EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103004/I/2017/1407499467

Nástroje využívané pre analýzu zákazníckeho sentimentu na sociálnych sieťach

Diplomová práca

2017 **Bc. Miroslav Mihalík**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Nástroje využívané pre analýzu zákazníckeho sentimentu na sociálnych sieťach

Diplomová práca

Študijný program: Informačný manažment

Študijný odbor: Kvantitatívne metódy v ekonómii

Školiace pracovisko: Katedra aplikovanej informatiky

Vedúci záverečnej práce: Ing. Mária Szivósová, PhD.

Bratislava 2017 **Bc. Miroslav Mihalík**

Čestné vyhlásenie

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne a že som uviedol všetku použitú literatúru.

Dátum: **.** **.....**

Poďakovanie

Touto cestou by som sa chcel poďakovať vedúcej diplomovej práce Ing. Márii Szivósovej, PhD. za odbornú pomoc, cenné rady a pripomienky, ktoré mi poskytla pri vypracovaní diplomovej práce.

Dátum: **.....**

ABSTRAKT

MIHALÍK, Miroslav:Nástroje využívane pre analýzu zákazníckeho sentimentu na sociálnych sieťach

– Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra aplikovanej informatiky. – Vedúci záverečnej práce: Ing. Mária Szivósová, PhD. – Bratislava: FHI EU, 2017, 57 s.

Náš každodenný život bol stále ovplyvnený tým, čo si ľudia myslia. Myšlienky a názory ktoré majú ostatný vždy ovplyvnia naše vlastné názory. Explózia Web 2.0 má za následok zvýšenú aktivitu podcastingu, blogovania, značkovania čo prispieva k RSS, sociálnemu bookmarkingu a sociálnym sieťam. V dôsledku toho došlo k erupcií záujmu čerpania zdrojov z obrovských dát pre určenie stanoviska. Analýza sentimentu alebo čerpanie je výpočtová úprava stanovísk, emócií a subjektivity textu. V tejto práci sa pozrieme na rôzne problémy pri aplikácií analýzy sentimentu. Budeme diskutovať o detailoch rôznych prístupov na vykonanie výpočtovej úpravy stanovísk a emócií.

Kľúčové slová : sentiment, analýza sentimentu, Twitter.

ABSTRACT

MIHALÍK, Miroslav: The tools used for analysis of customer sentiment on social networks– University of Economics in Bratislava. Faculty of Economic Informatics; Department of Applied Informatics. – Adviser: Ing. Mária Szivósová, PhD. – Bratislava: FHI EU, 2017, 57 p

Our day-to-day life has always been influenced by what people think. Ideas and opinions ofothers have always affected our own opinions. The explosion of Web 2.0 has led to increasedactivity in Podcasting, Blogging, Tagging, Contributing to RSS, Social Bookmarking, andSocial Networking. As a result there has been an eruption of interest in people to mine thesevast resources of data for opinions. Sentiment Analysis or Opinion Mining is thecomputational treatment of opinions, sentiments and subjectivity of text. In this report, wetake a look at the various challenges and applications of Sentiment Analysis. We will discussin details various approaches to perform a computational treatment of sentiments and opinions.

Key words: sentiment, sentiment analysis, Twitter.

Obsah

Úvod

Neodmysliteľnou súčasťou dnešného internetu sú sociálne siete. Do popredia záujmu sa dostali iba pred pár rokmi, ale je veľmi pravdepodobné, že sa tam usadia. Ľudia používajú rôzne sociálne siete k zdieľaniu názorov, obrázkov, fotiek alebo ku konverzácií s priateľmi. Vďaka vzniku online marketingu, firmy prezentujúce sa na sociálnych sieťach došli k názoru, že potrebujú sledovať o čom sa ich zákazníci bavia a aké majú názory. Dôležitou súčasťou prieskumu je analýza sentimentu v písaných prejavoch, ktorá môže odhaliť posuny v postojoch, ktoré v anketách, prieskumoch verejnej mienky alebo v dotazníkoch nie sú viditeľné.

V kapitole Súčasný stav riešenej problematiky práca definuje pojmy sentiment a analýza sentimnetu. Charakterizuje hlavné súčasti analýzy sentimentu a teoretické návrhy pre praktické prevedenie analýzy sentimentu.

