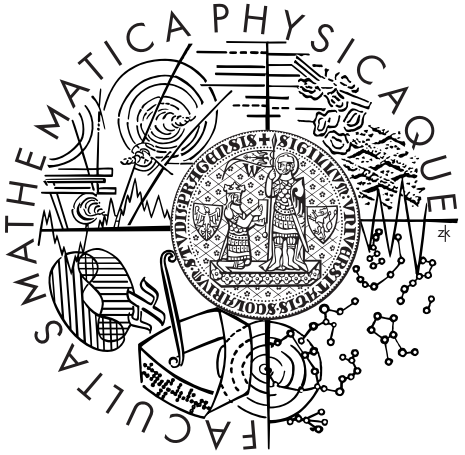
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jan Hajič

Popularita osob automaticky

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar Ph.D.
Studijní program: Informatika (B1801)
Studijní obor: obecná informatika

Praha 2011

Chci poděkovat Kateřině Veselovské a Janě Šindlerové za anotování, Václavu Novákovi ze společnosti Captaworks CZ s.r.o. za poskytnutí dat, Ondřeji Bojarovi za trpělivé a důsledné vedení a především svojí rodině za vše, díky čemu jsem tuto práci vůbec mohl psát.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenù, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona c. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Popularita osob automaticky

Autor: Jan Hajič

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar Ph.D., ÚFAL

Abstrakt: Možnost automaticky sledovat popularitu osob v novinách by jistě uvítaly nejen tyto osoby samotné. Počítačové zpracovávání subjektivity je sice rychle se rozvíjející podobor komputační lingvistiky, v češtině ovšem vůbec pro analýzu subjektivity a polarity v publicistice neexistují data. Začali jsme tedy s tvorbou ručně anotovaného korpusu polarity z českých publicistických textů, které se ovšem pro takové zpracování ukázaly jako krajně nevhodné. Dále jsme navrhli klasifikátor založený na statistických metodách, který by měl na základě tohoto korpusu popularitu sledovat, a otestovali jsme ho na korpusu recenzí bílého zboží a orientačně na zárodku našeho korpusu vět z novinových článků. Jako model jsme použili automaticky extrahovaný unigramový slovník, tři příbuzné metody pro zjišťování polárních lemmat a množství filtrů pro selekci relevantních lemmat. Na recenzích bílého zboží jsme dosáhli výsledků srovnatelných se světovým výzkumem už se základním modelem, naopak u českých publicistických textů vidíme kvůli jejich charakteru možný příslib až u více lingvisticky orientovaných metod.

Klíčová slova: subjektivita, anotace polarity, strojové učení, klasifikace textů

Title: Automatically Measuring Popularity of Persons

Author: Jan Hajič

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Ondřej Bojar Ph.D., ÚFAL

Abstract: Having the possibility of automatically tracking a person's popularity in the newspapers is an idea appealing not just to those in the media spotlight. While sentiment (subjectivity) analysis is a rapidly growing subfield of computational linguistics, no data from the news domain are yet available for Czech. We have therefore started building a manually annotated polarity corpus of sentences from Czech news texts; however, these texts have proven themselves rather unwieldy for such processing. We have also designed a classifier which should be able to track popularity based on this corpus; the classifier has been tested on a corpus of product reviews of domestic appliances and some introductory testing has been done on the nascent news corpus. As a model, we simply extract a unigram polarity lexicon from the data. We then use three related methods for identifying lemma polarity and a number of simple filters for feature selection. On the domestic appliance data, our simplest model has achieved results comparable to the state of the art, however, the properties of Czech news texts and preliminary results hint a more linguistically oriented approach might be preferrable.

Keywords: subjectivity, annotating polarity, machine learning, text classification

Obsah

1. Úvod

Automatické měření popularity osob je úloha počítačové lingvistiky z oboru Sentiment Analysis (analýza „sentimentu“, podrobněji v kapitole 2). Směřujeme k systému, který si „přečte“ noviny a „udělá si názor“ na vybrané osoby, o kterých se píše — přesněji řečeno bude odhadovat, zda budou případnému čtenáři spíše sympatické nebo nesympatické na základě klasifikace článků pro dané osoby.

Klasifikace textů podle polarity je úloha, která se již několik let řeší pro data jako recenze zboží či filmů. Takové texty mají zásadní výhodu v tom, že přichází rovnou označená (pisatel prakticky vždy k textu přidává nějaké hodnocení typu počet hvězdiček) a navíc jsou většinou — až na delší filmové recenze — lingvisticky nekomplikovaná. Oproti tomu publicistické texty před nás staví množství dalších překážek: jednak větší syntaktickou sofistikovanost, jednak výrazně složitější strukturu polarity — množství hodnocených a hodnotících entit, nepřímá hodnocení vyplývající ze znalosti světa apod. Především ale nepřichází s žádným kvantitativním ohodnocením od autora — je nutné na nich udělat anotace.

Tvorba dat pro analýzu sentimentu pro jazyky s bohatou morfologií (a vlastně všechny kromě angličtiny) je v současnosti v začátcích. (Téměř stejnou úlohu jako my řeší např. [8] na arabštině.) Pokud je nám známo, pro češtinu ještě žádná taková data neexistují. Postup, kterým jsme data anotovali, a potíže, na které jsme při nich naráželi, popíšeme v kapitole 3.

Pro klasifikaci jsme použili slovníkový klasifikátor založený na přesnosti jednotlivých lemmat (tak, jak je popsána v [2]). Srovnali jsme jeho výsledky při klasifikaci recenzí bílého zboží získaných ze serveru Mall.cz a námi oanotovaných segmentů a článků. Jako alternativu jsme použili naivní Bayesův klasifikátor. Podrobně jsou klasifikátory a jejich výsledky popsány v kapitole 4.

Zatímco na recenzích bílého zboží jsme dosáhli v současnosti běžných výsledků pro srovnatelný model (srovnání s [5]), klasifikace novinových článků výrazně nepřekonala jednoduchou baseline. Toto přičítáme především malé velikosti dat (viz sekce 4.6.2), která zase plyne z náročnosti anotací a problémů zvoleného anotačního schématu (sekce 3.2.1, 3.2.2). Kvůli příliš malému množství dat navíc nebylo možné zapojit latentní sémantickou analýzu (LSA) [1] pro automatické vyhledávání polárních lemmat v neoanotovaných datech.

2. Subjektivita a polarita v textu

Abychom mohli úspěšně celý experiment navrhnout, potřebujeme základní porozumění tomu, jak se pro nás relevantní jevy v jazyce realizují. Načrtneme tedy nyní stručný lingvistický model toho, jak v textu může být někdo hodnocen.

2.1 Subjektivita

Anglický výraz „sentiment“ nelze do češtiny dokonale přeložit. Slovník Lingea uvádí jako překlad „citový náboj“, blíží se také výrazy „postoj“, „smýšlení“, „nálada (vůči něčemu)“, „náhled“. Analýza smýšlení, jak jsme se rozhodli Sentiment Analysis překládat, se zabývá subjektivitou, subjektivními stavy v promluvách — momenty, kdy je vyjadřován osobní názor.

Obecnou charakteristiku subjektiviy a terminologii z velké části čerpáme z [2].

Subjektivním stavem tedy rozumíme takové místo v jazykovém projevu, kde mluvčí (nebo zprostředkovaný mluvčí) vyjadřuje svůj osobní názor. Takto vymezujeme subjektivitu vůči projevu objektivnímu, kde se pouze sděluje informace. Je zřejmé, že hranice mezi subjektivními a objektivními stavy není zdaleka ostrá; taková už je ovšem povaha subjektivity v jazyce. Vágnost provází subjektivitu na všech úrovních: byť existují jednoznačně subjektivní slova (explicitně hodnotící výrazy: výborně, špatné, líbit se, fuj), daleko častěji je subjektivita přítomná skrytě: v předpokladech („Tam policie postupem času zadržela na sedmdesát lidí.“ – případná subjektivita, resp. polarita — viz níže — muže pramenit např. z nedůvěry v policii), v „utvrzující predispozici“ (confirmation bias; „Podle Řápkové jde zřejme o chybu univerzity.“ — příznivce Řápkové souhlasí, odpůrce ji spíše obviní ze lži), ve zvolené slovní zásobě („Kmotr Mrázek…“). Navíc častejší než slova ryze citová jsou slova citově zabarvená, slova s pozitivními/negativními konotacemi apod. — schopnost vyjadřovat a správně interpretovat subjektivitu je tak úzce spjata se znalostí světa (world knowledge). Dále záleží na účastnících promluvy, zda a jak subjektivitu zpracují: kolikrát se stalo, že co bylo myšleno jako nevinná poznámka bylo pochopeno jako hrubá urážka. Do této směsi lidských a počítačům silně cizích jevů navíc přistupuje ještě ironie, která už na textových prostředcích nezávisí prakticky vůbec. Analýza subjektivity musí nějak zohlednit všechny tyto momenty vágnosti a nepolapitelnosti (případně i tím, že je bude ignorovat).

Navzdory těmto výzvám stále předpokládáme, že subjektivita je vyjadřována strojově zpracovatelným způsobem: že pro její vyjadřování používáme popsatelné, v textu (třeba nepřímo) pozorovatelné a počítači tak srozumitelné prostředky.

Zadání této práce nám navíc dovoluje omezit se na analyzování jedné podmnožiny subjektivních stavů: budou nás zajímat pouze momenty, kdy je vyjadřováno nějaké hodnocení (pozitivní či negativní): tzv. polární stavy.

2.2 Polarita

V polárních stavech rozeznáváme strukturu mluvčí — polární prvek — cíl. Mluvčím rozumíme entitu hodnotící, cílem entitu hodnocenou a polárním prvkem ty jazykové prostředky, které hodnocení vyjadřují. Lexikální polární prvky pak nazýváme polární výrazy.

Polární stavy dále mají orientaci: zda je cíl hodnocen pozitivně či negativně, a ev. i intenzitu: jak silně pozitivně či negativně je hodnocen. Budeme se dále zabývat pouze orientací.

Chceme hledat takové prvky polarity, které jsou co nejméně vázány na osobu mluvčího a adresáta; byť z povahy subjektivity není možné komunikující osoby zcela vyloučit, nebudeme se na tuto rovinu vágnosti v subjektivní komunikaci nijak ohlížet. Dále se vůbec nebudeme věnovat ironii (z podstaty statistických metod buď ironii stejně odhalí anotátoři, nebo není důležitá, nebo stejně není rozumně klasifikovatelná bez prosodických rysů; v českých publicistických textech — vyjma komentářů — se navíc prakticky nevyskytuje).

Alternativní model polarity můžeme vidět např. u [4], kde se pracuje se čtveřicí [téma,mluvčí,výrok,orientace] (přičemž, aby nebylo terminologického zmatku málo, pro orientaci používají termín „sentiment“).

