VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

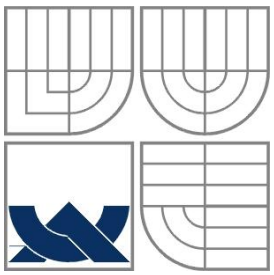
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

METODY STROJOVÉHO UČENÍ VE ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA

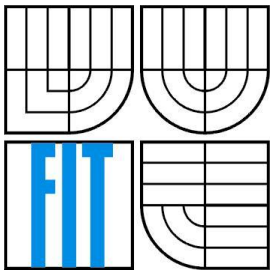
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE JAN VODIČKA
AUTHOR

BRNO 2012



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

METODY STROJOVÉHO UČENÍ VE ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA

MACHINE-LEARNING METHODS IN NATURAL LANGUAGE PROCESSING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

JAN VODIČKA

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2012

Abstrakt

Bakalářská práce se zajímá hodnocením sentimentu textu v českém jazyce za pomocí metod strojového učení, hlavně za použití naivního bayesovského klasifikátoru. Členění probíhá do dvou kategorií – pozitivní, negativní zprávy. Jako datové zdroje pro automatické vytvoření korpusu jsou použity zprávy ze sociální sítě Twitter, zbožového porovnávače Heuréka, filmové databáze ČSFD a restauračního portálu Scuk. Jsou porovnány z hlediska výkonnosti při hodnocení sentimentu. Následně je sestavena výsledná tréninková sada, která je použita při hodnocení zpráv z Twitteru v téměř reálném čase.

Abstract

Bachelor's thesis deals with sentiment analysis using machine learning methods, mainly naive bayes classifier. Input text can be classified as positive or negative message. There are used several data sources for create of automatic annotated corpus – social network Twitter, price comparator Heureka, movie database ČSFD and restaurant portal Scuk. These sources are compared in terms of performance in assessing the sentiment. Consequently, the final training dataset is created and used at almost real-time Twitter sentiment analysis.

Klíčová slova

Strojové učení, naivní bayesovský klasifikátor, hodnocení sentimentu, dolování názorů, Twitter

Keywords

Machine learning, naive Bayes classifier, sentiment analysis, opinion mining, Twitter

Citace

Vodička Jan: Metody strojového učení ve zpracování přirozeného jazyka, bakalářská práce, Brno, FIT VUT v Brně, 2012

Metody strojového učení ve zpracování přirozeného jazya

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Doc. RNDr. Pavla Smrže, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jan Vodička
16.5.2012

Poděkování

Rád bych poděkoval vedoucímu mé bakalářské práce Doc. RNDr. Pavlu Smržovi, Ph.D. za odborné vedení mé práce, Ing. Lubomíru Otrusinovi, Bc. Stanislavu Hellerovi a Mgr. Petru Škodovi za technickou výpomoc. Také bych rád poděkoval členům mé rodiny, kteří byli trpěliví a účastnili se na ruční anotaci zpráv. Děkuji.

Obsah

1 Úvod

S rostoucím počtem obyvatel na planetě jakožto potenciálních zákazníků, zvyšujícím se počtem produktů a služeb je vhodné vědět „co si lidé myslí“. Dostupnost výpočetní techniky a obrovský rozmach internetu v posledních 15 letech učinil svět rychlejším. Informace, které je možné získat během několika okamžiků je nepřeberně mnoho a neustále se jejich množství a rychlost, kterou přibývají, zvyšuje. Nedílnou součásti lidského chování a myšlení je nutnost nejen vstřebávat cizí názory a zážitky, ale zároveň je dávat najevo. S tím roste potřeba tyto informace sbírat a třídit tak, aby budoucí produkty výrobců plnily požadavky zákazníků.

Jedním z hlavních faktorů, které ovlivňují úspěšnost produktu, je fakt, zda se lidem daný produkt jednoduše líbí anebo nelíbí. Určitě každý z nás před samotným nákupem někdy zjišťoval od lidí ve svém okolí, kteří daný výrobek či službu používají, jejich zkušenosti. Web, jaký dnes známe, nám dává mnohem širší možnosti, a to, zjistit si o produktu detailní informace a názory od lidí, které ani neznáme.

Tato práce se zabývá možnostmi automatizovaného hodnocení sentimentu (nálady) textových zpráv z různých internetových zdrojů výhradně v českém jazyce. Ačkoliv hlavní náplní mělo být zjišťování nálady zpráv čistě ze sociální sítě Twitter, pro zvýšení přesnosti a generalizace klasifikátoru je použit další obsah z prostředí českého internetu. Konkrétně se jedná uživatelské recenze z následujících zdrojů:

- Filmový portál Česko-Slovenská filmová databáze (www.csfd.cz)
- Zbožový porovnávač cen Heuréka (www.heureka.cz)
- Databáze českých restaurací Scuk (www.scuk.cz)

Shromážděná data poskytují relativně velký rozsah témat, která zajímají většinu z nás. O to se zasloužil hlavně porovnávač cen Heuréka, jenž obsahuje nejen desítky tisíc recenzí samotných produktů, ale i internetových obchodů.

2 Pojmy

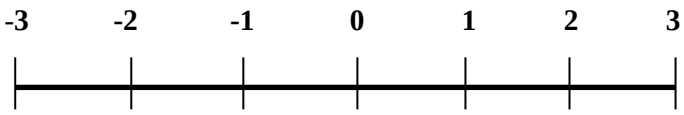
Většina z nás instinktivně chápe, co znamená, že je něco „pozitivní“ nebo „negativní“. Není však na škodu si některé termíny upřesnit.

2.1.1 Sentiment

Současné numerické počítače umí velmi dobře počítat s nespojitými hodnotami. Jak ale postihnout myšlenky, náladu (sentiment) člověka?

V roce 1932 se objevuje tzv. *Likertova škála*, jež je dodnes nejpoužívanější a nejspolehlivější technikou pro měření postojů [1]. Jedná se o pětibodovou škálu od „absolutně nesouhlasím“, přes středovou hodnotu „nevím“ ke „zcela souhlasím“. Vzhledem k tomu, že poskytuje jenom jednu dimenzi odpovědí, získané reakce od respondenta vedou pouze ke hrubým znalostem. Později, v roce 1957, vzniká termín *sémantický diferenciál* [2]. Popisuje náladu jako stupnici, která je diskrétní a pochopitelná pro většinu kultur. Hlavními nositeli sentimentu jsou přídavná jména.

Stupnice je rozdělená na 7 dílů. Vypadá následovně:



Obrázek 1 – stupnice sémantického diferenciálu

Hodnocení je očíslované a znamená:

- 0 = ani trochu
- 1 = trochu
- 2 = dostatečně
- 3 = velmi

Záporné hodnoty vyjadřují negativní postoj.

Autoři také objevili, že lidé v drtivé většině případů používají tři opakující se dimenze k vyhodnocení pojmů a frází:

1. Hodnocení (posouzení pomocí dvoupólových přídavných jmen, např. dobrý versus špatný)
2. Sílu výpovědi (účinek výpovědi je silný, slabý), která se dál dělí:
    a. vztah autora k tématu
    b. specifičnost (forma formulace: jasná, vágní)
    c. určitost (autor si je jistý nebo je na pochybách?)
3. Intenzita (emotivnost výpovědi)

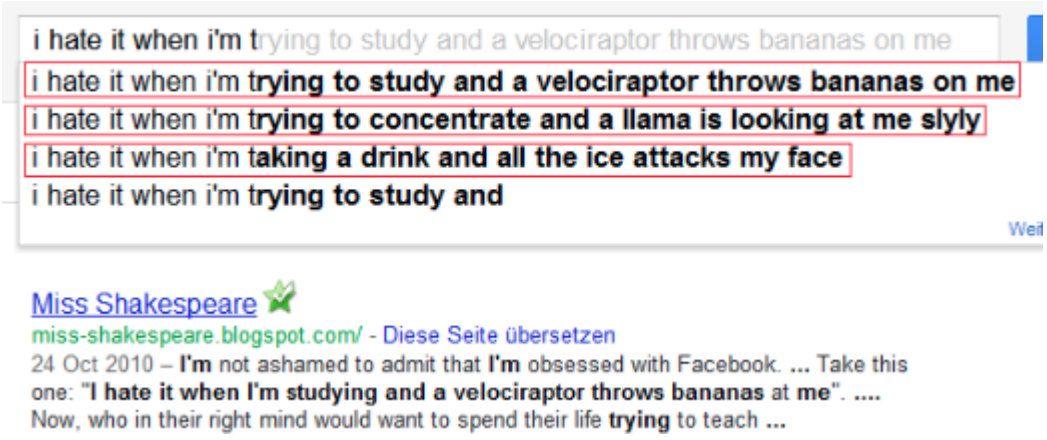
Výše uvedené předpisy lze již poměrně dobře aplikovat pro automatizovanou analýzu, kde vycházíme z diskrétních dat.

2.1.2 Analýza sentimentu

Samotná analýza sentimentu je tedy odhalování nálad jisté skupiny lidí se společným zájmem. Využívá sociologické, psychologické a marketingové techniky. Typickými příkladem jsou zákazníci vyjadřující se ke kvalitě konkrétního výrobku nebo obchodníci na burze. Na to může například navazovat predikce jejich budoucího chování, zda si spotřebitel koupí budoucí verzi produktu nebo zda obchodníci budou své cenné papíry prodávat nad cenou či pod cenou.

Již desítky let se s úspěchem používají dotazníky všech druhů, ať už se jedná o online dotazníky, vedené rozhovory apod. Většinou jsou zaměřené na konkrétní témata s konkrétními otázkami [3]. Podávají velmi dobré kvalitativní výsledky a s jejich využitím můžeme počítat i nadále. Jejich problémem je, že lze velmi snadno zkreslit některé odpovědi a často se tak v určitých oblastech děje. Typickým příkladem jsou hlavně otázky zasahující do osobního života, kde respondenti zkreslují své odpovědi tak, aby „lépe vyzněly“ pro okolí [4]. Je důležité podotknout, že tento neduh se promítá do automatizovaných analýz. Odpovědi dotazníků nejsou většinou bipolární, ale pro získání větších znalostí je možností několik.

S tím příliš nekorespondují dosavadní pokusy o systémy samočinné analýzy. Ty nejsou schopné poskytnout dostatečně přesné a detailní odpovědi jen na základě dat z blogů, internetových fór a recenzí, ale zaměřují se hlavně na rychlé, masové vyhodnocení čítajících až několika miliónů odpovědí s výstupem hodnot směřující k odpovědím typu „líbí/nelíbí“. Není tedy možné zjistit *„proč se výrobek líbí lidem ve věku…, kteří …, za podmínek, že … ale jen když …"*. Následující obrázek výstižně tuto skutečnost dokumentuje, jak konvenční systémy založené na strojovém učení fungují:



Obrázek 2 – Automatické doplňování dotazu na www.google.com (zdroj: googlefail.net)

2.1.3 Valence (psychologie)

Termín *valence* v oboru psychologie se většinou váže k diskuzi o pocitech člověka. Říká nám, zda nějaká událost, situace, objekt vyvolává přitažlivé pocity (pozitivní valence) nebo odtažité pocity, odpornost (negativní valence) [5]. Pojmy jako „strach“, „násilí“ mají negativní valenci, zatímco „smích“, „štěstí“ mají pozitivní valenci.

3 Historie a současné nástroje

První práce související s tímto tématem se zabývají detekcí, zda je článek subjektivní či objektivní [6]. Kolem roku 2001 se automatická analýza sentimentu dostává do popředí zájmu [7]. Vyhledávač Google Scholar však na dotazy *„opinion mining“* a *„sentiment analysis“* zaznamenává nízký počet prací až do roku 2005 (včetně), a to 96. Nyní (6. 4. 2012) jich indexuje přibližně 3640.

Důvodů vzestupu je několik (volně parafrázováno z [7]):

1. Dostupnost dat, díky rozvoji internetu, na kterých lze klasifikátory trénovat. Začíná se rozmáhat fenomén, kdy obsah samotných webových stránek je tvořen uživateli (Web 2.0). Jedná se hlavně o diskuzní fóra, chaty, blogy, agregátory uživatelských recenzí a hodnocení.
2. Zvýšení zájmu o metody strojového učení ve zpracování přirozeného jazyka, což úzce souvisí s prvním bodem.
3. Snaha tyto zdroje dat nejen komerčně využít.

Množství z nich se zaměřuje právě jen na samotné hodnocení sentimentu a nálady. Jejich výstupem je většinou škála tří kategorií „negativní“, „neutrální“, „pozitivní“, číselná hodnota udávající pravděpodobnost, ke které třídě zkoumaný objekt (slova, věty, odstavce, články nebo dokonce i celé weby) patří nebo širší číselná škála -5 až +5. Je jasné, že první zmíněné dělení ovlivňuje práh, který nám bude dané názory rozdělovat. Jeho nastavení je poté zásadní a je velmi ovlivněn kvalitou trénovacích dat. Přesnost těchto systémů kolísá v průměru od 65% do 90%. Je zajímavé, že větší přesnosti dosahují systémy, které pracují s datasety hodnocení automobilů než recenzí filmů [8].

Některé nástroje se snaží hodnocení rozšířit a vystihnout autorovu náladu jako „naštvaný“, „veselý“, „smutný“ apod. Tyto pokusy však většinou končí s relativně horšími výsledky v porovnání s bipolárním dělením na zprávy pozitivní a negativní. Jejich přesnost se pohybuje kolem 60% [9]. V drtivé většině totiž využívají kontroverzní rozdělení dat do tříd většinou dle charakteristických smajlíků a unigramů či bigramů, což se ukazuje pro jemnější dělení nálad jako nedostatečné.

Kromě naivního bayesovského klasifikátoru bylo studováno množství jiných jako je Support Vector Machines (SVM), rozhodovací stromy, metoda maximální entropie apod., nicméně téměř ve všech případech podával nejlepší výsledky právě ten bayesovský, viz [10], [11].

Převážná část prací pracuje s anglickým jazykem hlavně díky dostupnosti nástrojů a dat, které lze při řešení toho problému využít.

