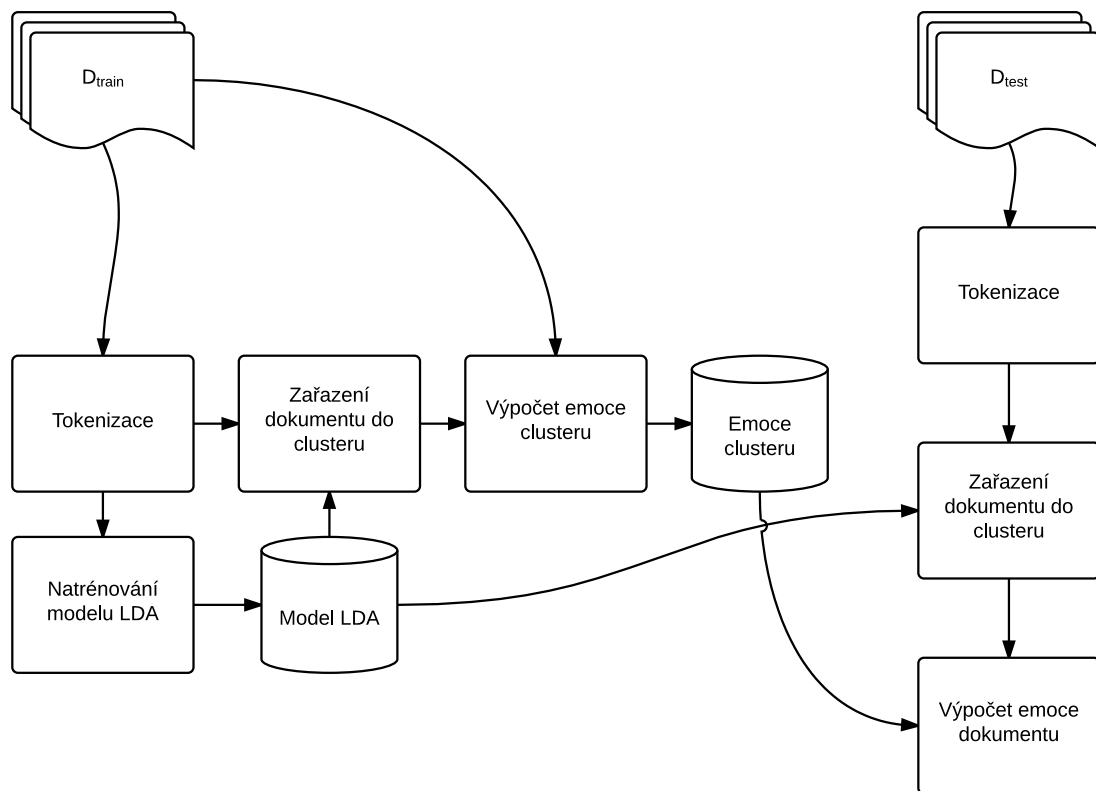
Výstupem metody je matice $\mathbf{LDA}_{m,n}$, kde m značí počet unikátních slov v korpusu a n počet clusterů, přičemž prvek této matice reprezentuje, s jakou „intenzitou“ je slovo zastoupeno v určitém clusteru.

Jelikož se jedná o metodu učení bez učitele, model byl vytvořen na základě podobnosti dokumentů bez informace o tom, jakou emoci reprezentuje. K tomu slouží

následující krok, ve kterém jsou dokumenty z korpusu D_{train} porovnávány s modelem. Výsledkem je informace o podobnosti (normovaná od 1 do 0) s jednotlivými clustery. S informací o podobnosti clusterů a emoci lze určit emoci clusteru a to pomocí různých metod, jejichž popis a výsledná přesnost je uvedena v kapitole 6.3.

V posledním kroku jsou načteny dokumenty z korpusu D_{test} a zpracovány obdobně jako dokumenty z D_{train} s tím rozdílem, že slova, která se nenacházela v D_{train} jsou ignorována, jelikož nejsou obsažena ani v matici $\mathbf{LDA}_{m,n}$, se kterou je dokument porovnáván. Výstupem porovnání je opět vzdálenost od jednotlivých clusterů a jelikož z předchozího kroku je známa informace o emoci, kterou vyjadřuje celý cluster, můžeme určit i emoci zpracovávaného dokumentu.

Celý proces je zobrazen na schématu 6.1.