Nastíníme nyní, jak se k jednotlivým částem polárních stavů budeme chovat.

2.2.1 Mluvčí a cíle

Vzhledem k tomu, že se zabýváme polaritou pouze z hlediska působení na čtenáře a neanalyzujeme názory jednotlivých mluvčích, nebudeme se zabývat identitou mluvčího vůbec. Tímto sice ztrácíme další úroveň přesnosti — orientace čtenáře k mluvčímu je zásadní pro evaluaci orientace polárního stavu — avšak dále nám to dovoluje omezit problém do zvladatelných mezí. Navíc nás nezajímá působení na jednoho konkrétního čtenáře, nýbrž průměrné působení ve skupině všech čtenářů. (Abychom tuto „průměrnou“ perspektivu zachovali, bylo nutné anotátorům přikázat neutralitu. Viz kapitola Data, sekce Anotace.) Tímto zjednodušením bychom tedy neměli nic ztratit, pokud platí, že důvěra čtenářů k danému mluvčímu se posčítá na stejné číslo pro všechny mluvčí.

Stejnou úvahu provedeme pro cíle a také je takto anonymizujeme.

2.2.2 Polární výrazy

Polární výrazy považujeme za primární nositele polarity, včetně orientace a intenzity. Na základě polárních výrazů budeme rozpoznávat, zda je v textu polarita přítomná, a v anonymizovaném modelu také polární výraz sám určuje orientaci stavu. Platí ovšem, že určitý výraz nemusí nést polaritu, kdykoliv se v textu vyskytne. Každému výrazu (slovu, lemmatu, n-gramu…) tak můžeme přiřadit pravděpodobnost, s jakou je polárním výrazem v nějakém (nějak orientovaném) polárním stavu.

Je třeba rozlišit dva pohledy na detekci polarity: zda je daný výraz nositelem polarity, tedy přímo polární výraz, a zda je výraz indikátorem polarity, tedy zda je pravděpodobné, že se v nějakém jeho okolí vyskytuje nositel polarity. Zde děláme most mezi lingvistickým a matematickým aspektem detekce polárních výrazů: zatímco nositel polarity je lingvistický pojem, indikátor polarity už je pojem čistě matematický.

2.2.3 Příklady polárních stavů

Uvedeme nyní některé typické (tzn. anotátoři se shodli) příklady polarity z našich dat (viz kapitola 3), na kterých ilustrujeme základní chování polarity:

(1) Podle ekologů je nicméně dobře, že Zeman dnes uviděl skutečný stav na Šumavě.

S touto větou máme lehkou práci. Mluvčí: „ekologové“, polární prvek: „dobře“, cíl: „že Zeman dnes uviděl skutečný stav na Šumavě“, orientace: pozitivní.

(2) Lidovci podle něj dále preferují termín voleb v listopadu.

Cílem je „termín voleb v listopadu“, polárním prvkem „preferují“, potenciální mluvčí jsou v této větě dva: jednak „on“, jednak „Lidovci“. Kdo je mluvčím v polárním stavu? Lidovci, záhadný „on“ žádné vlastní hodnocení neříká.

Jak by se situace změnila, kdybychom znali kontext?[1]

(3) Předseda poslaneckého klubu ČSSD Rostislav N. novinářům odpovídal na otázky o právě ukončeném jednání za zavřenými dveřmi. Lidovci podle něj dále preferují termín voleb v listopadu. „Tuto variantu pořádání voleb v současné situaci ale považuji za velmi nešťastnou,“ rozhořčeně dodal.

Do jaké míry hodnotí smyšlený funkcionář v druhé větě i Lidovce? Do jaké míry jeho přímé hodnocení „této varianty pořádání voleb“ ovlivní výsledné sympatie čtenáře k Lidovcům? (Do naší interpretace polárních stavů v této promluvě

[1] Vymyšlený pro tento příklad.

zvláštním způsobem přimíchává výrazně citově zabarvené „rozhořčeně“, které nasvědčuje tomu, že zmíněný R. N. ve skutečnosti chce hodnotit Lidovce. Co je na větě podivného, že ji nečteme v doslovném znění, že vyhodnotíme polaritu jinak, než jak nám syntaktická struktura naznačuje? Možná diskrepance nábojů — „sentimentů“ — mezi citově zabarvenými slovy.)

(4) Podle ve dení parku je podobný odhad nesmyslný. „Přemnožení kůrovce je problémem člověka [...protože odvykl pohledu na suché stromy].“

V první větě je mluvčí „vedení parku“, polární prvek „nesmyslný“ a cíl „podobný odhad“. Druhá věta vypadá na první pohled také jasně: zůstává mluvčí „vedení parku“, cíl je „přemnožení kůrovce“ a polární prvek „problémem“. Poněkud zvláštní je ovšem role slova „přemnožení“ — samo má negativní konotace (pravděpodobně spojené s předponou pře–), takže je vlastně také nějak polární prvek. Ovšem: není jasné, jestli se z věty samotné vůbec nějakou polární informaci dozvídáme — z aktuálního členění je vidět, že mluvčímu o kůrovce nejde. Slovo „problémem“ pak používá ne proto, že by chtěl hodnotit přemnožení kůrovce, nýbrž pouze jako predikaci (v přísudku jmenném se sponou: „je problémem“) odpovídající předpokladu věty: že přemnožení kůrovce je negativní jev. Větu můžeme přepsat jako: „Přemnožení kůrovce je vnímáno negativně pouze z perspektivy člověka.“ Mluvčí se spíše snaží očistit „přemnožení kůrovce“ od nespravedlivých (z jeho pohledu) negativních konotací. Nicméně skutečnost, že obě anotátorky zvolily první interpretaci, možná naznačuje, že člověk vnímá polaritu vlastně poměrně prvoplánově, což by mohlo automatickému zpracovávání nahrávat.

Vidíme, že polarita představuje svébytnou vrstvu jazyka přístupnou lingvistické analýze. Ta nám dává teoretické východisko pro automatické zpracování polarity.

3. Data

Použili jsme v našem experimentu dva charakterem odlišné zdroje dat: korpus stažených článků ze serveru Aktuálně.cz z období 2009–2010 a balík uživatelských recenzí bílého zboží ze serveru Mall.cz. Pro evaluaci jsme připravili 66 článků z Aktuálně.cz, aslespoň 10 pro každou vybranou cílovou osobu (Topolánek 20, Paroubek 10, Klaus 11, Fischer 11, Janota 14).

3.1 Články z Aktuálně.cz

Data z Aktuálně.cz mají celkem cca 560 000 slov v 1 661 článcích. Články ručně třídíme podle toho, jak jsou subjektivní, tzn. jestli je jejich celkové sdělení nějakým způsobem polární. V současnosti je ovšem velká vetšina stále neroztříděná. Polárních článků jsme identifikovali 175 (89 932 slov), nepolárních 188 (45 395 slov) a nerozhodných 90 (77 918 slov). Neroztřídených zůstává 1208 (333 601 slov). Jedná se o články z rubriky Domácí, tedy materiál zabývající se českou politickou scénou.

Tato skupina dat je pro systém zásadní, jelikož reprezentuje právě cílové texty — tématikou, a tedy i slovní zásobou a vyjadřovacím stylem. Ze své podstaty jsou však tyto texty obtížně analyzovatelné: česká publicistika se snaží o důslednou povrchovou neutralitu, tedy snaží se vystříhat veškerých explicitně hodnotících výrazů. České novinové texty jsou tedy i pro automatické zpracovávání polarity velmi náročné.

Ruční anotace prováděli celkem tři lidé, budeme je značit KV, JS a JH.

Ruční anotace segmentů proběhly zatím pouze na balíčku dvanácti polárních článků, celkem 410 segmentů (6 868 slov). Shoda, měřená Cohenovou kappou, se pohybovala lehce nad 0.63. Výrazně nižší shodu než [2], [8] připisujeme zvolenému anotačnímu schématu. Více o anotacích v další části.

Nepočítali jsme samozřejmě s tím, že bychom anotace udělali pro celých více než půl milionu slov, většinu článků jsme zamýšleli použít jako trénovací data pro LSA — avšak z takto malého označkovaného korpusu nemáme šanci získat pro automatické hledání polárních výrazů pomocí LSA dostatečně dobré výchozí „jádro“ a kvůli úzkostné novinářské neutralitě nemůžeme jako toto jádro spolehlivě použít standardní explicitně polární výrazy (dobrý, špatný…), jelikož se v textech dostatečně nevyskytují.

3.2 Anotace segmentů článků z Aktuálně.cz

Segmenty anotovaly JS a KV. V současnosti je označkován balíček 410 segmentů z 12 náhodně vybraných polárních článků (celkem 6 868 slov, 1 935 různých lemmat). Segmenty jsou především věty, avšak je mezi nimi i malé množství nevětných titulků a podnadpisů. Zajímala nás polarita z hlediska čtenáře — jak sympatické mu budou cíle polárních stavů v textu poté, co si segment přečte (podrobně popsáno v anotačních instrukcích v příloze 1).

Vyzkoušeli jsme tři způsoby anotace: vybraných cílů, všech cílů a všech polárních prvků. Při anotování všech cílů anotátoři hledali všechny cíle polárních stavů a značili je podle polarity (pozitivní/negativní, značit navíc „přítomná, avšak nejasná“ anotátoři považovali za zbytečné), při anotování vybraných cílů anotátoři značili všechny polární výskyty dané osoby a při anotování polárních prvků hledali naopak slova, díky kterým větě přisuzovali polaritu. Součástí anotací byly i ruční opravy segmentace textu.

Metoda značkování vybraných cílů se navíc stejně ukázala jako příliš řídká, takže jsme od ní upustili (KV z 267 označila jen 43, JS ze 451 segmentů označila pouze 42). Navíc, jak počty segmentů napovídají, dělala kupodivu potíže i segmentace. Anotace všech polárních prvků trpěla stejnými neduhy jako anotace všech cílů, navíc je časově ještě o něco náročnější, takže v tuto chvíli ještě není hotová.

Anotace všech cílů dosáhla shody měřené Cohenovou kappou přes 0.63 pro polární, negativní i pozitivní věty. Celkem bylo označeno okolo 30 % segmentů (JS 122, KV 139).

Poznámka ke značení: *kurzívou* budou v příkladech značeny cíle negativního hodnocení, **tučně** jsou značeny cíle hodnocení pozitivního.

3.2.1 Problémy s anotováním cílů

Potíže, na které jsme při značení cílů naráželi, spadají do několika kategorií. Nejsnáze se opravují potíže technického charakteru, které pramení z nedostatečně přesných instrukcí (např. zda má být předložka na začátku cílové fráze označena také jako cíl). Dále jsme naráželi na legitimně problematické jazykové jevy a chyby vzniklé špatnou interpretací instrukcí. Vzhledem ovšem k vágnosti úkolu, se kterým se potýkáme, hranice mezi těmito dvěma kategoriemi není úplně jasná.

