Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Automatická sumarizace názorů

Plzeň 2014 Bc. Petr Zápotocký

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 14. května 2014

Bc. Petr Zápotocký

Abstract

This work is focused on usage and development of automatic methods for opinion summarisation. Opinion summarisation could be used as a helpful tool for processing larger quantities of input data from electronic media. The first part of this work covers theoretical knowledge, including summarisation algorithms.The second part is an implementation and experimentation with automatic opinion summarisation. At the end, the reader should be able to construct own software for automatic opinion summarisation.

Abstrakt

Tato práce je zaměřena na použití a vývoj automatické sumarizace názorů. Sumarizace názorů může být použita jako pomocný nástroj pro zpracování velkého množství vstupních dat z elektronických médií. První část této práce je zaměřena na teoretické znalosti zahrnující algoritmy sumarizace. Druhá část obsahuje implementaci a experimentální část automatické sumarizace názorů. Po přečtení by měl být čtenář schopen naimplementovat vlastní software pro automatickou sumarizaci názorů.

Obsah

1 Úvod

V dnešní technologicky vyspělé době zaměřené z hlediska médií zejména na jejich elektronickou podobu je čím dál častější potřeba pracovat s textem na vyšší úrovni. Nekončíme tedy jen u pouhého čtení, ale například z hlediska efektivity potřebujeme v textu vyhledávat souvislosti, třídit důležité informace nebo vyhodnocovat stručné závěry. Do tohoto odvětví také v posledních pár letech velice důrazně promlouvají sociální sítě, z jejichž dat se dá vypozorovat opravdu velké množství důležitých informací.

S takovými daty tedy můžeme nakládat opravdu mnoha způsoby a můžeme je využívat za různým účelem. Tato práce je však zaměřená principiálně na sumarizaci názorů. Znamená to, že data z výše zmíněných zdrojů jsou zpracovávána tak, abychom jednoduchým způsobem zjistili hlavní myšlenky a názory. Standardně byste museli fyzicky pročíst veškerá zdrojová data tak, abyste vyvodili vlastní závěry. To však může být u rozsáhlých dat opravdu náročná metoda. Tato práce je tak zaměřena na způsob, jakým tuto sumarizaci můžeme provádět automaticky za použití speciálních algoritmů.

Druhým cílem této práce je implementace automatického sumarizátoru s následným testováním, experimentováním a samozřejmě analýzou dosažených výsledků, porovnáváním s reálně vytvořenými závěry. V tomto případě je automatický sumarizátor názorů zaměřen na zpracování zdrojových dat ze sociální sítě Facebook. Není však problém upravit software tak, aby byl schopen analyzovat data z jiných sociálních sítí nebo elektronických médií.

Po přečtení této práce by měl mít čtenář přehled o automatické sumarizaci názorů, jejím využití a také o její efektivnosti v daném algoritmickém zpracování. Dále by měl být schopen upravit sumarizátor pro analýzu dat z jiných zdrojů a vylepšovat celkovou efektivnost algoritmu. Přiložena je také demo aplikace pro kompletní dokreslení využití v praxi.

1.1 Zdroje pro automatickou sumarizaci

Jak již bylo výše zmíněno, sumarizátor by měl být schopen zpracovat jakákoliv textová data s jakýmkoliv objektivním názorem. Navíc již bylo zmíněno, že sumarizace je vhodná zejména pro odvětví, kde mohou výsledky poskytnout

kromě pouhé informace i něco navíc. Obchod, reklama, marketing všeobecně. To jsou přesně ta odvětví, kde se mohou výsledky využít s největším užitkem. Díky správné a rychlé analýze určitých zdrojových dat může firma flexibilně reagovat na jakékoliv problémy, které se k ní od nespokojených klientů nebo zákazníků dostanou. Dostáváme se tak do úrovně zásadních odlišujících faktorů úspěchu a neúspěchu firem.

Mezi hlavní zdroje těchto dat počítáme tedy taková místa, kde větší množství lidí ventiluje svůj osobní názor a tím tak dává najevo svůj postoj k určité situaci, službě nebo výrobku. Patří mězi ně zejména diskuze nebo sociální sítě.

Facebook je v našem okolí momentálně nejrozšířenější sociální síť, která se pro automatickou sumarizaci názorů hodí opravdu hodně. Její největší výhodou jsou tzv. facebookové stránky různých firem, společností nebo výrobků, pod kterými mohou dané subjekty přímo vystupovat a tvořit svojí vlastní sociální identitu. Díky vlastnostem velice rozšířené sociální sítě se v ideálním případě může jednat o velice efektivní nástroj pro komunikaci s klienty nebo zákazníky. Z tohoto důvodu se tak takovéto facebookové stránky stávají přímo ideálním zdrojem pro automatickou sumarizaci názorů, jelikož se zde v ideálním případě může nashromáždit opravdu veliké množství příspěvků, reakcí a názorů na výrobek, službu, společnost nebo dění kolem ní. Díky dobrému automatickému sumarizátoru by tak mohly tyto společnosti analyzovat vlastní přednosti nebo naopak nedostatky.

Twitter je sociální síť zaměřená zejména na reakce a glosování. V ČR zatím ne tak rozšířená jako například v USA, ale její obliba neustále roste. Proces automatické sumarizace by se zde dal využít například pro zpracování příspěvků s daným hashtagem, tedy s určitým atributem vyjadřující dané téma. Dají se tak opět analyzovat data, která jsou přidružena k nějaké značce výrobku nebo společnosti.

Sociální síť **Google+** vznikla jako reakce společnosti Google na vysoce se rozšiřující trend sociálních sítí. Má za úkol stejně jako ostatní sociální sítě podporovat zejména komunikaci a společenskou interakci. Jedná se o nejmladší sociální síť z výše zmíněných a je proto také zatím mezi uživateli nejméně rozšířená, ačkoliv tomu tak úplně říkat nemůžeme, jelikož díky rozšířenému využívání emailové služby Gmail společnosti Google je každý takovýto uživatel automaticky vlastníkem profilu v sociální síti Google+.

Obecně jakékoliv **diskuzní fórum** se stává ideálním předmětem pro ana-

lýzu pomocí automatického sumarizátoru názorů, jelikož přesně proto jsou diskuzní fóra využívána. Stačí jen zvolit vlákno nebo soubor vláken k tématu, které chceme analyzovat a máme také velice hodnotná zdrojová data.

Analyzovat se však dá téměř jakýkoliv text, jehož sumarizace by nás potenciálně mohla zajímat.

Blogy jsou v původní podstatě jakési deníčky, kam mohou uživatelé psát jakékoliv svoje myšlenky nebo zážitky, které se postupem času přetransformovali také na pomocný prostředek větších firem k informování kolem dění ve společnosti nebo kolem novinek z daného odvětví, které nemusí ani přímo s činností firmy souviset. Výhodou těchto blogů je také to, že se mohou v mnoha případech komentovat jednotlivé články a vzniká tak další prostor pro názorovou analýzu a následnou sumarizaci.

Recenze je další specifický typ textu, jež vyjadřuje vlastní názor autora na určité dílo, výrobek nebo službu. Jedná se zde opravdu o subjektivní pocity a názory a je tak možnost, že bychom tato data chtěli nějakým způsobem analyzovat, zejména pak v případě, že vznikne na daný výrobek nebo dílo recenzí desítky až stovky a my nejsme schopni z nich za rozumné náklady vyvodit reprezentativní výsledky.

Ačkoliv by se v rámci **zpravodajství** mělo jednat o objektivní a podložené zprávy, zkušenosti nám ukazují, že tomu tak být nemusí, ba naopak. Je to tak další z mnoha možností, kdy využívat automatickou sumarizaci názorů a to zejména, pokud se zpráv o dané věci sejde vícero z různých zdrojů.

2 Předzpracování textu

Při jakékoliv analýze textu, ať už českého nebo cizojazyčného, musíme brát v úvahu, že jakýkoliv text, který použijeme, nemusí být pro danou analýzu zcela ideální a měli bychom se tak zamyslet nad jeho předzpracováním, tedy přípravě textu k dané analýze tak, abychom dosáhli kvalitních výsledků. Příkladem uveďme nástroj Although Ark-tweetnlp tool [Gimpel et al.(2011)], jehož významnou vlastností je zaměření na emotikony a další speciální sekvence symbolů. Ačkoliv byl tento nástroj vyvinut a optimalizován pro anglický jazyk, [Habernal et al.(2013)] uvádí, že podle testování na českém jazyce je tento postup úspěšný i tam. Mezi další široce používané metody, které zapříčiní zmenšení slovníku jsou například lematizace nebo odstranění stopwords (tzv. stopslova, což jsou slova, která žádným způsobem nevyjadřují subjektivitu).

2.1 Vlastnosti českého jazyka

Pokud se v rámci této práce chceme zaměřit na automatickou sumarizaci názorů zejména v českém jazyce, měli bychom mít základní znalosti o tomto jazyku a o tom, jak která slova vznikají nebo jak s nimi nakládat. Základem je tedy znalost slovotvorby, což je podle [Marek Nekula(2003)] odvětví zabývající se tvorbou jednoslovných pojmenování a to buď na základě pojmenování názvů již existujících nebo názvů odvozených, respektive jejich vznikem. Obecně můžeme říci, že slovotvorba se dělí na dvě metody, derivaci (odvozování nových slov od již existujících) a kompozici (skládání nových slov ze slov již existujících). Mezi základní informace o českém jazyce, který je flektivní (můžeme slova časovat a skloňovat) bychom měli zařadit také definované části českých slov.

- **Kořen** - tvoří základ každého slova a je dále nedělitelný. Kořen určuje základní význam slova a díky předponám, příponám a koncovkám je dále rozvíjen, čímž vznikají nové a nové tvary slov [Marek Nekula(2002)].
- **Předpona** - v [Marek Nekula(2002)] je uvedeno, že předpona (prefix) je rozšíření základového slova, je připojena před toto slovo a mění jeho význam, ale slovní druh zůstává nepozměněn.

- **Přípona** - neboli sufix je opakem předpony a díky sufixaci (metoda tvořící nová slova pomocí přidávání přípon [Marek Nekula(2002)]), vznikají podobně jako u předpon slova nových významů, ale v tomhle případě dochází navíc také ke změně slovního druhu.
- **Koncovka** - podle [Marek Nekula(2002)] může být součástí přípony (na jejím konci), ale může být však napojena přímo na slovní základ.

2.2 Základní vlastnosti příspěvků

Pokud se v rámci této diplomové práce zaměříme přednostně na analýzu příspěvků z nějaké facebookové stránky, máme ideální zdrojová data pro určování sentimentu. Ačkoliv se může zdát, že se jedná o zcela ideální případ pro zkoumání sentimentu s následnou sumarizací, měli bychom se zaměřit na určité aspekty, které mohou klasifikaci sentimentu zkreslit.

- Předně musíme říci, že výsledný sentiment daného příspěvku nemusí být zcela totožný se sentimentem každého určitého slova z příspěvku. Jinak řečeno například slova s kladným sentimentem mohou být obsažena v příspěvku s výsledným záporným sentimentem nebo naopak.
- Jako u většiny textů se i ve většině příspěvků budou vyskytovat nezvýznamová slova, tzv. stopwords, která mohou analýzu zkreslovat.
- Mezi velké překážky ve správně klasifikaci sentimentu patří například ironie. Příspěvek s ironickým podtextem může být klasifikován zcela opačně.
- Mezi ostatní vlastnosti příspěvků na sociálních sítí patří primárně lidský faktor. Kvůli tomu se v textu vyskytují různé překlepy, přesmyčky nebo dokonce pravopisné chyby, které nelze předem nijak predikovat a vzniká tak další problém se správnou klasifikací.

V ideálním případě bychom se měli všech těchto výše zmíněných problémů zbavit a odstranit je, což však může být opravdu složité.

2.3 Stopwords

Jako stopwords označujeme slova, která nám do oblasti zkoumání sentimentu nepřináší žádný význam a jsou tudíž zbytečně rozšiřujícím faktorem zdrojových dat a tím i jeho zbytečného zkreslení. Je tedy velice žádoucí, abychom se takovýchto před samotnou klasifikací zbavili a celý proces tak značně zjednodušili. Při detailnějším zkoumání lze totiž vypozorovat, že se právě zmíněné stopwords vyskytují ve většině textech nefrekventovaněji a při zkoumání sentimentu i se stopwords by tak docházelo opravdu ke zbytečnému a rozsáhlému ovlivňování klasifikace. Z pohledu slovních druhů můžeme říci, že se jedná zejména o spojky nebo zájmena a jak je známo, tato slova žádný význam z hlediska sentimentu nenesou.

Jako referenční a zároveň reálný seznam stopwords byl použit seznam poskytnutý Ing. Josefem Steinbergerem, Ph.D., obsahující přes 800 těchto sentimentově nevýznamných výrazů. Kompletní seznam těchto slov naleznete v příloze A.

2.4 Lematizace

Pomocí lematizace dostáváme základní tvary slova daného významu, který se však v textu může vyskytovat v různých pádech nebo časech. Příkladem buď například sloveso *zaváhat*, které se po procesu lematizace ustálí ve tvaru *váhat*, tedy v infinitivu původního slovesa. Přesně takovýmto způsobem dojde ke sjednocení slov stejného významu, avšak rozdílných tvarů. Potenciálně tak zužujeme seznam slov použitých v analyzovaném textu.

2.5 POS tagging

Metoda zjišťující slovní druh a tvar analyzovaného slova pojmenovaná podle anglického významu part of speech (POS). Tento algoritmus generuje ke zkoumaným slovům pravidly definované řetězce a tím tak poznáme původní slovní druh, případně tvar. Díky těmto řetězcům máme k lematizovaným slovům původní informace a můžeme je v případě potřeby využít. Jedná se tak například o zjištění příslovcí určujících míru, které se po lematizaci zkrátily na příslušný základ. Při zdokonalování klasifikace sentimentu tak můžeme

využít i těchto informací a výsledky analýzy zpřesnit.