Obr. 6.1: Blokové schéma funkce algoritmu

6.3 Metody pro získání emoce clusteru

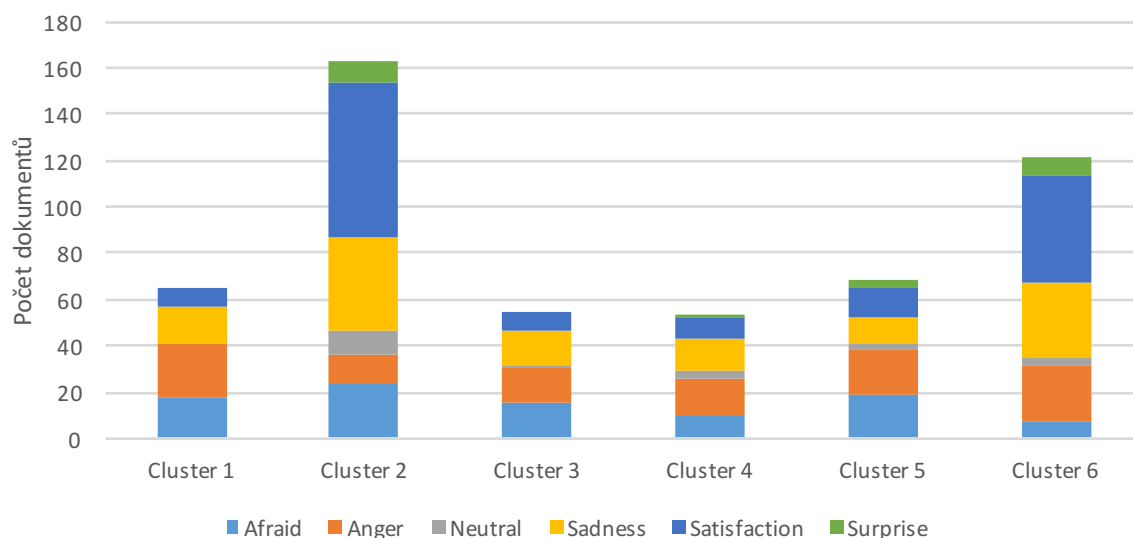
K získání těchto výsledků byla využita data popsaná v kapitole 6.1. Parametry algoritmu byly nastaveny empiricky a to následovně:

- Počet clusterů: 6

- Přestat počítat parametry dokumentu při změně menší než 1 %
- Velikost minidávky: 10

Po natrénování modelu bylo tedy vytvořeno 6 clusterů, v tabulce přílohy A je uvedeno ke každému clusteru deset slov reprezentující cluster spolu s jejich normalizovanou intenzitou (součet intenzit všech slov v clusteru se rovná 1, v tabulce pro přehlednost uvedenou v procentech).

Vůči tomuto modelu byly do jednotlivých clusterů rozřazeny dokumenty z D_{train} , počet dokumentů na základě emoce v clusteru je zobrazen v grafu 6.2. Dokument byl vždy přiřazen ke clusteru, k němuž měl nejnížší vzdálenost.



Obr. 6.2: Počet dokumentů náležící clusteru dle emoce

Jak bylo uvedeno v kapitole 6.2, k tomu, abychom získali informaci o emoci z korpusu D_{test} , musíme znát, jakou emoci reprezentuje cluster, ke kterému dokument náleží. Jelikož známe také informaci o vzdálenosti dokumentu od clusteru, je možné využít více způsobů, jakými lze emoci clusteru spočítat.

6.3.1 Maximální počet dokumentů

Nejjednodušší možností je považovat emoci celého clusteru podle toho, které dokumenty se v něm nejčastěji vyskytují. Dokument byl vždy přiřazen ke clusteru, k němuž měl nejnížší vzdálenost. Při použití této metody budou tedy emoce clusterů dle tabulky 6.2.

Nevýhoda této metody je zřejmá: potlačuje dokumenty s emocemi, které se v trénovací množině vyskytují s nižší intenzitou a taktéž celý cluster je reprezentován jen

Tab. 6.2: Emoce clusteru při použití metody Maximální počet dokumentů

| Cluster | Bez normalizace | S normalizací |
|---------|-----------------|---------------|
| 1 | Anger | Afraid |
| 2 | Satisfaction | Neutral |
| 3 | Sadness | Afraid |
| 4 | Anger | Anger |
| 5 | Afraid | Afraid |
| 6 | Satisfaction | Suprise |

jednou emocií, i když by například obsahoval stejný počet dokumentů s rozdílnými emocemi.

První problém můžeme eliminovat využitím normalizace, která omezuje vliv dokumentů s emocemi, které se v D_{train} vyskytují častěji. Počet dokumentů v clusteru je vynásoben proměnnou $k(e)$ podle vzorce 6.1.

$$k(e) = \frac{\sum D_{train}}{\sum d \in D_{train,e}} \quad (6.1)$$

V tabulce 6.2 lze vidět, že nyní jsou sice ke clusterům přiřazeny i emocionální kategorie, které byly v korpusu D_{train} zastoupeny méně než ostatní, stále však přetrvává problém v tom, že některé emoce byly kompletně potlačeny.