4 Analýza datových zdrojů

Velmi podstatnou částí v každé klasifikační úloze, kde se používají metody strojového učení s učitelem, je kvalita trénovacích dat [12]. Toto tvrzení dokládá i tato práce.
Následuje popis zdrojů, ze kterých jsem čerpal trénovací a testovací data.

4.1 Twitter

Sociální síť Twitter (www.twitter.com) vznikla v roce 2006. Aktuálně má síť 200 miliónu registrovaných uživatelů. Aktivních, těch, kteří se připojí, je zhruba polovina z nich. Každý den se na Twitter přihlásí 50 milionů uživatelů a ti denně napíšou přibližně 230 milionů tweetů [13]. *Tweet* je textová zpráva, která je limitována na 140 znaků. Twitter byl totiž koncipovaný v době svého založení jako SMS chat.

Nepřihlášený návštěvník může sledovat pouze příspěvky registrovaných uživatelů. V případě registrovaného uživatele obdržíte tzv. Timeline (počeštěnou jako „zeď"). Je to jednoduše místo, kde se objevují publikované zprávy ostatních registrovaných uživatelů, které sledujete, chronologicky seřazené od nejnovějších po nejstarší. Síla Twitteru a spousty dalších sociálních sítí spočívá v tom, že vám umožňují sledovat příspěvky jen těch lidí, kteří vás zajímají. Twitter nazývá uživatele, kteří sledují vaše příspěvky, jako „followers" (nepřekládá se, zažilo se jako „*folouveři*") a „following", což jsou uživatelé, jejichž příspěvky sledujete vy.

Veřejné příspěvky[1] mohou být tří (čtyř) typů:

1. Normální zprávy
2. Mentions (česky „zmínky") – zprávy, které zmiňují některé jiné uživatele pomocí označení se zavináčem: @jmeno_uzivatel.
3. Replies (česky „odpovědi") – odpovědi na jiné zprávy
4. Retweets – zpráva napsána jiným uživatelem, která je přeposlána. Ačkoliv Twitter řadí tyto zprávy mezi normální, existuje možnost, jak programově zjistit, že jde o retweet.

V obsahu zprávy se může objevovat ještě další entita, které Twitter rozeznává a umožňuje dle nich hledat další zprávy:

- Hashtag – slovo, které začíná mřížkou. V případě, že se ve zprávě vyskytuje, zmiňuje většinou subjekt, o kterém její obsah pojednává. Uživatelé jím však často nahrazují slova ve větě. Příklad: „*nový layout #skype klienta se mi teda fakt nelíbí.... #fail*"

[1] Twitter poskytuje i prostředí pro odesílání soukromých zpráv.

4.1.1 Český obsah

Vody českého Twitteru mapuje nejlépe asi služba Klábosení (www.klaboseni.cz). Přes vysoký počet uživatelů je těch českých a slovenských odhadován pouze na 96000. Kritériem pro indexaci však je pouze 15% tweetů v národním jazyce nebo přihlášení se k České či Slovenské republice v nastavení časového pásma.

Při rozhodování, zda vůbec použít jako jeden z datových zdrojů Twitter, jsem provedl analýzu 3000 náhodných Tweetů a dospěl k následujícím zjištěním. Dle analýzy je česky pouhých 1830 (61%) zpráv. Zbytek je psán anglicky, rusky či německy, i když uživatel je očividně dle jména a občasných příspěvků Čech. Dalším hojně zastoupeným jazykem je slovenština. Zbytek jsou pouze URL adresy či smajlíci. Z těchto 1830 zpráv je (rozřazení dle mého subjektivního usouzení):

- Reklamní sdělení (až spam) – 3%
- Retweety – 6%
- Zprávy (informace o teplotě, články z novin) – 6%
- Konverzace – 30%
- Ostatní jako jsou pozdravy, sdělení a „*social grooming*“ (překládáno jako utužování sociálních vztahů, přesný termín jsem nebyl schopný najít) – 55%

Tyto čísla hrubě odpovídají jiným studiím, které byly zaměřeny na anglický obsah [14]. Analýza 2000 tweetů:

- Spam – 4%
- Novinové zprávy – 3%
- Reklamní sdělení – 6%
- Retweety – 9%
- Konverzace – 38%
- Social grooming – 40%

Je příjemné vidět, že český Twitter není místem, kde narazíte na každém rohu na komerci či spam. To je důležité zjištění i pro další projekty, které budou chtít využít Twitter jako zdroj dat. Nicméně s přibývajícím počtem uživatelů (dle [15] devadesát místních účtů denně, tempo se navíc zrychluje) lze očekávat, že se poměr změní a vyrovná tomu anglickému v horizontu tří až čtyř let. Výskyt těchto sdělení je zde z důvodu, že lidé často sledují firemní účty společností, jejichž služby používají. Nejen že sledují novinky, ale dostávají tak občas slevy nebo speciální nabídky, které jsou vyhrazené právě pro tyto „věrné“ zákazníky.

Často jejich prostřednictvím lze také s firmou komunikovat na osobnější úrovni a řešit některé problémy. Je zde totiž velmi účinné „si stěžovat“ [16]. Tyto zprávy jsou pro analýzu sentimentu

poměrně vděčné, jelikož často obsahují smajlíky, přídavná jména a v drtivé většině jsou velmi upřímné, jasné a negativního ražení. Příklad (odstraněna vulgární slova):

*Podruhý ve dvou dnech odcházim z Vodafonu *** do nepříčetna. To snad neumí ani Český dráhy. (Pošta jo, to už vim.)*

Nebo

@Vodafone_CZ doporučuji přejmenovat Vodafone Samoobsluhu na Vodafone čekárnu. Net mám rychlej a stejně pořád čekám, grrrrr

Překvapením je podíl novinových článků, resp. jejich úryvků + odkaz na celé znění. Velká část uživatelů očividně svoji timeline používají podobně jako RSS čtečku, jelikož počet followerů, kteří sledují zpravodajské účty, se pohybuje okolo čísla 15000. Obsah tweetu je tak neúplný, často negativní. Pro samotnou analýzu jsou vhodné v případě, že budou ručně anotovány. Příklad zprávy:

„Zpravy: ParlamentniListy.cz: Exprimátor Liberce promluvil: Jedni tunelují, druzí se uskromňují. Nesmysl: Přesně před rok..... „

Menšinou mezi zprávami samotných uživatelů jsou samotné konverzace, které jsou velmi krátké. Poměrně často se jedná o otázku hozenou takzvaně *„do pléna“*, kdy si uživatel neví rady a potřebuje pomoc. Odezvou je často jedna odpověď od jednoho nebo více uživatelů. Je zajímavé, že odpovědi jsou většinou věcné a obsahují buď doporučení, nebo varování, jedná se ale většinou o osobní zkušenost samotného uživatele doplněnou URL adresou.

Nejpočetnější skupinou jsou oznámení, zážitky z cest, restaurací, kdy jsou zprávy psané prostřednictvím mobilních telefonů, nebo se chce autor s něčím pochlubit. Odpovědi na ně buď neexistují, nebo jsou typu: *„to je hezké“*, *„gratuluji“*.

4.1.2 Vytvoření korpusu

Vytvoření anotovaného korpusu, kde budou dvě třídy – negativní a pozitivní, jsem provedl dle výskytu smajlíků s pozitivní a negativní valencí, tak jak je nejspíše vnímá většina lidí. Vycházel jsem z několika seznamů, které jsem našel pod hesly „seznam smajlíků“ apod. a vybral ty s velkým počtem výskytů.

Příklady:

Pozitivní: :) :-) :]}] ;D ;==))

Negativní: :(:-(:[[:===[[[

Z příkladu jde vidět, že jsou smajlíci velmi proměnliví, je časté střídání znaků a jejich délka je též variabilní. Pro správnou identifikaci je použito regulárních výrazů, viz Příloha 1. Regulární výrazy

pro klasifikaci zpráv dle smajlíků. Zohledňují proměnlivou délku, střídání znaků, otočení smajlíků o 180 stupňů a do určité míry i kontext – aby neproběhla záměna s běžnou interpunkcí. Například (všimněte si „smajlíku“ za slovem „poradte“):

dilema (poradte): nainstalovat na ntb windows 7 32 nebo 64 bit?

4.1.3 Ruční anotace

Pro ruční anotaci vznikla jednoduchá webová aplikace (viz kapitola 6.6.5 Anotační aplikace pro zprávy z Twitteru), které hodnotiteli předkládá pouze české ještě neanotované tweety. Mohou být označeny jako pozitivní, negativní, neutrální (není ale v práci použito) nebo je přeskočit v situaci, kdy je anotátor na pochybách. Tyto tweety budou sloužit výhradně jako testovací sada.

Ručně bylo ohodnoceno zhruba 2000 zpráv mnou a dalšími dvěma příslušníky mé rodiny. Zprávy byly hodnoceny bez znalosti kontextu okolních zpráv - například rozhovoru, i když aplikace zobrazení tweetu, na něž se odpovídá, umožňuje. V případě, že se názor shodoval alespoň ve dvou případech, byl označen jako konečný. To ale nekoresponduje s fungováním anotační aplikace – ukázalo se totiž, že je vhodnější zaměřit se na kvantitu ohodnocení, protože buď se na sentimentu zprávy shodla většina, nebo každý zvolil jinou možnost. Bylo by proto vhodné, v rámci dalšího vývoje, zapojit větší počet lidí z různých kulturních prostředí.

Takovýto experiment [17] v českém prostředí byla provedena agenturou Ataxo, jež provozuje službu Klábosení zmiňovanou výše. V prvním kole předložila 30 lidem, využívající Twitter, náhodně vybraných zpráv a měli určit její celkový sentiment (pozitivní, neutrální, negativní), v druhém se měli zaměřit na určité klíčové slovo a poté ohodnotit. Výsledky jsou následující a to (převzato z komentáře ke studii [18]):

„V prvním kole totiž byla 70% shoda o sentimentu zmínky pouze v 30 případech z 90, tedy v jedné třetině! V druhém kole, kdy respondenti znali monitorované klíčové slovo, byla shoda o něco lepší, 70 % uživatelů se shodlo v 43 případech z 90. Tedy ani ne v polovině případů.“

Nicméně po střízlivém zhlédnutí předložených zpráv je třeba podotknout, že asi jejich polovina byla neúčelná a znalost jejího sentimentu neměla žádný užitečný a praktický dopad. Tento tweet ohodnotilo 50% anotátorů jako pozitivní, 39% negativní a 11% neutrální:

„Co to, žes přijela? Já myslel, že při stávce busy nejedou… tak já v to doufala…přijel řidič, tak jsem nastoupila a ptám se ho proč taky nestávkuje? A co on na to? Prej: „Já stávkuju!“ A dal mi to zadarmo :D“

Naproti tomu tento tweet označilo všech 100% lidí jako negativní:

> *„Česká spořitelna – super banka, kde nikdo není ochotný a možná nemá potuchy co vám dělá na účtě“*

4.2 Heuréka

Dalším místem, které je velmi vyhovující pro automatickou analýzu sentimentu je zbožový porovnávač cen Heuréka (www.heureka.cz). Spousta prací na toto téma, které se začaly objevovat po roce 2001, pracovaly právě s těmito typy serverů nebo podobnými, kdy je v internetovém obchodě pod prezentací produktu možnost výrobek hodnotit. Důvodů lze nalézt hned několik:

- Většinou pevná struktura samotné stránky s výrobkem, kterou lze dobře strojově a bezchybně zpracovávat
- Obsahuje hodnocení samotnými uživateli většinou ve formě pěti až šesti stupňové škály, která je vyjádřena hvězdičkami nebo procenty
- Krom bodového hodnocení je většinou dostupné textové hodnocení, které je navíc rozděleno na dvě části – pozitivní a negativní zkušenosti. To je obrovská výhoda, jelikož nám tak automaticky vzniká anotovaný korpus.

Server Heuréka poskytuje prostor nejen pro hodnocení výrobků, ale i samotných obchodů, kde zákazník provede nákup. Databázi výrobků dodávají samotné obchody, které se musí zaregistrovat, spolu se svým popisem a parametry zboží. V době stahování (3-5. března 2012) čítala cca 200 000 výrobků v 20 000 obchodech. Nabízené zboží eshopů je poloautomaticky agregováno a v databázi v podstatě neexistují duplicity, které by vznikaly z chybného zařazení.

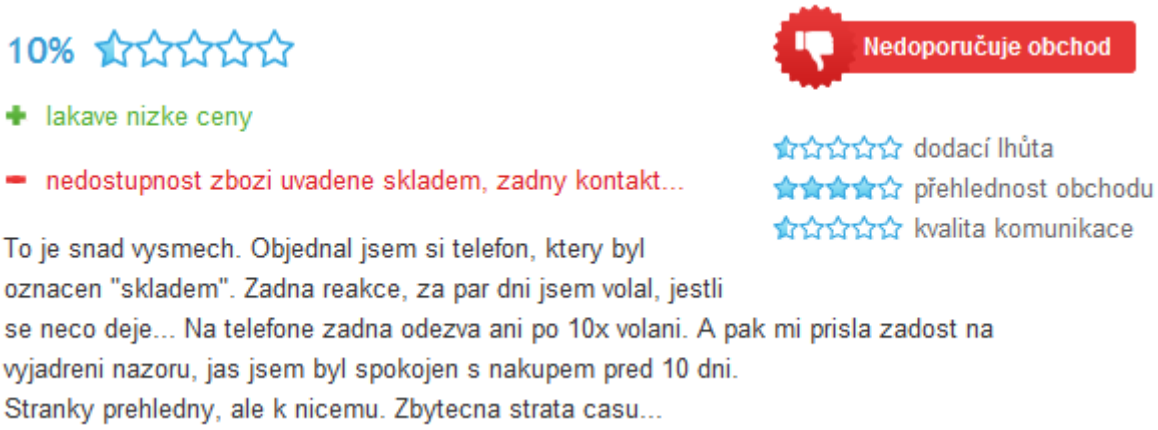
Škála hodnocení je desetistupňová (10-100 procent po 10). Zvlášť může uživatel recenzovat spokojenost (nespokojenost) s obchodem a výrobkem. K dispozici je prostor pro vyjádření se slovem pro „plusy“, „mínusy“, celkové zhodnocení včetně bodů a další. Pouze u recenze samotného e-shopu je možnost navíc bodově hodnotit rychlost dodání, přehlednost obchodu a kvalitu komunikace. Zvlášť se hodnotí obchod a zvlášť výrobek.