2.6 N-Gramy

Tato konstrukce slouží zejména pro určení sentimentu tam, kde bychom ho pouze za pomoci zkoumání jednotlivých slov nikdy neodhalili. N-gram nám znázorňuje dvě a více slov, jdoucí po sobě v přesně daném pořadí, čímž můžeme definovat sentiment i u slovních spojení, jejichž samostatná slova sama o sobě žádný sentiment neobsahují. Konkrétně v našem případě následného praktického využití můžeme nejvíce používaná slovní spojení (například bigramy obsahující dvě slova) s daným sentimentem přidat do používaných slovníků a zvýšit tak přesnost výsledků analýzy.

3 Analýza názorů

V analýze názorů je cílem rozpoznat a klasifikovat subjektivní obsah textu. Rozpoznáváme tedy subjektivní názory, jejichž analýza nám může pomoci k rychlému pochopení zásadních myšlenek autora/autorů. Zdrojová data mohou být klasifikována jako celek a z hodnocení se tak můžeme dozvědět celkový názor na výrobek nebo službu. Například v rámci zpravodajství pak můžeme klasifikovat jednotlivé zprávy a rychlým způsobem je rozlišit na dobré a špatné. Vše se ovšem odvíjí od subjektivních názorů autora textu a nemusí to tak jednoznačně znamenat, že obecně dobrá zpráva bude popsána v superlativech. Za určitých okolností pak může nastat situace, že ve vašich očích dobrá zpráva bude sentimentově vyhodnocena jako špatná nebo naopak. Takovéto problémy mohou vznikat především u názorů na konkrétní osoby, celebrity nebo politické dění. Je proto nutné zaměřit se na celkový kontext zprávy, což kompletní klasifikaci zdrojových dat často velice znesnadňuje. Pokud se chceme těmto problémům vyhnout a dosáhnout opravdu hodnotného výsledku, měli bychom na danou situaci sehnat co nejvíce zdrojů v ideálním případě od různých autorů, z různých zemí nebo v různých jazycích. Poté bychom měli mít mnohem větší jistotu, že výsledky klasifikace nebudou zkreslené a budou mít dostatečnou vypovídající hodnotu.

Co se týče zdrojů v různých jazycích, je nutné přiznat, že není zcela snadné vytvořit analyzátor pro více jazyků, který by navíc pracoval nad obrovským množstvím zdrojových textů.

Problém také nastává v případě diverzifikace zdrojových textů. Ve výsledku můžeme prohlásit za použitelné pouze recenze a to ještě ve většině případů pouze v angličtině. Sentimentové slovníky jsou také ve většině případů pouze anglické. V případě ostatních jazyků nemůžeme slovníky považovat za adekvátní slovníkům anglickým, jelikož nejsou vytvářeny pro zcela stejné účely podle naprosto stejných pravidel.

3.1 Slovníkové metody

Jedná se o jeden ze dvou základních přístupů, jak sentimentově analyzovat zdrojový text [Taboada et al.(2011)]. K určování sentimentu se využívá ručně nebo automaticky vytvořených slovníků, na jejichž bázi dochází k dané ana-

lýze a to tak, že se prochází slovo od slova a určuje se celkové sentimetnové skóre na základě výskytu daného slova ve slovníku. Velkou část každého slovníku pro analýzu sentimentu zahrnují přídavná jména, jelikož právě ta jsou ve velké části nositeli sentimentu.

3.1.1 Slovníky

Jak již bylo napsáno, slovníky pro slovníkové metody mohou být vytvořenu buď ručně nebo automaticky pomocí základních slov a jejich dalším rozšiřováním tak, že každé nové přídavné jméno je dopočítáno podle frekvence přidružení ke stávajícím základním slovům. V principu je totiž na věc nahlíženo tak, že kladná přídavná jména se častěji objevují v blízkosti kladných základních slov a opačně. Velká část výzkumu slovníků a slovníkových metod se zaměřuje na použití přídavných jmen jako indikátorů sémantické orientace textu. Takovéto slovníky tedy tvoří seznam přídavných jmen s určením jejich sémantické orientace, tedy s informací, zda se jedná o přídavné jméno kladné nebo záporné.

[Taboada et al.(2011)] však zmiňuje, že nejen přídavná jména jsou nositeli sémantické informace a proto důležité zaměřit se také na slovníky podstatných jmen, sloves nebo příslovcí. Konkrétně v tomto experimentu byla pak jednotlivým slovům přidána váha od -5 do 5 bodů, která vyznačuje intenzitu a polaritu slova. Slovníky podstatných jmen a sloves byly tvořeny ručně a slovník příslovcí byl vytvořen ze slovníku přídavných jmen s tím, že daným příslovcím byla ponechána váha jeho vzorového přídavného jména. Samozřejmě existuje několik zvláštních případů, kdy je do slovníku přidáno příslovce ručně tak, aby byl výsledný slovník co nejkvalitnější.

Další možnou součástí systému slovníku jsou slova, která zvyšují (např.: ”velmi”) nebo snižují (např.: ”mírně”) intenzitu slov vedle sebe. Problémem je však různá míra ovlivnění jednotlivých slov a navíc konkrétní změna je také závislá na slově u kterého intenzita mění. Mezi slovní druhy těchto zesilujících slov patří většinou příslovce nebo přídavná jména. Není však na škodu do slovníků zařadit také další informace, jako například používání velkých písmen, které lidi používají pro větší zdůraznění nebo používání vykřičníků.

3.1.2 Negace

Jako důležitou vlastnost příspěvků, respektive názorů zmiňme negaci. Tato vlastnost totiž může z potenciálně kladného příspěvku vytvořit názor zcela opačný, tedy záporný. Zejména v anglickém jazyce je pak velice důležité tyto záporné výrazy, které negují jiná slova vyhledávat. Často se však stává, že tyto negace nejsou přímo spojené s daným slovem, které ovlivňují a je tak často nutná rozsáhlejší analýza k nalezení těchto návazností. V experimentu [Taboada et al.(2011)] byly použity dvě metody. Hlavním problémem je však negace polarity daného výrazu, jelikož není zcela správný přístup otáčet skóre jen pomocí znaménka. Mnohem je lepší je zavedení principu posunu negace, což v základu znamená posun skóre o určitou hodnotu. Ze skóre 5 nám tak nevznikne -5, ale například 1.

3.2 Algoritmy strojového učení

Strojové učení můžeme správně z určitého pohledu považovat za podobné tomu lidskému. Klasifikátor založený na strojovém učení totiž funguje tak, že se nejprve naučí danou věc (analýzu sentimentu) na správně označkovaném datasetu a následně použije tyto znalosti k roztřídění zkoumaných dat do daných kategorií.

3.2.1 Support Vector Machines

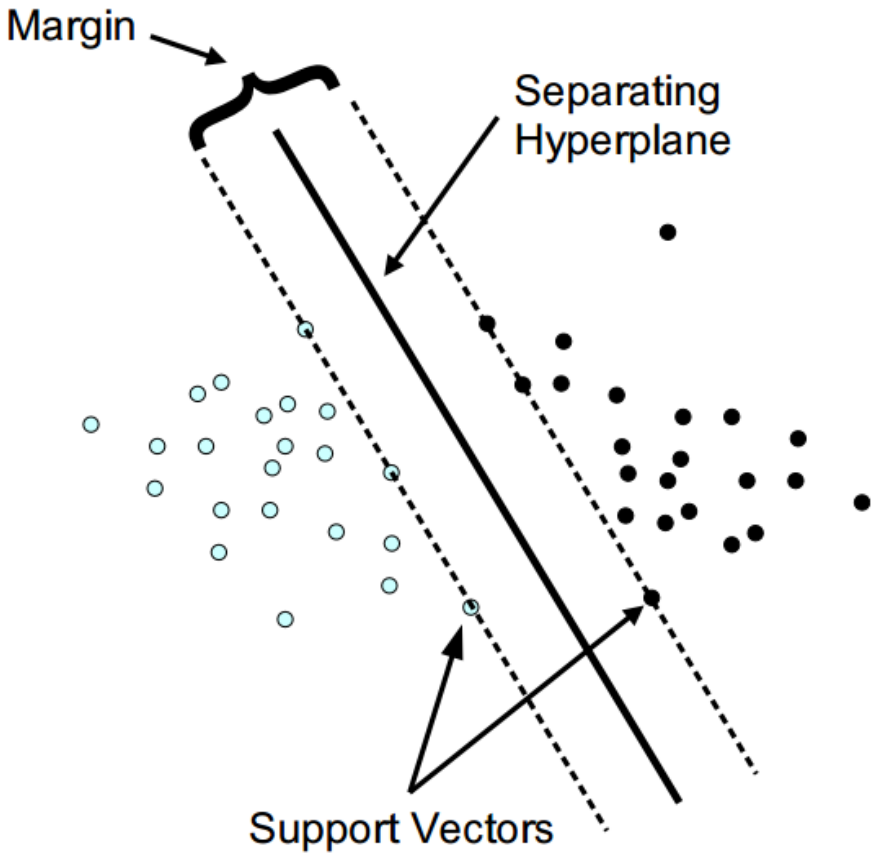
Přístup původně založený na binární klasifikaci, tedy rozděluje analyzovaná data do dvou tříd [Corinna et al.(1995)]. Jak nastiňuje [Meyer(2001)] můžeme tento přístup popsat několika základními aspekty:

- **Rozdělení tříd** - v podstatě hledáme optimální dělící nadrovinu mezi dvěma třídami s maximálním rozpětím nejbližších bodů *(Obr. 3.1)*.
- **Překrývající se třídy** - příkladové body vyskytující se na špatné straně nadroviny jsou váhově omezovány tak, aby došlo ke snížení jejich vlivu.
- **Nelineárnost** - pokud nastane problém s lineární separovatelností dané úlohy, pomocí jádrové transformace (kernel transformation) do-

chází k transformaci dat do vyšší dimenze, kde již není problém danou úlohu separovat lineárně.

- **Řešení problému** - celá úloha může být formulována jako problém kvadratické optimalizace, který může být známými technikami vyřešen.

Program schopný provádět všechny zmíněné body se nazývá Support Vector Machine.



Obrázek 3.1: Dělící nadrovina mezi dvěma třídami (separating hyperlane) obklopena dvěma vektory (support vectors), které jsou tvořeny příklady nejblíže nadrovině. Vidět je také maximální vzdálenost vektorů (margin)

3.2.2 Naive Bayes

Jedná se o jeden z nejznámějších algoritmů strojového učení založený na principu nezávislého výskytu slov v dokumentu v závislosti na bayesovskému teorému. Znamená to, že se jedná o metodu předpokládající podmíněnou nezávislost třídy. Jednotlivá slova dokumentu D jsou postupně analyzována a tříděna do určených tříd C. I když je tento přístup velice jednoduchý, jeho účinnost je vcelku dobrá, ale zásadní výhodou je jeho malá výpočetní složitost [Zhang(2004)].

Bayesův teorém

Berme v úvahu, že X je případ u kterého neznáme zařazení do třídy C. Poté mějme hypotézu H, že patří do určité třídy C_i. Účelem je tedy zjistit podmíněnou pravděpodobnost $P(X|H)$. Na tomto základě a znalostech pravděpodobností $P(H)$ a P(X) můžeme podle Bayesova teorému vypočítat požadovanou pravděpodobnost $P(H|X)$ (3.1):

$$(P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{3.1}$$

Naivní Bayesův klasifikátor

Na začátku musí dojít k naučení, tedy musíme mít trénovací množinu D, ze které se klasifikátor naučí zařazování do správných tříd. Každý příklad je totiž v trénovací množině přiřazen do správné třídy a popsán n-rozměrným vektorem atributů. Nyní můžeme předpokládat, že máme m klasifikačních tříd C a chceme každý příklad přiřadit do třídy s maximální pravděpodobností (3.2):

$$P(C_i|X) > P(C_j|X) \tag{3.2}$$

Třída C_i pro kterou je pravděpodobnost v určené třídě maximální, je podle Bayesova teorému (3.3):

$$P(C_i|X) = \frac{P(X|C_i)P(Ci)}{P(X)} \tag{3.3}$$

Zjednodušeně můžeme říci, že příklad X je zařazen do třídy, pro kterou je pravděpodobnost $P(X|C_i)P(C_i)$ maximální, jelikož pravděpodobnost $P(X)$ je konstantní pro všechny případy [Manning et al.(2008)].

3.2.3 Modelování maximální entropií

Maximum entropy modeling, česky řečeno modelování s maximální entropií, je způsob, pomocí kterého se dají klasifikovat informace z různorodých zdrojů. Problém s touto klasifikací je založen na potenciálně velkém počtu příznaků, které mohou být celkem složité a využívají předem získaných znalostí o důležitosti a očekávaném výskytu jednotlivých specifických vlastností při následné klasifikaci. Každý tento příznak následně odpovídá určitému specifickému omezení v rámci specifikace. Ze všech modelů vyhovujících daným omezením se vybere model s maximální entropií. V ideálním případě by měly být zadány všechny potenciálně důležité informace před začátkem klasifikace a následně ponechány trénovacímu procesu pro vytvoření nejlepšího modelu.

Pro každou danou sadu příznaků se počítá očekávaný výskyt jednotlivých příznaků na základě trénovací množiny. Příznaky jsou tedy binární funkce v následujícím tvaru *(3.4)*.

$$p(x, y) = \begin{cases} 1 & \text{když } x > 0 \text{ a } y = 1 \\ 0 & \text{jinak} \end{cases} \tag{3.4}$$

kde x je vektor vstupního prvku a y je označení třídy.

Nyní je zapotřebí zjistit rozdělení pravděpodobnosti $p(y|x)$, kde x je aktuální slovo a y je označení třídy. Potřebujeme tedy zjistit maximální entropii všech rozdělení v souvislosti s $p(y|x)$.

Nejlepší rozdělení pravděpodobnosti se pak počítá pomocí následujícího tvaru *(3.5)*.

$$p(y|x) = \frac{1}{Z(x)} \exp \sum_{i=1}^{n} \lambda_j f_j(x, y) \tag{3.5}$$

$$Z(x) = \sum_{y} \exp \sum_{i} \lambda_i f_i(x, y)$$

kde Z je pouze normalizační konstanta, která zajišťuje rozdělení pravděpodobnosti $p(y|x)$, n značí počet příznaků a λ je váha daného příznaku ([Zapotocky(2012)]).