6.3.2 Procentuální zastoupení

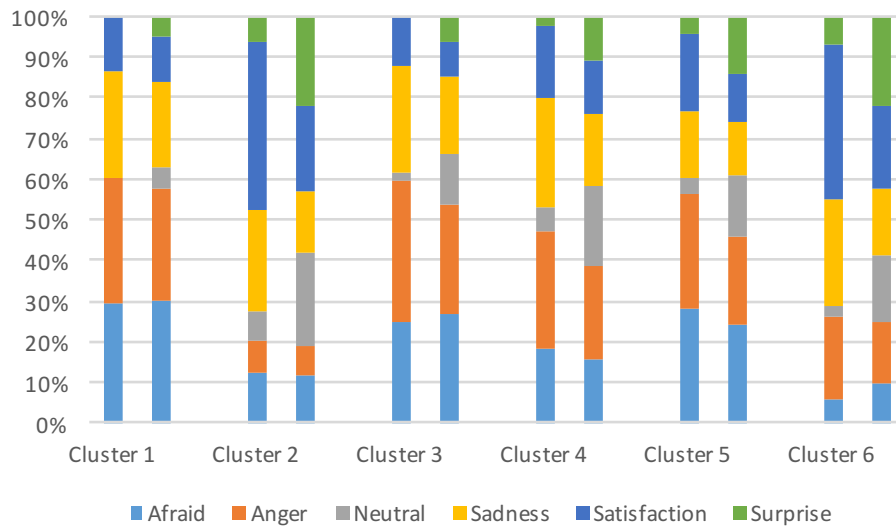
Další z možností je neposuzovat cluster jako jednoznačného zástupce určité kategorie, ale jako procentuální vyjádření reprezenzujících, s jakou intenzitou cluster emoci reprezentuje. Graf 6.3 uvádí použití této metody bez normalizace, obdobně jako u předchozí metody lze však použít i normalizovanou metodu.

6.3.3 Vzdálost od clusteru

Předchozí metody využívaly jen ty dokumenty, jejichž vzdálenost ke clusteru byla co nejmenší. Je možné však brát v úvahu i dokumenty vzdálené více. Vzdálenost dokumentu je definována od 0 do 1, kde 1 značí nejmenší vzdálenost.

$$E_{C_j} = \sum_{d \in D_{train,e}} distance(d, C_j) \quad (6.2)$$

Výsledky metody bez normalizace jsou uvedeny v grafu 6.3.



Obr. 6.3: Procentuální zastoupení a vzdálenost dokumentů v clusteru

6.3.4 Porovnání metod

K porovnání přesnosti výše zmíněných metod byly použity dokumenty z korpusu D_{test} a to tak, že byl ke každému dokumentu nalezen nejbližší cluster a vybrán procentuální podíl emocí clusteru, kterou byl dokument označen.

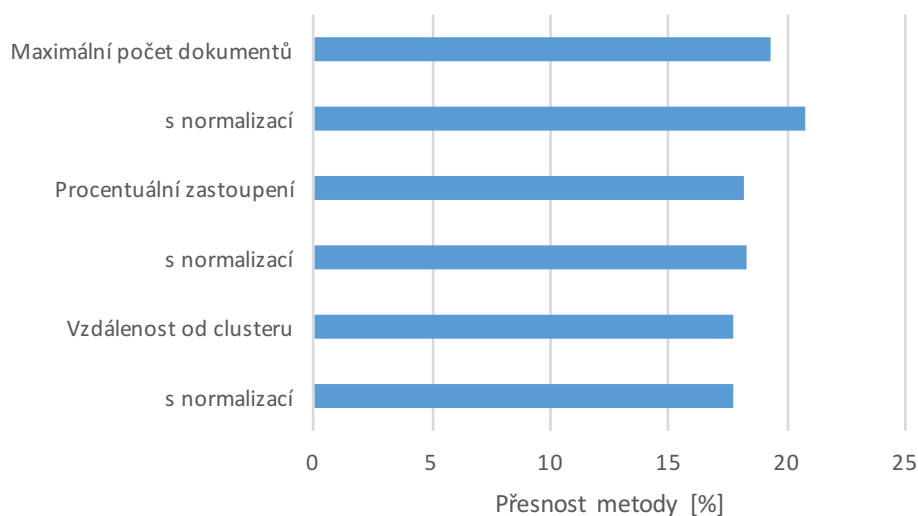
Jelikož jsou dokumenty z D_{test} s různými emocemi rozloženy nerovnoměrně, což by ovlivnilo výsledky měření ve prospěch těch, které jsou zastoupeny častěji, jak v D_{train} , tak v D_{test} (viz tabulka 6.1). Proto byl vytvořen korpus D'_{test} , jenž je tvořen náhodně vybranými 300 dokumenty z korpusu D_{test} tak, že dokumenty s rozdílnými emocemi jsou zastoupeny ve stejné míře.