V kapitole Výsledky práce sa práca zaoberá sociálnou sieťou Twiiter a nástrojmi, ktoré sú využívané pre analýzu sentimentu na tejto sociálnej sieti. V tejto časti sa práca venuje rozboru výstupov vopred vybraných nástrojov pre analýzu sentimentu a opisu ich funkcií. Táto analýza ma za účel poukázať na silnú a slabú stránku týchto nástrojov. V závere tejto kapitoly je opísaná budúcnosť a problémy ktoré musia byť vyriešené pre efektívnejšie výsledky dosiahnuté pomocou analýzy sentimentu.

1 Súčasný stav riešenej problematiky doma a v zahraničí

V dnešnej dobre už nie je efektívne získavať názory a pocity zákazníkov starou formou dotazníkov. Najefektívnejšou formou ako získavať ich názory je analyzovať sentiment na sociálnych sieťach ktoré zákazníci vo veľkej miere využívajú.

1.1 Sentiment

Každý človek využíva rôzne emócie a zastáva si svoje názory. Sú to psychicky a sociálne konštruované procesy, ktoré zahrnujú subjektívne zážitky. Tieto zážitky môžu byť tak pozitívne ako aj negatívne. Človek hodnotí udalosti podľa subjektívneho vzťahu a to vedie k jeho záverečnému postoju k problematike. Sentiment je postoj mysle, náklonnosti alebo odporu, ktorá závisí na rozumovej úvahe. Takýto postoj môže vyvolávať kladné alebo záporné emócie. Postoj môže byť aj trvalý, a nepodlieha vedomej kontrole. Už v stredoveku bol pojem sentiment známy. Pokladal sa za prejav duše ktorý človeku spôsoboval radosť ale tak isto aj utrpenie. Postupom času aj vďaka kresťanstvu ktoré oceňovalo lásku ,začal napomáhať k hlbšiemu hodnoteniu cieľov. Začiatkom 18. storočia pár filozofov začalo vyzdvihovať city a sentiment. Podľa týchto filozofov boli city a sentiment prejavom úprimného prejavu ľudskosti a považovali to za pravý základ morálky. Práve takéto ohodnotenie citov dalo podnet na vznik romantizmu. V 19. storočí bol romantizmus na vrchole a zaznamenal veľkú popularitu aj ľudových vrstvách. Avšak neskôr v Strednej Európe dostal sentimentalizmus negatívny význam ako len prejav neúčinného pestovania citov.[1] Takáto dvojznačnosť sentimentu a citov ostala v našej spoločnosti dodnes. Z jednej strany sú city pozitívnym zdrojom a motívom jednania, či už dobrého alebo zlého. Avšak na druhú stranu si uvedomujeme ich nestálosť a tak isto aj nespoľahlivosť. Problém spočíva v rozdelení ľudských schopností na rozumovej a citovej úrovni. Analýza sentimentu je v počítačovej vede klasifikovaná ako jeden z problémov spracovania prirodzeného jazyka. Analýza a algoritmy, ktoré sa zaoberajú problematikou, majú za cieľ zistiť, aký je postoj človeka ktorý daný text píše k určitej téme. Prvou úlohou je zistiť, či analyzovaný text obsahuje vôbec nejaký názor. Rozlišujeme dve kategórie:

- objektívny
- subjektívny

1

Objektívny nazývame text, ktorý neobsahuje žiadnu subjektivitu. To znamená že neobsahuje žiadny názor na ktorý by sme mohli využiť analýzu.[2] Naopak subjektívny text obsahuje subjektivitu. Na zisťovanie či je text objektívny alebo subjektívny exituje veľké množstvo algoritmov. Ako príklad si môžeme uviesť napríklad AdaBoost, BoosTexter. Tieto algoritmy využívajú ukážkové dáta, vďaka ktorým sa následne učia a robia binárnu klasifikáciu. Rozhodujú či text patrí alebo nepatrí do danej množiny. Následne môžeme rozlišovať druh a rozsah sentimentu. Výsledok a presnosť sa odvíja od vzorky dát. Môžeme analyzovať blogy, príspevky v diskusných fórach, články, statusy na sociálnych sieťach alebo len samotné vety. Existujú tri druhy sentimentov a jeden špeciálny.