Na serveru mohou recenzovat výrobky a obchody anonymní nebo registrovaní uživatelé. V drtivé většině se tomu děje tak, že asi 10 dní po nákupu produktu v daném obchodě obdrží uživatel email se žádostí, aby provedl své hodnocení na serveru Heuréka. Konverze, kdy uživatel hodnocení skutečně vyplní, se pohybuje kolem 15 procent, což je poměrně vysoké číslo. Recenze jsou většinou věcné s bohatým zastoupením přídavných jmen. Možnost podvržení recenzí tak samozřejmě existuje, ale je poměrně pracný.

4.2.1 Vytvoření korpusu

Do pozitivních zpráv jsem započítal všechny proslovy, které byly v plusových hodnoceních a pak ty, které se nacházeli ve shrnutí recenze, přičemž celkové zhodnocení muselo mít minimálně 85%. Do negativních jsou započítány negativní proslovy a ty shrnutí, kde celkové zhodnocení je maximálně 50%. Určení těchto prahů jsem provedl na základě hrubé ruční analýzy, kdy ve většině případů uživatelé shrnovali své hodnocení samými superlativy nebo naopak popisovali své špatné zkušenosti. Mezi tímto intervalem se často objevovali asi z poloviny kladné a z poloviny záporné názory, proto nejsou započítávány.

Následující obrázek příklad tuto skutečnost dokumentuje, všimněte si barvitého a poměrně rozsáhlého popisu zhodnocení, což je velmi výhodné z pohledu obsahu trénovacích dat:



Obrázek 3 - ukázka hodnocení uživatele (zdroj: heureka.cz)

4.3 ČSFD

Československá filmová databáze (www.csfd.cz) je největší český portál zabývající se čistě filmovou tématikou. Počet filmů, seriálů, dokumentů apod. je k datu 21. 4. 2012 roven číslu 287 103. Registrovaní uživatelé, kteří mohou komentovat a hodnotit filmy, jsou velmi aktivní a doposud recenzovali přesně 1 949 486 krát. Stupnice hodnocení je šestistupňová, procentuální a její škála je od 0 do 100 procent (0, 20, 40, 60, 80, 100).

4.3.1 Vytvoření korpusu

Mezi množinu pozitivních jsou zařazeny všechny recenze, jejichž hodnocení je větší nebo rovno 80%. Mezi negativní pak ty, které dosahují maximálně 40%. Interval mezi těmito skupinami je poměrně široký, ale zajišťuje, že obsažená slova budou poměrně disjunktní.

Jak vypadá recenze, jejíž hodnocení dosáhlo 60 procent:

> *Spousta gagů.. některých povedených, některých míň.. spousta známých tváří v malých roličkách většinou sebe sama.. milý film.. ale jedno shlédnutí stačí..*

Většina z nás jistě usoudí, že její sdělení nevyznívá ani oslavně, ale ani hanebně. Není proto vhodné ji zařazovat do trénovacích dat.

Samotný text recenzí neobsahuje tolik přídavných jmen jako například u Heuréky.

4.4 Scuk

Posledním místem, které posloužilo jako zdroj dat, je gastronomický portál Scuk (www.scuk.cz) obsahující hlavně recenze restaurací, kaváren atd.

Tento server je charakteristický tím, klade požadavky na délku recenzí, jejich obsah a jejich autory. Obsah musí být nezaujatý, recenzovat zde nemůže každý, ale pouze vybraní jedinci, kteří *„projeví dobré znalosti o jídle a pití“* [19]. Server byl vybrán záměrně, aby se odlišil od tradičních online hodnocení čítajících několik vět. Průměrná délka recenze je totiž cca 2200 znaků (viz 4.5 Statistiky), věty jsou rozvité, obsah je velmi barvitý s častými odbočkami mimo téma. Stupnice hodnocení je pětihvězdičková.

4.4.1 Vytvoření korpusu

Recenzí není mnoho, a proto jsem nechal velmi malou mezeru mezi pozitivními a negativními recenzemi, aby bylo dat co nejvíce. Všechny příspěvky, jejichž hodnocení je větší nebo rovno 70% jsou zvoleny jako pozitivní a všechny pod 60% (včetně) jsou negativní.

4.5 Statistiky

Následující tabulka porovnává zdroje z pohledu množství zpráv, které byly staženy pomocí crawlerů, a některých charakteristik, které můžou ovlivnit natrénování klasifikátorů. Všechny procentuální hodnoty jsou počítány v poměru k českým zprávám.

	Twitter	Twitter ruční anotace	Heuréka	ČSFD	Scuk
Počet zpráv	4 985 343	2480	2 263 901	1 025 739	2215
Z toho českých	3 005 357 60%	2005 80%	2 253 310 99,5%	1 011 680 98%	2215 100%
Bez diakritiky	623 042 20%	581 28%	246 760 10%	72 584 7%	0 0%
Pozitivní	592 953 19%	683 34%	1 578 484 69%	618 574 61%	1481 66%
Negativní	47 639 1,6%	521 25%	406 834 18%	164 971 16%	734 33%
Průměrná délka pozitivních (znaky)	73	79	83	343	2211
Průměrná délka negativních (znaky)	78	88	110	298	2260

Tabulka 1 - statistiky datových zdrojů

Zbytek zpráv u Twitteru byl ohodnoceno následovně:

- Neutrální – 590 (29%)
- Přeskočeno (nejednoznačné zprávy) – 214 (10%)

Procentuální zastoupení českých tweetů koresponduje s ruční kontrolou (viz 4.1.1).
Počet textových hodnocení u Heuréky je zkreslen tím, že se jedná už o rozdělená hodnocení na plusy a mínusy. Skutečný počet stažených celistvých komentářů bude odhadem asi poloviční. Nelze si nevšimnout, že délka negativních komentářů je o 25 procent delší než těch pozitivních, uživatelé očividně popisují svoje špatné zkušenosti podrobněji.

Dobrou zprávou ale je, že pozitivních komentářů kolem nás je mnohonásobně větší než těch negativních. To značí, že většina internetových obchodů, které přežívají, vyhovují požadavkům zákazníků, filmová tvorba je poměrně kvalitní a události, které se dotýkají nás všech, jsou veselé povahy. I když v případě Twitteru je třeba brát tyto výsledky s rezervou, spousta zpráv je sarkastických:

> *bych se mohl zivit zachranou dat z poskozenejch notebooku, za posledni mesic uz patej :-)*

Nebo pozbývají významu:

@Elyse_W Cože?! =D

Proto procentuální rozložení nesouhlasí s ruční anotací, tam podobné zprávy nebyly anotovány nebo byly označeny za neutrální.

4.5.1 Srovnání SUBPOS

Pro zvýšení přesnosti klasifikátoru lze využít i znalosti z rozdílu distribuce slovních druhů. Statistiky nicméně hlavně v případě Twitteru neodpovídají skutečnosti, jelikož POS tagger v PDT (viz 6.1 - Funkce pro práci s nástroji PDT 2.0) si neporadí s nespisovnou češtinou, zvlášť bez diakritiky.

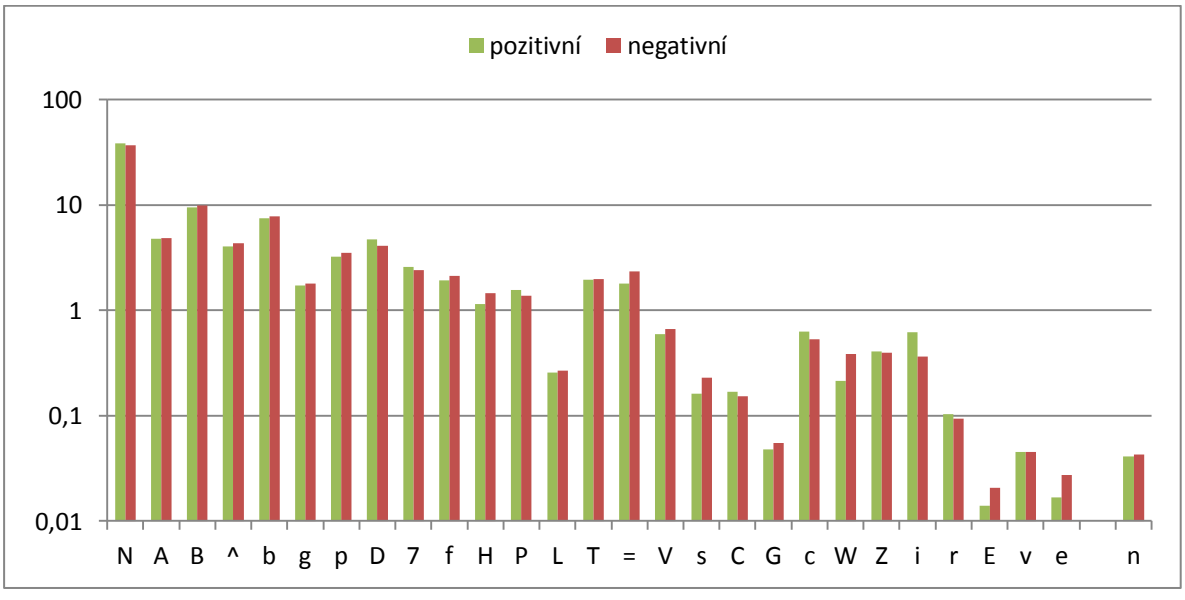
V následujících grafech jsou vyneseny relativní počty jednotlivých slovních druhů v dané třídě pro každý použitý zdroj, dle (pro každý tag):

$$P^t_{poz,neg} = \frac{N^t_{poz,neg}}{N^t_{poz} + N^t_{neg}}$$

Kde N je četnost výskytů tagu v pozitivní nebo negativní třídě.

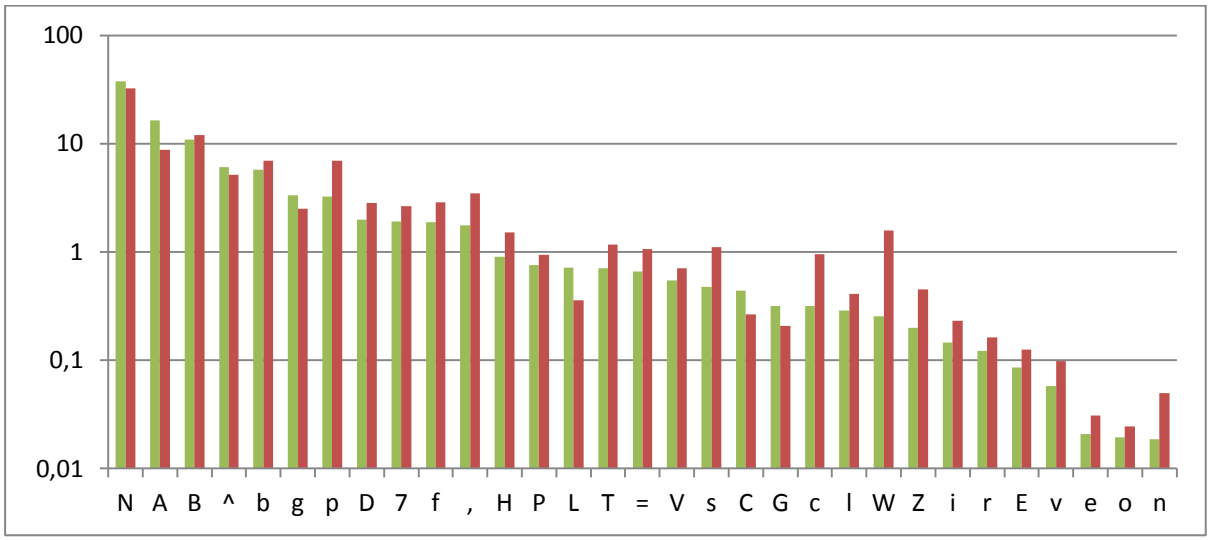
Na vodorovné ose jsou vybrané slovní druhy dle [20], na svislé ose je použito logaritmické měřítko.