4 Automatická sumarizace

V dnešním světě využívání moderních technologií a zejména internetu v rámci komunikace a sdílení textových informací v elektronické podobě nastává problém velkého množství umístěných dat a jejich čím dál tím složitější a udržitelnější struktury a přehled. To má za následek motivaci kondenzace textů do jednodušších informací se základní vypovídající hodnotou[Jezek et al.(2008)].

Existuje spousta různých definic, které popisují význam sumarizace:

- "Stručné, ale přesné vyjádření obsahu dokumentu."
- "Destilace nejdůležitější informace ze zdrojových dat."

A kvantitativní aspekty, kterými lze výsledný souhrn charakterizovat:

- Sémantická informativnost - lze charakterizovat jako míru schopnosti rekonstruovat původní text ze souhrnu
- Koherence - způsob, jakým části souhrnu vytvoří integrovanou souvislou sekvenci
- Kompresní poměr - podíl délky souhrnu oproti délce vstupních dat

Jak již bylo řečeno výše, sumarizátor zpracovává již předpřipravená data ze zdrojových textů. Tím rozumíme data (příspěvky, věty, text), která obsahují negativní, pozitivní nebo v krajním případě neutrální názor.

4.1 Metody sumarizace

4.1.1 Průkopnické metody

První algoritmus, který byl využíván pro automatickou sumarizaci textu byl založen pouze na jednoduchých principech indikující části vybrané do výsledné sumarizace, byl implementován již v roce 1958 [Luhn(1958)]. Jak uvádí

[Jezek et al.(2008)], tak to byl přístup založený na principu zjišťování frekvence slov v dokumentu s myšlenkou, že často používaná slova jsou vypovídající pro téma daného článku. U dalších tipů algoritmů se využívá například pozic daných slov nebo vět v dokumentu a jejich následné analýzy.

4.1.2 Statistické metody

Jak bylo dokázáno v [Salton(1988)], význam výrazů použitých v daném dokumentu je nepřímo úměrný počtu dokumentů v korpusu obsahující daný termín. V [Kupiec et al.(1995)] byla popsána více důmyslná metoda založená na bayesovském klasifikátoru výpočtu pravděpodobnosti, jestli by věta ze zdrojového dokumentu měla být vybrána do výsledného souhrnu. K učení klasifikátoru autoři využili 188 korpusů obsahují vždy pár dokumentu s jeho souhrnem. Jako charakteristické příznaky baysovského vzorce využívali například frekvenci použití slova, délku věty nebo pozici slova v odstavci.

4.1.3 Metody založené na textovém propojení

Obecně se jedná o metody založené myšlence propojení textu nebo jeho kontextu. Jako příklad můžeme uvést metodu lexikálních řetězců [Barzilay(1997)], která je založena na stanovení určitých vztahů mezi jednotlivými slovy (opakování, synonyma, antonyma, atd.) a skládání závislých slov do řetězců. Věty se do souhrnu vybírají podle obsahu nejsilnějších řetězců, tedy řetězců, jejichž váha je podle počtu a typu vztahů nejdůležitější.

Do skupiny těchto metod zahrnujeme také přístupy založené na teorii rétorické struktury (RST - Rhetorical Structure Theory)[Steinberger(2009)]. Jedná se o metodu zakládající se na principu rétorických vztahů struktur textu, které dohromady spojují textové jednotky. Snažíme se o práci a propojení jádra (hlavní myšlenka autora textu) a satelitu, což je méně důležitá až okrajová část textu. Takovéto vztahy vytváří stromovou strukturu, na jejímž základě se vytváří výsledný souhrn nejčastěji tak, že jsou jednotlivé věty dokumentu penalizovány podle toho, jako rétorickou roli hrají v celkové stromové struktuře.

4.1.4 Aspektově řízená sumarizace

V [Steinberger et al.(2011)] je navržen přístup pracující s vícero jazyky. Tento systém je zaměřen na podobné otázky, jakými jsou definovány kategorie TAC'10 a pro zachycení dalších aspektů využívá automatického učení pojmů významově souvisejících s ručně vytvořeným datasetem. Cílem bylo vybrání nejčastěji zmiňované informace a zároveň okamžitá kontrola zachycení požadovaných aspektů. Na tomto základě byla navržena kombinace společného výskytu výrazů a aspektech vycházejících z daného systému.

V experimentech byl využit systém NEXUS ([Tanev et al.(2008a)]), který analyzuje zprávy o násilných činech a přírodních nebo lidských katastrofách. Jako příklady můžeme zmínit určování a identifikaci vražd, povodní nebo únosů. Systém NEXUS analyzuje jednotlivé zprávy a vrací podstatná data v jednotlivých parametrech, jakými jsou například: *typ události, počet obětí, škoda, atd.*. Celý proces je závislý na kombinaci ručně vytvořených pravidel a výsledků získaných ze strojového učení. U experimentů sumarizace pomocí tohoto systému byly sumarizovány novinové články z korpusu a výsledné atributy byly přiřazeny k aspektům sumarizace.

Na základě výše zmíněných experimentů bylo zjištěno, že některé aspekty z aspektově řízené sumarizace odpovídají informacím získaným ze systému NEXUS. Zejména si pak odpovídají aspekty typu události parametrům ve strukturách událostí NEXUS.

V experimentech [Steinberger et al.(2011)] byly sumarizovány novinové články z korpusu a výsledné parametry byly navázány na aspekty sumarizace. Například typ události byl navázán na aspekt toho, co se stalo.

Pro ostatní aspekty byl použit systém Ontopopulis ([Tanev et al.(2008b)]), který je určen pro automatické učení sémantických tříd založený na distribuční sémantice. Principiálně je tento algoritmus založen na seznamu slov z určité sémantické třídy a poté probíhá učení dalších slov ze třídy stejné. Je zřejmé, že pro vytvoření přesnějšího slovníku je nutný ruční zásah a vyčištění výsledku. Seznam je seřazený podle spolehlivosti výrazů a pokud kontrolujeme seznam shora dolů, můžeme si určit přesnost tak, že proces zastavíme na určité úrovni seznamu. Pokud bychom výsledky ručně neupravili, nemuseli bychom dosáhnout očekávané přesnosti.

4.1.5 Časová analýza

Časová analýza může být velice důležitá při určitých dílčích úlohách sumarizace ve spojitosti se zjišťováním data a času událostí. Pomocí těchto dat můžeme dosáhnout lepší sumarizace zejména u příběhově orientovaných analýz. Časová analýza byla například integrována do procesu zabývající se detekcí a normalizací tzv. časových výrazů (Timex) [Steinberger et al.(2012)], jejichž rozsah a klasifikace jsou definovány v částečném souladu se standardem TIMEX2 [Ferro et al.(2005)].

Časové výrazy jsou rozděleny do krátkých seznamů časových typů (datum, čas, ...), zahrnují číselné i nečíselné formáty dat, absolutní nebo relativní výrazy a nakonec taktéž jejich kombinace. Timex analýza se rozděluje na dvě fáze, rozpoznávání a normalizaci. Fáze rozpoznávání je založena na detekci a segmentaci textu částečně závislých na pravidlech použitého jazyka. Pravidla určují struktury reprezentující časové výrazy, které jsou následně používány normalizačním modulem. Ten provádí výběr, který určuje a udržuje referenční čas pro relativní rozlišení, jež je na začátku brán jako datum vytvoření článku a je postupně aktualizován podle jednoduché heuristiky. Nejprve se nalezne nejbližší dřívější analyzovaný časový výraz v téže větě se stejnou úrovní granularity. Znamená to, že časové údaje v rámci dnů nemohou být spojeny s datovými údaji v rámci let. Poté se používá referenční čas na určení relativních časových údajů (timex) tak, že probíhá výpočet přesných kalendářních hodnot a nakonec dochází k normalizaci, díky které se jednoduše dostaneme k času cílové události. Tato metoda je silně ovlivněna pouze časovými výrazy a nemůžeme tak očekávat správnou funkci u textů s jinou časovou souvislostí.

4.1.6 Kompresní a parafrázové techniky

[Turchi et al.(2010)] empiricky dokazuje, že lidské výsledky sumarizace obsahují v průměru více kratších vět než výsledky automatické sumarizace. Kompresí nebo přeformulováním bychom tak mohli prvotně zkrátit celkový výsledek sumarizace a tím by vznikl prostor pro další rozšíření výsledků s obsahem dalších potenciálně důležitých vět. [Turchi et al.(2010)] také zkoumal možnosti jazykově nezávislých postupů při dosahování výše zmíněných cílů. Experimenty ukázali, že takovýto přístup je možný.

Systém je založen na výběru nejvíce charakteristického výrazu z každé

věty. Pro každý výraz je počítáno skóre z LSA, kde matice **U** obsahuje znázornění pojmu v tématech a matice **S** obsahuje důležitost těchto témat. Proto matice $T = U \cdot S$ představuje prostor ohodnocený podle důležitosti témat. Pro každý výraz i je vypočteno skóre z $||t_i||$. Kromě toho je vypočítán jazykový model pravděpodobnosti pro 4-gramy. Skóre by mělo odrážet lokální důležitost termínu v rámci datasetu a jazykový model pravděpodobnosti udává globálně důležité termíny, například slovesa. Po normalizaci hodnot jednotlivých příznaků a jejich kombinování dostaneme hodnoty, které reprezentují význam daného termínu ve větě. Finální sekvence výrazů se skládá přibližně ze 70% pojmů a pokud dodáme například stopwords, dostaneme celou sekvenci v rozumně čteném tvaru.

4.1.7 Sumarizace založená na LSA

Přístupy založené na společném výskytu výrazu (jako např.: LSA) představují dobrý základ pro budování jazykově nezávislého (nebo vícejazyčného) sumarizátoru. Přístup LSA (Steinberger a Ježek, 2009) nejprve vytvoří matici jednotlivých vět a obsažených slov ze zdrojového textu, poté se aplikuje singulární rozklad (SVD – Singular Value Decomposition) a nakonec se výsledné matice použijí k identifikaci a extrahování nejdůležitějších vět. SVD nachází skryté (ortogonální) dimenze, které, zjednodušeně řečeno, odpovídají různým tématům zdrojového textu.

Detailněji můžeme říci, že konstruujeme matici **A** vět a slov v nich obsažených. Každý prvek odpovídá vážené frekvenci daného výrazu v dané větě. Zabýváme se tedy m různými výrazy a n různými větami ze zkoumaného dokumentu tvořící matici **A** o velikosti $m \times n$. Prvek $a_{i,j}$ matice **A** představuje váženou frekvenci termínu i ve větě j a tato váha je definována jako (4.1):

$$a_{i,j} = L_{i,j} \cdot G_i, \tag{4.1}$$

kde $L_{i,j}$ je lokální váha termínu i ve větě j a G_i je globální váha výrazu i ve zdrojovém dokumentu. Bylo zjištěno, že nejlépe fungují váhové systémy (Steinberger et al, 2007.) používající binární lokální váhu a globální váhu založenou na entropii (4.2):

$L_{i,j} = 1$ když se slovo i vyskytuje ve větě j alespoň jednou; jinak $L_{i,j} = 0$,

$$G_i = 1 - \sum_{j=0}^{j<n} \frac{p_{i,j} \log p_{i,j}}{\log n}, p_{i,j} = \frac{t_{i,j}}{g_i}, \tag{4.2}$$

kde $t_{i,j}$ je frekvence termínu i ve větě j, g_i je celkový počet daného výrazu vyskytujícího se ve zdrojových datech a n je počet zdrojových vět.

Po tomto kroku je na výše uvedenou matici aplikován singulární rozklad (SVD). Singulární rozklad matice $m \times n$ je definován jako (4.3):

$$A = U \cdot S \cdot V^T, \tag{4.3}$$

kde $\mathbf{U}(m \times n)$ je sloupcově ortonormální matice, jejíž sloupce nazýváme levé singulární vektory. Matice obsahuje znázornění termínů vyjádřených v nově vytvořených dimenzích. $\mathbf{S}(n \times n)$ je diagonální matice, jejíž diagonální prvky jsou nezáporné singulární hodnoty řazené v sestupném pořadí. $\mathbf{V}^T(n \times n)$ je řádková ortonormální matice, která obsahuje znázornění vět vyjádřených v nově vytvořených dimenzích. Rozměry matic jsou redukovány na r nejdůležitějších dimenzí a dostáváme tak matice $\mathbf{U}'(m \times r), \mathbf{S}'(r \times r), \mathbf{V}'^T(r \times n)$. Hodnota r lze nastavit v závislosti na funkčnosti sumarizátoru. Další možností je strojové učení r na tréninkových datech (Steinberger a Ježek, 2009).

Z matematického pohledu SVD mapuje m-rozměrný prostor určený maticí **A** do r-rozměrného singulárního prostoru. Z pohledu NLP SVD odvozuje latentní sémantickou strukturu dokumentu reprezentovaného maticí **A**.: tj rozdělení původního dokumentu do r lineárně nezávislých základních vektorů, které vyjadřují základní témata ze zdrojových dat. SVD umí zachytit vztahy mezi termíny tak, že pojmy a věty mohou být rozděleny na sémantické základy spíše než na základě pouhých slov. Kromě toho v případě, že se v datovém zdroji opakuje kombinace slov, bude tento jev zachycen a reprezentován jedním ze singulárních vektorů. Velikost odpovídající singulární hodnoty určuje míru důležitosti tohoto vzoru ve zdrojových datech. Všechny věty obsahující danou kombinaci slov budou promítány do jednoho singulárního vektoru a věta, která nejlépe reprezentuje tuto kombinaci bude mít největší hodnotu indexu vektoru. Za předpokladu, že každé konkrétní slovní spojení popisuje určité téma v dokumentu, může být každý singulární vektor vnímán jako reprezentant takového tématu (Ding, 2005). Velikost jeho singulární hodnoty představuje míru důležitosti takového tématu.