- pozitívny
- negatívny
- neutrálny
- bipolárny

Pozitívny a negatívny sentiment je zjavný z názvu. Neutrálny sentiment algoritmy zjednodušujú a považujú ho za subjektívne vyjadrenie. Najviac komplikovaný je bipolárny sentiment. Je to sentiment ktorý je negatívny aj pozitívny zároveň. Analýza sentimetu nie je triviálny problém. V dnešnej dobre neexistujú nástroje ktoré by ju zvládali na požadovanej úrovni v angličtine nie to ešte v slovenskom jazyku. Nástroje preto pracujú s chybovosťou. Preto je potrebné pracovať s veľkou vzorkou dát, aby sa chybovosť čo najviac minimalizovala.

1.2 Analýza sentimentu

Analýza sentimentu je spracovanie prirodzeného jazyka a extrakcia informácií . Cieľom je získať informácie o pocitoch človeka ktorý daný text napísal. Tieto informácie môžu byť pozitívne ale aj negatívne. Všeobecne povedané, analýza sentimentu si kladie za cieľ zistiť postoj autora, pokiaľ ide o určitú tému . V posledných rokoch exponenciálny nárast využitia internetu je hnacou silou pre analýzu sentimentu. Internet je obrovské úložisko štruktúrovaných a neštruktúrovaných dát. Analyzovať takéto dáta a extrahovať z nich užitočné informácie verejnej mienky je náročná úloha. Analýza emócií je založená na tom, že určí či daný text obsahuje negatívny alebo pozitívny sentiment. Môže sa to určovať na základe viet, pokiaľ sú tieto vety klasifikované. Tak isto aj slová v daných vetách musia byť klasifikované. Analýza sentimentu tieto slová identifikuje, a určí ktoré slovo má akú

[2]

emóciu. Auto môže písať o objektívnych skutočnostiach alebo svojich subjektívnych názoroch a postojoch k danej problematike.[3] Je nutné rozlišovať predmet ku ktorému je sentiment smerovaný. Text môže obsahovať veľa subjektov ale je nutné nájsť entitu, ku ktorej bol sentiment smerovaný. Emócie sú klasifikované ako objektívne (fakty), pozitívne (označujú stav šťastia, blaženosti alebo pochádzajú zo strany spisovateľa), alebo negatívne (označujú stav smútku, skľúčenosti či sklamanie zo strany spisovateľa.

1.3 Aplikácie analýzy sentimentu

"Word of mouth" (WOM) je proces odovzdávania informácií od človeka k človeku a hrá dôležitú úlohu pri rozhodovaní zákazníkov pri nákupe. V komerčných situáciách , WOM obsahuje zdieľanie spotrebiteľských postojov, názory a reakcie o firmách, produktoch alebo službách s ostatnými zákazníkmi. WOM komunikačné funkcie sú založené na sociálnych sieťach a na dôvere. Ľudia sa spoliehajú na názory rodín s deťmi , priateľoch , známych osobností a iných ľudí na sociálnych sieťach . Výskum poukázal , že ľudia ľahko uveria zdanlivo nezaujatým názorom iných ľudí na takzvaných on-line hodnotiacich stránkach. Toto je miesto kde prichádza na rad Analýza Sentimentu. Rastúca dostupnosť bohatých zdrojov spotrebiteľskej mienky ako napríklad blogov, statusov na sociálnych sieťach im zjednodušujú "rozhodovací proces". S explóziou Web 2.0 platformy majú spotrebitelia veľký priestor pre zdieľanie názorov. Významné firmy si uvedomujú akú silu majú v dnešnej dobe spotrebiteľské hlasy a snažia sa formovať ich názory. Analýza sentimentu tak nájde uplatnenie v oblasti spotrebiteľského trhu pre hodnotenie tovarov alebo služieb. Pomocou sociálnych sietí je možné napríklad zistiť názor na nové trendy. Tak isto je možné zistiť či nedávno zverejnený film je hit. Všeobecne klasifikujúce aplikácie je možné rozdeliť do niekoľkých kategórií

- Applications to Review-Related Websites

(sú to napríklad filmové recenzie, recenzie výrobkov atď.)

- Applications as a Sub-Component Technology

(napríklad detekcia spamu, kontextová detekcia informácie atď.)