Twitter



Obrázek 4 - rozložení slovních druhů napříč pozitivními a negativními zprávami z Twitteru

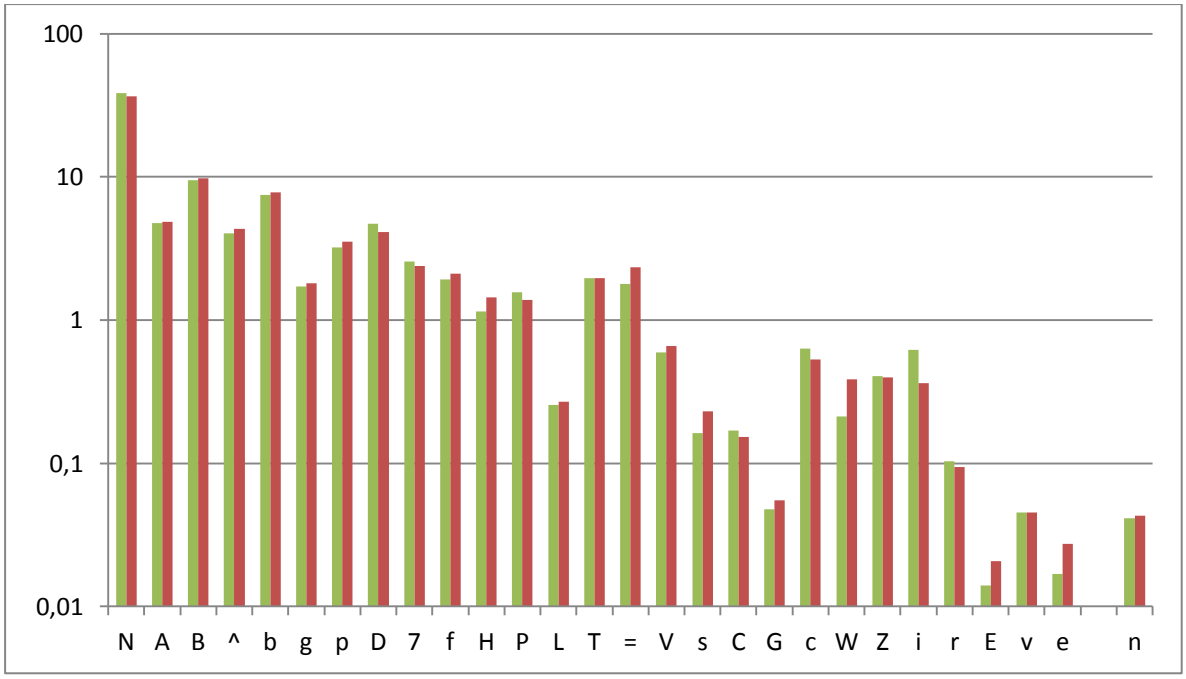
Heuréka



Obrázek 5 - rozložení slovních druhů napříč pozitivními a negativními recenzemi z Heuréky

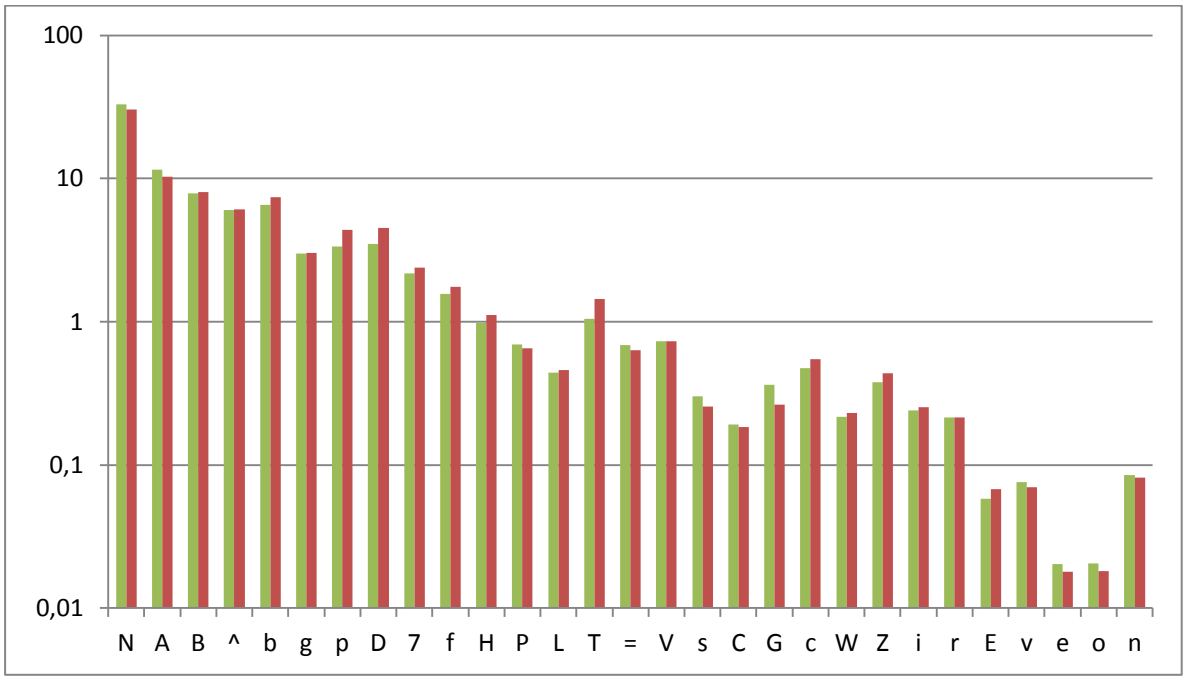
Všimněte si, rozdílu mezi relativním počtem přídavných jmen (sloupec A) a příslovci (g) mezi Twitterem a Heurékou.

ČSFD



Obrázek 6 - rozložení slovních druhů napříč pozitivními a negativními recenzemi z ČSFD

Scuk

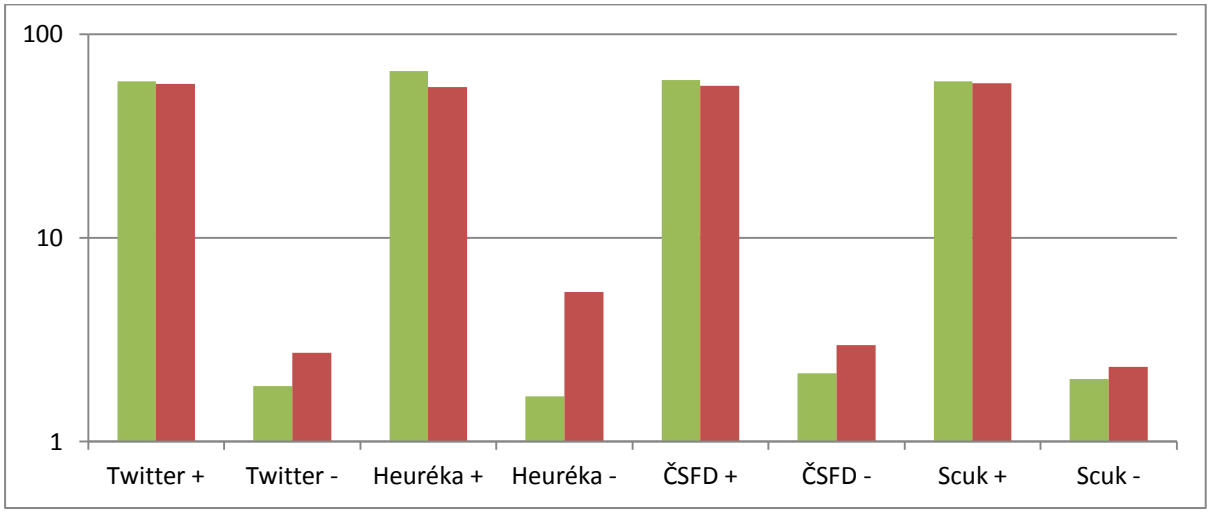


Obrázek 7 - rozložení slovních druhů napříč pozitivními a negativními recenzemi ze Scuku

Rozložení záporných podob slov

Významnější měrou však přispějí k lepším výsledkům příznaky, zda jsou slova v záporném tvaru (předpona „ne“ apod.).

V tomto grafu je srovnání rozložení negovaných slov napříč všemi datovými zdroji a třídami. Plusové znaménko značí korpus pozitivních zpráv, negativní znaménko korpus negativních zpráv.



Obrázek 8 - rozložení záporných podob slov

5 Klasifikace

5.1 Naivní bayesovský klasifikátor

Hlavním použitým klasifikátorem v této práci je naivní bayesovský klasifikátor založený na aplikaci bayesova teorému. Pracuje tak s pravděpodobnostním rozložením použitých příznaků (features), které slouží k trénování a vyhodnocování, a využívá předpokladu, že všechny použité příznaky jsou na sobě nezávislé. To samozřejmě v případě skladby vět přirozeného jazyka neplatí, ale ukazuje se, že je tento přístup možný používat s dobrými až s velmi dobrými výsledky nejen v případě rozpoznávání sentimentu. Tento předpoklad nám navíc dává možnost hodnotit proslovy, které nejsou poskládány z gramaticky správně tvořených vět, ale dovoluje klasifikovat i tzv. „bag of words“ (neuspořádaná skupina slov), aniž by trpěla jeho výsledná úspěšnost.

Asi nejznámější použití toho klasifikátoru najdeme při detekci nevyžádané pošty, kde se často vyskytují nesmyslné, náhodně generované věty a podobně.

Bayesův teorém říká, jaká je pravděpodobnost výskytu jevu A v případě, že nastal jev B [21]:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Aplikace pro náš problém klasifikace sentimetu pak vypadá následovně:

$$P(s|x) = \frac{P(x|s) * P(s)}{P(x)}$$

s … je sentiment (v našem případě pozitivní nebo negativní)
x … je zkoumaná zpráva
$P(s|x)$ … je posteriorní pravděpodobnost (posterior probability)
$P(x|s)$ … je věrohodnost (likelihood), kterou získáme z trénovacích dat
$P(s)$ … je apriorní pravděpodobnost (prior probability), tedy předem známá pravděpodobnost výskytu třídy. Ta lze spočítat z četnosti výskytů příznaků v jednotlivých třídách:

$$P(s) = \frac{počet\ zpráv\ ve\ třídě}{počet\ všech\ zpráv}$$

$P(x)$ … je evidence. Tu lze spočítat pomocí sum rule $P(x) = \sum_s P(s, x)$

5.1.1 Použití

Pro klasifikace neznámé zprávy je nutné zjistit $P(s|příznak_1, příznak_2 \dots příznak_n)$, kde $příznak_i$ je jeden z příznaků neznámé zprávy.

$$P(s|příznak_1, \dots příznak_n) = \frac{\prod_{i=1}^{n} P(příznak_i|s) * P(s)}{\sum_s P(s, příznak_1, \dots příznak_n)}$$

Hypotézu, která tento předpis maximalizuje, pak považujeme za nejpravděpodobnější.

Evidence je vždy konstantní a proto je možné ji vynechat vzhledem k tomu, že nás nezajímá konkrétní hodnota výsledné pravděpodobnosti, ale poměr vzájemných hodnot. Stejně tak je možné zanedbat apriorní pravděpodobnost, jelikož je vhodné pro trénování používat stejný počet zpráv, pak:

$$P(s|příznak_1, \dots příznak_n) = \prod_{i=1}^{n} P(příznak_i|s)$$

Protože při klasifikaci dlouhé zprávy můžeme snadno při násobení ztratit přesnost, je dobré rovnici zlogaritmovat:

$$P(s|příznak_1, \dots příznak_n) = \sum_{i=1}^{n} log\big(P(příznak_i|s)\big)$$

5.1.2 Laplaceovo vyhlazování

Je možné, že se vyskytne případ, kdy příznak není obsažen v trénovacích datech a jeho věrohodnost $P(x|s)$ je nulová. Jelikož ale tato pravděpodobnost figuruje v součinu (když není použit zlogaritmovaná verze), získáváme tak nulovou pravděpodobnost pro celou zkoumanou zprávu. Aby se tomuto nežádoucímu chování zamezilo, je možné buď příznak ignorovat, nebo použít Laplaceovo vyhlazování, které mu přiřadí nenulovou pravděpodobnost [22]:

$$P\_odhad(příznak|s) = \frac{počet\ výskytů\ příznaku\ ve\ třídě + \alpha}{počet\ výskytů\ všech\ příznaků\ ve\ třídě + \alpha * počet\ všech\ příznaků}$$

α je vhodně zvolená kladná konstanta, v našem případě zvolíme $\alpha = 1$. Ta nám zaručí, že nově přidělená pravděpodobnost neznámému příznaku nebude nulová, ale zároveň bude dostatečně

malá, aby ve velkém množství vzorků, které máme k dispozici pro trénování, výrazně neovlivnila celkový výsledek. Tato hodnota byla zvolena na základě testování na korpusu Heuréky. Nicméně je zde prostor pro experimentování, jelikož při použití jiného korpusu nebo při testování na reálných datech může podávat lepší výsledky menší hodnota.

5.1.3 Výběr příznaků a zvyšování přesnosti

K dispozici máme nástroje, které umožňují získat lemma a další informace o jednotlivých slovech ve větě, proto se nemusíme omezovat pouze na sestavování N-gramů jenom ze samotných slov (příznaků). V kapitole 7 Vyhodnocení jsou komentovány výsledky právě s použitím i těchto informací.

Pro zvýšení přesnosti je experimentováno s vynecháním slov kratších než N znaků (typicky jde o spojky, předložky apod.), které nemají důležitou informační hodnotu, ořezáním diakritiky a vynecháním některých stopslov (viz Příloha 3. Použitá stopslova

), ačkoliv spousta z nich už je odstraněna pomocí omezení délky.

Distribuce „rozšířených“ slovních druhů (SUBPOS, viz [20]) a negovaných slov se napříč negativními a pozitivními množinami zpráv také liší, a proto je možné zakomponovat do vyhodnocení i tyto znalosti:

$$P(s|x) = P(G|s) * P(POS|s) * P(NEG|s)$$

G je vektor N-gramů příznaků reprezentující zprávu
POS je vektor POS tagů reprezentující zprávu, opět se předpokládá, že jsou vzájemně nezávislé
NEG je vektor příznaků určující, zda jsou slova v záporném tvaru či ne

V případě, že se používají jako příznaky lemmata slov, je nutné tokenu přidat příznak, zda bylo původní slovo v negativním tvaru, protože slovo jako „nejsem“ se vyskytuje mnohem častěji v negativně laděných zprávách a slovo „jsem“ je naopak hlavně v těch pozitivních. Obě jsou ale pozměněna na „být“. Tak bychom zbytečně ztrácely cenné příznaky, hlavně při použití bigramů a více – „jsem spokojený“ vs. „nejsem spokojený“.

Dále si je nutné uvědomit, že PDT také nepodává stoprocentně přesné výsledky, zvláště při použití nespisovné češtiny. Proto ve výjimečných případech nám může při klasifikaci jeho použití uškodit.

6 Implementace

6.1 Funkce pro práci s nástroji PDT 2.0

V průběhu řešení této práce vzniklo několik funkcí pro Python 2.7 zapouzdřující práci s tokenizérem a morfologickým analyzátorem pražského závislostního korpusu (PDT) [2] , který pracuje bohužel jenom na úrovni souborů. Na serveru *athena3,* kde byla práce vyvíjena, sice byla část těchto funkcí dostupná, avšak byly nekompatibilní, jelikož jsou psány pro Python verze 2.5.

Motivací pro použití těchto nástrojů je použití jejich výstupů jako příznaků pro klasifikátor. Vzniklé funkce umožňují spouštět i dávkové operace a minimalizují tak výkonnostní propady opakovaným spouštěním analyzátoru, což trvá asi 3 sekundy. Je možno zvolit jenom spuštění tokenizéru bez morfologické analýzy.