Objektivita anotace

Především může být velmi obtížné odfiltrovat nežádoucí vlivy osoby anotátora: na rozdíl od neosobního značení např. tektogramatických rolí se z principu potká-

váme se subjektivitou — dostáváme anotátora vlastně do paradoxní pozice, kdy se snaží objektivně značit subjektivní působení textu na sebe. Prakticky se nelze vyhnout osobním sympatiím a antipatiím anotátorů k entitám vyskytujícím se ve zpracovávaných textech, jelikož tyto se týkají vypjaté politické situace v letech před volbami 2010. (V této situaci by zřejmě pomohlo najít anotátory zahraniční, kteří dostatečně dobře umí česky, případně texty přeložit do angličtiny.) Ukázalo se však, že když na anotátory zmínění aktéři nehledí denně z titulních stránek, zapomenout na tehdejší emoce vcelku obstojně jde.

Příliš nadšení lingvisté

Potíží specifickou pro náš experiment byla skutečnost, že všichni anotátoři jsou lingvisté seznámení s dalším postupem, takže občas značili věty, ve kterých našli zajímavý polární výraz, i když ve větě samotné polarity příliš nebylo (nebo byla úplně jinde).

(5) A. Vláda schválila něco jiného, než co slibovala.
B. Vláda schválila **něco jiného**, než co slibovala.

Pokud je v této větě co k označení, pak je to vláda - negativně.

Naopak se výjimečně chovali lingvisté příliš málo nadšeně:

(6) A. Z 60 prací jich v knihovně fakulty chybí 35.
B. Z 60 prací jich v knihovně fakulty chybí 35.

Zde bychom spíše rádi viděli negativně označenou fakultu, byť zde žádný explicitně hodnotící výraz není (možným kandidátem na dobrý polární indikátor je leda tak „chybí“).

„Víceúrovňové“ cíle

Více potíží způsobila nejasná interpretace toho, co je vlastně cílem v polárním stavu: při rozboru polárního stavu přes jazykové roviny je totiž možné se „zastavit“ a cíl určit na různých úrovních.

(7) A. Dům *byl* před sedmi lety neúspěšně *dražen*, nyní je v zástavě banky.
B. *Dům* byl před sedmi lety neúspěšně dražen, nyní je v zástavě banky.

Anotátorka A říká, že je jí na základě věty méně sympatický proces dražby domu (pravděpodobně proto, že byl neúspěšný), zatímco pro B je entitou, o níž přijímá polární informaci, dům samotný (protože byl dražen neúspěšně). Zde preferujeme druhou variantu — těžko si můžeme myslet, že si čtenář spíše odnese dojem z dražby než z domu. Tyto potíže jsou možná způsobeny hledáním struktury

mluvčí — polární prvek — cíl jako analogické tektogramatické struktury actor — predikát — patiens: byť se obě struktury shodovat můžou (a často se shodují), není to pravidlem.

(8) A. *Tímto prohlášením* ovšem rozezlil šéfa partaje Mirka Topolánka, který *něco takového* zásadně odmítl.

B. Tímto prohlášením ovšem rozezlil šéfa partaje Mirka Topolánka, který něco takového zásadně odmítl. *EXTERN*

Zde A zaznamenává hodnocení „tohoto prohlášení“, resp. jeho obsahu („něčeho takového“), zatímco B hodnotí externího autora tohoto prohlášení.

Rozsah polarity

Dále se často objevovala diskrepance mezi lokální a globální polaritou — zatímco určitá část věty mohla samotná nést polární informaci, v kontextu celé věty se polarita ztrácí (nebo dokonce opačně orientuje).

(9) A. V případě jeho kandidatury na tento post by **jej** podporovalo pouze 13 % dotázaných, a to z řad voličů ČSSD a KSČM.

B. V případě jeho kandidatury na tento post by *jej* podporovalo pouze 13 % dotázaných, a to z řad voličů ČSSD a KSČM.

Zatímco B „jemu“ přiznává podporu a dál větu nezkoumá, A bere v potaz „pouze“, což interpretuje tak, že zmíněný „on“ příliš podporován není, a tedy si sám dělá spíše negativní názor. (Ještě se tedy naskýtá možnost, že nízkou podporu jeden anotátor interpretuje pozitivně a druhý negativně…) Preferujeme širší záběr. Je možné, že v tomto konkrétním případě k chybě A došlo také kvůli zmiňované „lingvistické deformaci“ ve snaze zachytit typickou orientaci „podporovalo“.

Rozdílná ochota souhlasit s mluvčím

Nejčastějším zdrojem neshod byla rozdílná ochota anotátorů souhlasit s mluvčím z polárního stavu, tedy přijmout jeho hodnocení za vlastní, nechat jeho názor působit na sebe. Vysoká frekvence těchto případů je dána také novinářským stylem, kdy se naprostá většina alespoň trochu explicitně hodnotících výrazů nachází v pasážích citujících.

(10) A. „Nikdo z nás nepodporuje bezzásahovost v celém parku.“

B. „Nikdo z nás nepodporuje *bezzásahovost* v celém parku.“

Tato příčina neshod přímo pramení z perspektivy, ze které anotátoři pracují, když značí působení textu na čtenáře. Nemusí se nutně jednat o problém: ve větě může být slabý polární stav, který na čtenáře působit může a nemusí, a pak prostě při výpočtu modelu tuto nejistotu zohledníme. Na druhou stranu se také může jednat o hlubší problém, neshodu nepřímo plynoucí z požadavku na neutralitu vůči mluvčím: anotátor nemusí být schopen působení polárního stavu na sebe vyhodnotit, pokud identitu mluvčího nezná (či podle instrukcí „zapomene“). Připomínáme nezájem anotátorek o kategorii „orientace nejasná“ — nejedná se zde o zorientování polárního stavu z hlediska čtenáře (zda souhlasí s mluvčím či mu oponuje), nýbrž o to, že čtenář bez znalosti identity mluvčího nedokáže v podstatné části případů vůbec říci, jestli polární stav přítomný ve větě na něj působí.

3.2.2 Jak anotace vylepšit?

Ukázalo se, že anotovat polaritu v českých publicistických textech je náročný úkol a naše anotace naráží na řadu prakticky neřešitelných problémů. Co tedy můžeme udělat?

Většina polarity, kterou jsme identifikovali v novinovém textu, je souhlas či nesouhlas s názorem nějakého mluvčího v novinách citovaného. Je tedy pravděpodobné, že za velkou část obtíží a neshod může kombinace požadované neutrality a perspektivy, kterou jsme pro anotace zvolili: působení textu na čtenáře. Z hlediska úlohy, kterou řešíme — simulování vlivu publicistického textu — je takový pohled jistě opodstatněný, v praxi je ovšem zjevně náročné na tak malé ploše jako je věta s rozumnou jistotou určit, jaký si odneseme dojem. Dokázali bychom asi dosáhnout lepší shody (a spolehlivějších anotací), kdybychom anotovali místo působení na čtenáře to, jak se hodnotí entity v textu navzájem.

Abychom ovšem dokázali z takových anotací splnit zadání, bylo by pro klasifikační systém nutné přidat vrstvu „důvěry“: identifikovat nějaké entity, u kterých je důležité znát důvěru čtenáře k nim, a přiřadit jim koeficient, kterým bychom modifikovali orientaci (a intenzitu) polárních stavů, kde jsou mluvčími či cíly. Navíc by pak při klasifikaci samotné bylo nutné automaticky mluvčí identifikovat, což je další ne zcela snadná úloha (zahrnuje např. řešení koreference).

Doufali jsme také, že perspektiva čtenáře bude výhodná pro identifikování těch polárních stavů, které jsou pro čtenáře relevantní. Ukazálo se však, že pokud anonymizujeme mluvčí a cíle, je obtížné vůbec polární stav identifikovat a zorientovat, a že úlohu identifikace polárního stavu a přiřazení relevance vůči čtenáři bude tedy lépe od sebe odlišit.

Jinou variantou, jak výsledky anotací vylepšit, by bylo zrušit anonymizaci,

požadavek na neutralitu — a drasticky rozšířit počet anotátorů, ideálně na reprezentativní vzorek čtenářské obce, abychom tak vyloučili zkreslení vzniklé názory anotátorů. Toho by se dalo dosáhnout zveřejněním nějakého anotačního nástroje na internetu (což s sebou ponese zase řadu dalších obtíží, například spolehlivost anotátorů, nehledě na náklady a propagaci).

3.3 Značkování celých článků

Pro účely evaluace jsme označili 66 článků (Topolánek 20, Paroubek 10, Klaus 11, Fischer 11, Janota 14). Značení prováděl anotátor JH podle instrukcí pro anotace vybraných cílů adaptovaných pro články (co článek, to „segment“). Stejným postupem jsme ještě označili 12 článků, které jsme značili po segmentech.

3.4 Recenze ze serveru Mall.cz

Ze serveru Mall.cz jsme získali balík 10 177 recenzí (158 955 slov, 13 473 unikátních lemmat) roztříděných do skupiny pozitivních (6 365) a negativních (3 812) samotnými autory recenzí.[1] Tato data je výrazně snazší zpracovat, jelikož ze své podstaty jsou zcela nepokrytě polární a navíc se v nich prakticky nevyskytuje komplikovanější syntaktická struktura než souvětí ve slučovacím poměru. Často v nich figurují explicitně polární výrazy a vzhledem k tomu, že v recenzích téměř není ironie, jejich užití je zcela prototypické. Navíc se v nich díky roztřídění nemusí nic značkovat, abychom měli anotaci gold-standard.

Prezentují ovšem zase jinou sadu obtíží (značně menších, než publicistické texty) — pravopisné chyby vnášejí do dat šum v podobě lemmat navíc (tagger objevil např. krásné české příslovce *lediodama*), některé recenze jsou špatně zatříděné, občas se obzvlášť v negativních recenzích vyskytují variace na téma „Nic“ či „Nevím“. (Těchto se ovšem většinu podařilo jednoduše odstranit.)

[1] Uživatelé serveru zanášejí recenze do formuláře, který obsahuje zvlášť pole pro negativní a zvlášť pro pozitivní recenze.

4. Experiment

Jak tedy zjistit, jak hezký si čtenář daného článku udělá o cílové osobě obrázek?

Abychom dokázali přesně určit, jaký dojem si o cílové osobě čtenář textu odnese, potřebovali bychom identifikovat, kde je cílová osoba cílem polárního stavu (pro čtenáře relevantního) a jakou orientaci tento stav má.