- Aplikácie v obchodných a vládnych spravodajstvách

(zistenie postojov a spotrebiteľských trendov)

- Aplikácie medzi rôznymi doménami

3

(zistenie verejnej mienky o vládnych predstaviteľoch alebo mienka o zákonoch, pravidlách a nariadeniach).[4]

1.4 Výzvy pre analýzu sentimentu

Prístupy sentiment analýzy majú za cieľ extrahovať pozitívny a negatívny sentiment ktorý nesie slovo v určitom texte. Musí klasifikovať text ako pozitívny, negatívny, objektívny alebo neurčitý ak nie je schopný nájsť určité slovo vo svojom ložisku. V tomto ohľade môžeme využiť postup ktorý sa nazýva klasifikácia textu. V klasifikácií tetu exituje veľa tried, ktoré zodpovedajú rôznym témam, zatiaľ čo v analýze sentimentu máme len tri široké kategórie. Zdá sa teda že analýza sentimentu je jednoduchšie ako klasifikácia textu. Avšak nie je to celkom pravda. Všeobecné problémy možno zhrnúť do týchto tried.

1.4.1 Implicitný sentiment a sarkazmus

Veta môže obsahovať implicitný sentiment aj bez prítomnosti akéhokoľvek sentimentu uloženého v slove. Ako príklad si môžeme uviesť nasledujúci príklad.

"Ako môže niekto sedieť pri tomto filme ?"

"Človek by mal pochybovať o psychickom zdraví spisovateľa, ktorý napísal túto knihu."

Oba vyššie uvedené texty nie sú explicitné. Nemajú žiadny negatívny sentiment ktorý nesie slovo, avšak obe tieto vety sú negatívne. A tak je dôležitejšie určiť sémantiku v analýze sentimentu ako ju zisťovať v syntaxe.

1.4.2 Doménová závislosť

Exituje veľa slov , ktorých polarita sa mení z domény do domény. Ako príklad si môžeme uviesť nasledujúce príklady:

"Príbeh bol nepredvídateľný."

"Riadenie vozidla je nepredvídateľné."

"Choď si prečítať knihu."

V prvom príklade je emócia kladná, zatiaľ čo emócia v druhom príklade je negatívna. Tretí príklad má pozitívny sentiment k doméne knihy, ale negatívny k doméne čitateľa. (čitateľ je požiadaný aby si šiel prečítať knihu.)

[4]

1.4.3 Zmenené očakávania

Niekedy autor zámerne nastaví kontext tak, aby ho na konci vyvrátil. Môžeme si uviesť nasledujúci príklad:

" Tento film mal byť brilantný. Hrajú v ňom prvotriedny herci, režisér je dobrý a Stallone sa snaží podávať dobrý výkon. Avšak ani to nestačí."

Napriek prítomnosti slov, ktoré sú pozitívne, celkový sentiment je negatívny pretože zásadná je posledná veta . Zatiaľ čo v tradičnom texte by bol pri textovej klasifikácií tento text klasifikovaný ako pozitívny v analýze sentimetu sa kladie dôraz na poslednú vetu ktorá je negatívna. Celkový sentiment tohto textu je negatívny.

1.4.4 Pragmatika

Je dôležité odhaliť pragmatiku užívateľského stanoviska, ktorá môže zmeniť celkový sentiment. Pozrime sa na nasledujúce príklady:

" Práve som dopozeral zápas, kde Barca ZNIČILA AC Miláno."

" Koniec seriálu ma úplne ZNIČIL."

Veľké písmená môžu byť použité so zámerom naznačovať emóciu. Prvý príklad naznačuje pozitívny sentiment, zatiaľ čo druhý naznačuje negatívny sentiment. Existuje mnoho ďalších spôsobov ako vyjadriť pragmatizmus.

1.4.5 Svetové poznanie

Často musí byť začlenený do systému pre detekciu citov aj svetové poznanie.

Pozrime sa na nasledujúce príklady:

"Je to Frankenstein."

" Táto žena je ako Matka Tereza"

Prvá veta ukazuje negatívny sentiment, zatiaľ čo druhá zachytáva pozitívny sentiment. Človek musí vedieť informácie o Frankenstainovi a Matke Tereze aby vedel určiť aký je to sentiment.

1.4.6 Subjektívna detekcia

Pod pojmom subjektívna detekcia rozumieme to, že je potrebné rozlišovať medzi objektivitou a subjektivitou v texte. Detekčný modul musí vedieť odfiltrovať objektívnu skutočnosť. Ale je to zložité. Môžeme sa pozrieť na nasledujúce príklady:

Neznášam romantické príbehy.