Každý token nese krom termu i jeho lemma a poziční tagy vyprodukované jako výsledek morfologické analýzy, jejichž popis lze nalézt zde [20].

6.2 Čištění textů

Použité datové zdroje dávají tušit, že je užívána čeština, kde pravidla nehrají žádnou roli. Nejčastějším prohřeškem je chybějící diakritika, nedodržování interpunkce, slangové a nespisovné výrazy, častý výskyt URL adres apod. Aby klasifikátor podával co nejpřesnější výsledky, je nutné texty jak před trénováním, tak před samotnou klasifikací neznámého obsahu, normalizovat.
Prováděny jsou následující úkony:

1. Převedení na malá písmena
2. Zkrácení slov s opakujícími se písmeny („Anoooo“ na „ano“)
3. Odstranění smajlíků, URL adres apod. Hashtagy a mentions jsou záměrně zachovány, aby ovlivňovaly výsledné skóre.
4. Odstranění všech interpunkčních znamének krom čárky, tečky a dalších znaků mimo českou abecedu.

6.3 Tokenizace a vytváření N-gramů

Aby bylo možné počítat pravděpodobnosti výskytu slov či sousloví, je nutné tyto termy nejdříve získat. Postačující je dělit věty podle nealfanumerických znaků, přičemž i ty se stávají tokeny. To

[2] URL: http://ufal.mff.cuni.cz/pdt2.0/index-cz.html

v podstatě činí tokenizér v PDT, navíc však respektuje i desetinná čísla čárkami a tečkami, čísla, které jsou zapsány s oddělením tisíců pro lepší čitelnost apod.

Jakmile máme k dispozici seznam tokenů, je vhodné si z jejich posloupnosti vytvořit izolované věty (samozřejmě stále prezentovanou tokeny). Jelikož je ve fázi výběru příznaků a testování nutné často s tokeny experimentovat – filtrovat je, provádět nad nimi operace jako je odstranění diakritiky, spojovat je apod., zapouzdřuje věty tokenů třída *NGramStack*. Ta dovoluje registrovat tzv. filtry a modifikátory. Filtr je třída, která přijímá token a jeho typ (původní slovo, lemma, tag) a vrací příznak, zda má být slovo vymazáno a jestli dále rozděluje větu na další samostatné proslovy. Modifikátor pracuje podobně, avšak vrací upravený vstupní symbol – například spojení lemmatu a některé části SUBPOS tagu. Pak je již možné získávat požadované N-gramy. Pro tento účel je použita funkce *util.ngrams()* z balíku Natural Language Toolkit (NLTK) [3].

6.4 Identifikace češtiny

Možnost výskytu cizojazyčných textů, je nejvíce eliminována výběrem datových zdrojů. Avšak i tam lze nalézt poměrně velké množství slovenských a anglických proslovů. Ty totiž můžou ovlivnit chování klasifikátoru. U bayesovského přístupu to vzhledem k jeho povaze zas tak nevadí, nicméně u použití Support vector machines (SVM) klasifikátoru, jenž se snaží optimálně rozdělit trénovací množiny pomocí nadroviny můžou tyto šumová data zavazet.

Identifikace je prováděna pomocí nástroje „Identifikace jazyka“ od Vojtěcha Mrázka [4]. Hrubým testováním se ukázalo, že úspěšnost při identifikaci kratších textů, je velmi nízká a navíc program není schopen rozpoznat slovenštinu či ruštinu, kterou ve většině případů zaměňoval za češtinu. Množina textů, které jsou označeny jiným jazykem než je čeština je proto v druhém průchodu testován naivním způsobem na výskyt všech běžných českých písmen s diakritikou a nejčastějších slov nebo aspoň jejich prefixů (aby se započítala tvary v množném čísle apod.) a zároveň je vyloučena ruština pomocí testu na výskyt azbuky a angličtina, pomocí charakteristických bigramů (I am, in any, …). Tímto však dojde k výběru i slovenských proslovů. Ty jsou poté v třetím průchodu vyloučeny obdobným způsobem – testována je přítomnost častých slovenských slov, které se nevyskytují v češtině, například: „isté“, „čo“ apod. Funkce je implementována jak pro Python, tak pro PostgreSQL jako uložená procedura, kvůli výkonnostním důvodům. Regulární výrazy provádějící tyto kontroly jsou v Příloha 2. Regulární výrazy pro zpřesnění detekce češtiny

[3] URL: http://www.nltk.org
[4] Dostupné na serveru minerva1 v /mnt/minerva1/nlp/projects/lang_id3

6.5 Bayesovský klasifikátor

Třída *Bayes* implementuje naivní bayesovský klasifikátor, který umí pracovat N třídami. Není tedy omezen jenom na dvě, jež jsou použity v této práci. Poskytuje metody pro iterativní trénování včetně přidávání nových tříd za běhu, kdy je možné předat jako argumenty jenom samotné příznaky nebo pouze text či množinu textů, ze kterých si metoda sama příznaky získá na základě zvolených parametrů, předané třídě *NGramStack* a dalších atributů, které mění výsledky klasifikátoru jako je připojení POS tagů apod.

Pro vnější komunikaci byla navržena třída *BayesXmlRpc*, jež klasifikátor zapouzdřuje a běží jako server, s nímž je možno komunikovat pomocí volání vzdálených procedur (XML-RPC). Je tak zajištěno API pro aplikace třetích stran. Toho využívá prezentační webová aplikace, kdy si návštěvníci mohou nechat vyhodnotit svoji zprávu a dokonce v případě, že nesouhlasí s výsledkem, klasifikátor přetrénovat. Tyto hodnocení se ukládají i do databáze, což je vhodné pro budoucí analýzy. Při ukončení běhu je celý klasifikátor persistován na disk, včetně nově přidaných hodnocení a během opětovného spuštění není nutné ho trénovat znovu.

6.6 Stahování z datových zdrojů

Tato část popisuje návrh crawlerů dat a databáze pro skladování zdrojových dat. Jako systém řízení báze dat (SŘBD) je použito PostgreSQL 9. Navržené schéma je v příloze . Crawlery jsou implementovány v jazyce Python 2.7 za pomocí knihovny Beautiful Soup 4, která parsuje HTML stránky do DOM dokumentu, přičemž se umí částečně zotavit z chyb, pokud nejsou www stránky validní.

6.6.1 Twitter

6.6.1.1 Twitter API

Twitter poskytuje dvě programová rozhraní pro přístup k jeho veřejnému obsahu. Obě jsou vázána na registrovaný účet a umožňují filtrovat zprávy dle konkrétních uživatelů. To se totiž ukázalo jako jediný způsob jak získat české tweety. Twitter sice umožňuje výběr dle jazyka a lokace v podobě souřadnic, ale čeština chybí a zprávy buď neobsahují informaci o poloze, nebo neodpovídají - většina tweetů z pražského prostředí má lokaci v Berlíně.

Obě rozhraní vracejí data v JSON[5] formátu a poskytují kompletní informace, jak o zprávě, tak o uživateli.

Informace, které je vhodné, krom samotného obsahu a identifikačních čísel (dále ID) pro implementační účely, uchovávat pro eventuální potřeby rozpoznávání sentimentu:

- Zpráva:
  - Datum publikování – pro případnou agregaci
  - Jméno autora
  - ID zprávy, na kterou tweet odpovídá – pro případnou agregaci sentimentu. Je totiž pravděpodobné, že pokud má zpráva pozitivní valenci, odpověď na ní bude též pozitivní, pokud se pojednává o nějakém termínu.
  - ID uživatele, kterému odpovídá – ze stejného důvodu jako výše
  - Počet retweetů – pokud byl Tweet někým přeposlán, vyplatí se mu dát větší váhu
- Uživatel:
  - Počet followerů
  - Počet following

API poskytuje další informace, které by se mohly dále hodit pro heuristické zpřesňování výsledků. Krom výše zmiňované polohy existuje příznak, zda je uživatel označil tweet jako oblíbený, avšak po stažení několika miliónů zpráv se ukázalo, že těchto zpráv je jenom několik.

Název	Data v reálném čase	Omezení	Historie
REST API	Dle dokumentace ne, praxe ale ukázala, že téměř.	350 požadavků za hodinu, v jednom požadavku lze žádat maximálně o 100 tweetů	Dostupných posledních 3200 zpráv
Streaming API, základní role	Téměř.	Umožňuje sledovat 5000 uživatelů nebo 400 klíčových slov	Ne, jenom v případě, že nejsou žádné nové

Tabulka 2 - srovnání Twitter API

Z výše uvedených údajů jde vidět, že API je poměrně restriktivní. Je totiž určené hlavně pro běžné klientské aplikace. U Streaming API existuje několik rolí, které se liší hlavně v limitu sledovaných zpráv či klíčových slov a o jejich přidělení můžete požádat.

Kontaktoval jsem technickou podporu se žádostí na navýšení sledování uživatelů na počet 15 tisíc, čímž by se velmi zjednodušilo sledování uživatelů. Nicméně žádost zamítly s odůvodněním, že

[5] Popis formátu: http://www.json.org/

neposkytují tyto role pro výzkumné účely a sami doporučili postavit aplikaci nad REST API s použitím několika účtů, které budou stahovat data paralelně. To výrazně zkomplikovalo implementaci crawleru, kde je nutné shromažďovat data velkého počtu vymezených uživatelů včetně historie.

6.6.1.2 Seznam českých uživatelů

Možnost, jak získat přezdívky nebo ID českých uživatelů, existuje prakticky jenom jedna. Bylo by sice možné, ač náročné, napsat crawler, který shromažďuje uživatele ze zmiňované služby Klábosení, jelikož nejspíše neexistuje žádný jiný způsob jak k nim přistupovat strukturovaně jinak než hledáním. Navíc není pravděpodobně možné takto využívat tuto službu vzhledem k licenčním podmínkám.

Naštěstí existuje služba Czechia Twitter [6], která byla donedávna (prosinec 2011) aktivní, soudě dle posledních přidaných uživatelů (celkem 14000), a ručně shromažďovala „seznam českých Twitteristů“. Podmínkou pro přidání do seznamu je nutné, aby uživatel nebyl pasivním, ale aby aktivně přispíval převážně českými zprávami. Přesné podmínky pro zařazení do seznamu autor nezmiňuje. Nicméně po hrubé kontrole tweetů jsem uvážil, že tento seznam bude dostatečný. Výhodou také je, že je již částečně eliminován problém Čechů publikujících cizojazyčně. Seznam je k dispozici ve formátu CSV pro libovolné použití za cenu 2 Eura.

6.6.1.3 Crawler pro Twitter

Zásadním problémem, který ovlivnil celou implementaci je přísný limit síťových požadavků (viz Tabulka 2 - srovnání Twitter API). Následující pseudokódy demonstrují, jak pracuje stahování tweetů zohledňující historii. Pro komunikaci s Twitter API je použita knihovna python-twitter [7], která zapouzdřuje HTTP požadavky do funkcí.

Pseudokód:

1. Načti účty, jež slouží pro připojení k Twitter API z konfiguračního souboru
2. Vytvoř takový počet procesů kolik je dostupných účtů
3. V každém procesu inicializuj crawler, připojení k databázi a k Twitter API
4. Vytvoř další proces, který se stará o dávkové čištění textů, jazykovou identifikaci pro další eliminaci cizojazyčných textů a klasifikaci. Dávka je spouštěna jednou za minutu a zpracovává maximálně 100 000 tweetů. Paměťová náročnost je při tomto počtu okolo 250MB, v běžném provozu ale bude zanedbatelná.

[6] URL: http://www.czechiatwitter.cz
[7] URL: http://code.google.com/p/python-twitter/

Pseudokód samotného crawleru:

1. Načti nové uživatele z tabulky *users_to_add*. Načti informace o těchto účtech, v případě, že nejsou označeni jako smazané, vlož je do seznamu sledovaných
2. Načti sledované uživatele, seřaď je dle data posledního stáhnutí sestupně. První ty, jejichž zprávy ještě stáhnuty nebyly
3. Pro každého uživatele:
    a. Stáhni první stranu tweetů, začni však od posledního staženého. Pokud je odpovědí od serveru překročení limitu, uspi se do té doby, než se limit obnoví, pokud je odpovědí chyba, uspi se na předem definovanou dobu dle typu chyby
    b. Pro všechny stáhnuté zprávy:
        i. Ulož ji do databáze, aktualizuj u uživatele ID posledního staženého tweetu
        ii. V případě, že se jedná o odpoveď uživateli, který není v seznamu, přidej ho do tabulky *users_to_add*
    c. V případě, že se žádná nestáhla, aktualizuj čas posledního stažení uživatele
    d. Uspi se na 15 sekund (empiricky zjištěná minimální doba, kdy často nedochází k odmítnutí spojení)

Funkce jsou ve skutečnosti implementovány a agregovány tak, aby docházelo k co nejmenšímu zatížení komunikace mezi serverem a databází. Aby stažení velkého množství tweetů trvalo pouze krátkou dobu a následné sledování uživatelů nenabíralo velkého zdržení, je registrováno 10 účtů. Není ale problém účty přidávat nebo je naopak zakázat pomocí konfiguračního souboru. Deset instancí crawleru tak zkontroluje 14 000 uživatelů za 6 hodin v případě, že uživatelé nebudou psát více jako 100 tweetů za tuto dobu, což je v reálném provozu více než dostatečné. Maximální rychlost stažených zpráv za hodinu v případě, že je crawler plně vytížen je tedy: 350*100*počet crawlerů.

6.6.2 Crawler pro Heuréku

Portál nabízí [8] mapu webu (sitemap) v XML formátu, který jsem objevil náhodou náhodným zkoušením běžných adres, jelikož v době, kdy byl crawler psán nebyl k dohledání ani pomocí Google. Nejspíše však šlo o náhodný výpadek, jelikož nyní k nalezení je.

Vzhledem k velikosti portálu je samotný soubor pouze seznamem, jehož položky jsou adresy na další komprimované XML sitemapy, kde lze nalézt skutečné umístění obchodů a výrobků. Jelikož největší české obchody mají i více jak 40000 recenzí, je použito stránkování.