Výsledky jsou uvedeny v grafu 6.4 a byly získány jako průměr z pěti průběhů algoritmu (i když model LDA byl při každém spuštění stejný, při porovnání se počítá distribuční funkce gama z náhodných čísel, což ovlivňuje výsledky). Z grafu je patrné, že nejlepších výsledků dosahuje metoda, kdy celý cluster považujeme za zástupce jedné emoce.

I tak je přesnost této metody nízká, konkrétně přibližně jen 20 %. Pokud by byla emoce dokumentu určena náhodně, přesnost by byla přibližně 16 %.

6.4 Vliv parametrů na přesnost

Předchozí měření byla provedena pro model s nastavenými parametry tak, jak jsou uvedeny v 6.3. V této kapitole bude provedena změna těchto parametrů a za použití



Obr. 6.4: Přesnost použitých metod

metody pro určení emoce clusteru *Maximální počet dokumentů s normalizací dle počtu dokumentů* určen jejich vliv na přesnost.

6.4.1 Počet clusterů

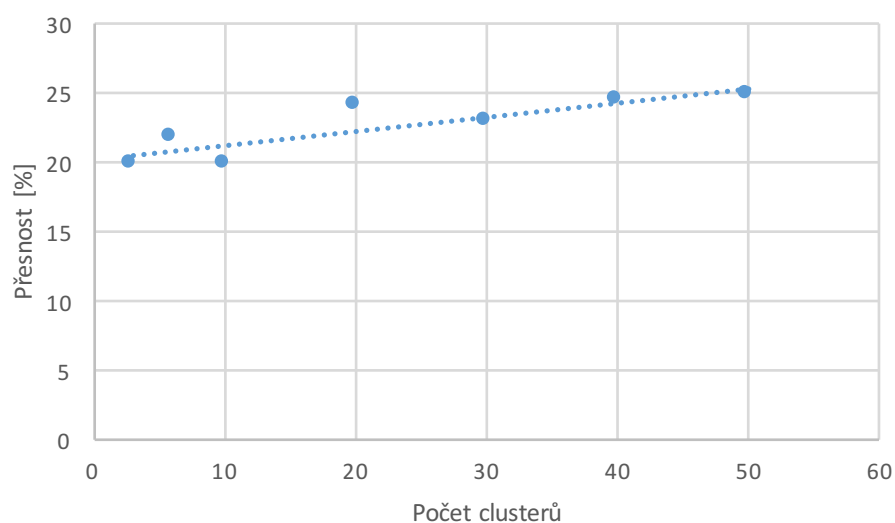
Jak je zřejmé z grafu 6.2, clustery neobsahují dokumenty pouze jedné emoce. Z tabulky 6.1 vyplývá, že pokud se pro učení emoce clusteru používá metoda maximálního počtu dokumentů, některé emoce při ní nejsou zastoupeny vůbec. Proto byl sledován vliv počtu clusterů na přesnost modelu.

Z grafu 6.5 plyne, že přesnost narůstá s rostoucím počtem clusterů, od 20 clusterů se pohybuje kolem 25 %. Vyšší počet clusterů má však vliv na rychlost a paměťovou náročnost.

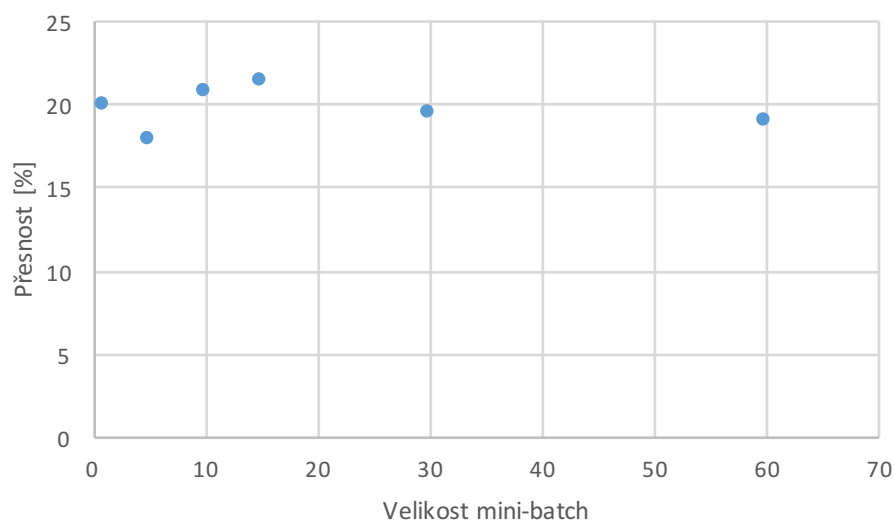
6.4.2 Velikost minidávky

Minidávka určuje počet dokumentů, které jsou zpracovány v jedné dávce bez provedení kroků M a E, které synchronizují parametry modelu minidávky s parametry celého modelu. Zvýšení počtu dokumentů v minidávce tedy vede ke snížení časové náročnosti výpočtu a také k omezení šumu vznikajícího z unikátních slov v dokumentu – přílišné potlačení již ale vede ke snížení přesnosti modelu.