Pro účely našeho experimentu uděláme několik zjednodušení. Nebudeme zjišťovat, zda je cílová osoba skutečně cílem polárního stavu: předpokládáme, že každý výskyt cílové osoby v polárním stavu ji orientuje podle orientace stavu. Dále nehledáme hranice polárních stavů uvnitř segmentů — pokud v segmentu identifikujeme polární stav, předpokládáme, že se týká všech entit přítomných ve větě. Tato zjednodušení nám umožní pracovat s větou (příp. s kontextovým oknem nějaké šířky) podle modelu bag-of-words, tedy nebrat v potaz pořadí slov. Zároveň s sebou samozřejmě také nesou výrazné riziko nepřesností.

Dělení na trénovací a testovací data provádíme náhodně s opakováním (v náhodném pořadí procházíme segmenty a při poměru trénovacích a testovacích dat n:1 každý $n+1$ vybraný segment zařadíme do testovacích dat).

4.1 Lemmatizace

Všechna data prošla lemmatizací. Použili jsme Hajičův statistický tagger [3]. Z morfologických značek jsme zachovali POS (Part of Speech, slovní druh) a negaci. Kdykoliv tedy dále budeme mluvit o lemmatech, máme na mysli trojici $(lemma, POS, negace)$; množině všech těchto trojic vyskytujících se v daných datech budeme říkat slovník a značit ji L_{Data}.

4.2 Detekce polárních výrazů

Při detekci polárních stavů se zaměříme na polární výrazy. Potřebujeme najít funkci, která řekne, jak dobré je dané lemma jako indikátor existence polárního stavu, resp. negativně (pozitivně) orientovaného polárního stavu. Na základě zjednodušení, která jsme provedli, považujeme každé slovo v anotované větě obsahující polární stav za indikátor s orientací podle daného stavu. Řešíme tedy dvojí klasifikaci: za prvé, zda je segment polární, a za druhé, jak je orientovaný, pokud polární je.

Za hodnotící funkci pro lemma zvolíme jeho *přesnost*: odhad pravděpodobnosti (na základě dat), že segment obsahující dané lemma bude polární (s danou orientací). Přesnost spočteme takto:

$$prec(l, X) = \frac{freq(l, X)}{freq(L)} \tag{4.1}$$

kde l je dané lemma, $X \in \{\, nonpolar, polar, negative, positive \,\}$ zastupuje polaritu, resp. její orientaci, $freq(l)$ je počet výskytů l v datech a $freq(l, X)$ je počet výskytů l v segmentech dat s polaritou X. [2]

Říká nám $prec(l, X)$ skutečně něco o indikační síle? Vidíme, že

$$prec(l, X) = P_{data}(X \mid l), \tag{4.2}$$

kde $P_{data}(X \mid l)$ je pradvěpodobnost, že náhodně vybraný výskyt l se v našich datech bude nacházet v segmentu s polaritou X. S rostoucí velikostí dat tedy $prec(l, X) \to P(X \mid l)$. To je přesně to, co po indikační funkci chceme: aproximovat pravděpodobnost, že segment obsahující dané lemma obsahuje polární stav.

Baseline (pravděpodobnost, že náhodně vybrané slovo indikuje X) spočítáme potom jako [2]

$$baseline(X) = \frac{\sum_{l \in L_{Data}} freq(l, X)}{\sum_{l \in L_{Data}} freq(l)} = \frac{|Data_X|}{|Data|} \tag{4.3}$$

pro slovník L_{Data} a dostaneme tak hodnoty, proti kterým můžeme měřit, zda je dané l lepší indikátor než náhodné slovo; například od $prec(l, X)$ můžeme jednoduše $baseline(X)$ odečíst (nebo vydělit). Budeme značit $aprec(l, X) = prec(l, X) - baseline(X)$ („additive precision“), $mprec(l, X) = \frac{prec(l,X)}{baseline(X)}$ („multiplicative precision“). Všimněme si, že PMI$(l, X) = log\frac{p(X|l}{p(X)} = log(mprec(l, X))$. [1].

Pro baselines dále platí:

$$baseline(nonpolar) + baseline(polar) = 1 \tag{4.4}$$

$$baseline(negative) + baseline(positive) = baseline(polar) \tag{4.5}$$

Výchozím modelem je tedy rozšířený slovník $L^{\star}{}_{Data}$, který se skládá z uspořádaných 14-tic

[1] PMI značí Pointwise Mutual Information — často používaná míra asociovanosti, která mívá dobré výsledky. Příklad použití třeba viz `http://www.ling.uni-potsdam.de/~gerlof/docs/npmi-pfd.pdf` Najít vztah mezi PMI a *aprec* by mohlo být zajímavé, viz 4.6.1

$$(\, l, freq(l), prec(l, \dots), aprec(l, \dots), mprec(l, \dots)\,). \tag{4.6}$$

Pro každé lemma tedy máme takto spočtený odhad indikační síly na základě dat. (Obdobně lze slovník vybudovat i pro n–gramy, kde však zase více narazíme na řídkost dat. Ukazuje se ovšem, že n–gramy mohou výsledky vylepšit [2] [5].)

Přesnosti lemmat do rozšířeného slovníku z oanotovaných dat od jednotlivých anotátorů kombinujeme váženým průměrem podle frekvencí v anotátorských slovnících, frekvence samotné pak průměrem aritmetickým (čímž mohou být frekvence ve zkombinovaném slovníku necelá čísla). Srovnání kvality rozšířených slovníků pro naše dvě skupiny dat jsou v příloze 2

4.3 Klasifikace recenzí bílého zboží

Přistoupíme nyní k popisu postupu klasifikace recenzí bílého zboží.

4.3.1 Slovníkový klasifikátor

Při klasifikaci recenzí bílého zboží ze servery Mall.cz postupujeme v souladu s předchozí sekcí: vytvoření rozšířeného slovníku pro trénovací data představuje trénování klasifikátoru. Jak již bylo řečeno, zásadní výhody recenzí oproti novinovým textům spočívají v tom, že a) víme, že každý segment v datech je polární, b) pisatel je jediným mluvčím a především c) pisatel používá explicitně polární výrazivo. Navíc odpadla nutnost ručních anotací, a tedy šum vzniklý neshodami. Tím pádem je zde náš úkol výrazně jednodušší, než co nás bude čekat u novinových článků

Nepolární segmenty sice neexistují, ovšem některé segmenty se vyskytují jak v pozitivních, tak v negativních recenzích (typicky se jedná o krátké „výkřiky do tmy“: „Nevím“, „Zboží ještě nedorazilo“, „Cena“.) Takových segmentů je zhruba 6 ze 100. Tyto klasifikujeme hlasováním s prostou většinou a dál se jimi nezabýváme (pokud hlasování dopadne nerozhodně, zatřídíme je podle vyšší baseline — takových bývá ale pouze okolo jednoho z 500).

Díky takto vynucené absenci nepolárních segmentů pro tato data navíc platí:

$$baseline(polar) = 1 \tag{4.7}$$

$$baseline(negative) + baseline(positive) = 1, \tag{4.8}$$

a tedy

$$aprec(l, positive) + aprec(l, negative) = 0 \tag{4.9}$$

$$mprec(l, positive) * mprec(l, negative) = 1 \tag{4.10}$$

Samotná klasifikace položky r probíhá takto:

$$c(r, positive) = \sum_{\substack{l \in r \\ l \in L^\star{}_{Data}}} aprec(l, pos) \tag{4.11}$$

$$c(r, negative) = \sum_{\substack{l \in r \\ l \in L^\star{}_{Data}}} aprec(l, neg) \tag{4.12}$$

$$C(r) = c(r, positive) - c(r, negative) \tag{4.13}$$

$$class(r) = \begin{cases} -1, & \text{if } C(r) < 0 \quad \text{(negative)} \\ 1, & \text{if } C(r) > 0 \quad \text{(positive)} \end{cases} \tag{4.14}$$

Pokud v r není žádné slovo ze slovníku $L^\star{}_{Data}$, zvolíme odpověď s vyšší baseline. Místo *aprec* samozřejmě můžeme použít i *mprec*, PMI nebo případně další funkce pro měření indikační síly.

Pokud klasifikátor narazí na slovo, které se v datech nevyskytuje, použije speciální lemma UNSEEN. Hodnoty $prec(\text{UNSEEN}, X)$ se spočítají jako pro „průměrné slovo“:

$$prec(\text{UNSEEN}, X) = \frac{\mathbf{E}(freq(l, X))}{\mathbf{E}(freq(l)} \tag{4.15}$$

$$= \frac{1}{\mathbf{E}(freq(l)) * |L_{Data}|} \sum_{l \in L} freq(l, X) \tag{4.16}$$

$$= \frac{|Data_x|}{|Data|} \tag{4.17}$$

$$= baseline(X) \tag{4.18}$$

a tedy $aprec(\text{UNSEEN}, X) = 0$ a $mprec(\text{UNSEEN}, X) = 1$ pro všechny možné polarity X.

4.3.2 Naivní Bayesův klasifikátor

Jako alternativní klasifikátor jsme zvolili naivní Bayesův klasifikátor (NB) s lemmaty jako features. Tato metoda strojového učení je dobře známá, popíšeme ji tedy jen velmi stručně. Třídu c segmentu S, který se skládá ze slov $l_1, l_2 \ldots l_{|S|}$, najdeme podle

$$c = argmax_{c \in X} P(c \mid l_1, l_2 \ldots l_{|S|}), \tag{4.19}$$

kde X je již známá množina možných tříd. Z Bayesovy věty:

$$c = argmax_{c \in X} P(c \mid l_1, l_2 \ldots l_{|S|}) \tag{4.20}$$

$$= argmax_{c \in X} \frac{P(l_1, l_2 \ldots l_{|S|} \mid c) * P(c)}{P(l_1, l_2 \ldots l_{|S|})} \tag{4.21}$$

Zde NB předpokládá nezávislost[2] $P(l_i \mid X) \quad \forall i \ \in 1 \ldots |S|$, navíc $P(l_1, l_2 \ldots l_{|S|})$ je stejná $\forall c \in X$, takže můžeme psát:

$$c = argmax_{c \in X} P(c) * \prod_{l \in S} P(l \mid c). \tag{4.22}$$

Při implementaci NB se kvůli podtékání běžně dělá ještě logaritmická transformace a dostáváme

$$c = argmax_{c \in X} logP(c) + \sum_{l \in S} logP(l \mid c). \tag{4.23}$$

Za odhad podmíněných pravděpodobností $P(l \mid X)$ vezmeme pravděpodobnosti $P_{data}(l \mid X)$, které vypočítáme jako

$$P_{Data}(l \mid X) = \frac{freq(l, X)}{|Data_X|} \tag{4.24}$$

a jelikož se může stát, že některý odhad bude nulový a na slovo pak v datech narazíme, používáme při výpočtu $P_{Data}(l \mid X)$ Laplaceovo vyhlazování s parametrem s:

$$P_{Data}(l \mid X) = \frac{freq(l, X) + s}{|Data_X| + |L_{data}| * s}. \tag{4.25}$$

Podobně jako u lexikálního modelu spočteme ještě hodnoty $P_{Data}(\text{UNSEEN} \mid X)$, tentokrát jako

[2] Což samozřejmě neplatí, avšak přesto NB klasifikuje překvapivě dobře.

$$P_{Data}(\text{UNSEEN} \mid X) = \frac{\sum_{l \in L_{Data}} |\{S \mid l \in S, S \in Data_X\}|}{|Data_X| * |L_{Data}|}. \tag{4.26}$$

4.4 Klasifikace novinových článků

Podobné postupy jako pro recenze jsme použili i pro klasifikaci novinových článků, ovšem vzhledem k nedostatečnému množství dat nemůžeme čekat příliš relevantní výsledky. Pokusili jsme se tedy alespoň vyzkoušet různé modely a klasifikátory na trénovacích datech (12 článků, ze kterých pochází oanotované segmenty), abychom získali alespoň nějakou představu o možnostech těchto modelů. Dále jsme zkoumali, jak scestný je náš model přechodu od klasifikace segmentů ke klasifikaci celých článků. Zde je nutno mít opět na paměti, že pro stejnou úroveň jistoty jako při klasifikaci segmentů je třeba při klasifikaci článků mít nezanedbatelně více označkovaných segmentů; je možné, že rozdíly ve výsledcích klasifikace článků a segmentů mohou být způsobeny prostě tím, že se pro články statistické metody klasifikátoru nebudou „chytat“ — budou klasifikovat v podstatě náhodně.