[8] URL: http://www.heureka.cz/sitemap_index.xml

Vypreparované adresy jsou jednorázově uloženy do tabulky *urls_to_crawl* s příznakem *is_crawled=0*. Struktura obchodu a produktové stránky je mírně odlišná, nicméně samotná oblast recenzí je identická, tudíž lze použít stejnou kód pro jejich uložení.

Při samotném parsování je nutné se vyrovnat s občasnými nejednotnostmi, aby crawler nehavaroval a mohl spolehlivě běžet několik dní bez dozoru – někdy chybí bodové hodnocení, někdy textová recenze, zřídka je namísto recenze umístěná reklama apod. Konstrukce byla tedy poměrně časově náročná. Jakmile je stránka úspěšně rozparsována, ukládají se shrnutí, plusy a mínusy napsané uživatelem, jméno uživatele, celkové hodnocení do tabulky *heureka* a *heureka_ratings* a nastaví se příznak *is_crawled=1*, což indikuje úspěšné stažení a při eventuálním přerušení crawleru se již nezačne stejná stránka stahovat znova. Riziko ale spočívá v tom, že pokud pauza potrvá dlouho dobu, ztratí se vlivem přibývání příspěvků a s tím souvisejícím stránkování, některé recenze.

Hrozba duplicit je zažehnána, jelikož samotné recenze mají jedinečné ID. Heuréka neomezuje počet navázání spojení za určitý časový limit ze stejné IP adresy, a proto nebylo nutné implementovat umělé čekání. Stahování trvalo asi 3 dny.

6.6.3 Crawler pro ČSFD

Nejjednodušší strukturu umístění má filmový portál ČSFD. Všechny recenze jsou na jedné adrese i přes jejich velké množství. Jednotlivé filmy mají URL ve tvaru http://www.csfd.cz/film/[ČÍSLO_FILMU]/?all=1, první film má identifikační číslo 1 a každý další má o jedničku vyšší. Do tabulek *csfd_films*, *csfd_people*, *csfd_ratings* se ukládá text recenze, bodové hodnocení, přezdívka uživatele, jména herců a tvůrců, kteří se na filmu podíleli, což může být užitečné pro potenciální rozšíření hodnocení sentimentu zohledňující určité osoby apod.

Tabulka *csfd_people_film* slouží k propojení tvůrců a filmů, na kterých se podíleli. V případě, že se stahování přeruší, začíná se pak znovu od filmu s nejvyšším číslem. Vznik duplicit u tvůrců, herců, filmů a recenzí nehrozí - každá entita má svoje jedinečné ID. Mezi jednotlivými připojeními je nutné udělat pauzu alespoň 5 sekund. Stahování trvalo asi 14 dní.

6.6.4 Crawler pro Scuk

Portál Scuk poskytuje sitemap na standardní adrese www.scuk.cz/sitemap.xml.

Po chvíli prohlížení lze zjistit, že recenze podniků se nacházejí na adrese s určitým prefixem. Všechny stránky mají stejnou strukturu a nejsou stránkované. Vzhledem k počtu (~2000) není

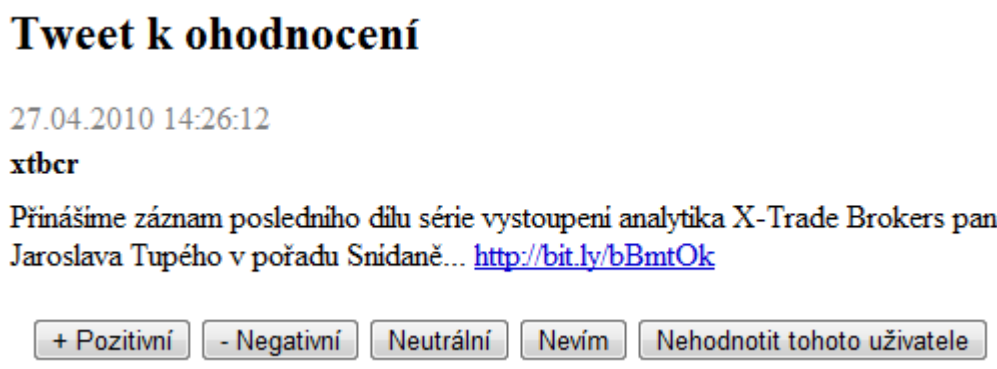
implementováno žádné navazování přerušeného stahování a je tak nutné v případě přerušení stahovat vše znova.

Duplicitní záznamy se však v databázi neobjeví, jelikož každou entitu lze rozeznat dle unikátního identifikačního čísla. Ukládána je samotná recenze, bodové hodnocení, přezdívka uživatele do tabulek *scuk_ratings* a *scuk_restaurants*. Aby server nezablokoval určitou IP adresu, je nutné vložit čekání mezi připojeními minimálně 10 sekund. Stahování trvalo asi 6 hodin.

6.6.5 Anotační aplikace pro zprávy z Twitteru

Jednoduchá, ale rychlá, webová ajaxová aplikace napsaná v jazyce PHP využívající framework Nette a pro práci s databázovou vrstvou knihovnu dibi. Zprávy jsou načítány z tabulky *tweets*. Aby se zamezilo opakovanému hodnocení již anotovaných tweetů, kontroluje se výskyt ID zprávy v tabulce *sentiments*, kde jsou uloženy hlasy anotátorů a jejich IP adresy kvůli možnosti jejího zablokování, kdyby aplikace byla využita veřejností. Jelikož jsou zprávy načítány z databáze sekvenčně, je možné, že se anotátorům ve stejný čas mohou objevit stejné zprávy. Nejjednodušší řešení toho problému je načítat vždy tweety pseudonáhodně. Většina SŘBD nabízí náhodné seřazení řádků, to ale trvá příliš dlouho (i několik hodin při více jak 5M řádcích). Proto je do relace *tweets* přidán indexovaný sloupec *seq_id*, kde je uloženo jedinečné celé číslo a řádky pak tvoří vzestupnou posloupnost čísel bez mezer. Je tak potom možné vygenerovat náhodné číslo v rozsahu MIN(seq_id) až MAX(seq_id) a vybrat řádek s tímto číslem.

Ovládat a hodnotit zprávy je pro urychlení práce možné i klávesovými zkratkami (Q, W, E, R).



Obrázek 9 - anotační aplikace tweetů

7 Vyhodnocení

K dispozici jsou čtyři různé zdroje, jejichž větné skladby jsou i poměrně různorodé, když porovnáme například proslovy z Twitteru a Scuku. Je proto vhodné porovnat, jak si různě natrénovaný klasifikátor stojí oproti dalším zdrojům.

V následujícím obsahu je použito těchto pojmů, které slouží pro vyhodnocení systémů [23] a vycházejí z matice záměn (confusion matrix):

Správná třída \ klasifikace	+	–
+	TP – true positive (správně určen jako pozitivní)	FN – false negative (falešně nagativní)
–	FP – false positive (falešně pozitivní)	TN – true negative (správně určen jako negativní)

Tabulka 3 - matice záměn

$$celková\ správnost, úspěšnost\ (accuracy) = \frac{TP + TN}{TP + TN + FP + FN}$$

$$přesnost\ (precision) = \frac{TP}{TP + FP} \qquad sensitivita\ (recall) = \frac{TP}{TP + FN}$$

7.1 Výběr modifikátorů a filtrů

Nejdříve bylo provedeno automatizované testování, kdy se vyzkoušely všechny možné kombinace modifikátorů, filtrů a různých trénovacích sad lišících se jak ve velikosti (od 2000 až po 140 000 zpráv, pokud byl tento počet k dispozici), tak ve zdroji dat. Velikosti trénovacích a testovacích sad byly vždy v poměru 0.75/0.25 a počet pozitivních a negativních zpráv byl ekvivalentní. Těchto testů bylo provedeno 80 000, z nich jsem vybral horních 25% s největší úspěšností a vybral tu kombinaci modifikátorů a filtrů, jejichž výskyt byl nejčastější:

- Odstranění slov kratších jak 2 znaky (včetně) a stopslov – zvýšení správnosti klasifikace v průměru o 2% oproti ponechání těchto slov. Odstranění delších slov přinášelo zhoršování výsledků napříč téměř všemi testy.
- Odstranění diakritiky – zvýšení o 1%

- Použití lemmat namísto původních slov, kde byl připojen příznak, zda jde o znegované slovo – zvýšení o 2%. Pokud se tento příznak nepřipojil, přesnost byla většinou stejná nebo menší. Bylo ozkoušeno i použití pouhých slovních druhů (POS) a rozšířených slovních druhů SUBPOS namísto slov, ale tam úspěšnost byla velmi nízká (zhruba 55%) nezávisle na velikosti sady nebo použití trigramů a více.
- Jako další samostatné příznaky posloužily i rozšířené slovní druhy a příznaky znegování – zvýšení o 2%. K většímu zpřesnění výsledků docházelo hlavně při nízké velikosti trénovací sady.

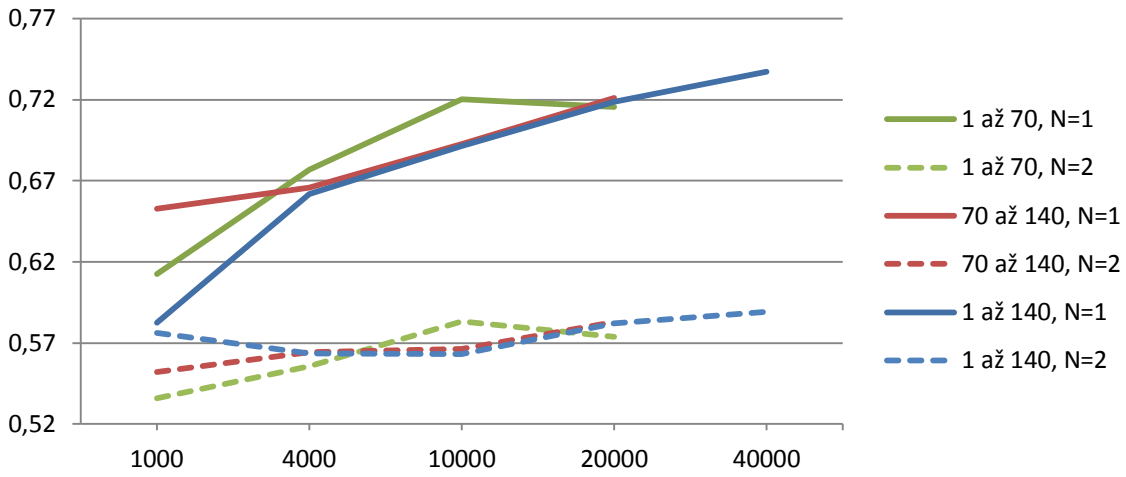
Dále jednoznačně vyplynulo, že se vyplatí používat unigramů namísto a bigramů a více a to i při velké testovací sadě v případě všech zdrojů a při použití jakékoliv délky proslovů. Ve většině případů byla přesnost (recall) a sensitivita (precision) podobná, lišila se maximálně o 1-4% a to jak v případě detekcí u pozitivních i negativních zpráv. Při praktickém použití, například v marketingu, by bylo vhodnější, aby měl systém větší sensitivitu, protože cena ztráty správného názoru je vyšší než cena ztráty při falešné pozitivitě.

Tento způsob testování byl zdlouhavý a neefektivní z hlediska časových a výpočetních prostředků – testování trvalo asi 4 dny v 10 paralelních běžích a při průměrné spotřebě paměti 5GB (použit klasifikátory z NLTK), nicméně dovolil odhalit nuance, které ovlivňovaly úspěšnost klasifikace.

7.2 Vyhodnocení datových zdrojů

Následující grafy ukazují výsledky klasifikátorů při použití různých délek proslovů (např. 1-140) a použití unigramů (v grafu N=1) a bigramů (N=2) v závislosti na počtu zpráv. Na ose X je počet použitých zpráv pro jednu třídu, celkový použitý počet je pak dvojnásobný, jelikož je použit stejný počet pozitivních i negativních proslovů. Na ose Y je úspěšnost. Pokud je použita množina s velkým rozmezím a malým limitem, např. 20-900, limit 1000, je poskládána co nejrovnoměrněji z hlediska délky.

7.2.1 Twitter



Obrázek 10 - úspěšnost klasifikace při použití Twitteru jako trénovací i testovací sady

Klasifikátor založený na principu maximální entropie [24] (dále maxent) podával velmi podobné výsledky (±2%), ale čas potřebný pro natrénování byl při 10 iteracích desetinásobný oproti bayesovskému. Byl zkoušen i klasifikátor založený na rozhodovacích stromech [25], ale jelikož trénování trvalo velmi dlouho dobu, nebylo možné ověřit všechny velikosti sad, nicméně při použití malého počtu podával výsledky zhruba o 10% horší.

Použití bigramů bohužel nepřineslo očekávané zlepšení statistik. S přibývajícím počtem byl ale trend lineárně vzestupný, a pokud by byl k dispozici dostatečný počet dat (>500 000 zpráv), byly by výsledky jistě zajímavější. Takové množství nicméně nelze na českém Twitteru asi sehnat.