Matice $\mathbf{V}^T$ obsahuje znázornění vět v tématech LSA a **S** obsahuje dů-

ležitost těchto témat. Proto je jejich výsledkem matice $\mathbf{F} = \mathbf{S} \cdot \mathbf{V^t}$, která reprezentuje větný latentní prostor vážený podle důležitosti tématu. Výběr věty začíná měřením délky vektoru věty v matici $\mathbf{F}$. Délka vektoru může být vnímána jako měřítko významu této věty v rámci LSA témat. Věta s nejvyšším skóre je vybrána jako první do celkového výsledku (odpovídající vektor v $\mathbf{F}$ je označován jako $\mathbf{f_{best}}$). Po jeho umístění ve výsledcích je zobrazení tématu/vět v matici $\mathbf{F}$ změněno odečtením informace obsažené v této větě (4.4):

$$F^{(it+1)} = F^{(it)} - \frac{f_{best} \cdot f_{best}^T}{||f_{best}||^2} \cdot F^{(it)}, \tag{4.4}$$

Délky vektorů podobných vět jsou anulovány pro případ, aby nebyly do souhrnu vybrány znovu a předešlo se tak ke zbytečné redundanci. Po anulování informací ve vybrané větě pokračuje proces u věty, která má největší skóre vypočtené na základě aktualizované matice $\mathbf{F}$. Tento proces je iterativně opakován do té doby, než výsledný souhrn obsahuje požadovaný počet výsledků.

4.2 Sumarizace názorů

Jak uvádí [Balahur et al.(2012)], dnešní svět je velice ovlivněn sociálními sítěmi a weby a lidé na jejich základě produkují obrovské množství dat vhodných k analýze a sumarizaci. Vhodnost je zřejmá například při marketingových průzkumech apod. Jedním z velkých problémů je opravdu velké množství dat, které můžeme získat po zaměření na konkrétní téma, které nás zajímá. standardně by tak musel uživatel probírat tisíce a tisíce úryvků nebo příspěvků a utvořit si sám vlastní souhrn hlavních myšlenek a poznatků. Je známo několik přístupů jak se zaměřit konkrétně na sumarizaci názorů.

V experimentu popsaném v [Balahur et al.(2012)] je určování sentimentu, tedy určování, zda příspěvek nebo věta obsahuje nějaký názor, rozdělování do čtyř skupin. Tyto skupiny znázorňují míru názoru. Můžeme je jednoduše pojmenovat jako *pozitivní, silně pozitivní, negativní a silně negativní*. K této analýze jsou používány kombinace různých slovníků. Hlavní motivací pro tuto metodu je její jednoduchost a snadné použití.

4.2.1 Data

Základem pro takovou analýzu a následnou sumarizaci je vhodný výběr dat. V experimentu byly využity dva datové soubory. První z nich obsahoval blogové příspěvky a jejich komentáře a druhý datový soubor obsahoval recenze uživatelů britských bank.

Co se týče blogů a blogových příspěvků, tak můžeme říci, že se mnohdy jedná o činnost vzdělaných lidí a navíc je většinou takovýto blog zaměřen na konkrétní téma. To je ideální příležitost pro jejich analýzu tak, abychom dostali relevantní souhrn z daného tématu od několika odborníků. Prvotní motivací a úspěšným použitím taky bylo nasazení ke studiu výsledku voleb nebo ke studiu kolísání tržních cen akcií. K danému experimentu bylo využito 51 blogových vláken psaných v anglickém jazyce. Jejich struktura byla vždy tvořena hlavním příspěvkem autora a následnými reakcemi ostatních čtenářů, čímž byla určena výsledná intenzita dané příspěvku (nízká, střední nebo vysoká). Datový soubor 89 bankovních hodnocení byl pak u tohoto experimentu použit zejména ke kontrola správnosti přístupu k dané problematice.

4.2.2 Základní analýza názorů a sumarizace

Hlavním cílem bylo navržení fungujícího systému, který vytvoří kvalitní souhrn z dvou výše zmíněných typů dat. Standardní přístup je založen na klasifikaci jednotlivých příspěvků podle názorů a jejich následné sumarizaci. Blogové příspěvky a komentáře byly postupně tříděny do tří samostatných skupin: věty obsahující pozitivní názor, věty obsahující negativní názor a neutrální nebo objektivní věty. Poté následuje činnost sumarizátoru, který samostatně provede sumarizaci pozitivních a negativních vět. Takto oddělená sumarizace se provádí z důvodu zajištění dvou výsledných souhrnů. V opačném případě by sumarizátor mohl určit jako důležité pouze věty z jedné nebo druhé skupiny.

Z pohledu analýzy názorů bylo tedy prvotním úkolem jejich roztřídění na negativní a pozitivní a následné numerické ohodnocení důležitosti (čím vyšší negativní skóre, tím více negativní věta a naopak). Jelikož tato klasifikace nebyla zaměřena na konkrétní téma, ale byla prováděna v obecném kontextu, bylo zapotřebí použít více různých slovníků tak, aby byly výsledky co možná nejlepší. Každý z použitých zdrojů byly namapovány do čtyř skupin s různým skóre: pozitivní (1), negativní (-1), hodně pozitivní (4) a hodně negativní (-

4). Výhoda tohoto přístupu je jeho jednoduchost a možnost využití i pro další světové jazyky. Navíc bylo prokázáno, že tato metoda zmůže u novinových citací dosáhnout úspěšnosti až 82

Nyní přichází na řadu část sumarizace, pro kterou byla v experimentu využita LSA (více v kapitole Sumarizace založená na LSA).

4.2.3 Sumarizace založená na intenzitě názoru

Oproti výše zmíněné metodě je tato část zaměřena na zkoumání vlivu intenzity názoru na výslednou sumarizaci. Jinými slovy řečeno je cílem zkoumání, jestli velmi pozitivní nebo velmi negativní věty charakterizují výsledný souhrn nebo nikoliv. Pro testování tohoto jednoduchého principu založeného na myšlence, že velice intenzivní názory jsou také velice vypovídající je zapotřebí tří důležitých částí: systém pro analýzu názorů produkující také jednotlivé intenzity, sumarizátor zohledňující tyto intenzity a referenční korpus anotovaných dat.

U experimentu [Balahur et al.(2012)] byla využita výš zmíněná základní analýza názorů s tím, že je algoritmus doplněn o řazení komentářů podle intenzity pro danou polaritu a následně se vybírá vždy komentář s nejvyšším skóre.

Po otestování tohoto přístupu bylo zjištěno, že tento přístup není zcela reprezentativní a proto je dobré ho kombinovat s dalšími funkcemi.

4.2.4 Sumarizace založená na tématu analýzy názorů a sémantických informací

Pokud se zaměříme na sumarizaci v obecném kontextu, nejsou výsledky až tak dobré ([Balahur et al.(2012)]). Nicméně, pokud se jedná o kladný nebo záporný názor, v celkovém hodnocení se objevil. Objevovaly se ale také příspěvky, které byly zcela irelevantní. Je pro nás proto důležité, aby každý příspěvek, který se dostane do výsledné sumarizace byl zaměřen na konkrétní téma. K analýze byla požita LSA (latentní sémantická analýza) a byly v úvahu oproti jiným experimentům sémantické informace.

Analýza názorů je zde prováděna stejným způsobem jako výše. Znamená

to, že byly vybrané zdroje mapovány do čtyř skupin (pozitivní, negativní, vysoce pozitivní a vysoce negativní). Každé této skupině pak byly přiřazeny číselné hodnoty (tedy ve stejném pořadí 1, -1, 4, -4). Ve druhé fázi byly pomocí LSA odfiltrovány pouze příspěvky spojené z daným tématem a postupně jako v předchozím postupu byly určeny skóre těchto vybraných příspěvků.

Pro systém sumarizace byla použita stejná metoda jako výše, tedy sumarizace založená na LSA. Zde se však navíc počítá se sémantickými informacemi kombinující několik vlastností zdrojů jako jsou lexikální informace nebo informace o jednotlivých subjektech. Oproti základní metodě je tedy vstupní matice do LSA rozšířená o tyto informace, jakými jsou například synonyma a další vlastnosti.

[Balahur et al.(2012)] také uvádí, že tento obecný postup, kdy nejprve dochází k analýze příspěvků v závislosti na obsahu subjektivního názoru a následné sumarizaci je motivován lepšími výsledky. V opačném případě se totiž tak dobrých výsledků nedosahuje.

5 Sumarizátor názorů na Facebooku

5.1 Architektura

Celý software je rozdělen do pěti hlavních částí (balíků - packages).

- App - obecná část reprezentující hlavní běh celé aplikace.
- Facebook - balík obsahující veškerou komunikaci s facebookem a třídu reprezentující získaná data.
- GUI - zajištění grafického rozhraní aplikace.
- Sent - balík obsahující třídy zaměřené na část aplikace zajišťující výběr sentimentu. S tím souvisí také reprezentace jednotlivých slovníků.
- Summ - část zaměřená na samotnou sumarizaci. Řešen je zde celý algoritmus sumarizátoru i s jeho pomocnými třídami.

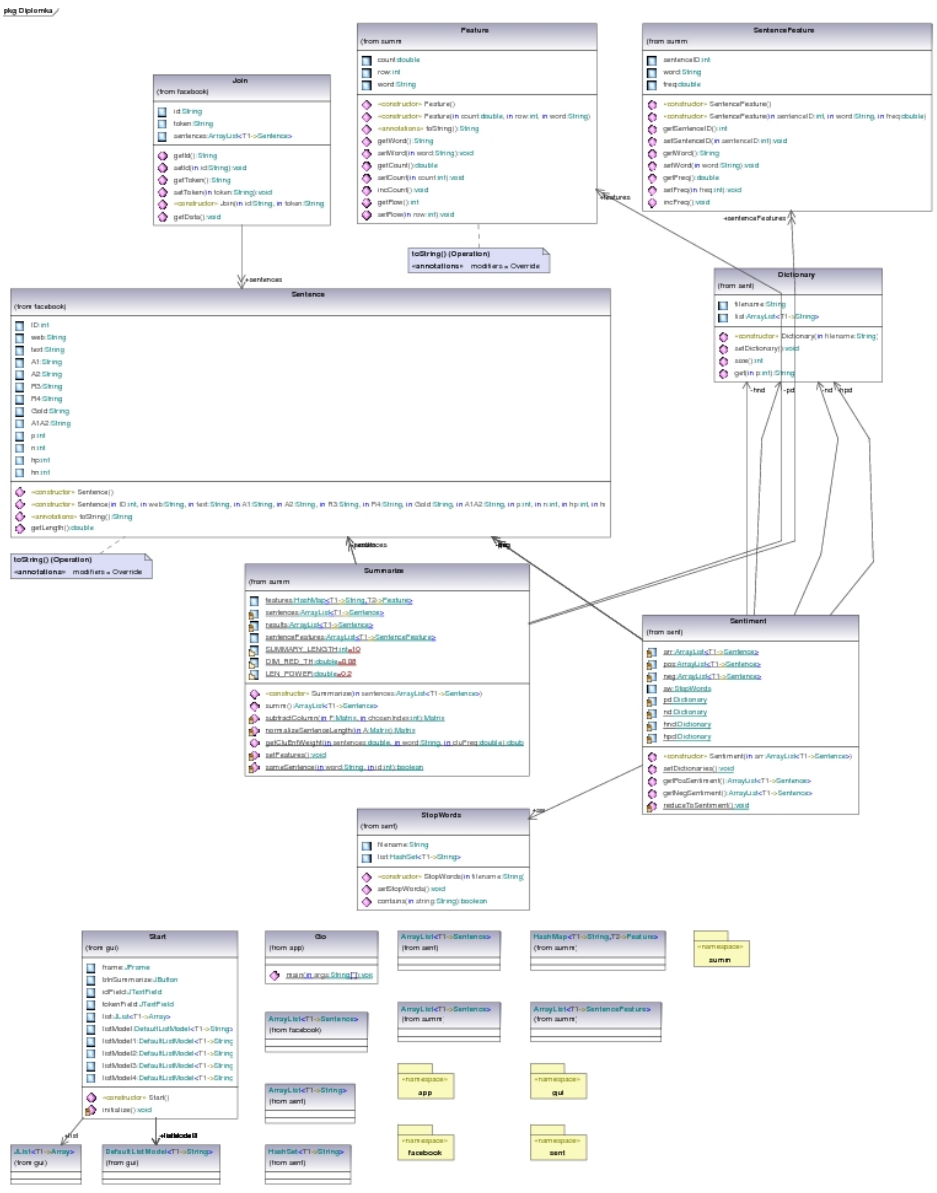
Na obrázku 5.1 je vidět náhled UML diagramu architektury.

5.2 Implementace

Celou aplikaci můžeme rozdělit na méně zajímavé a obecné části a na zajímavější a důležité kusy, které detailněji rozebereme. Mezi obecné části patří balík *App*, ve kterém je implementována metoda *Main()* a volají se zde další důležité metody z ostatních balíků. Zde je dobré zmínit, že pro vývoj celé aplikace bylo použito vývojové prostředí Eclipse KEPLER za použití knihovny SWING pro vytvoření grafického rozhraní aplikace. Vše je implementováno v jazyce Java 1.7. Pro vytvoření diagramu UML byla použita aplikace Altova UModel.

5.2.1 Implementace analýzy sentimentu

Veškeré třídy a použité algoritmy pro analýzu sentimentu obsahuje balík *Sent*. Jako hlavní musíme zmínit třídu *Sentiment.java* definující algoritmy



Obrázek 5.1: Náhled UML diagramu reprezentující architekturu celé aplikace

a postupy analýzy sentimentu. První důležitou bodem je volání metody *setDictionaries*, která má za úkol inicializovat veškeré slovníky potřebné pro analýzu sentimentu. V našem případě se jedná o inicializaci slovník s kladnými výrazy, zápornými výrazy, extra kladnými výrazy, extra zápornými výrazy a se stopwords. Slovníky mají svoji strukturu a samozřejmě také implementovanou zmíněnou metodu pro jejich inicializaci. Metoda je velice jednoduchá, postupně projde textový soubor daného slovníku a vloží každý prvek do vytvořeného pole typu *ArrayList*. Od této doby tedy můžeme pracovat s těmito slovníky v rámci určování sentimentu.

Nyní se dostáváme k využití čtyř proměnných třídy Sentence.java. Ve třídě *Sentiment.java* je využívána metoda *reduceToSentiment*, která má za úkol projít celé pole objektů reprezentující jednotlivé příspěvky a inkrementovat danou proměnnou při každém výskytu nějakého slova ve slovníku. Takto se nám zaktualizují hodnoty těchto čtyř proměnných, které budou rozhodující pro konečnou analýzu sentimentu.