Výsledky měření jsou uvedeny v grafu 6.6. Z něj vyplývá, že nejlepších výsledků bylo dosaženo pro minidávku s 15 dokumenty, s vyšší velikostí přesnost mírně klesala,



Obr. 6.5: Závislost přesnosti modelu na počtu clusterů



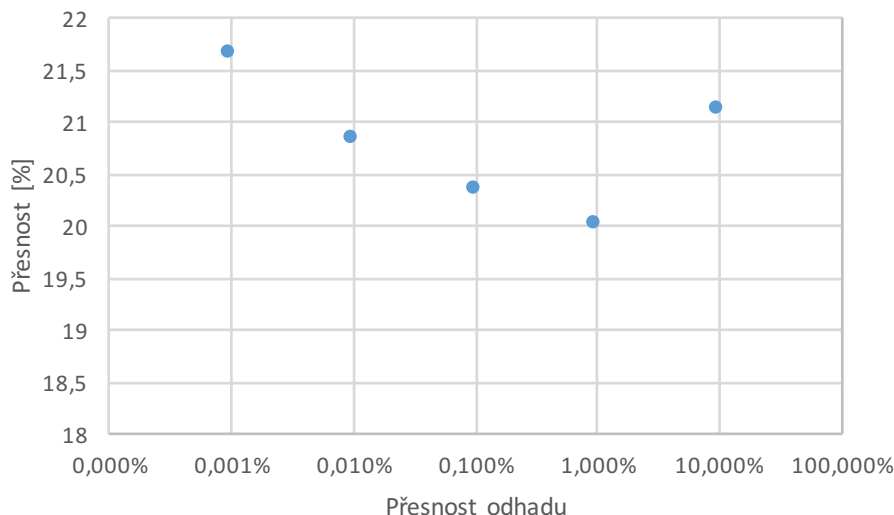
Obr. 6.6: Závislost přesnosti modelu na velikosti minidávky

nejmenších hodnot dosahovala pro 5 dokumentů, naopak pokud byl v minidávce jen jeden dokument, byla přesnost vyšší.

6.4.3 Přesnost odhadu parametrů

Tento parametr určuje, k jaké nejmenší změně musí dojít mezi dvěma iteracemi při zpracování jednoho dokumentu, aby se ve zpracování dokumentu pokračovalo.

Z grafu 6.7 vyplývá, že tento parametr má jen malý vliv na přesnost celého modelu, nejmenší přesnosti dosahuje při hodnotě minimální změny 1 % a narůstá jak pro větší, tak i pro nižší hodnoty.



Obr. 6.7: Závislost přesnosti modelu na přesnosti výpočtu modelu

6.4.4 Optimalizace parametrů

Dle předchozích bodů byly upraveny parametry modelu, aby bylo dosaženo co největší přesnosti. Konkrétně byly nastaveny na:

- Počet clusterů: 50
- Přestat počítat parametry dokumentu při změně menší než 0,0001 %
- Velikost minidávky: 15

Po této změně parametrů došlo ke zvýšení přesnosti odhadu emoce dokumentu o 4 procentní body na 24 %.

6.5 Sloučení trénovacího a testovacího korpusu

Předchozí měření byla provedena tak, že pro natrénování modelu bylo využito jen dokumentů z korpusu D_{train} a dokumenty z D_{test} byly použity jen pro ověření přesnosti. Tento postup je typický pro metody strojového učení s učitelem, neboť dokumenty musí být předem označovány příslušnou emocí. U sémantického modelování však dochází jen ke vzniku clusterů pouze na základě slov, které dokument obsahuje a tedy bez nutnosti dodání informace o jeho emoci.

Při tomto měření byl model LDA vytvořen jak z dokumentů korpusu D_{train} , tak i D_{test} za použití optimalizovaných parametrů. Pro určení emoce clusteru byla využita metoda *Maximální počet dokumentů*.

Tento způsob však nepřinesl měřitelné zlepšení přesnosti.

6.6 Snížení počtu kategorií

V dalším kroku bylo zjišťováno, jak se projeví snížení počtu emocionálních kategorií, a to sloučením pozitivních (surprise, satisfaction), neutrálních a negativních (afraid, anger, sadness). Opět byly použity optimalizované parametry.

V tomto případě se přesnost byla těsně pod 50 %, v případě náhodného výběru kategorie by byla přibližně 33 %.