Na čem klasifikátor selhává? Znaménka plus a mínus znamenají měření na pozitivní a negativní množině:

	1000	**4000**	**10 000**	**20 000**	**40 000**
+ Recall	0,710	0,645	0,605	0,609	0,636
– Recall	0,453	0,677	0,777	0,828	0,837
+ Precision	0,565	0,667	0,731	0,780	0,796
– Precision	0,610	0,656	0,663	0,679	0,697

Tabulka 4 - srovnání přesnosti a sensitivity, Twitter, délka zpráv 1 až 140

Následuje porovnání při natrénování klasifikátoru z různých zdrojů při užití unifikovaných testovacích sad s následujícími parametry:

	Twitter	Twitter manuální anotace	Heuréka	ČSFD	Scuk	Mix
Počet zpráv	20 000	1204 (+ 683, – 521)	40 000	20 000	450	Twitter: 6000 + manuál. Heuréka: 6000 ČSFD: 6000 Scuk: 450

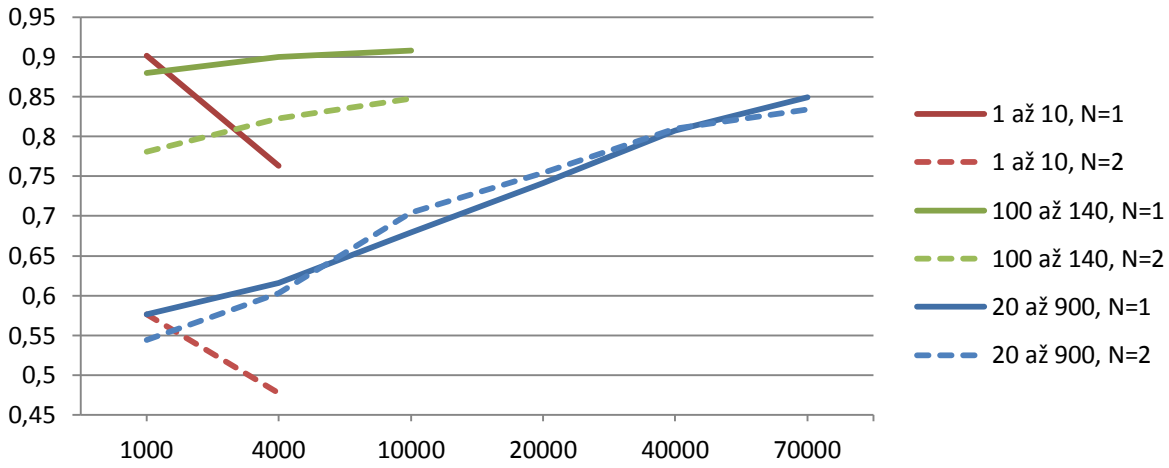
Tabulka 5 - parametry testovacích sad

Poznámka: počet pozitivních a negativních zpráv je ekvivalentní s výjimkou ručně anotované sady u Twitteru, průměrné délky zpráv jsou obdobné jako v Tabulka 1 - statistiky datových zdrojů

	Úspěšnost	+ Recall	– Recall	+ Precision	– Precision
Twitter	66%	0,490	0,835	0,748	0,621
Twitter man.	55%	0,361	0,837	0,767	0,469
Heuréka	52%	0,072	0,974	0,742	0,512
ČSFD	56%	0,231	0,900	0,698	0,539
Scuk	50%	0,013	0,986	0,5	0,5
Mix	55%	0,241	0,873	0,662	0,529

Tabulka 6 - výsledky testování Twitteru (délka zpráv 1-140, limit 40000) jako trénovací sady oproti ostatním zdrojům

7.2.2 Heuréka



Obrázek 11 - úspěšnost klasifikace při použití Heuréky jako trénovací i testovací sady

Při použití délek zpráv 30-60 a 60-100 byl trend úspěšnosti stejný jako u 100-140, ale horší o zhruba -5%. Maxent klasifikátor podával stejné výsledky s výjimkou délek 20-9999, kdy byla úspěšnost stále na zhruba 55% nezávisle na počtu zpráv.

Úspěšnost při použití bigramů je při velkém rozptylu délek (modrá čára) téměř stejná jako při použití unigramů. V porovnání s Twitterem se zde jedná totiž o užší témata s menší dimenzionalitou použitého slovníku recenzentů.

Na čem klasifikátor selhává?

	1000	**4000**	**10 000**	**20 000**	**40 000**	**70 000**
+ Recall	0,152	0,232	0,359	0,486	0,621	0,717
– Recall	1	0,998	0,998	0,996	0,992	0,980
+ Precision	1	0,995	0,995	0,992	0,988	0,973
– Precision	0,541	0,565	0,609	0,659	0,724	0,776

Tabulka 7 - srovnání přesnosti a sensitivity, Heuréka, délka zpráv 20 až 900

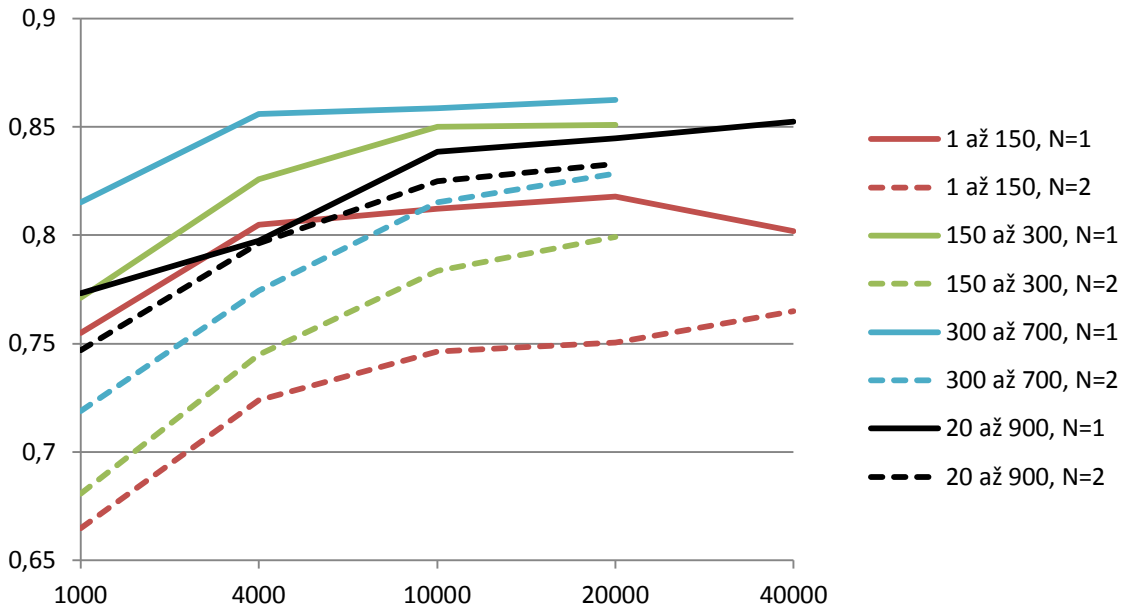
Tato sada je výhodná v tom, že skutečně u negativních zpráv jsou pouze a jenom věcné věty. Získáváme tak velmi význačné příznaky pro detekci hlavně negativních proslovů.

Porovnání oproti ostatním zdrojům:

	Úspěšnost	**+ Recall**	**– Recall**	**+ Precision**	**– Precision**
Twitter	52%	0,113	0,941	0,659	0,515
Twitter man.	50%	0,206	0,942	0,842	0,445
Heuréka	78%	0,588	0,979	0,965	0,704
ČSFD	60%	0,246	0,970	0,891	0,562
Scuk	56%	0,160	0,960	0,805	0,533
Mix	68%	0,461	0,915	0,848	0,624

Tabulka 8 - výsledky testování Heuréky (délka zpráv 20-900, limit 70000) jako trénovací sady oproti ostatním zdrojům

7.2.3 ČSFD



Obrázek 12 - úspěšnost klasifikace při použití ČSFD jako trénovací i testovací sady

Maxent klasifikátor podával horší výsledky zhruba o 5-10%.

	1000	4000	10 000	20 000	40 000
+ Recall	0,935743	0,921922	0,942377	0,941188	0,934393
– Recall	0,610442	0,672673	0,734294	0,74835	0,770377
+ Precision	0,706061	0,737981	0,78006	0,789032	0,802732
– Precision	0,904762	0,896	0,927236	0,927138	0,921522

Tabulka 9 - srovnání přesnosti a sensitivity, ČSFD, délka zpráv 20 až 900

Porovnání oproti ostatním zdrojům:

	Úspěšnost	+ Recall	– Recall	+ Precision	– Precision
Twitter	54%	0,478	0,612	0,552	0,540
Twitter man.	58%	0,625	0,521	0,659	0,484
Heuréka	62%	0,576	0,665	0,633	0,611
ČSFD	85%	0,906	0,806	0,824	0,895
Scuk	60%	0,846	0,353	0,566	0,697
Mix	69%	0,693	0,702	0,704	0,691

Tabulka 10 - výsledky testování ČSFD (délka zpráv 20-900, limit 40000) jako trénovací sady oproti ostatním zdrojům

7.2.4 Scuk

Počet dostupných proslovů je velmi malý, proto je úspěšnost pouze 59%, maxent klasifikátor byl pod hranicí 50%. Sensitivita na množině pozitivních zpráv dosahovala 75%, na negativních 43%. To je projev malé mezi „vzdálenosti“ množinami, viz 4.4.1 Vytvoření korpusu.

7.2.5 Výsledná sada

Výsledná trénovací sada byla na základě těchto výsledku sestavena následovně:

	Twitter	**Heuréka**	**ČSFD**	**Scuk**
Počet zpráv (pro každou třídu stejný počet)	56 190 (všechny dostupné)	100 000	70 000	584 (všechny dostupné)
Délka zpráv	1 - 140	1 - 300	1 - 300	Není omezeno
Průměrná délka zprávy	75	73	138	2270

Tabulka 11 - parametry unifikované testovací sady

Jako příznaky byly použity unigramy s modifikátory a filtry uvedenými výše. Z nich jsem následně vybral pouze ty, jejichž četnosti byly větší jak jedna. Bylo experimentováno s odstraněním nevýznačných příznaků, resp. bylo zanecháno 75% s největší významností, nicméně nedošlo ke kladnému ovlivnění výsledků. Významnost $c(příznak)$ je počítána jako podíl nejvyšší a nejnižší pravděpodobnosti pro každou třídu:

$$c(příznak) = \frac{P_{max}(příznak|s_1)}{P_{min}(příznak|s_2)}$$

Výsledky jsou zde:

	Úspěšnost	**+ Recall**	**– Recall**	**+ Precision**	**– Precision**
Twitter	66,8%	0,494	0,832	0,745	0,608
Twitter man.	61,1%	0,481	0,831	0,818	0,532
Heuréka	87,4%	0,775	0,964	0,952	0,805
ČSFD	85,5%	0,817	0,904	0,893	0,844
Scuk	73,9%	0,680	0,770	0,745	0,705
Mix	81,3%	0,727	0,881	0,871	0,761

Tabulka 12 - výsledky testování finální sady oproti ostatním zdrojům

Pokud vezmeme v úvahu průměrnou úspěšnost, je lepší oproti natrénování jenom z jednoho zdroje, sice nedosahuje takových maximálních úspěšností, nicméně ostatní parametry jsou vyrovnanější.

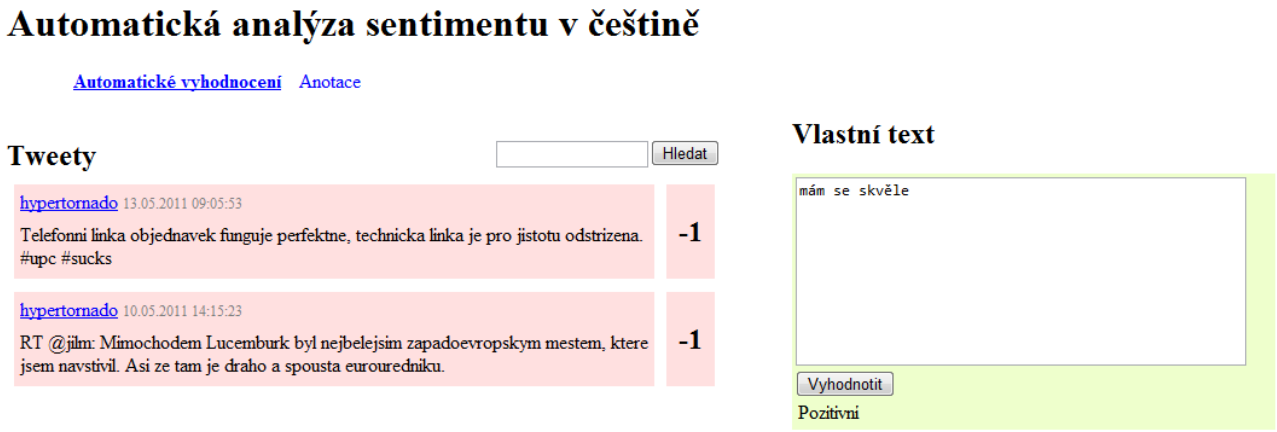
Největší problémy činí opět Twitter, kde systém selhává při klasifikaci pozitivních zpráv. Důvodem je především to, že lidé k sarkastickým a špatným zprávám připojují veselé smajlíky, aby zmírnili rozhořčení. Naopak při detekci negativních jsou výsledky dobré a myslím, že umožňují užití tohoto systému v praxi. Pokud by se ale zkoumaly zprávy s výskytem klíčového slova, které by vyjadřovalo značku, firmu nebo určitý typ výrobku, úspěšnost by byla vyšší, zvláště u krátkých zpráv, protože tam jsou lidé ve svých výpovědích mnohem konkrétnější.

Bylo otestováno použití slov jenom s určitými slovními druhy (slovesa, přídavná jména a podstatná, přídavná jména, slovesa, příslovce), ale většina statistik byla o 10% horších.

Zkoušel jsem otestovat systém novinovými titulky a abstrakty s proslovy se slovy „hoří“, „zavraždil“, „zemřel“ apod., ale vzhledem k tomu, že v trénovacích datech je výskyt velmi malý, nebyl systém často úspěšný. V rámci rozšíření by bylo vhodné přidávat těmto význačným slovům větší váhu na základě slovníku.

7.2.5.1 Prezentační aplikace

Jak implementovaný bayesovský klasifikátor natrénovaný na výsledné sadě funguje v praxi, je prezentováno na webové aplikaci dostupné na http://athena3.fit.vutbr.cz:30101/stud/xvodic03. Je možné si nechat vyhodnotit libovolnou zprávu nebo sledovat tweety, jak je hodnotí sám systém. Rozhodnutí klasifikátoru je možné upravit a tím tak zpřesnit jeho budoucí vyhodnocování.