Konečně se dostáváme k nejdůležitějším metodám a to k *getPosSentiment* pro vyjádření kladných názorů a *getNegSentiment* pro vyjádření záporných názorů. Určování takto názorově zabarvených příspěvků je použito právě hodnot daných čtyřmi proměnnými *(p, n, hp, hn)* třídy *Sentence.java* reprezentující jednotlivé příspěvky. Metoda zkoumá všechny čtyři hodnoty a určuje tak podmínky, za kterých obsahuje příspěvek kladný nebo záporný názor. Tyto podmínky byly v rámci experimentu obměňovány a laděny. Příkladem však můžeme ukázat jednu z použitých podmínek výběru příspěvků s negativním názorem.

```
public ArrayList<Sentence> getNegSentiment(){
  neg = new ArrayList<Sentence>();
  for (int i=0; i<arr.size(); i++){
       if(((arr.get(i).n > 1)||(arr.get(i).hn > 1)) &&
       ((arr.get(i).p < 3)||(arr.get(i).hp < 3))){
           neg.add(arr.get(i));
       }
  }
return neg;
}
```

5.2.2 Implementace automatické sumarizace

Celkově balík *summ* obsahující vše potřebné pro samotnou sumarizaci obsahuje tři třídy, *Feature.java*, *SentenceFeature.java* a nejdůležitější *Summarize.java*. Z pohledu sumarizace je tedy nedůležitější metoda *summ()*, ve které je kompletně obsažen algoritmus sumarizace založené na LSA. Tato metoda se volá v hlavní třídě a konstruktorem tohoto objektu sumarizátoru je *ArrayList* naplněný jednotlivými příspěvky.

V první fázi se v této metodě volá pomocná metoda *setFeatures*, která vytvoří seznam unikátních výrazů (třída *Features.java*) vyjma slov ze slovníku stopwords a zároveň u každého takového slova při procházení aktualizuje čítač reprezentující celkový počet výskytu daného slova v analyzovaném datasetu. Tento seznam nám poté tvoří y-ovou souřadnici potřebné matice. Druhý úkol této metody je inicializace a naplnění pole typu *ArrayList* objekty třídy *SentenceFeatrue.java* vyjadřující jednotlivé příspěvky na x-ové souřadnici potřebné matice a tato matice nám tak ukazuje výskyt jednotlivých slov v daných příspěvcích.

Následuje vypočtení hodnoty každého prvku matice. Znamená to, že počítáme váhu (důležitost) jednotlivých slov. K tomu je určena metoda *getCluEntWeight*, která je vytvořena v návaznosti na adekvátní vzorec 4.2.

```
private static double getCluEntWeight(double sentences,
                                       String word, double cluFreq) {
  double sum = 0;
  for (int i=0; i<sentenceFeatures.size();i++){
    if (sentenceFeatures.get(i).getWord().equals(word) == true){
         double p = sentenceFeatures.get(i).getFreq() / cluFreq;
         sum = sum + (p*Math.log(p));
    }
  }
return(1 - sum/Math.log(sentences));
}
```

Následuje normalizace vzniklé matice a následný singulární rozklad, jenž je implementován v dodané knihovně *jama.jar* poskytnuté Ing. Josefem Steinbergerem Ph.D. Vzniklé matice následně zredukujeme.

Posledním velkým krokem je samotné vybrání vět do výsledné sumari-

zace. To probíhá tak, že se vybírá věta s nejdelším vektorem. Následně musíme u této věty (tématu) zaručit, aby se nám do výsledku nedostala znovu. K tomu slouží pomocná metoda *substractColumn*, která je založena na vzorci 4.4.

```
private static Matrix subtractColumn(Matrix F, int chosenIndex) {
   Matrix newF = new Matrix(F.getRowDimension(),
                                        F.getColumnDimension());

  // výpočet druhé mocniny délky nejlepšího vektoru
  double len = 0;
  for (int i=0; i<F.getRowDimension(); i++)
      len += F.get(i, chosenIndex)*F.get(i, chosenIndex);

  // vytvoření matice vbest * vtbest (obě normalizované)
  Matrix VV = new Matrix(F.getRowDimension(), F.getRowDimension());
      for (int i=0; i<F.getRowDimension(); i++)
         for (int j=0; j<F.getRowDimension(); j++)
            VV.set(i, j, F.get(i, chosenIndex) *
                                  F.get(j, chosenIndex) / len);

  // vytvoření matice VV * F
  Matrix VVF = new Matrix(F.getRowDimension(),
                                          F.getColumnDimension());
  VVF = VV.times(F);

  // odečtení matice VVF z matice F
  newF= F.minus(VVF);

  return newF;
}
```

5.2.3 Implementace rozhraní pro Facebook

Vzhledem k tomu, že v našem případě je automatický sumarizátor názorů zaměřen na analýzu příspěvků z facebookových stránek, jedná se o velice důležitou část.

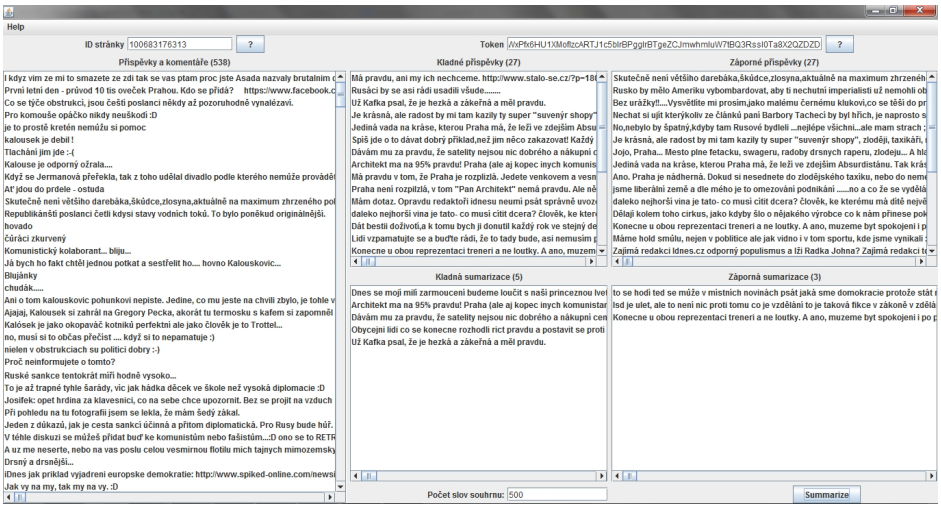
Pro komunikaci a extrahování příspěvků z facebookových stránek bylo

použito rozhraní **Facebook4J API** (http://facebook4j.org/). Toto rozhraní je zcela vyhovující a připravené pro veškeré potřeby automatického sumarizátoru. Balík *facebook* tedy obsahuje třídu *Join.java* reprezentující samotné rozhraní a komunikaci s Facebookem.

Zjednodušeně můžeme říci, že k navázání komunikace je zapotřebí tzv *tokenu*, což je unikátní identifikace uživatele nebo aplikace, jejíž pomocí dochází k extrahování příspěvků a komentářů. Druhou podstatnou věcí je identifikace facebookové stránky, která může být implementačně řešena přímo pomocí názvu stránky (URL) nebo jako v našem případě pomocí identifikačního čísla. Následně už jen dochází k navázání komunikace a za pomoci metod *getFeed* a *getMessage* dostaneme určený počet hlavních statusů, ze kterých můžeme extrahovat ještě komentáře a odpovědi metodou *getComments*.

Druhou důležitou třídou je *Sentence.java*, která reprezentuje strukturu příspěvku i s jeho náležitými vlastnostmi. Obsahuje tedy hlavně *ID*, což je unikátní identifikační číslo, díky kterému můžeme daný příspěvek v jakémkoliv stavu aplikace jednoznačně identifikovat. Toto číslo je také velice důležité pro analýzu sentimentu a následnou sumarizaci. Samozřejmostí třídy *Sentence.java* je proměnná *text*, která je typu *String* a reprezentuje samotný text příspěvku. Nesmíme zapomenout také na čtyři proměnné *(p, n, hp, hn)*, které jsou typu *Integer* a při extrahování příspěvku vytváříme objekt této třídy, jehož startovní hodnoty zmíněných čtyř proměnných jsou nulové. Značí totiž počet slov z každého ze čtyř slovníků se kterými se ovšem pracuje až při určování sentimentu. Na závěr ještě musíme připomenout dalších několik proměnných, které jsou v této struktuře připraveny pro práci s korpusem, ale ve standardním režimu je nepoužíváme.

Tato implementace je z uživatelského pohledu úzce spojená s uživatelským rozhraním celé aplikace. Ta byla navržena tak, aby byla co možná njesnadněji ovladatelná a hlavně přehledná. Uživateli byla poskytnuta také možnost nastavení velikosti výsledné sumarizace. Náhled celé aplikace můžete vidět na obrázku 5.2. Je vidět, že rozmístění jednotlivých výčtů je velice logické a navíc jsou uživateli dodávány statistické informace z jednotlivých stádií procesu analýzy. Rozhraní tak zcela přehledně ukazuje veškeré příspěvky extrahované z dané facebookové stránky, které jsou pak analyzovány z pohledu obsahu subjektivního názoru. Tyto názory jsou pak v aplikaci přehledně rozděleny na kladné a záporné. Nakonec aplikace ukazuje skutečnou výslednou kladnou i zápornou sumarizaci opět v oddělených listech.



Obrázek 5.2: Náhled automatického sumarizátoru názorů na Facebooku

5.3 Testování aplikace

Celá aplikace je velice jednoduše připravena k okamžitému použití a byla v reálu otestována na několika facebookových stránkách s různým zaměřením a různým stylem testování. Jako uživatel sociální sítě Facebook mám manažerská práva k několika veřejným i testovacím stránkám. Veškerá funkčnost tak mohla být testována na veřejných stránkách z příspěvky různých uživatelů, ale mohla být testována také na soukromých testovacích facebookových stránkách s možností ovlivnění obsahu. Explicitně pak byly otestovány facebookové stránky iDnes.cz, Nokia.cz) a

5.3.1 Nokia

Stránky byly vybrány zejména z důvodu velkého počtu názorů, jelikož je známe, že společnost Nokia se v očích uživatelů nejeví zcela důstojně. V tabulce 5.1 můžeme vidět statistiky z konkrétního použití při zaměření na posledních 40 statusů a nastavení délky sumarizace na 500 slov.

Konkrétně byly do výsledné kladné sumarizace vybrány tyto příspěvky:

- ”Oblíbených aplikací je několik. Vzhledem k tomu, že mám Lumii 1020, tak zejména fotoaplikace. Krom těch od Nokie jsem si oblíbil HDR

Photo Cameru, která zvládá skvělé HDR snímky a po vyzkoušení vícero aplikací pro HDR ji mohu směle označit za jednu z nejlepších na Storu v tomto oboru a v tuto dobu. Dále jsem si oblíbil fotoeditory Fhotoroom a Fotor. V tom prvém si člověk může velmi dobře vyhrát s nastavením parametrů a krom různých obligátních efektů může směle upravovat i parametry jako vyvážení stínů a světel, korekce barevnosti a expozice, atd., takže fotky si často upravíte velmi dobře přímo v telefonu a nemusíte je už dále upravovat v počítači. Fotor se zaměřuje zejména na efekty, ale pro určité operace je rychlejší, než Fhotoroom a navíc umí i hezké koláže. Další oblíbenou aplikací se stalo YouRadio, které umožnuje zdarma streamovat muziku dle vybraných "nálad"a ještě si uložit až dvě hodiny pro off-line poslech, když je člověk třeba venku a nemá zrovna ideální FUP na datovém balíčku. Nálady si lze tvořit vlastní a muziku si tak naladit přesně dle mého gusta. Výbornými aplikacemi jsou české Bouřky s vždy aktuálním stavem meteoradarů nad Českem, tak předpovědní Meteoservis. Pro čtení novinek používám už od WP7 skvělou čtečku Nextgen Reader s níž se člověk v RSS neztratí a má stále aktuální přehled o dění v tom, co jej zajímá. Pro sledování YouTube videa dobře poslouží aplikace MetroTube, nebo YouTube HD, byt' v IE11 ve WP8.1 už lze video přehrávat skvěle i přímo v prohlížeči. Ve světě meziměstské dopravy vždy spolehlivě poradí WMM Jízdní řády a v terénu pak zase Mapy.cz od Seznamu. Před spaním si rád pustím nějaké ty uklidňující tóny a skvěle se usíná třeba s aplikacemi Sleep, nebo Sleep Bug. Zkrátka a dobře - není pravda, že by na Storu bylo málo aplikací. Najít velmi dobré aplikace se dají a kdyby hoši v MS ještě trochu více mákli na odstranění, nebo alespoň odfiltrování "balastu", bylo by hledání ve Storu opravdu radostí. Těším se, že s oficiálním příchodem WP8.1 se počet kvalitních aplikací zase zvedne a padnou tak i poslední argumenty odpůrců Windows Phone :)."

- "Dobrý den, dobre pokud to dobre chápu, tak by mela byt záruka platna. Podle toho imei se mi navíc na Vašich stránkách objeví v zaruce v této zemi. Ale potřeboval bych vedet, jak bych postupoval při pripadne reklamaci, když nemam zarucni list ani doklad o koupi. Jestli si jen v servise podle čísla najdou jestli je v zaruce, a tak to stačí. Dekuji za odpověd'."

- "Prosim Vas, nehrajte si tu na chytreho. Psal jsem, ze je mnoho vsriant. Chtel bych Vas videt vyvijet aplikace. To by asi byla sranda co? Ja to delam uz dost dlouho. Nokia je dobra a se svoji Lumii 1020 i 1520 jsem spokojen."

- "hezké lumie :) snad poprvé v životě něco vyhraju :) máte moc hezkou a informativní stránku :) a pěknou soutěž kterou jsem udělal :)"

Můžeme vidět, že do kladné sumarizace se dostaly vcelku zajímavé a z velké části opravdu kladné názory, které nám zjednodušeně mohou pomoci udělat si obrázek o aktuálních pozitivech značky.