6.7 Zhodnocení výsledků

Při klasifikaci emocí do šesti kategorií byla při optimalizaci parametrů modelu LDA dosažena maximální přesnost 24 %. Při snížení počtu kategorií na tři byla přesnost určení emoce přibližně 50 %. Tato metoda tedy nedosahuje výsledků, které byly dosaženy například pomocí SVM v ostatních výzkumech, jak bylo uvedeno v kapitole 4. Z nich také vyplývá, že použití samotné metody LDA není ke klasifikaci emocí vhodné a pro zvýšení přesnosti je potřeba použít metodu rozšířenou o vrstvu emocionální klasifikace. Ta však vyžaduje již označené slova příslušnou emocí, podobně jako u slovníkových metod popsanych v kapitole 3.1, ovšem s tou výhodou, že pro určení emoce dokumentu se na rozdíl od slovníkové metody se v něm toto slovo nemusí vyskytovat.

Použití obdobné metody bylo plánováno pro klasifikaci emocí v této práci, pro češtinu ovšem neexistuje takový volně dostupný slovník a jeho vytvoření by bylo nad rámec této práce.

Přesnost by mohla být zvýšena větší velikostí trénovacího korpusu a to nejenom v počtu dokumentů, ale také v počtu slov na jeden dokument, aby se dosáhlo větší shody mezi dokumenty na základě použitých slov.

7 ZÁVĚR

V průběhu této práce bylo teoreticky shrnuto, jaké procesy a postupy je nutné použít pro klasifikaci textů dle emocí pro strojové metody bez učitele. Z těchto metod byla vybrána skupina sémantického modelování obsahující běžně používané algoritmy – LSA, pLSA a LDA. Z těchto algoritmů byla dle dostupné literatury zvolena metoda LDA, která by měla v dosažených výsledcích převyšovat jak metodu LSA, tak i pLSA.

Metoda LDA byla implementována v jazyce Java k volnému použití bez omezení a tato aplikace byla použita pro klasifikaci emocí korpusu 860 česky psaných dokumentů rozřazených do šesti emocionálních kategorií.

Měřením byla určena nejvhodnější metoda, pomocí níž je možné určit emoci, kterou vyjadřuje cluster modelu LDA. Tato metoda byla nazvána jako *Maximální počet dokumentů s normalizací* a označuje celý cluster za zástupce jedné emoce dle počtu dokumentů, které měli od daného clusteru nejmenší vzdálenost.

Za pomoci této metody byly určeny parametry modelu, které vedly k nejlepší přesnosti odhadu emoce neznámého dokumentu. Největší vliv na přesnost mělo nastavení počtu clusterů, do kterých byly dokumenty přiřazeny, a velikost minidávky, tedy počtu dokumentů zpracovaných dohromady v jedné dávce. Naopak parametr přesnosti výpočtu modelu zásadní vliv na přesnost neměl.

Experimentálně bylo vyzkoušeno i sloučení trénovacího a testovací korpusu, což však nemělo na přesnost detekce vliv.

Maximální přesnost, tedy počet dokumentů, u kterých byla správně určena emoční kategorie, dosáhla 24 %. Při snížení počtu emocionálních kategorií přesnost narostla na 50 % a dosahovala výsledků, které byly dosaženy v ostatních výzkumech.

Pro zvýšení přesnosti by bylo vhodné využít předem připravený slovník emocionálně zabarvených slov přiřazených k jedné z emocionálních kategorií.

LITERATURA

- [1] NAKONEČNÝ, Milan. *Emoce*. Vyd. 1. V Praze: Triton, 2012, 501 s. ISBN 9788073876142.
- [2] VYBÍRAL. Přestaňme šířit hlouposti. *Psychologie dnes* [online]. 2002, roč. 8, č. 10 [cit. 2014-12-14]. Dostupné z: <http://psych.fss.muni.cz/vybiral/storage/2002d-Mehrabian.doc>
- [3] DALE, Robert, Hermann MOISL a Harold SOMERS. *Handbook of natural language processing*. New York: Marcel Dekker, 2000, xviii, 943 p. ISBN 08-247-9000-6.
- [4] FELDMAN, Ronen a James SANGER. *The text mining handbook advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press, 2007. ISBN 05-113-3507-5.
- [5] RYCHLÝ, Pavel. CzAccent – Simple Tool for Restoring Accents in Czech Texts. In: HORÁK, Aleš a Pavel RYCHLÝ. *RASLAN 2012: Recent Advances in Slavic Natural Language Processing*. Brno: Tribun EU, 2012, s. 85-89. ISBN 978-80-263-0313-8. Dostupné z: <http://nlp.fi.muni.cz/raslan/raslan12.pdf>
- [6] STROSSA, Petr. *Počítačové zpracování přirozeného jazyka*. Vyd. 1. Praha: Oeconomica, 2011, 314 s. ISBN 9788024517773.
- [7] AGRAWAL, Ameeta a Aijun AN. *Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations*. [online]. [cit. 2014-12-14]. Dostupné z: http://www.cse.yorku.ca/~aan/research/paper/Emo_WI10.pdf
- [8] SOWMIYA, J.S. a S. CHANDRAKALA. Joint Sentiment/Topic extraction from text. *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*. IEEE, 2014, s. 611-615. DOI: 10.1109/ICACCCT.2014.7019160. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7019160>
- [9] WAILA, P., MARISHA, V. K. SINGH a M. K. SINGH. Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews. *2012 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2012, s. 1-6. DOI: 10.1109/ICCIC.2012.6510235. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6510235>