Rozšířený slovník pro klasifikátor jsme vybudovali z oněch 410 označkovaných segmentů. Klasifikátor samotný jsme upravili takto: našli jsme všechny segmenty, kde se cílová osoba nachází (případné vícevětné segmenty — chyby segmentace — jsme automaticky odřezávali na nejbližší tečce; chyb takto upravené segmentace bylo naštěstí zcela zanedbatelně). Tyto jsme „slili“ podle modelu bag-of-words do jednoho, který jsme pak klasifikovali dvakrát: poprvé, zda je segment polární, a podruhé (pokud polární je), jak je orientován.

4.4.1 Segmenty

Nejprve jsme zjišťovali, jak dobrá je klasifikace samotných anotovaných segmentů, abychom získali představu o tom, jak je náš jednoduchý model přechodu od segmentů k článkům dobrý. „Skutečnou“ třídu segmentu jsme dostávali z anotací, v případě neshod v přítomnosti polarity jsme větu označili za neutrální, pokud se neshodovala orientace, označili jsme ji jako „obousměrnou“.

Použili jsme stejný klasifikátor jako pro recenze. Do slovníku jsme navíc dopočítali již zmiňovanou PMI. Pustili jsme klasifikátor trénovat a testovat napřed na všechy označené segmenty, abychom dostali teoretickou horní hranici modelu; toto jsme udělali pro *aprec*, *mprec* a PMI. Zdaleka nejlépe z tohoto srovnání vyšla PMI, takže jsme nadále používali ji.

Na označkovaných segmentech jsme pak (opět stejně jako u recenzí) provedli

desetinásobnou cross-validaci s poměrem 1:1. [3]

4.4.2 Články

Stejně jako u segmentů jsme napřed pustili klasifikátor na trénovací data. Tato ovšem představují pouhých 12 článků a nemůžeme tedy vyloučit, že se za každým zlepšením skrývá overfitting.

Dále jsme ručně označili 66 jiných článků pro naši pětici vybraných osob (Fischer, Janota, Klaus, Paroubek, Topolánek) a spustili jsme klasifikátor na ně. Kvůli dominanci dlouhých vět při sčítání přes všechny segmenty jsme normalizovali součty jednotlivých segmentů délkou daného segmentu.

4.5 Vylepšování modelu

Pro vylepšování základního modelu jsme použili následující postupy:

4.5.1 Filtrování slovníku

V rozšířeném slovníku se může objevit šum: nízkofrekvenční lemmata, která se náhodou vyskytla v recenzích s opačnou polaritou, než jaká by jim příslušela. Nejsnazší způsob, jak se takového šumu zbavit, je filtrovat slovník podle frekvence: nastavili jsme filtry na 100, 50, 10 a 5 výskytech. Kvůli různým baselines ovšem jednoduchý frekvenční filtr vylučuje indikátory nerovnoměrně — je pravděpodobnější, že bude vyloučen indikátor třídy s nižší baseline. Použili jsme tedy ještě parametrizovaný frekvenční filtr s parametrem t:

$$t_X = t * baseline(X) \tag{4.27}$$

$$freq(l, X) = freq(l) * prec(l) \tag{4.28}$$

a v kroku klasifikátoru 4.11, kde přičítáme $aprec$ pro $l \in L$, navíc přibude podmínka

$$freq(l, X) \geq t_X. \tag{4.29}$$

[3] Malé množství dat sice svádí ještě k vícenásobné cross-validaci, avšak tím se spolehlivost výsledků příliš nezvýší, jelikož se klasifikační položky budou jen více opakovat. Overfittingu se tedy takto nevyhneme.

Tedy: u každého lemmatu rozlišíme, pro které třídy polarity je kvůli frekvenci málo spolehlivé. Tímto bychom mohli navíc dosáhnout jistého pročištění slovníku u dobrých vysokofrekvenčních indikátorů: pokud se omylem několikrát vyskytly v recenzi opačné polarity, např. adjektiva kvůli negativnímu slovesu (vazby typu „není dobrý“), neprojdou frekvenčním filtrem pro počítání $c(r, negative)$ v algoritmu 4.11.

Jiný způsob, jak filtrovat, je otestovat statistickou signifikanci rozdílu mezi $aprec(l, neg)$ a $aprec(l, pos)$ a použít jako features pouze slova, u kterých rozdíl signifikantní bude. Použili jsme dvoustranný jednovýběrový t-test na různých hladinách (0.999, 0.95, 0.8) s nulovou hypotézou, že je dané lemma rovnoměrně rozložené přes všechny třídy.

Dále jsme zkusili ve slovníku ponechat pouze autosémantické slovní druhy (slovesa, adjektiva, substantiva a adverbia), případně předložky.

4.5.2 Negace

Další modifikací základního modelu bylo zohlednění větné negace: pokud jsme v segmentu našli záporné sloveso, zorientovali jsme všechny členy segmentu opačně. (Recenze produktů byly v naprosté většině případů jednovětné.) Zjemněná verze této feature přeorientovávala pouze specifikované slovní druhy následující za záporným slovesem.

4.5.3 Lepší vážení podle baselines

Z definice $aprec(l, X)$ vyplývá, že $aprec(l, X) \in [-baseline(X), 1 - baseline(X)]$ a tedy *aprec* „diskriminuje“ ve prospěch třídy s nižší baseline. Kdy se tato nerovnováha projeví?

Nechť $baseline(pos) = \frac{2}{3}, baseline(neg) = \frac{1}{3}$. Představme si položku $S = (l_1, l_2, l_3)$, kde l_1 a l_2 jsou slova čistě pozitivní ($prec(l_{1,2}, pos) = 1, prec(l_{1,2}, neg) = 0$) a l_3 je naopak čistě negativní. ($prec(l_3, neg) = 1, prec(l_3, pos) = 0$). Potom $\sum_{l \in S} aprec(l, pos) = 2 * (1 - baseline(pos)) = \frac{2}{3}$, $\sum_{l \in S} = 1 - baseline(neg) = \frac{2}{3}$. Jedno dokonale negativní slovo tak vyvažuje efekt dvou dokonale pozitivních. To má své dobré opodstatnění — odpovídá to skutečnosti, že jelikož je v datech dvakrát méně negativních segmentů, negativní slovo je dvakrát vzácnější a tedy má dvakrát větší váhu — avšak baselines jsou pouze *odhad* toho, jaké je skutečné zastoupení negativních a pozitivních slov, na základě trénovacích dat, takže je možné, že skutečná baseline leží jinde. Proto zavádíme do modelu možnost posunout baselines pro výpočet *aprec* pomocí parametru p_{shift}:

$$aprec'(l,X) = aprec(l,X) + \frac{\frac{baseline(pos)+baseline(neg)}{2} - baseline(X)}{\frac{p_{shift}}{2}}) \qquad (4.30)$$

Parametr p_{shift} má ten smysl, že posune hranice intervalů možných hodnot $aprec(l,X)$ o $\frac{1}{p}$ rozdílu mezi $baseline(pos)$ a $basline(neg)$ k sobě. Pro $p_{shift} = 2$ tedy vycentrujeme $aprec$ na průměr z obou baselines, pro $p_{shift} > 2$ je posun menší, takže rozdíl v možných hodnotách mezi $aprec(l,pos)$ a $aprec(l,neg)$ pouze o něco zmenšíme, naopak pro $p_{shift} < 2$ můžeme začít „diskriminovat" opačně — pro záporné hodnoty parametru dokonce děláme rozdíl vah extrémnější. Nejlepší hodnotu najdeme experimentálně.

4.6 Výsledky

Pro všechnu evaluaci jsme použili desetinásobnou cross-validaci s poměrem trénovacích a testovacích dat 1:1. Uvádíme pro daný model vždy accuracy, recall, precision a f-score pro klasifikované třídy a průměr recall, precision a f-score přes všechny třídy vážený skutečným zastoupením tříd v testovacích datech. Jako baseline jsme vždy zařadili všechny segmenty do nejfrekventovanější třídy.

Poznámka ke značení:

- Sloupce tabulek: Acc značí accuracy, R značí recall, P je precision, F je f-score. R/P/F bez závorek značí průměrné hodnoty, v závorkách pak danou hodnotící míru pro příslušnou třídu (– je negativní, 0 nepolární, + pozitivní).
- $freq(x)$ a $pfreq(x)$ značí frekvenční filtr na úrovni x, resp. parametrizovaný f.f. pro $t = x$,
- POS značí filtrování podle slovních druhů,
- v bez závorky značí základní model negace,
- v se závorkou značí zjemněný model negace,
- shift(x) značí posun baselines s p_{shift} = x (všechny posuny v tabulkách jsou mezi pozitivní a negativní baseline),
- t-test(x) značí filtrování statistickou signifikancí na hladině $1 - x$.
- Co se značení slovních druhů týče, A jsou adjektiva, D příslovce, N substantiva, V slovesa a R předložky (podle [3]);
- samotné X značí obsah předchozího řádku.
- Zajímavá čísla a nejlepší model pro danou úlohu jsou značeny **tučně**, čísla zajímavě nízká *kurzívou*.