Nyní se podívejme na výsledky záporné sumarizace:

- "Jedním z důvodů, proč jsem si pořídil Lumii byla offline navigace HERE Drive. Většinou stačí, ale bohužel k dokonalosti jí chybí ještě hodně (nejvíce mi vadí chybějící navigace v pruzích, podivné hlášky držte se vlevo/vpravo a nemožnost zadat průjezdní místa). Skvělá je aplikace Lidé, která je v podstatě automaticky provázaná s mým online adresářem a sociálními sítěmi. To pro mne bylo velmi příjemné překvapení. Velmi také oceňuji fotoaplikaci NOKIA Camera, Její ovládání a funkce nemají chybu a neznám nic lepšího. Mít toto i na svém fot'áku, to by bylo něco :-) A OneDrive, to asi ani nemusím rozvádět. Pořídil jsem si i tablet Nokia Lumia 2520 (trochu krkolomně ze zahraničí) právě kvůli provázanosti s PC a mobilem a Nokia aplikacím."
- "Dobrý den, tyto potíže se týkají účtu Windows Live, proto může být nutné upravit nastavení propojených služeb Live pro Facebook: Přejděte na web http://Live.com a přihlaste se do svého Windows Live učtu.Vyberte propojené služby Windows Live.Vyberte službu Facebook a podle pokynů službu propojte. Zkuste v nastavení účtů v telefonu zapnout účet služby Facebook. Zvolte možnost Start > Nastavení > e-mail+účty Co se týče přerušování hovorů, doporučujeme vyzkoušet, zda přistroj bude normálně fungovat s jinou SIM kartou. V případě, že je to potíž na straně SIM karty, obrat'te se na svého telefonního operátora, aby Vám kartu vyměnil. Nokia Česká republika"
- "Dobrý den, chci si koupit bluetooth headset Nokia BH-112U. Jak dlouho vydrží headset spárovaný s mobilem? Jde mi o to, aby když ho ráno spáruji s mobilem (Lumia 620), tak aby vydržel spárovaný celý den, resp. pracovní dobu. Když mi bude někdo volat, aby pokaždé šlo hovor přijmout headsetem a nemusel jsem šahat na mobil. Známý má headset HB-108 na Lumii 820 a po nějaké době od spárování se toto zruší a musí headset znovu s mobilem spárovat, což je problém. Tak jestli to je chyba jeho kusu nebo to je dáno technologií a má to tak každý headset? Děkuji."

- ”Dobrý den mám tady takovou pikantnost.... Jde o reklamaci Nokia Lumia 1520.. byl jsem v obchodě kde jsem ji koupil. je na ni vadný pixel.... 1 . No a oni mě řekli že do tří pixelů je to unosné... :(:(Moc dobře věděli že je to Nokia z Polska.... poslali fotku do Olomouce a dneska mě řekli že prostě neuznají reklamaci..... Mrzí mě to dost zklamala mě Nokia..... :(:(”

Opět je možno vidět, že se do výsledné záporné sumarizace dostaly z velké části opravdu záporné názory, které nám naznačují hlavní problémy dnešních uživatelů produktů společnosti Nokia. Můžeme zde také pozorovat, že zde správně funguje nastavení analýzy názorů, jelikož u příspěvků můžeme vidět oslovení ”dobrý den”, jenž ovšem na zařazení mezi záporné názory nemělo vliv.

5.3.2 iDnes.cz

Zaměření portálu iDnes je hlavně na aktuální dění u nás i ve světě a je tak možné v sumarizaci očekávat příspěvky zaměřené právě na témata posledních dní. Při tomto testování byla nastavena délka výsledné sumarizace na 200 slov. Statistické údaje můžete vidět v tabulce 5.1.

Kladná sumarizace:

- ”Je evidentní, že máme příliš mnoho politických stran, hnutí a různých uskupení. Bylo by více než žádoucí aby si strany na svojí existenci začaly vydělávat bez účasti daňových poplatníků. Kvalita našeho života totiž rozhodně neodpovídá počtu stran ani jejich volebním programům. Až na prvním místě jsou peníze, i takto lze chápat množství politických subjektů. Zvláště, když žijí a fungují na základě nikoli svých finančních prostředků. Eurovolby jsou pro ně zvláště vítaným prostředkem k získání financí. V případě minimálního úspěchu dostanou slušné peníze a navíc v případě zvolení do Evropského parlamentu i možnost zašít se do instituce, kde na ně nebude z Čech až tolik vidět. Navíc, když se dá předpokládat nízká volební účast. Je to od stran moudré a chytré, ne však už tak pro občany této země. Ale ptá se jich tu vůbec ještě někdo na něco? Myslím, že ne.”
- ”Bylo to samoúčelné a tudíž to nemělo s karikaturou nic společného, jen s pomluvou. V tom má Petra Paroubková naprostou pravdu! Myslím

ten obrázek v Reflexu!"

- "A 1000x víc havárii zavinili lide co byli střízliví a nebyli pod vlivem. Tudíž je statisticky lepší jezdit ozralej a nafetovanej pak je menší pravděpodobnost havarie :-D"

Můžeme vidět, že jednoznačnost kladných názorů není zcela tak jasná jako u stránek Nokia. Je to dáno především tím, že zde lidé reagují opravdu na velké množství aktuálního dění a tak nemůžeme očekávat, že se v sumarizaci objeví nějaký názor přímo na portál iDnes. I přesto je však vidět, že se do sumarizace dostaly příspěvky zcela tématicky různorodé, což naznačuje správnost funkce sumarizátoru z pohledu jeho funkčnosti.

Záporná sumarizace:

- "Je to vlaštni jak kdejaké neřadstvoo ppostupňé ziska vá i přiznivce nebo prosté tolerovani..Dokonce by někteři sohlasli s legalisaci. Ale otazka je proč zdravi člověk potřebujé drogu? Odpověd je prosta je to jedinec,ktery ztratil soudnost sam nad sebou a pak je mu jedno jak jeho jednani ovlivni třeba i život jiného.Jak takoveto jedince přivest na normalni uvažovani,prosté. Za smrt smrt! Ostatni by pochopili velice rychle že pit nebo drogovat se prpstě při řizeni nesmi. Když v Norsko jede opilec a bez nehody skonči na 3 roky v base. V českе rep. Jede na rekraci na svou chalupu v zamoři a nejde ani k soudu.(Janoušek a pod). Tak že se dál budeme těšit na dalši silnični tragedie.věc mě přijde tako á že si nej ic ožrali a sfetovany jezdi zakonodarci proto se zakon na tvrdé tresty nedostavaji na program dne"
- "Kdyby se už konečně začalo něco dělat se zloději a podobnými, kterých je náš stát plný. Miliardy jsou v pr...i, ale chytají se malé čudly a velryby si klidně plavou dál. Je mi z toho našeho Česka na blití."
- "Já si prvně myslel, že je to přepadení nebo teroristickej útok."

Zde můžete jasně opět pozorovat různorodost příspěvků a je zde více markantní výběr opravdu záporných názorů, které shrnují určitý pohled na danou věc. Zřejmé je to zejména u druhého příspěvku ze záporné sumarizace.

	Nokia	**iDnes**
Adresa	facebook.com/nokia.cz	facebook.com/iDNES.cz
Celkem příspěvků	219	431
Kladných názorů	16	20
Záporných názorů	14	24
Kladná sumarizace	4	3
Záporná sumarizace	4	3

Tabulka 5.1: statistické vyhodnocení reálného testování aplikace na třech facebookových stránkách

6 Experiment

Celý experiment je založen na zjištění a porovnání účinnosti automatického sumarizátoru názorů oproti manuální subjektivní sumarizaci. K tomuto účelu bylo zapotřebí získání testovacích dat, která by se mohla použít pro oba typy sumarizace a jejiž výsledky by se dali reálně porovnávat.

6.1 Získání datasetu

Jako testovací data byly použity označkované korpusy poskytnuté Ing. Josefem Steinbergerem, Ph.D. Tyto korpusy jsou všechny ve stejné formě ve formátu CSV. Jedná se o posbírané příspěvky z facebookových stránek:

- Kofola - www.facebook.com/kofolaceskoslovensko
- McDonald's - www.facebook.com/McDonalds
- Milka - www.facebook.com/Milka.cz.sk
- O2 - www.facebook.com/o2cz
- Slevomat - www.facebook.com/slevomat
- T-Mobile - www.facebook.com/TmobileCz
- Vodafone - www.facebook.com/vodafoneCZ
- Xparfemy - www.facebook.com/Xparfemy.cz
- ZOO Praha - www.facebook.com/zoopraha

Každý z uvedených korpusů obsahuje přibližně 1100 příspěvků z výše uvedených facebookových stránek.

6.1.1 Formát dat

Jedná se o soubory ve formátu CSV obsahující tyto sloupečky:

- ID - jedná se o unikátní klíč daného příspěvku (pořadové číslo)
- Web - jméno facebookové stránky
- Text - konkrétní text daného příspěvku
- A1 - ohodnocení příspěvku první osobou (0, n, p, b)
- A2 - ohodnocení příspěvku druhou osobou (0, n, p, b)
- R3 - ohodnocení příspěvku třetí osobou (0, n, p, b) - v případě neshody A1 a A2
- R4 - ohodnocení příspěvku čtvrtou osobou (0, n, p, b) - v případě neshody A1 a A2
- Gold - výsledný názor na příspěvek
- A1=A2 - porovnání názorů osob A1 a A2

Příklad části korpusu můžete vidět na obrázku 6.1.

ID	web	Text	A1	A2	R3	R4	GOLD	A1=A2
1332	vodafoneCZ	Dobrý den, vašeho kontaktu, rozhodně ráda využiji. Nyní jsem dorazila zpět do kanceláře z Vaší pobočky, kde nemají pra	0	n	0		0	NEPRAVDA
1333	vodafoneCZ	a vyměnit se dá na pobočce?	0	0			0	PRAVDA
1334	vodafoneCZ	mě nevadí když nevolají,alespoň vím, že se nic neděje	0	0			0	PRAVDA
1335	vodafoneCZ	Vypraveni pohadky o tom, jak nam jednou vodafone CZ bude uctovat stejne poplatky jako vodafone UK...:-P	0	n	0		0	NEPRAVDA
1336	vodafoneCZ	no myslím, že se kampan vodafone s fail tarify moc nepodařila. ono se nelze nedivit...kde nic není ani smrt nebere.	n	n			n	PRAVDA
1337	vodafoneCZ	Já jezdím každý den na kole do práce a i dálky.	0	0			0	PRAVDA
1338	vodafoneCZ	asi se jich ještě nesešlo sedm, natož 17 a 37 :)	0	0			0	PRAVDA
1339	vodafoneCZ	A taky je tam free mobile. Fr už aby to bylo i tady!	0	0			0	PRAVDA
1340	vodafoneCZ	to je super,že nejrychlejší, ale mě by se hodil taky mimo velká města :D myslím, že ani 61% nebude pravda .)	0	n	n		n	NEPRAVDA
1341	vodafoneCZ	VT: přesně. Kvůli němu bych tam nešel. Ještě bych pak musel jít do mobil pohotovosti a to raději zůstanu doma.	0	n	n		n	NEPRAVDA
1342	vodafoneCZ	No jo. Aby to pak nebyla typicka: Nechci slevu zadarmo a v konecnem dusledku jeste zaplati i tu "slevu":-D	n	n			n	PRAVDA
1343	vodafoneCZ	Kam se hrabou reklamy na T Mobile :D	p	n	p	p	p	NEPRAVDA
1344	vodafoneCZ	Do tří týdnů? :-/ Co se dá dělat, no. I tak děkuji :)	0	n	0		0	NEPRAVDA
1345	vodafoneCZ	Dobrý den, tak to není opravdu dobrá zpráva, škoda. Prosím upravte si tedy mapu pokrytí, protože není pravdivá. Díky a r	n	n			n	PRAVDA
1346	vodafoneCZ	Děkuju moc.... ale tohle sem zkoušel ... nevím kde je chybka :) ale nikdy to nedělalo. Až z ničeho nic :)	0	0			0	PRAVDA
1347	vodafoneCZ	jasně no proč ne... každopádně až donedávna sem byla s vodafonem spokojená... ted už uvažuju že odejdu jinam	n	n			n	PRAVDA
1348	vodafoneCZ	Dobré ráno. Prosím o odpověď na mou včerejší otázku. Děkuji	0	0			0	PRAVDA
1349	vodafoneCZ	děkuji za odpověď. Už se nám podařilo dovolat.	p	p			p	PRAVDA
1350	vodafoneCZ	Super:-))) asi nejvtipnejsi ze vsech:)	p	p			p	PRAVDA
1351	vodafoneCZ	Dnes jsem byla u Vás na prodejně, kde mi paní řekla, že je to možné, že je to simkartou a duplikovala mi ji. Tak uvidím. Ji	0	0			0	PRAVDA
1352	vodafoneCZ	Ondro, to vám ještě nedošlo, že z ho.na opravdu bič neupletete?	0	n	n		n	NEPRAVDA
1353	vodafoneCZ	telefon se mi za cenu líbí . když se rozbije tak šup s ním do opravy .)	p	p			p	PRAVDA
1354	vodafoneCZ	takže ta aktualizace zmizela	0	0			0	PRAVDA

Obrázek 6.1: Ukázka označkovaného korpusu facebookových stránek společnosti Vodafone

6.2 Testování

V rámci testování bylo tedy zapotřebí získat subjektivní výsledky na základě anotace a sumarizace fyzických osob a anotace a sumarizaci automatické.