Obrázek 13 - ukázka prezentační aplikace

8 Závěr

Cílem práce bylo realizovat systém, který za pomoci pokročilých metod strojového učení v oblasti zpracování přirozeného jazyka, vzhledem k šířce záběru bylo téma po domluvě s mým vedoucím upřesněno na analýzu zpráv ze sociální sítě Twitter. Použitý bayesovský klasifikátor sice nepatří mezi vysoce sofistikované metody rozpoznávání, nicméně jeho vhodné použití pro tento problém, kde je k dispozici velké množství dat, se ukázalo být jako dobré rozhodnutí.

8.1 Shrnutí vlastní práce

Byl vytvořen systém pro základní analýzu sentimentu pracující s českým jazykem, která může být reálně použita při zpracování velkého množství neznámých dat například pro marketingové nebo sociologické studie.

Pro prezentaci funkčnosti byla vytvořena webová aplikace, která graficky zastřešuje systém, jenž v téměř reálném čase sleduje nové příspěvky (a obchází tak přísné limity Twitter API) vybraných uživatelů na Twitteru a určuje jejich sentiment. V případě, že návštěvník nesouhlasí s výsledkem, může zprávu ručně ohodnotit a tím dojde k přetrénování klasifikátoru za běhu. Uživatel si též může vyhledat určité zprávy dle klíčového slova a také je možno si nechat vyhodnotit vlastní zprávu.

Dále byly naimplementovány skripty pro stažení věcného obsahu ze serveru Heuréka, ČSFD a Scuk. Vhodnost všech zdrojů dokládají provedené analýzy. Pro pokusy o zvýšení přesnosti byly vytvořeny funkce v jazyku Python, které umožňují práci dávkovou práci s PDT. Pro potřeby testování bylo ručně ohodnoceno přes 2000 českých tweetů třemi anotátory a pro tento účel vznikla jednoduchá anotační webová aplikace.

Jako vedlejší produkty několika slepých uliček, vznikl crawler tweetů pro službu Klábosení, který nebyl použit z důvodů licenčních podmínek a crawler uživatelů pro službu Czechia Twitter, jejichž seznam jsem nakonec obdržel po dlouhé době ve formátu CSV. Snaha ukládání N-gramů a POS tagů do databáze, kdy se ukázalo, že výkon klesá až moc exponenciálně s počtem přibývajících řádků a nakonec čtení z ní trvalo déle než samotné vyhodnocení pomocí PDT i při řádném užití indexů.

Práce pro mě měla osobní přínos z hlediska práce s „většími“ daty, což byla moje první větší zkušenost s takovým objemem, kdy je snadné rychle vyčerpat všechny dostupné systémové prostředky. Dále seznámení se s problematikou zpracování přirozeného jazyka a klasifikace.

8.2 Další vývoj a rozšíření

Největším přínosem pro budoucí rozvoj, který by zlepšil generalizaci klasifikátoru, by bylo doplnění dat z dalších specifických zdrojů zaměřující se na pohostinství, dovolené, automobily apod. Přesnost při testování současných dat by se nezvýšila, a když, tak jen nepatrně, ale byl by to velký přínos pro zvýšení generalizace. Přibyla by hlavně některá význačná podstatná slova a slovesa, jiné slovní druhy by nejspíše zasaženy nebyly, jelikož většina už je obsažena v současných datech.

Velkým, obtížně řešitelným neduhem je, že klasifikátor je binární a nerozpoznává neutrální zprávy. Příčinou je, že jsem nebyl schopen najít dostatečně kvalitní zdroj dat, který by tyto texty značil. Práce, s nimiž jsem se setkal, řešily tento problém použitím článků z novin. S tímto dělením však zásadně nesouhlasím, protože i jen po letmém zhlédnutí je hned jasné, že novinové články a titulky novin našich i zahraničních deníků rozhodně nenesou žádné neutrální zprávy, spíše naopak. Jedním z možných řešení se možná nabízí v použití encyklopedických článků. Tam sice nenajdeme citově zabarvené věty, ale obsah i přesto může značit pozitivní či negativní dopady. Bylo by tedy nutné použít obezřetného manuálního výběru.

Bylo by vhodné otestovat více klasifikátorů, zda by nepodávaly lepší výsledky, nicméně při studování rešerší a dostupných materiálů se jeví použitý naivní bayesovský klasifikátor jako nejlepší a to nejen z důvodu dobré přesnosti, ale i implementační jednoduchosti.

Pro zvýšení výkonu aplikace by bylo dobré, kdyby morfologický analyzátor v PDT byl nahrán v operační paměti (běžel by tedy například jako serverová aplikace) a nemusel by se při každém požadavku znovu spouštět.

Bibliografie

1. HAYES, N. *Základy sociální psychologie*. Praha: Portál, 1998, 112 s.. ISBN 80-7178-198-3. Kapitola Měření postojů.
2. OSGOOD, SUCI a TANNENBAUM. *The Measurement of Meaning*. Board of Trustees of the University of Illinoas, 1957. ISBN 0-252-74539-6.
3. RESEARCH, M. *Carl McDaniel; Jr. and Roger Gates*. 8. Wiley, 2009. ISBN 9780470087022.
4. WEISS, P. a J. ZVĚŘINA. *Sexuální chování v ČR situace a trendy*. Portál, 2001. ISBN 80-7178-558-X.
5. FRIJDA, N. H. *The Emotions*. Cambridge, UK: Cambridge University Press, 1986, 207 s..
6. WIEBE, J. M. Identifying subjective characters in narrative. In: *International Conference on Computational Linguistics*. 1990, s. 401-08.
7. *Foundations and Trends® in Information Retrieval*. 2008, č 2. ISBN: 978-1-60198-150-9. Strany 1-135.
8. TURNEY, P. D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, USA: 2002, s. 417-24.
9. VASSERMAN, S. a S. OWSLEY. In: *Exploring Mood on the Web* [online]. 2009 [cit. 2012-04-16]. Dostupné z: http://www.icwsm.org/2009/data/sood_vasserman_icwsm09_final.pdf
10. PANG, B. a L. LEE. In: *A Sentimental Education: Sentiment Analysis Using Subjectivity* [online]. [cit. 2012-Duben-30]. Dostupné z: http://acl.ldc.upenn.edu/acl2004/main/pdf/319_pdf_2-col.pdf
11. PAK, A. a P. PAROUBEK. In: *A Sentimental Education: Sentiment Analysis Using Subjectivity* [online]. [cit. 2012-Duben-30]. Dostupné z: http://deepthoughtinc.com/wp-content/uploads/2011/01/Twitter-as-a-Corpus-for-Sentiment-Analysis-and-Opinion-Mining.pdf
12. *AI Magazine*. AAAI, 1997, **4** (18), 45-64 s..
13. Mashable Social Media. *Twitter Has 100 Million Active Users* [online]. 11. září. 2011 [cit. 2012-duben-16]. Dostupné z: http://mashable.com/2011/09/08/twitter-has-100-million-active-users/
14. PEAR ANALYTICS. [online]. 2009 [cit. 2011-Listopad-20]. Dostupné z: http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf
15. APPELTAUEROVÁ, L. et al. In: *H1* [online]. 8. červenec. 2011 [cit. 2012-Duben-17]. Dostupné z: http://www.h1.cz/studie-cesko-v-socialnich-sitich
16. *LUPA.cz* [online]. 13. Březen. 2012 [cit. 2012-Duben-17]. Dostupné z: http://www.lupa.cz/

clanky/vodafone-pacha-datovou-sebevrazdu-v-primem-prenosu-na-socialnich-mediich/

17. ATAXO. In: *Sentiment analýza bez sentimentu, zdrojová data ke studii* [online]. [cit. 2012-Duben-22]. Dostupné z: http://www.slideshare.net/josefslerka/sentiment-analza-bez-sentimentu-8570897
18. Lupa.cz. *O sentiment analýze bez sentimentu aneb jeden malý experiment* [online]. 2011. Červenec. 14 [cit. 2012-Duben-22]. Dostupné z: http://www.lupa.cz/clanky/o-sentiment-analyze-bez-sentimentu-aneb-jeden-maly-experiment/
19. *Scuk - Hodnotitelé* [online]. [cit. 2012-Duben-22]. Dostupné z: http://www.scuk.cz/hodnotitele/
20. *Positional Tags: Quick Reference (Czech "HM" Morphology)* [online]. 2000 [cit. 2012-Duben-20]. Dostupné z: http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html
21. *Wikipedia, Bayes Theorem* [online]. [cit. 2012-Duben-07]. Dostupné z: http://en.wikipedia.org/wiki/Bayes'_theorem
22. *Wikipedia, Additive Smoothing* [online]. [cit. 2012-Duben-08]. Dostupné z: http://en.wikipedia.org/wiki/Additive_smoothing
23. VOMLELOVÁ, M. In: *Slajdy k předmětu Strojové učení na matematicko-fyzikální fakultě Univerzity Karlovy v Praze* [online]. [cit. 2012-Duben-07]. Dostupné z: http://kti.mff.cuni.cz/~marta/sliMarts.pdf
24. Multinomial logit. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikipedia Foundation, 2011- [cit. 2012-05-08]. Dostupné z: http://en.wikipedia.org/wiki/Maximum_entropy_classifier
25. Decision tree learning. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikipedia Foundation, 2011- [cit. 2012-05-08]. Dostupné z: http://en.wikipedia.org/wiki/Decision_tree_learning
26. ComScore. *Online Consumer-Generated Reviews Have Significant Impact on Offline Purchase Behavior* [online]. 29. listopad. 2007 [cit. 2012-duben-16]. Dostupné z: http://www.comscore.com/Press_Events/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior
27. *Klábosení* [online]. [cit. 2012-Duben-16]. Dostupné z: http://www.klaboseni.cz

Seznam příloh

Příloha 1. Regulární výrazy pro klasifikaci zpráv dle smajlíků

Pozitivní: ([xXoO8:;=*\+-]+[D>\)\]\}]+)|(D{2,})|([\)]{2,})|((\^|)[\[\(\{]+[8:;=\+-]+($|))
Negativní: ([^h][^t]{2}[^p][^s]?([xo8:;=\+-])+**(([\[\{[^Đ]])+|([\(])+|(/+)))|([\(]{2,}[^:])|(:+\-+<+)|([\)]\-?:)|(8+[\-]+[\(]+)|(:+[*]+[\(]+)

Příloha 2. Regulární výrazy pro zpřesnění detekce češtiny

„Detekce“ češtiny
([ěščřžýáíéúťňůďó]|(^(super|hura)$)|((^|["%:\\.\\,\\!\\?])(ne|nic|ve?|ano|ty taky|vse ok|narozdil|neni|pani|promin|(ne)?vidim|dobra|casu?|uspora|to te|alternativa|vyse|vse|viz|jo|neda|to j(sou|e)|problemy?|hodiny?|pry|na|tom|katalog|deje|co|se?|nic|moc|oci|proc|je|kdyz|nebo|ani|totiz|proto|a tak|kvuli|jenomze|avsak|primo|nadto|jak|lenka|jednak|hned|bud|nebot|tedy|tudiz|nejen|(do)?konce|nybrz|to by|naopak|jenze|sice|ale|jiste|ze?|jenomze|prestoze|trebaze|jelikoz|jestlize|jestli|kdyz|vzdyt|totiz|nadto|lec|nybrz|vodka|krutej|prima|tohle|si?|cas|setri|za|tak)(^|["%:\\.\\,\\!\\?]))|vypn|urcit[eya]?|delaj|funguj|sledujem|problemo|beztobav|velmi|rychl|procpak|prijed|doporuc(uj|il)|inspirativn|zavir|hezk|mil[ae]|hnusn|pripravi|vyzkouseno|spatn|odporn|hloupost|oskliv|originaln[ei]|shit|nuda|naprost|hudba|meli|vubec|bohuzel|jeden|zklamani|velik[ey]|slabsi|slab[ye]|nekolik|zajimav[ye]|hodne|mal(icka)?[oyea]|nema|mene|prehledn|perfektn|zmakno?ut|vhodn|mozno|zkouset|slabin|pohod|snad|trapn|nudn|stihnul|drah[ya]|bordel|v t[ye])

Vyloučení slovenštiny, angličtiny a ruštiny
([бвгджзийклнптфцчшщъыьэюя]|i\'m|too|doesn\'t|i did|does|to get|how can|is the|in any|longlife|wanna|((^|["%:\\.\\,\\!\\?])(ťažk[eo]|isté|čo|tie|solídn[eoy]|som|niekd[ye]|aj|iný|páči[tl]|ako|pre|so|bol|nie|ktorom|prečo|scenár|načo|mňa|najkrajší|keď|ak[eú]|typicku|inak|bšie|diel|tých|lepšie|znenia|ktorú|dlhšia|geniálny|rokoch|som|dajú|pozerat|najlepší|iný)($|["%:\\.\\,\\!\\?]))|upratovač|písať|súhlasím|vobec|nebolo|hviezdičiek|našiel|príbeh|zaspať|naozaj|komédia|zab[uú]dnut[aá]|požičanú|kópia|hviezdičky|found|looking|pozerám|stiahnut|obchádza|povedal|baviť|väčšinu|ďalšia|poňaté|zdlhave|zaujímavým|vynikajúci|povinnosť|môžem|rýchlosť|opäť|štastie|televízna|réžia|nudn[ya]|vyzera|nezmysl|tazko|ã)

Příloha 3. Použitá stopslova

a, tímto, můj, svým, po, jak, si, s, v, ve, za, z, na, jako, ta, před, jejich, ještě, až, také, vaše, která, náš, pokud, pro, u, na, její, jej, tomto, tohle, tuto, tyto, jej, zda, proč, máte, tato, kam, tohoto, tohodle, tamtoho, kdo, mi, nám, své, tom, tomuto, proto, který, nebo, co, ten, ta, tak, při, po, od, toho, by, protože, asi, ho, svých, aj, tu, tedy, této, kde, naší, což, tím, takže, tím, takže, k, ke, nad, čí, pod, těma, mezi, přes, ty, pak, vám, ani, když, však, tento, aby, nebo

Příloha 4. Zdrojové texty a data na přiloženém CD/DVD

Příloha 5. Databázové schéma