4.6.1 Klasifikace recenzí bílého zboží

Slovníkový klasifikátor dosáhl na recenzích bílého zboží velmi slušných výsledků, Naivní Bayes se neosvědčil (nejlépe fungoval se zcela minimálním vyhlazováním, $s = 0.0001$, avšak výsledků slovníkového klasifikátoru nedosahoval). Základní srovnání klasifikátorů na trénovacích a testovacích datech viz tabulka 4.1:

Model	Acc	R(-)	P(-)	F(-)	R(+)	P(+)	F(+)	R	P	F
baseline	0.630	0	0	0	1	0.630	0.773	0.370	0.233	0.286
Sl.k.,train	0.960	0.964	0.935	0.949	0.958	0.977	0.967	0.960	0.961	0.960
Sl.k.,test	0.889	0.907	0.821	0.862	0.878	0.939	0.908	0.889	0.894	0.890
Bayes,train	0.864	0.717	0.901	0.798	0.955	0.849	0.899	0.803	0.879	*0.833*
Bayes, test	0.827	0.630	0.872	0.730	0.947	0.811	0.874	0.745	0.847	*0.781*

Tabulka 4.1: Baseline, chování na trénovacích datech a klasifikace bez filtrů.

Vyzkoušeli jsme pro slovníkový klasifikátor všechny tři indikační síly bez filtrování slovníku: *aprec*, *mprec* i PMI. Nejlépe dopadla *aprec*, PMI kupodivu skončila poslední. Podle vývoje precision/recall se zdá, že PMI na recenzích poněkud nepatřičně favorizuje třídy s vyšší baseline, *aprec* má naopak o dost mírnější tendenci opačnou. (Viz tabulka 4.2.)

Model	Acc	R(-)	P(-)	F(-)	R(+)	P(+)	F(+)	R	P	F
PMI	0.878	0.810	0.852	0.83	0.918	0.891	0.904	0.850	0.867	0.858
mprec	0.887	0.898	0.815	0.855	0.88	0.937	0.907	0.892	0.860	0.874
aprec	0.889	0.907	0.821	0.862	0.878	0.939	0.908	0.889	0.894	0.890

Tabulka 4.2: Slovníkový klasifikátor, porovnání aprec/mprec/PMI.

Jako částečně správná se ukázala úvaha o nepřesnosti odhadu baselines. Nastavení p_{shift} na 25 (připomínáme, že vyšší hodnoty p_{shift} znamenají menší korekci) sice vylepšilo výsledky lexikálního klasifikátoru pouze o zanedbatelných 0.02, takže samotná chyba v odhadu baselines není významná, avšak dobře nastavený posun dokáže kompenzovat nevyváženosti v chybně klasifikovaných instancích vzniklé použitím nějakého jiného filtru či modifikace (srov. precision a recall v tabulce 4.3), což může vylepšit i celkové hodnocení. Všimněme si, jak zvýšení p_{shift} dokáže kompenzovat část ztrát způsobených frekvenčním filtrem.

Model	Acc	R(-)	P(-)	F(-)	R(+)	P(+)	F(+)	R	P	F
shift(25)	0.895	0.851	0.866	0.858	0.925	0.912	0.918	0.895	0.892	0.893
shift(30)	0.894	0.864	0.875	0.86	0.916	0.917	0.916	0.894	0.892	0.893
shift(25)+freq(20)	0.878	0.798	0.867	0.831	0.929	0.883	0.906	0.878	0.875	0.875
shift(35)+freq(20)	0.882	0.817	0.860	0.838	0.924	0.893	0.908	0.882	0.879	0.880

Tabulka 4.3: Efekt posunu baselines.

Filtrování podle frekvence žádné zlepšení nepřineslo, od spodní hranice 10 výskytů se naopak výsledky začaly zhoršovat. Především začly přibývat pozitivně klasifikované negativní segmenty — toto přičítáme skutečnosti, že jelikož $baseline(neg)$ byla asi poloviční oproti $baseline(pos)$, ze slovníku vypadlo víc dobrých indikátorů negativity než pozitivity. Poté, co jsme filtrování upravili, aby na různé baselines ohled bralo, se chovalo lépe. (Co se skutečných filtrovaných frekvencí týče, $pfreq(2n)$ odpovídá v součtu přes všechny třídy $freq(n)$).

Model	Acc	R(-)	P(-)	F(-)	R(+)	P(+)	F(+)	R	P	F
freq(100)	0.813	0.753	0.755	0.753	0.849	0.849	0.849	0.813	0.813	0.812
freq(50)	0.850	0.820	0.796	0.807	0.869	0.886	0.877	0.85	0.851	0.851
freq(10)	0.887	0.885	0.829	0.856	0.889	0.927	0.908	0.887	0.890	0.888
freq(5)	0.889	0.901	0.827	0.862	0.882	0.934	0.908	0.889	0.893	0.890
pfreq(100)	0.886	0.904	0.810	0.855	0.875	0.940	0.906	0.893	0.858	0.874
pfreq(50)	0.889	0.903	0.818	0.858	0.881	0.939	0.909	0.895	0.863	0.877
pfreq(20)	0.887	0.902	0.814	0.856	0.878	0.938	0.907	0.893	0.860	0.875
pfreq(10)	0.888	0.905	0.815	0.858	0.879	0.940	0.908	0.895	0.861	0.877

Tabulka 4.4: Slovníkový klasifikátor, frekvenční filtry.

Filtrování podle slovních druhů také nemělo vliv na celkové hodnocení, dokud jsme nezačali odebírat autosémantické slovní druhy — pak se výsledky dle očekávání zhoršovaly. Měnil se ovšem charakter chyby: klasifikátor měl výrazně větší tendenci klasifikovat při ponechání samotných autosémantických slovních druhů negativní recenze pozitivně než naopak, zatímco jinak recall na pozitivních recenzích vždy precision předběhl. Není jasné, zda se jedná o záležitost dat samotných, či zda je skutečně negativita spíše než pozitivita indikována dalšími slovními druhy. (Viz tabulka 4.5).

Jednoduchý model negace (přeorientování celé věty) vedl k výraznému zhoršení. Zjemněné zacházení s negací výsledkům ani nepomohlo, ani neublížilo. (Viz tabulka 4.5.)

Model	Acc	R(-)	P(-)	F(-)	R(+)	P(+)	F(+)	R	P	F
t-test(99.9)	0.879	0.881	0.817	0.848	0.877	0.922	0.899	0.878	0.882	0.879
t-test(80)	0.888	0.901	0.821	0.859	0.880	0.935	0.907	0.888	0.892	0.889
vneg	0.746	0.565	0.690	0.621	0.851	0.770	0.809	0.671	0.719	0.690
vneg(ADN)	0.887	0.901	0.813	0.855	0.878	0.938	0.907	0.892	0.859	0.874
POS(ADVN)	0.885	0.860	0.843	0.852	0.901	0.913	0.906	0.885	0.886	0.885
X+freq(10)	0.879	0.811	0.862	0.836	**0.921**	*0.888*	0.904	0.879	0.878	0.878
POS(ADV)	0.855	0.818	0.797	0.807	0.877	0.891	0.884	0.840	0.832	*0.836*
POS(ADVNR)	0.887	0.880	0.832	0.855	0.892	0.924	0.907	0.887	0.889	0.888
X+freq(5)	0.887	0.869	0.840	0.854	0.898	0.917	0.907	0.887	0.888	0.887
X+freq(50)	0.846	0.772	0.812	0.791	0.891	0.865	0.878	0.846	0.845	0.845

Tabulka 4.5: Slovníkový klasifikátor, aprec, rozličné filtrování

4.6.2 Klasifikace článků a segmentů z článků

Kvůli velmi nízké frekvenci pozitivních segmentů v datech se je prakticky vůbec nepodařilo identifikovat — a když nějakým nastavením p_{shift} pro identifikaci polarity ano, pak jejich počet zcela zastínil počet pozitivně klasifikovaných neutrálních segmentů (nehledě na to, že tím trpěly celkové výsledky).

Co se vlastností *aprec*, *mprec* a PMI týče, ukázalo se, že *aprec* i *mprec* jsou výrazně náchylnější klasifikovat neutrální segmenty jako polární, přičemž ovšem u polárních segmentů nezvyšují recall. To je opět zřejmě důsledek řídkosti a rozložení dat.[4] (Srovnání, ještě s Naivním Bayesem vyhlazovaným na $s = 0.0001$,

[4]Dobré indikátory polarity se nevyskytují v neutrálních segmentech, a jelikož je polárních segmentů málo, je vyšší šance, že se v trénovacích datech nevyskytnou vůbec. Při použití funkce pro výpočet indikační síly, která je náchylnější klasifikovat daný segment jako polární, pak nabývají na skutečné indikační mohutnosti (tzn. reálném dopadu na klasifikaci) pro polaritu především středně- a vysokofrekvenční slova, která se vyskytují ve skutečnosti rovnoměrně, avšak kvůli malým datům se náhodou vyskytla ve více trénovacích než testovacích polárních segmentech. Tím pádem sice mají vyšší indikační sílu, avšak recall pro polární segmenty nezvýší, jelikož se v testovacích polárních segmentech vyskytují zase méně, než by příslušelo jejich skutečnému rozdělení — naopak zhorší recall na neutrálních segmentech a precision polárních, protože se častěji vyskytnou v neutrálních segmentech a kvůli příliš polárnímu odhadu indikační síly tak „tahají“ neutrální segmenty s sebou do polárních tříd. Nepomůžou nám ani slova, která se naopak náhodou vyskytují méně často, než by jim příslušelo, v trénovacích polárních segmentech — byť by měly mít zase vyšší indikační mohutnost na testovacích datech tím, že se častěji vyskytují v testovacích polárních segmentech, z trénování mají indikační sílu podhodnocenou. Toto je zcela standardní šum, ovšem při malých datech prudce roste pravděpodobnost, že bude dané slovo po rozdělení dat na trénovací a testovací náhodou ve své třídě přes tyto dvě skupiny nerovnoměrně rozložené. Nerovnoměrnost rozložení tříd na celých datech toto riziko ještě zvyšuje. Mimochodem, mohlo by být zajímavé použít nějakou indikační mohutnost pro feature selection.

v tabulce 4.6.)

Model	Acc	R(-)	P(-)	F(-)	R(0)	P(0)	F(0)	R(+)	P(+)	F(+)	R	P	F
Bayes	0.787	0.032	0.667	0.061	0.997	0.788	0.880	0	0	0	0.784	0.651	*0.694*
aprec	0.880	0.984	0.714	0.828	0.870	0.993	0.927	1	0.465	0.635	0.882	0.886	0.865
mprec	0.814	0.967	0.686	0.803	0.789	0.992	0.879	1	0.299	0.460	0.818	0.859	0.800
PMI	**0.976**	0.968	0.968	0.968	0.994	0.982	0.988	1	0.905	0.950	0.979	0.955	**0.967**

Tabulka 4.6: Aprec/mprec/PMI slovníkového klasifikátoru na trénovacích datech.