6.2.1 Manuální sumarizace

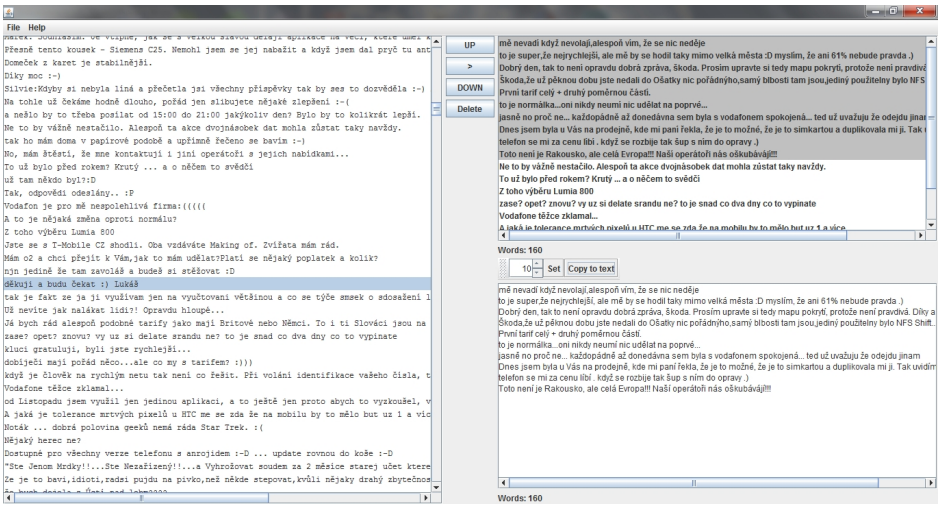
Tato část experimentu probíhala s pomocí pěti dobrovolníků, kteří měli za úkol pomocí doplňkové aplikace pro manuální výběr a export vybrat z dodaných korpusů příspěvky obsahující kladný nebo záporný názor. Pro maximální rozpětí témat bylo pro experiment vybráno těchto pět korpusů:

- McDonald's - www.facebook.com/McDonalds
- Milka - www.facebook.com/Milka.cz.sk
- Slevomat - www.facebook.com/slevomat
- Vodafone - www.facebook.com/vodafoneCZ
- ZOO Praha - www.facebook.com/zoopraha

Každý dobrovolník si za použití doplňkové aplikace vždy naimportoval jeden z korpusů a následně postupně zkontroloval každý příspěvek. Pokud uznal nějaký příspěvek jako názorový, mohl jej označit a přesunout do seznamu příspěvků, které podle něj názor obsahují. Takto vytvořený seznam je pak možné exportovat do souboru ve formátu CSV. Těchto celkem 25 souborů (5 souborů od 5 dobrovolníků) bylo použito k otestování úspěšnosti automatického sumarizátoru názorů. Ukázku aplikace pro manuální výběr příspěvků obsahující názor můžete vidět na obrázku 6.2.

Následně byli ještě tito dobrovolníci požádání o manuální sumarizaci těchto příspěvků obsahujících názor. Výsledky jednotlivých dobrovolníků (testerů) můžete vidět v další části této práce.

Rozdíly jednotlivých testerů nastávaly zejména při obecném pohledu na určování názorů. Jedním z problémů byla například analýza korpusu Milka, jelikož někdo považoval za názory pouze ty, které jasně reagovaly přímo na společnost, ale někteří braly jako názor na společnost také názory na její



Obrázek 6.2: Náhled doplňkové aplikace pro ruční výběr příspěvků obsahujících názor

produkty, což ve své podstatě není zcela špatný přístup, jelikož názory přímo na produkty firmy jsou názory, které nás zajímají. Tento problém je, jak se následně dozvíme, také problém automatického sumarizátoru, který má opačný problém u jiného typu dat. Statisticky se však dobrovolníci například počtem vybraných příspěvků lišili v rámci jednotek procent, což lze považovat za dobré.

6.2.2 Automatická sumarizace

Stejně jako u manuální sumarizace bylo k automatické sumarizaci použito vybraných pět korpusů. Do automatického sumarizátoru bylo postupně vloženo pět korpusů. Po automatické sumarizaci byl vždy vybrán určitý počet kladných i záporných příspěvků, které byly následně sumarizovány. Tyto výsledky pak byly použity pro porovnání výsledků manuální sumarizace.

Tester	Vybrané příspěvky	Procentuálně	Maximální váha
Tester 1	101	9%	505
Tester 2	168	15%	840
Tester 3	183	16,4%	915
Tester 4	146	13%	730
Tester 5	155	13,9%	775

Tabulka 6.1: Počet vybraných příspěvků jednotlivých testerů korpusu McDonald's

6.3 Výsledky

6.3.1 Manuální výběr subjektivních příspěvků

Nyní se dostáváme k vyhodnocení manuální sumarizace. Pět testerů postupně manuálně sumarizovalo výše zmíněných pět korpusů. Postupně se seznámíme s výsledky jednotlivých testerů a vzájemně porovnáme a vyhodnotíme jednotlivá témata (korpusy). V následujících tabulkách je možné vidět konkrétní čísla a procentuální vyjádření shodnosti anotací jednotlivých testerů.

Metoda určení míry úspěšnosti anotací je založena na určení váhy jednotlivých příspěvků. Nejprve byla tedy určena váha jednotlivých příspěvků podle toho, kolikrát byl daný příspěvek vybrán jako názor. Pokud tedy nějaký příspěvek označilo jako názor všech pět testerů, obdržel tento příspěvek váhu hodnoty 5. Pokud byl vybrán například dvěma testery, obdržel váhu hodnoty 2. Pokud nebyl příspěvek označen jako názor žádným testerem, zůstala jeho váha 0. Procentuální úspěšnost je pak určována pomocí součtu vah vybraných příspěvků, které tvoří určitou část z maximálního možného součtu.

McDonald's

Celkově tento korpus obsahoval 1118 příspěvků. Základní číselné vyjádření výsledků jednotlivých testerů vidíte v tabulce 6.1 a porovnání shody jednotlivých testerů vidíte v tabulce 6.2.

Tester	**Skóre**	**Procentuální úspěšnost**
Tester 1	415	81,6%
Tester 2	587	70%
Tester 3	654	71,5%
Tester 4	524	71,8%
Tester 5	598	77,2%

Tabulka 6.2: Porovnání shodnosti jednotlivých testerů u korpusu McDonald's

Tester	**Vybrané příspěvky**	**Procentuálně**	**Maximální váha**
Tester 1	75	6,7%	375
Tester 2	149	13,4%	745
Tester 3	123	11%	615
Tester 4	94	8,4%	470
Tester 5	104	9,3%	520

Tabulka 6.3: Počet vybraných příspěvků jednotlivých testerů korpusu Milka

Milka

Celkově tento korpus obsahoval 1117 příspěvků. Základní číselné vyjádření výsledků jednotlivých testerů vidíte v tabulce 6.3 a porovnání shody jednotlivých testerů vidíte v tabulce 6.4.

Slevomat

Celkově tento korpus obsahoval 1116 příspěvků. Základní číselné vyjádření výsledků jednotlivých testerů vidíte v tabulce 6.5 a porovnání shody jednotlivých testerů vidíte v tabulce 6.6.

Tester	**Skóre**	**Procentuální úspěšnost**
Tester 1	312	83,2%
Tester 2	590	79,2%
Tester 3	501	81,4%
Tester 4	367	78%
Tester 5	393	75,6%

Tabulka 6.4: Porovnání shodnosti jednotlivých testerů u korpusu Milka

Tester	**Vybrané příspěvky**	**Procentuálně**	**Maximální váha**
Tester 1	57	5,1%	285
Tester 2	76	6,8%	380
Tester 3	39	3,5%	195
Tester 4	78	7%	390
Tester 5	66	5,9%	330

Tabulka 6.5: Počet vybraných příspěvků jednotlivých testerů korpusu Slevomat

Tester	**Skóre**	**Procentuální úspěšnost**
Tester 1	156	54,7%
Tester 2	253	66,6%
Tester 3	135	69,3%
Tester 4	269	69%
Tester 5	229	69,4%

Tabulka 6.6: Porovnání shodnosti jednotlivých testerů u korpusu Slevomat

Vodafone

Celkově tento korpus obsahoval 1119 příspěvků. Základní číselné vyjádření výsledků jednotlivých testerů vidíte v tabulce 6.7 a porovnání shody jednotlivých testerů vidíte v tabulce 6.8.

Tester	**Vybrané příspěvky**	**Procentuálně**	**Maximální váha**
Tester 1	88	7,9%	440
Tester 2	146	13%	730
Tester 3	53	4,7%	265
Tester 4	106	9,5%	530
Tester 5	127	11,3%	635

Tabulka 6.7: Počet vybraných příspěvků jednotlivých testerů korpusu Vodafone

Tester	**Skóre**	**Procentuální úspěšnost**
Tester 1	295	67%
Tester 2	514	70,4%
Tester 3	198	74,7%
Tester 4	367	69,3%
Tester 5	436	68,7%

Tabulka 6.8: Porovnání shodnosti jednotlivých testerů u korpusu Vodafone

Tester	**Vybrané příspěvky**	**Procentuálně**	**Maximální váha**
Tester 1	75	6,6%	375
Tester 2	68	6%	340
Tester 3	54	4,8%	270
Tester 4	70	6,2%	350
Tester 5	76	6,7%	380

Tabulka 6.9: Počet vybraných příspěvků jednotlivých testerů korpusu ZOO Praha

ZOO Praha

Celkově tento korpus obsahoval 1132 příspěvků. Základní číselné vyjádření výsledků jednotlivých testerů vidíte v tabulce 6.9 a porovnání shody jednotlivých testerů vidíte v tabulce 6.10.

Tester	**Skóre**	**Procentuální úspěšnost**
Tester 1	288	76,8%
Tester 2	248	72,9%
Tester 3	193	71,5%
Tester 4	256	73,1%
Tester 5	280	73,7%

Tabulka 6.10: Porovnání shodnosti jednotlivých testerů u korpusu ZOO Praha

Korpus	**Vybrané příspěvky**	**Procentuálně**	**Maximální váha**
McDonald's	23	2%	115
Milka	27	2,4%	135
Slevomat	32	2,9%	160
Vodafone	34	3%	170
ZOO Praha	43	3,8%	215

Tabulka 6.11: Porovnání shodnosti anotace automatického sumarizátoru u jednotlivých korpusů

6.3.2 Automatický výběr subjektivních příspěvků

V tabulce 6.11 je vidět základní číselné vyjádření výsledků automatického sumarizátoru se zaměřením na část výběru příspěvků s názorem. Zde je nutné podotknout, že nastavení výběru záporných názorových příspěvků automatického sumarizátoru bylo provedeno následujícím způsobem:

```
((arr.get(i).n > 1)||(arr.get(i).hn > 1))
&&
((arr.get(i).p < 3)||(arr.get(i).hp < 3))
```

V překladu to znamená, že jako záporné názory byly vybrány příspěvky, jenž obsahovaly alespoň dvě slova ze slovníků záporných, ale zároveň však příspěvek nemohl obsahovat více než dvě slova ze slovníku kladných výrazů. U analýzy kladných názorů bylo použito principiálně stejné, ale samozřejmě zcela opačné nastavení.

Tabulka 6.12 pak vyjadřuje míru úspěšnosti automatického sumarizátoru z pohledu určování příspěvků obsahující názor.

Z výsledků můžeme vidět, že výběr názorů automatickým sumarizátorem se od ručního výběru liší. Z pozorování, testování a procházená příspěvků můžeme najít několik důvodů. Jako jeden z hlavních důvodů můžeme zmínit opravdu velké množství příspěvků, které ve své podstatě jsou názory, avšak ne přímo na danou společnost. Markantní je to například u Slevomatu nebo u ZOO Praha. U Slevomatu se objevuje opravdu velké množství názorů na služby nebo výrobky, které Slevomat v rámci své nabídky nabízí. Sumarizátor je ve své základní podstatě určuje jako názory, avšak jednotliví testeři správně tyto příspěvky neuznali jako názory na danou společnost. Myslím si,

Korpus	**Skóre**	**Procentuální úspěšnost**
McDonald's	68	59,1%
Milka	39	29%
Slevomat	33	21%
Vodafone	67	39,4%
ZOO Praha	46	21,4%

Tabulka 6.12: Porovnání shodnosti výběru příspěvků obsahující názor automatickým sumarizátorem u jednotlivých korpusů

že právě tohle je ten největší důvod, proč sumarizátor nedosahuje v oblasti určování názorů větší úspěšnosti.

Podobný případ je u korpusu ZOO Praha, kde se objevují spousty názorů na jednotlivá zvířata, avšak to ve své podstatě nejsou názory na ZOO jako takovou. Příkladem jeden příspěvek, který byl sumarizátorem určen jako kladný názor:"no to je něco nádhernýho, žirafátku přeji do života jen to nejlepší". Můžeme pozorovat, že se opravdu jedná o kladný názor. Ten je však zaměřen na mládě žirafy a ne na ZOO. A takovýchto případů se napříč všemi korpusy vyskytuje opravdu velké množství.

Na druhou stranu se musíme podívat na analýzu korpusu McDonald's, u kterého jsme se sumarizátorem dosáhli nejlepší shody. Zde je paradoxně důvod velice podobný, avšak efekt je zcela opačný. Na rozdíl od ostatních korpusů zde jednotlivý testeři nabyli dojmu, že názory na produkty společnosti McDonald's jsou názory přímo na společnost. Je to svým způsobem správná úvaha, jelikož tyto produkty společnost přímo ovlivňuje a vyrábí. Názory na kvalitu nebo chuť se tak zde objevují ve velkém množství a shoda testerů se sumarozátorem je větší.

6.3.3 Srovnání automatické sumarizace

Nyní se podíváme na výsledky automatického sumarizátoru a porovnáme jeho výsledky s výsledky jednotlivých testerů.

Co se týče konečné sumarizace automatického sumarizátoru v porovnání se sumarizací jednotlivých testerů, poslouží nám pro přehled následující tabulky. Pro experiment bylo použito dvojí nastavení automatického sumarizátoru. Jako první jsme použili omezení výsledné sumarizace na počet šedesáti

Korpus	**Příspěvků**	**Maximální váha**	**Skóre**	**Úspěšnost**
McDonald's	3	15	0	0%
Milka	3	15	2	13,3%
Slevomat	4	20	5	25%
Vodafone	4	20	0	0%
ZOO Praha	4	20	3	15%

Tabulka 6.13: Výsledky úspěšnosti sumarizátoru kladných názorů v porovnání s manuálně sumarizovanými korpusy s omezením sumarizace na délku 60 slov

Korpus	**Příspěvků**	**Maximální váha**	**Skóre**	**Úspěšnost**
McDonald's	3	15	1	6,7%
Milka	3	15	4	26,7%
Slevomat	3	15	4	26,7%
Vodafone	3	15	6	40%
ZOO Praha	3	15	0	0%

Tabulka 6.14: Výsledky úspěšnosti sumarizátoru záporných názorů v porovnání s manuálně sumarizovanými korpusy s omezením sumarizace na délku 60 slov

slov. Sumarizaci těchto kladných názorů můžete vidět v tabulce 6.13 a sumarizaci záporných příspěvků můžete vidět v tabulce 6.14. V druhém případě bylo nastaveno omezení na počet sta slov a výsledky můžete vidět pro kladné názory v tabulce 6.15 a pro záporné 6.16. Každý tester měl za úkol vybrat deset příspěvků (pět kladných a pět záporných), které dle jeho názoru nejvíce vystihují a sumarizují tak všechny příspěvky z daného korpusu.