- [10] DING, Wanying, Xiaoli SONG, Lifan GUO, Zunyan XIONG a Xiaohua HU. A Novel Hybrid HDP-LDA Model for Sentiment Analysis. *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. IEEE, 2013, s. 329-336. DOI: 10.1109/WI-IAT.2013.47. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6690033>
- [11] TURNEY, Peter. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* [online]. Philadelphia, 2002 [cit. 2014-12-14]. Dostupné z: <http://www.aclweb.org/anthology/P02-1053.pdf>
- [12] PANG, Bo, Lillian LEE a Shivakumar VAITHYANATHAN. Thumbs up? Sentiment Classification using Machine Learning Techniques. In: *Proceedings of EMNLP* [online]. 2002 [cit. 2014-12-16]. Dostupné z: <http://www.cs.cornell.edu/home/lllee/papers/sentiment.pdf>
- [13] DRUSA, Tomáš. *Detekce nepřirozených odkazů ve webových stránkách* [online]. Brno, 2014 [cit. 2014-12-14]. Dostupné z: http://is.muni.cz/th/256167/fi_m/. Diplomová práce, Fakulta informatiky. Masarykova univerzita. Vedoucí práce Jiří Materna.
- [14] MANNING, Christopher, Prabhakar RAGHAVAN a Hinrich SCHÜTZE. *Introduction to information retrieval* [online]. Cambridge [etc.]: Cambridge University Press, 2008, s. 117-119 [cit. 2014-12-14]. ISBN 9780511809071.
- [15] VEJVODA, Jiří. *Seskupování zpravodajských článků* [online]. Brno, 2014 [cit. 2014-12-14]. Dostupné z: https://is.muni.cz/th/256128/fi_m/. Diplomová práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Petr Sojka.
- [16] BAKER, Kirk. *Singular Value Decomposition Tutorial*. [online]. 2005, January 14, 2013 [cit. 2014-12-14]. Dostupné z: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf
- [17] MATERNA, Jiří. *Probabilistic Semantic Frames* [online]. 2014 [cit. 2014-12-14]. Dostupné z: http://is.muni.cz/th/98897/fi_d/. Disertační práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Karel Pala.
- [18] MATERNA, Jiří. Sémantická analýza textů (5). In: *Blog fulltextového týmu* [online]. 2011 [cit. 2014-12-14]. Dostupné z: <http://fulltext.sblog.cz/2011/12/04/semanticka-analyza-textu-5/>

- [19] MATERNA, Jiří. Sémantická analýza textů (6). In: *Blog fulltextového týmu* [online]. 2011 [cit. 2014-12-15]. Dostupné z: <http://fulltext.sblog.cz/2011/12/22/semanticka-analyza-textu-6/>
- [20] HOFFMAN, BLEI a BACH. *Online Learning for Latent Dirichlet Allocation* [online]. New York, 2010. Dostupné z: <http://www.cs.princeton.edu/~blei/papers/HoffmanBleiBach2010b.pdf>
- [21] ŘEHŮŘEK, Radim a Petr SOJKA. *Software Framework for Topic Modeling with Large Corpora*. [online] In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta, 2010. p. 46–50, 5 pp. ISBN 2-9517408-6-7. Dostupné z: <http://is.muni.cz/repo/884893/lrec2010-rehurek-sojka.pdf>

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

| | |
|------|-----------------------------------|
| A | matice slov a dokumentů v korpusu |
| d | dokument v korpusu |
| D | korpus |
| K | počet témat |
| M | počet unikátních slov v korpusu |
| N | délka korpusu |
| LDA | Latentní Dirichletova alokace |
| LSA | Latentní sémantická analýza |
| pLSA | Pravděpodobnostní LSA |
| t | slovo v dokumentu |
| z | téma |

SEZNAM PŘÍLOH

| | |
|----------------------------|----|
| A Clustery | 47 |
| B Ukázkový výstup programu | 48 |
| C Obsah přiloženého CD | 49 |