Kvůli těmto potížím, které neoddělitelně patří k malým a výrazně nerovnoměrně rozděleným datům, neměla žádná z filtrovacích a dalších metod navržených pro klasifikaci recenzí šanci se projevit. Malá data podle očekávání zabránila i použití „unikátu“: zástupného symbolu za všechna unikátní slova, který se osvědčil v [2] (když jsme ho vyzkoušeli, klasifikace se zhoršila — kvůli mohutnému šumu bylo unikátní slovo spíše indikátorem neutrality). Výsledky různých filtrů ukazuje tabulka 4.7.

Model	Acc	R(-)	P(-)	F(-)	R(0)	P(0)	F(0)	R(+)	P(+)	F(+)	R	P	F
Baseline	0.779	0	0	0	1	0.779	0.857	0	0	0	0.779	0.606	0.682
all	0.728	0.152	0.237	0.179	0.900	0.799	0.846	0.043	0.051	0.043	0.717	0.643	0.676
pfreq(20)	0.736	0.162	0.268	0.193	0.908	0.803	0.852	0.085	0.117	0.097	0.730	0.659	0.689
pfreq(10)	0.744	0.175	0.273	0.207	0.908	0.808	0.855	0.040	0.098	0.055	0.733	0.666	0.694
X+v(ADNV)	0.748	0.195	**0.328**	0.241	0.908	0.815	0.858	0.070	0.095	0.076	0.736	0.673	0.701
v(ADN)	0.741	0.163	0.282	0.202	0.906	0.810	0.855	0.067	0.081	0.073	0.731	0.664	0.694
pfreq(10)+X	0.755	0.196	**0.345**	0.243	0.918	0.814	0.862	0.084	**0.185**	0.106	0.744	0.684	0.706

Tabulka 4.7: Slovníkový klasifikátor, testovací data PMI, rozličné filtrování

Klasifikace článků dopadla ještě výrazně hůře než klasifikace segmentů. Navíc jestli malé množství dat bylo problematické u klasifikace segmentů, u článků to platí dvojnásob. Skutečnost, že se jednoduchou baseline podařilo výrazně překročit, připisujeme baseline samotné — na úrovni článků převládalo neutrální hodnocení výrazně méně než na úrovni segmentů, takže baseline byla nižší. Tato skutečnost také napovídá, že stačí relativně málo polárních segmentů na to, aby byl celý článek vůči sledované osobě polární. Jistou inspirací do budoucna je relativní úspěch kombinace *aprec* a PMI na trénovacích datech — *aprec* pro rozpoznávání polarity, PMI pro určování orientace polárních stavů. (Výsledky viz

tabulky 4.8 pro klasifikaci „trénovacích“ článků, 4.9 pro efekt p_{shift} na testovacích článcích.)

Model	Acc	R(-)	P(-)	F(-)	R(0)	P(0)	F(0)	R(+)	P(+)	F(+)	R	P	F
Baseline	0.333	0	0	0	1	0.333	0.5	0	0	0	0.417	0.139	0.208
shift(12)	0.583	0.600	1	0.750	1	0.500	0.667	0	0	0	0.617	0.542	0.528
aprec/PMI	0.667	0.800	1	0.889	1	0.500	0.667	0	0	0	0.683	0.542	0.574

Tabulka 4.8: Na „trénovacích“ datech.

Model	Acc	R(-1)	P(-1)	F(-1)	R(0)	P(0)	F(0)	R(1)	P(1)	F(1)	R	P	F
Baseline	0.514	0	0	0	1	0.514	0.678	0	0	0	0.14	0.07	0.1
all	0.439	0.4	0.222	0.286	0.686	0.632	0.658	0.077	0.1	0.087	0.331	0.233	0.268
shift(50)	0.439	0.6	0.25	0.353	0.629	0.647	0.638	0.077	0.125	0.095	0.429	0.255	0.303
shift(20)	0.485	0.5	0.25	0.333	0.743	0.65	0.693	0.077	0.167	0.105	0.393	0.264	0.302
shift(12)	0.545	0.6	0.375	0.462	0.829	0.63	0.716	0.077	0.25	0.118	0.459	0.344	0.377
shift(9)	0.561	0.5	0.385	0.435	0.886	0.62	0.73	0.077	0.333	0.125	0.415	0.364	0.366
shift(5)	0.515	0.2	0.286	0.235	0.886	0.554	0.682	0.077	0.333	0.125	0.255	0.301	*0.253*

Tabulka 4.9: Na testovacích datech — ilustrace efektu p_{shift}

Představu, proč výsledky klasifikací nad články zdaleka nedosahovaly úspěšnosti klasifikátoru recenzí, si snadno můžeme udělat z již zmiňovaných ukázek rozšířených slovníků vygenerovaných z těchto dat v příloze 2.

5. Závěr

Popsali jsme základní vlastnosti polarity v textu a na několika příkladech ukázali, jak se chová. Čím jazykově vyspělejší texty budeme pro automatické zpracovávání polarity používat, tím více lingvistického popisu a porozumění bude třeba. Novinové texty jsou spíše sofistikovanější a staví před zjednodušené modely polarity překážky jako polarita skrytá v předpokladech či obalená v citacích. Naproti tomu polaritu v textech jako uživatelské recenze, které svou snahu předat polární informaci nijak neskrývají, můžeme úspěšně modelovat velmi jednoduše.

Ukázalo se, že české novinové texty jsou kvůli novinářské kultuře pro takové zpracování obzvlášť nevhodné; je náročné vůbec vytvořit sadu anotátorských instrukcí pro značkování polarity a anotace sama je problematická a nepřesná. Tím pádem se bohužel nepodařilo vytvořit dostatečně dobrý lexikon polárních výrazů jako jádro pro automatické vyhledávání polárních výrazů pomocí latentní sémantické analýzy (LSA) [1].

Dále se ukázalo, že anotace polarity v českých novinových článcích tak, jak jsme je prováděli, nepovedou ke kvalitnímu označkovanému korpusu. V situaci, kdy se v datech prakticky nevyskytuje explicitní hodnocení jinak než v citacích (a kvůli politické kultuře vlastně také velmi opatrně formulovaných), se navíc stává zásadním problémem konflikt mezi zvolenou perspektivou čtenáře, ze které jsme polaritu anotovali, a požadavkem na neutralitu vůči entitám v anotovaných textech, který tato perskpektiva při malém počtu anotátorů vyžaduje — nejen že v této situaci lze jen obtížně značit orientaci polarity, nýbrž je problematické vůbec polární stav identifikovat. Nabízí se dvě cesty, kterými se s touto překážkou můžeme vypořádat: buď rezignovat na perspektivu čtenáře a anotovat podobně jako [2] a [8] všechny polární stavy v textu, nebo rezignovat na neutralitu, což ovšem, pokud máme mít spolehlivá data, obnáší změnu anotačního paradigmatu — místo několika stabilních a vyškolených mít mnoho různě spolehlivých a spíše příležitostných anotátorů — a kvůli různým osobním preferencím a „utvrzující predispozici“ je na novinových textech takováto anotace silně závislá na rozložení anotátorů přes politické a sociální spektrum. Při změně perspektivy bychom pro systém klasifikující cílové osoby v novinových článcích ovšem museli dodat velmi netriviální vrstvu „důvěryhodnosti“ jednak pro mluvčí, jednak pro polární stavy samotné, zatímco druhá varianta by zachovala čtenářskou perspektivu, čímž by byla zamýšlenému klasifikačnímu systému blíže. Nabízí se i možnost obě anotace kombinovat a na základě rozdílů v nich na stejných datech trénovat onu vrstvu důvěryhodnosti. Další variantou, jak zachovat neutralitu a nevzdat se perspektivy bližší zamýšlené aplikaci, může být anotace v jiném jazyce (kde ovšem

pak vznikají potíže s přenosem anotace zpět do zdrojového jazyka a navíc je také náchylná k potížím s neschopností identifikovat polární stav bez znalosti identity mluvčího).

Bylo vidět, že na vhodnějších středně velkých datech z recenzí produktů (srov. s [5]) dokáže náš jednoduchý slovníkový klasifikátor dosáhnout téměř state-of-the-art výsledků. Bohužel jsme kvůli nedostatku dat nemohli natrénovat podobný klasifikátor pro novinové články; výsledky na segmentech však naznačují, že dosáhnout u článků podobné úrovně bude obtížnější. Ověřit, jak se chová náš jednoduchý model přechodu od oanotovaných segmentů ke klasifikaci celých článků, se také kvůli malým datům nepodařilo.

Rozšíření modelu na n-gramy by mohlo výsledky ještě dále vylepšit, viz [5] a [2]. Stejně tak se zdá, že půjde využít i bohatá morfologie [8], kterou máme snadno k dispozici [3]. Další výzkum si zasluhuje i role negace — námi zvolené jednoduché modely jejího působení sice neměly na klasifikaci kladný vliv, avšak negace zcela nepochybně s polaritou souvisí. Nabízí se ještě například použití „contextual valence shifters“ [6] a pro hledání indikátorů polarity ve hře samozřejmě zůstává mocný nástroj vektorových sémantických modelů jako LSA [7]. Nakonec můžeme s tím, jak bude růst naše lingvistické porozumění pro polaritu v češtině, ještě využít existujících korpusů jako Pražský závislostní korpus, abychom měli k dispozici i spolehlivá data na syntaktické či tektogramatické rovině jazyka. Samostatnou kapitolou je pak ještě zapojování různých (ručně či automaticky vytvořených) lexikonů polárních výrazů, které výrazně pomohlo v [8].

Všechny tyto možnosti jsou ovšem podmíněny existencí kvalitního a dostatečně rozsáhlého korpusu polarity, jehož vytvoření zůstává jasnou prioritou; doufáme, že výsledky této pilotní práce tomu napomůžou nejen v našem pokračujícím výzkumu.[1]

[1] Data, klasifikátor a příslušnou technicku a uživatelskou dokumentaci dáváme k dispozici na přiloženém CD.

Seznam použité literatury

[1] Landauer, T. K., Foltz, P. W. and Laham, D. *Introduction to Latent Semantic Analysis.* Discourse Processes, 25:259-284, 1998

[2] Wiebe, Janice, et al. *Learning subjective language.* Computational Linguistics 30 (3):277–308, 2004.

[3] Hajič, Jan *Disambiguation of Rich Inflection (Computational Morphology of Czech).* Karolínum, Charles University Press, 2004.

[4] Kim, S.-M. and Hovy, E.H. *Determining the Sentiment of Opinions.* Proceedings of the COLING conference, Geneva, 2004.