Z analýzy všech výsledků automatického sumarizátoru můžeme hovořit o úspěšnosti maximálně 40%. Na druhou stranu musíme zdůraznit, že po procesu analýzy obsahu subjektivních názorů dosahovala aplikace úspěšnosti cca 75%. Je jasné, že z každým dalším procesním krokem aplikace dochází ke znepřesňování celkového výsledku a ačkoliv se můžou zdát naměřené hodnoty nízké, z osobního pohledu musím říci, že ruční výběr názorů do konečné sumarizace je opravdu velice subjektivní pohled na věc a navíc se může k danému tématu vyjadřovat několik podobných příspěvků. Nemusí tak být vůbec ojedinělý případ, kdy sumarizátor vybere do výsledného souhrnu jiný příspěvek než jednotlivý tester manuálně, avšak vypovídající hodnota může být hodně podobná. V našem měření se ovšem tyto výsledky počítají jako

Korpus	**Příspěvků**	**Maximální váha**	**Skóre**	**Úspěšnost**
McDonald's	5	25	1	20%
Milka	5	25	2	8%
Slevomat	5	25	4	16%
Vodafone	5	25	0	0%
ZOO Praha	3	15	2	13,3%

Tabulka 6.15: Výsledky úspěšnosti sumarizátoru kladných názorů v porovnání s manuálně sumarizovanými korpusy s omezením sumarizace na délku 100 slov

Korpus	**Příspěvků**	**Maximální váha**	**Skóre**	**Úspěšnost**
McDonald's	5	25	4	16%
Milka	3	15	4	26,7%
Slevomat	4	20	6	30%
Vodafone	5	25	9	36%
ZOO Praha	4	20	2	10%

Tabulka 6.16: Výsledky úspěšnosti sumarizátoru záporných názorů v porovnání s manuálně sumarizovanými korpusy s omezením sumarizace na délku 100 slov

chybné a tak mohou být z tohoto pohledu zkreslené.

7 Závěr

Výsledkem celé práce je fungující demo aplikace zaměřená na automatickou sumarizaci názorů na sociální síti Facebook. Tato aplikace byla důkladně testována v reálném prostředí a můžeme říci, že na první pohled působí důvěryhodně a výsledná sumarizace opravdu přibližně vypovídá o hodně probíraných tématech dané facebookové stránky. Ve výsledcích je také vidět správná vlastnost sumarizátoru o výběru příspěvků z různých témat.

Následně byl proveden experiment zaměřený na úspěšnost automatického sumarizátoru v porovnání se subjektivní anotací pěti testerů. Bylo vypozorováno, že ve fázi určování sentimentu dosahuje oproti manuální anotaci aplikace úspěšnosti přibližně 75% a úspěšnost výsledné sumarizace se pak pohybuje přibližně okolo 20%.

Při analýze výsledků experimentu také bylo vypozorováno, že určování sentimentu příspěvku je mnohem účinější u negativních příspěvků. Podobný trend je pozorován i u samotné sumarizace, i když ta je samozřejmě předchozím procesem ovlivněna. Důvodem může být například zcela jednoduché vysvětlení, že většina lidí se k příspěvku odhodlá zejména pokud se jim něco nelíbí nebo pokud mají nějaký problém. Také potom v takových případech používají více zabarvená slova, která pomáhají k lepší analýze jednotlivých příspěvků a tím i k lepší účinnosti celé aplikace.

7.1 Vylepšení výsledků sumarizace

V rámci algoritmu pro určování sentimentu a sumarizace názorů můžeme nejspíše dosáhnout dalšího vylepšení výsledků. Pro příklad zmiňuji několik možností, které by při zapracování do algoritmu měly pomoci:

1. **Klíčová slova** - před sumarizací dané facebookové stránky bychom mohli nastavit speciální klíčová slova specifická pro dané téma a přidělit jim speciální váhu. Základním klíčovým slovem by zajisté byl název firmy nebo výrobku, který stránka zastupuje. Tím bychom zajistili upřednostnění vět, které v sobě obsahují přímo název stránky, což již často znamená, že se jedná o příspěvek s názorem. Podobně můžeme zařadit další klíčová slova specifická například pro služby nebo výrobky

společnosti, což bude mít velice podobný efekt. Příkladem můžeme zmínit klíčové slovo "vstupné"u facebookové stránky ZOO Praha.

2. **Specifická nevýznamná slova** - kromě filtrace stopwords se můžeme u konkrétních stránek zaměřit na slova, která se nejspíše budou v příspěvcích vyskytovat často, nicméně nás tato slova přímo nezajímají a výslednou sumarizaci bychom zbytečně zkreslili. Pokud opět vezmeme v úvahu facebookové stránky ZOO Praha, můžeme se zaměřit na jména nejdiskutovanějších zvířat s tím, že příspěvky obsahující nějaké z daných jmen bude obsahovat názor pouze na dané zvíře, což nás primárně nemusí zajímat.

3. **Emotikony** - na sociálních sítích se velice často používají pro vyjádření názoru emotikony, což je specifická posloupnost symbolů vyjadřující náladu nebo aktuální postoj. Pokud bychom tyto speciální sekvence znaků zařadili do slovníku, mohli bychom dosáhnout přesnějších výsledků v hodnocení sentimentu.

4. **Specifické sekvence slov** - další možností zpřesnění výsledků je potlačení známých sekvencí slov, které obsahují slova z některého slovníku, ale již dopředu víme, že na výsledek nemají žádný vliv, ba naopak výsledky zkreslují. Příkladem může být například pozdrav "dobrý den", který v sobě obsahuje pozitivní přídavné jméno, nicméně na reálný výsledný sentiment příspěvku nemá žádný vliv. Zajisté lze však nalézt případy opačné, kdy bychom naopak měli určité sekvence slov, jejich slova se nevyskytují ve slovnících, zvýhodnit. Příkladem uveďme spojení "to je ono", které se skládá ze samých stopwords, ovšem dohromady v tomto znění může značit kladné zabarvení příspěvku.

5. **Překlepy a nespisovné výrazy** - To je další velice výrazný problém analýzy textů napsaných lidmi zejména na sociálních sítích, diskuzích a podobných méně kontrolovaných místech. Jedním způsobem pro odstranění těchto nepřesností je určitě analýza a následné zanesení přesných překlepů, zkratek nebo nespisovných výrazů. To nám může do určité míry pomoci a můžeme se tak vyvarovat zkreslení alespoň u velice známých překlepů nebo výrazů. Ostatní náhodné jevy ale mohou být pro správnou analýzu velice složité a někdy možná i nemožné, pokud jejich pravý význam nerozpozná ani člověk, který oproti počítači chápe kontext celého příspěvku.

6. **Negace** - Jako konkrétní příklad uvádím slovní spojení "není to špatné", což je v principu kladné hodnocení, nicméně analyzátor založený na slov-

níkových metodách vidí standardně dvě záporná slova a převážně vyhodnocuje příspěvek jako záporný, což je špatně. Pro vylepšení by mohl být do algoritmu analýzy sentimentu přidán algoritmus pro analýzu a váhování negací.

Literatura

[Balahur et al.(2012)] BALAHUR, A. et al. Challenges and Solutions in the Opinion Summarization of User-generated Content. *J. Intell. Inf. Syst.* October 2012, 39, 2, s. 375–398. ISSN 0925-9902.

[Barzilay(1997)] BARZILAY, E. M. R. Using Lexical Chains for Text Summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, s. 10–17, 1997.

[Corinna et al.(1995)] CORINNA, C. et al. Support-Vector Networks. *Mach. Learn.* September 1995, 20, 3, s. 273–297. ISSN 0885-6125.

[Ferro et al.(2005)] FERRO, L. et al. TIDES 2005 Standard for the Annotation of Temporal Expressions. In *Technical Report*. The MITRE Corporation, 2005.

[Gimpel et al.(2011)] GIMPEL, K. et al. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, s. 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6.

[Habernal et al.(2013)] HABERNAL, I. et al. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, s. 65–74, Atlanta, Georgia, June 2013. Association for Computational Linguistics. Dostupné z: `http://www.aclweb.org/anthology/W13-1609`.

[Jezek et al.(2008)] JEZEK, K. et al. Automatic text summarization. In *The state of the art 2007 and new challenges*. ZČU, 2008.

[Kupiec et al.(1995)] KUPIEC, J. et al. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, s. 68–73, New York, NY, USA, 1995. ACM. ISBN 0-89791-714-6.

[Luhn(1958)] LUHN, H. P. The Automatic Creation of Literature Abstracts. *IBM J. Res. Dev.* April 1958, 2, 2, s. 159–165. ISSN 0018-8646.

[Manning et al.(2008)] MANNING, C. D. et al. *Introduction to Information Retrieval*. New York, NY, USA : Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

[Marek Nekula(2003)] MAREK NEKULA, K. *Příruční mluvnice češtiny*. Brno : NLN, s.r.o., 2003. ISBN 80-7106-134-4.

[Marek Nekula(2002)] MAREK NEKULA, K. *Encyklopedický slovník češtiny*. Brno : NLN, s.r.o., 2002. ISBN 987-80-7106-484-8.

[Meyer(2001)] MEYER, T. U. W. D. Support Vector Machines. The Interface to libsvm in package e1071. Online-Documentation of the package e1071, 2001.

[Salton(1988)] SALTON, G. (Ed.). *Automatic Text Processing*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 1988. ISBN 0-2:1-1227-8.

[Steinberger(2009)] STEINBERGER, J. K. J. Evaluation Measures for Text Summarization. *Computing and Informatics*. 2009, 28, 2, s. 251–275.

[Steinberger et al.(2011)] STEINBERGER, J. et al. Aspect-Driven News Summarization, 2011.

[Steinberger et al.(2012)] STEINBERGER, J. et al. Towards language-independent news summarization. In *Proceedings of the Text Analysis Conference 2011*. NIST, 2012.

[Taboada et al.(2011)] TABOADA, M. et al. Lexicon-based Methods for Sentiment Analysis. *Comput. Linguist.* June 2011, 37, 2, s. 267–307. ISSN 0891-2017.

[Tanev et al.(2008a)] TANEV, H. et al. Real-Time News Event Extraction for Global Crisis Monitoring. In *Proceedings of the 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems*, NLDB '08, s. 207–218, Berlin, Heidelberg, 2008a. Springer-Verlag. ISBN 978-3-540-69857-9.

[Tanev et al.(2008b)] TANEV, H. et al. Weakly Supervised Approaches for Ontology Population. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*, s. 129–143, Amsterdam, The Netherlands, The Netherlands, 2008b. IOS Press. ISBN 978-1-58603-818-2.

[Turchi et al.(2010)] TURCHI, M. et al. Using Parallel Corpora for Multilingual (Multi-document) Summarisation Evaluation. In AGOSTI, M. et al. (Ed.) *Multilingual and Multimodal Information Access Evaluation*, 6360 / *Lecture Notes in Computer Science*, s. 52–63. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15997-8.

[Zapotocky(2012)] ZAPOTOCKY, P. Stemmer pro češtinu. ZČU, 2012.

[Zhang(2004)] ZHANG, H. The Optimality of Naive Bayes. In BARR, V. – MARKOV, Z. (Ed.) *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. AAAI Press, 2004.

A Uživatelská dokumentace

Aplikaci spustíte otevřením souboru Sumarizator.jar, ideálně ve složce, ve které je umístěn. Pokud budete chtít apliakci přesunout ze zdrojové složky jinam, musíte na stejné místo umístit i slovníky, které zajišťují správnou funkci aplikace (*cs_hn.txt*, *cs_hp.txt*, *cs_n.txt*, *cs_p.txt*, *stoplist.txt*). Pokud tyto slovníky nebudou ve stejné složce jako zdrojová aplikace, nebude její funkčnost správná. Po otevření okna je požadováno nastavení minimálně jednoho, maximálně však tří údajů.

Jediná povinná položka pro správnou funkčnost je vyplnění tzv. tokenu, což je časově omezený klíč pro aktivací rozhraní Facebook API. Vložte tedy token do textového pole v pravém horním rohu. Získáte ho zkopírováním z webé stránky https://developers.facebook.com/tools/explorer.

Druhou již nepovinně upravovanou položkou je ID stránky. Jedná se o unikátní číslo charakterizující danou facebookovou stránku, kterou chcete sumarizovat. Defaultně je v aplikaci nastaveno ID facebookové stránky portálu iDnes.cz. Pokud chcete sumarizovat jinou facebookovou stránku, zjistěte si její ID na webové stránce http://findmyfacebookid.com/. Provedete to tak, do jednoduchého formulářového pole vložíte domovskou URL vámi vybrané facebookové stránky a potvrdíte tlačítkem "Lookup numeric ID". Vygenerované ID pak vložte v aplikaci do textového pole "ID stránky"v levém horním rohu.

Poslední nastavitelnou položkou je omezení délky výsledného souhrnu v závislosti na počtu slov. Toto nastavení se provádí v pravém dolním rohu aplikace a defaultně je nastaveno na 500 slov. Vy toto nastavení můžete kdykoliv měnit.

Pokud jste si nastavili všechny možné parametry podle vašich představ, stačí stisknout tlačítko "Summarize"umístěné v pravém dolním rohu. Po chvilce se vám již ukáží dostupné výsledky a statistiky. V seznamu na levé části vidíte všechny extrahované příspěvky a komentáře. Navíc nad tímto seznamem můžete vidět stejně jako u ostatních seznamů počet. V prostřední části pak vidíte nahoře seznam kladných názorů a pod ním jejich sumarizaci. V pravé části je pak nahoře seznam se zápornými názory a pod ním seznam záporné sumarizace.