A CLUSTERY

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
|----------------------|--------------------|-------------------|-------------------|--------------------|-------------------|
| 0,84 %*klasičnost | 1,90 %*postrádat | 0,39 %*světelnost | 0,99 %*paráda | 1,06 %*legrace | 2,92 %*naléhavě |
| 0,84 %*docílený | 1,44 %*obdivuhodný | 0,39 %*cenový | 0,62 %*dílo | 0,67 %*doporučovat | 2,92 %*žádat |
| 0,84 %*nekřiklavost | 1,23 %*dobrý | 0,39 %*objektiva | 0,56 %*milý | 0,53 %*page | 2,65 %*kurevsky |
| 0,75 %*sofistikovaný | 1,14 %*atraktivní | 0,39 %*pořad | 0,54 %*abnormální | 0,42 %*vřele | 2,55 %*všudy |
| 0,65 %*negativní | 1,00 %*hodně | 0,39 %*krásna | 0,46 %*použití | 0,40 %*ago | 2,53 %*udivit |
| 0,57 %*nenajít | 0,96 %*dokonalý | 0,39 %*chandra | 0,41 %*pozitivní | 0,38 %*fascinovat | 2,47 %*slovo |
| 0,57 %*pokojit | 0,85 %*film | 0,39 %*sasanka | 0,33 %*dotaz | 0,36 %*kultovní | 2,45 %*pecko |
| 0,57 %*splňovat | 0,64 %*líbit | 0,39 %*poetický | 0,32 %*pochval | 0,35 %*primer | 2,05 %*léto |
| 0,57 %*nárok | 0,63 %*super | 0,36 %*proces | 0,31 %*nesnáz | 0,35 %*otec | 0,73 %*krásně |
| 0,55 %*spokojenost | 0,62 %*moci | 0,30 %*dosah | 0,28 %*odezva | 0,35 %*lba | 0,70 %*osvěžující |

B UKÁZKOVÝ VÝSTUP PROGRAMU

Loaded 521 documents from training corpus

Number of unique tokens: 3455

Number of average unique tokens per document: 15.289827255278311

Topic 1 with 68 documents, 17,00x afraid, 21,00x anger, 0,00x neutral, 21,00x sadness, 9,00x satisfaction, 0,00x surprise

By count: 0,25x afraid, 0,31x anger, 0,00x neutral, 0,31x sadness, 0,13x satisfaction, 0,00x surprise

By gamma: 0,29x afraid, 0,28x anger, 0,04x neutral, 0,24x sadness, 0,10x satisfaction, 0,05x surprise

Topic 2 with 54 documents, 17,00x afraid, 15,00x anger, 2,00x neutral, 9,00x sadness, 10,00x satisfaction, 1,00x surprise

By count: 0,31x afraid, 0,28x anger, 0,04x neutral, 0,17x sadness, 0,19x satisfaction, 0,02x surprise

By gamma: 0,28x afraid, 0,24x anger, 0,14x neutral, 0,13x sadness, 0,13x satisfaction, 0,09x surprise

Topic 3 with 47 documents, 8,00x afraid, 17,00x anger, 1,00x neutral, 8,00x sadness, 11,00x satisfaction, 2,00x surprise

By count: 0,17x afraid, 0,36x anger, 0,02x neutral, 0,17x sadness, 0,23x satisfaction, 0,04x surprise

By gamma: 0,16x afraid, 0,26x anger, 0,14x neutral, 0,12x sadness, 0,14x satisfaction, 0,18x surprise

Topic 4 with 64 documents, 11,00x afraid, 9,00x anger, 6,00x neutral, 19,00x sadness, 19,00x satisfaction, 0,00x surprise

By count: 0,17x afraid, 0,14x anger, 0,09x neutral, 0,30x sadness, 0,30x satisfaction, 0,00x surprise

By gamma: 0,16x afraid, 0,14x anger, 0,26x neutral, 0,20x sadness, 0,18x satisfaction, 0,07x surprise

Topic 5 with 240 documents, 26,00x afraid, 33,00x anger, 12,00x neutral, 57,00x sadness, 94,00x satisfaction, 18,00x surprise

By count: 0,11x afraid, 0,14x anger, 0,05x neutral, 0,24x sadness, 0,39x satisfaction, 0,07x surprise

By gamma: 0,11x afraid, 0,10x anger, 0,20x neutral, 0,15x sadness, 0,20x satisfaction, 0,25x surprise

Topic 6 with 48 documents, 10,00x afraid, 15,00x anger, 1,00x neutral, 15,00x sadness, 6,00x satisfaction, 1,00x surprise

By count: 0,21x afraid, 0,31x anger, 0,02x neutral, 0,31x sadness, 0,12x satisfaction, 0,02x surprise

By gamma: 0,21x afraid, 0,25x anger, 0,09x neutral, 0,22x sadness, 0,11x satisfaction, 0,11x surprise

Loaded 291 from test corpus

Number of orphan tokens: 2028

Overall quality: 19.93127147766323

C OBSAH PŘILOŽENÉHO CD

Přiložené CD obsahuje následující adresáře a soubory:

- `Diplomová práce.pdf` – elektronická podoba této práce
- `Latex` – adresář se zdrojovými kódy pro sázeací program \LaTeX
- `Aplikace` – zdrojové kódy aplikace včetně použitého korpusu