[5] Cui, Hang, Mittal, Vibhu, and Datar, Mayur *Comparative Experiments on Sentiment Classification for Online Product Reviews* Proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence, 2006

[6] Kennedy, A., Inkpen, D. *Sentiment Classification of Movie Reviews Using Contextual Valence Shifters.* Computational Intelligence 22(2):110-125, 2006

[7] Banea, C., Mihalcea, R. and Wiebe, J. *A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources.* Proceedings of the International Conference on Language Resources and Evaluations (LREC 2008):2764-2767, 2008.

[8] Abdul-Mageed, M., Diab, M. and Korayem, M. *Subjectivity and Sentiment Analysis of Modern Standard Arabic.* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:587–591, 2011.

Seznam tabulek

Přílohy

1 Pokyny pro anotaci

- Segmentace a čištění
  - Co věta, to řádek (tzn. odřádkování na konci)
  - Rozdělovat i vnitřky přímých řečí (á la WSJ)
  - Nechávat i nadpisy a různé další nevětné segmenty jako samostatné věty (nedoplňovat tečky na konci!)
  - Může se stát, že tam naopak budou jenom kusy vět (čistící skript není dokonalý), chovejte se k nim prostě jako k celým větám.
  - Jsou občas i minute-by-minute reportáže ze sněmovny á la sport - kdyby tam zůstaly ty časové údaje, tak je vyházejte.
  - Za nadpisem většinou zbývá taková část „Autor [jméno], Praha, [čas], větší obrázek“ — když na to narazíte, tak to odstraňte
  - Můžou se vyskytnout i zbytky HTML, kusy odkazů apod. — odstranit.
    - Pokud se nachází uprostřed slova, nenechat po nich nic. Pokud mají mezeru z jedné strany a z druhé navazují, mezeru zachovat. Pokud mají mezery z obou stran, jednu z mezer zrušit taky. (Důležité kvůli automatickému vyhledávání rozdílů…)
  - Stejně tak se občas vyskytují v článcích popisky obrázků a zbytky popisů obrázků, například kdo je vyfotil. Taky odstranit.
- Značkování — po cílech
  - Označit „cíl“ (tzn. entitu, o které věta sděluje polární informaci)
  - Zapomeňte při anotování na to, co se s tím pak bude dělat, a značte takové věci, na které má smysl si dělat dobrý nebo špatný názor…
    - pokud se jedná o pozitivní hodnocení, označit TUČNĚ
    - pokud se jedná o negativní hodnocení, označit KURZÍVOU
    - pokud se jedná o víceslovný výraz, označit ho (změna!) celý, tzn. i třeba celou vedlejší větu
  - Pokud se cíl v polární větě nevyskytuje, připište za konec věty EXTERN a odpovídajícím způsobem označit (bold/italic).
  - Zkuste ignorovat, kdo je zdrojem. Například jestli tam potkáte větu „Václav Klaus zveřejnil seznam osob, které podle něho škodí státu.“, neznačte hned „osob“ pozitivně. :)
  - Při hodnocení moc nepřemýšlejte a řiďte se prvním dojmem.
  - Koreference: značte výskyt cíle nejblíže k polární části.
  - Více cílů: označte všechny.
  - „Bad News“: neznačit (příklad: „Šumava skutečně není v pořádku, došlo tady k velké katastrofě.“ — katastrofu neznačit.)

- Značkování — po lidech
  - Označit každý polární výskyt cílového člověka — kdykoliv máte pocit, že je mu daná věta ke škodě nebo k užitku, tak označit (nezávisle na tom, kdo nebo co je v dané větě cílem).
    * pokud se jedná o pozitivní hodnocení, označit TUČNĚ
    * pokud se jedná o negativní hodnocení, označit KURZÍVOU
    * pokud se jedná o víceslovný výraz — např. celé jméno — označit ho celý (tzn. „předsedovi ODS Mirku Topolánkovi“ označit celé).
  - Zkuste prosím ignorovat osobní antipatie či sympatie k cílovým osobám.
    * Především se snažte vyvarovat confirmation bias. (Když je mi někdo předem sympatický, cokoliv se o něm dozvím interpretuju pozitivně (a naopak)…)
    * Možná dobrý způsob, jak simulované nestrannosti dosáhnout, je představovat si, že cílové osoby vůbec neznáte nebo že se jmenují jinak.
  - Při hodnocení moc nepřemýšlejte a řiďte se prvním dojmem.
  - Koreference uvnitř věty: značte ten výskyt, kde je celé jméno. (Věty jsem vybíral podle přítomnosti kořene jména, takže v každé alespoň jednou je, krom případů špatné segmentace.)
  - Pokud se ve větě vyskytuje více cílových osob, označte je nezávisle na sobě všechny.
  - Pokud kvůli nedokonalé segmentaci pronikla do balíčku věta, ve které se cílová osoba nevyskytuje ani přes koreferenci, tak ji vyhoďte.
    * Pokud se ale ve větě cílová osoba vyskytuje přes koreferenci, tak ji ale označte (tím se nic nezkazí a i kdyby, odfiltrovat to půjde kdyžtak snadno)
- Miscellaneous
  - Měřte si čas
  - Pište si případné další poznámky, připomínky atd.

2 Ukázky ze slovníku na recenzích a segmentech z článků

Vybrali jsme nejlepší rozumně frekventované indikátory převažující polarity z daných dat. (Nerelevantní sloupce jsme vynechali.)

Recenze

lemma	freq	pr(-)	pr(+)	apr(-)	apr(+)	mpr(-)	mpr(+)
dobrá—NA	107	0	1	-0.379	0.379	0	1.611
skvělý—AA	173	0	1	-0.379	0.379	0	1.611
výborná—NA	125	0	1	-0.379	0.379	0	1.611
výborný—AA	146	0.007	0.993	-0.372	0.372	0.018	1.6
perfektní—AA	150	0.013	0.987	-0.366	0.366	0.035	1.589
pěkný—AA	262	0.015	0.985	-0.364	0.364	0.04	1.586
snadný—AA	360	0.017	0.983	-0.362	0.362	0.044	1.584
super—AA	216	0.019	0.981	-0.361	0.361	0.049	1.581
spokojenost—NA	106	0.019	0.981	-0.36	0.36	0.05	1.58
výborně—DA	101	0.02	0.98	-0.359	0.359	0.052	1.579
jednoduchý—AA	526	0.021	0.979	-0.358	0.358	0.055	1.577
přehledný—AA	172	0.023	0.977	-0.356	0.356	0.061	1.573
vzhled—NA	277	0.025	0.975	-0.354	0.354	0.067	1.57
poměr—NA	112	0.027	0.973	-0.352	0.352	0.071	1.568
výkonný—AA	196	0.031	0.969	-0.349	0.349	0.081	1.561
spokojený—AA	128	0.031	0.969	-0.348	0.348	0.082	1.56
hezký—AA	156	0.032	0.968	-0.347	0.347	0.085	1.559
kvalitní—AA	276	0.033	0.967	-0.347	0.347	0.086	1.558
tichý—AA	939	0.033	0.967	-0.346	0.346	0.087	1.557
design—NA	502	0.036	0.964	-0.343	0.343	0.095	1.553
spokojený—AA	171	0.041	0.959	-0.338	0.338	0.108	1.545
chod—NA	325	0.043	0.957	-0.336	0.336	0.114	1.541
příjemný-1—AA	102	0.049	0.951	-0.33	0.33	0.129	1.532
lehký—AA	192	0.057	0.943	-0.322	0.322	0.151	1.518
výběr—NA	106	0.066	0.934	-0.313	0.313	0.174	1.504
dobrý—AA	721	0.076	0.924	-0.303	0.303	0.201	1.488
rychlý—AA	175	0.08	0.92	-0.299	0.299	0.211	1.482
obsluha—NA	246	0.081	0.919	-0.298	0.298	0.214	1.48
kvalita—NA	337	0.086	0.914	-0.293	0.293	0.227	1.472
dobře—DA	510	0.09	0.91	-0.289	0.289	0.238	1.465
prát—VA	164	0.098	0.902	-0.282	0.282	0.257	1.454
značka—NA	220	0.1	0.9	-0.279	0.279	0.264	1.45
spotřeba—NA	294	0.102	0.898	-0.277	0.277	0.269	1.446
nízký—AA	403	0.107	0.893	-0.272	0.272	0.281	1.439
+—Z-	156	0.109	0.891	-0.27	0.27	0.287	1.435

Tabulka 5.1: Recenze: 35 nejkladnějších slov podle *mprec* a frekvencí > 100

Články

lemma	freq	prec(n)	prec(p)	prec(np)	prec(pp)	mprec(p)	mprec(np)	mprec(pp)
proti-1—R-	12	0.292	0.708	0.667	0.083	2.426	3.176	0.895
park—NA	10	0.450	0.550	0.450	0.100	1.865	2.089	1.074
být—VN	18	0.500	0.500	0.389	0.167	1.744	1.881	1.744
zelený—AA	14	0.321	0.679	0.393	0.286	2.354	1.879	3.069
ústavní—AA	16	0.625	0.375	0.375	0	1.308	1.814	0
soud—NA	14	0.643	0.357	0.357	0	1.220	1.686	0
tak-3—D-	13	0.577	0.423	0.346	0.077	1.462	1.652	0.826
”—Z-	118.5	0.570	0.430	0.342	0.097	1.486	1.636	1.025
předseda—NA	14	0.500	0.500	0.321	0.250	1.744	1.534	2.715
prezident—NA	15	0.633	0.367	0.333	0.100	1.220	1.534	1.047
před-1—R-	13	0.615	0.385	0.308	0.077	1.341	1.488	0.826
pro-1—R-	17	0.500	0.500	0.324	0.235	1.733	1.479	2.625
podle-2—R-	45	0.522	0.478	0.311	0.167	1.639	1.479	1.763
poslanecký—AA	15	0.600	0.400	0.300	0.100	1.395	1.471	1.047
otázka—NA	10	0.700	0.300	0.300	0.050	1.046	1.451	0.496
demokrat—NA	15	0.567	0.433	0.300	0.200	1.499	1.432	2.148
člověk—NA	15	0.633	0.367	0.300	0.233	1.243	1.393	2.479
než-2—J-	10	0.650	0.350	0.300	0.100	1.203	1.393	1.074
být—VA	137.5	0.625	0.375	0.287	0.105	1.302	1.383	1.136
kůrovec—NA	11	0.773	0.227	0.273	0.045	0.809	1.319	0.451
za-1—R-	18	0.611	0.389	0.278	0.111	1.336	1.311	1.194
uvést—VA	13	0.692	0.308	0.269	0.038	1.046	1.280	0.381
ale—J-	30	0.600	0.400	0.267	0.167	1.371	1.270	1.763
dva‘2—C-	10	0.700	0.300	0.250	0.050	1.082	1.239	0.578
říci—VA	23	0.630	0.370	0.261	0.109	1.281	1.236	1.186

Tabulka 5.2: Lemmata z článků: 25 nejzápornějších podle *mprec*, prvních 25 s frekvencí > 10