Nepáči sa mi tento film “Láska v Seattli"
Prvý príklad obsahuje objektívnu skutočnosť. Táto skutočnosť hovorí o tom že autor nemá rád romantické filmy. Zatiaľ čo druhý príklad obsahuje subjektívny názor na konkrétny film a táto emócia je negatívna.

1.4.7 Entitná identifikácia

Text alebo veta môže obsahovať viacero subjektov. Je nesmierne dôležité zistiť, ktorý je subjekt ku ktorému sa stanovisko vzťahuje. Pozrime sa na nasledujúce príklady:
Samsung je lepší ako Nokia.
iOS je lepší ako Android.
Z uvedených príkladov je jasné, že sentiment pre Samsung a iOS je pozitívny a pre Nokiu a Android je sentiment negatívny.

1.4.8 Negácia

Manipulácia s negáciou je náročná úloha pre analýzu sentimentu. Negácia môže byť vyjadrená pomocou jemných spôsobov, dokonca bez využitia akéhokoľvek negatívneho slova. Spôsob často nasleduje v manipulácií s negáciou explicitne vo vetách "nemám rád film". Sentiment je vyjadrený pomocou slova "ne-mám". Ale to nefunguje pri vetách ako " Nemám rád ten film, ale páči sa mi réžia." Takže musíme vziať do úvahy rozsah negácie. Musíme zistiť či sa negácia vzťahuje na celý text, alebo len na jeho časť. Takže treba mať na pamäti že slová ako ne-mám a slová ako ale môžu byť v jednej vete a sentimet môže byť rôzny. S týmto treba počítať pri navrhovaní algoritmu.[5]

1.5 Funkcie pre analýzu sentimentu

Inžinierska funkcia je veľmi dôležitou a zásadnou úlohou pre analýzu sentimentu. Prevod časti textu do funkcie vektora je základným krokom pri riadení akýchkoľvek dát pri analýze zákazníckeho sentimentu. V nasledujúcej časti sa budeme venovať niektorým bežne používaným funkciám v analýze senzitivity.

1.6 Termín Prítomnosť vs. Termín Frekvencia

Termín frekvencia bola vždy považovaná za zásadnú v tradičnom vyhľadávaní informácií v klasifikácii textu. Neskôr sa však zistilo, že termín prítomnosť je viac dôležitá pre analýzu sentimentu ako termín frekvencia. To znamená, že binárne ocenené prvky vektora ktoré sa vyskytujú v texte sú ohodnotené hodnotou 1 a ak sa nevyskytujú sú

[5]

ohodnotené hodnotou 0. To však nie je efektívne ako sme videli na predchádzajúcich príkladoch. Taktiež bolo zistené, že vzácne slová ktoré sa vyskytujú v textoch môžu obsahovať viac informácií ako bežne používané slová. Tento jav sa nazýva Hapax Legomena.

1.7 Pozícia

Slová sa objavujú v určitých pozíciách v texte. Ak sa nachádzajú v určitej pozícií môžu niesť väčšiu emóciu ako keď sa nachádzajú inde. Môžeme uviesť príklad keď sú slová pozitívne po celú dobu. Prítomnosť negatívneho sentimentu na konci vety zohráva rozhodujúcu úlohu pri určované emócie. Takže slova objavené v prvých niekoľkých vetách a posledných pár vetách majú väčšiu váhu ako slová ktoré sú na iných miestach v texte.

1.8 N-gram funkcia

N-gramy sú schopné zachytiť kontext do určitej miery a sú široko používané v prírodných jazykoch pri spracovaní úlohy. Čím vyššieho rádu sú n-gramy tým sú užitočnejšie pre vec debaty. Zistilo sa že unigramy prekonajú bigramy pri klasifikácií recenzie filmu a naopak v niektorých prostrediach bigrami dosahujú naopak lepšie výsledky. Preto je potrebné vedieť načo sa daný algoritmus využije a podľa toho určiť aký n-gram sa využije a bude vhodnejší.[6]

1.9 Subsekvencia jadra

Väčšina prác pri analýze sentimetu používa slová alebo modely na úrovni vety. Výsledky sú potom spriemerované naprieč všetkými slovami/vetami/n-gramami za účelom jedného jediného modelu ktorý sa dá použiť pri každej kontrole. Algoritmus Bikel et al. používa subsekvencie. Intuícia je, že funkcia implicitne zaujatých subsekvenčných jadier je dostatočne bohatá, aby odstránila explicitné znalosti inžinierstva alebo modelovania textu. Sekvenčné slovo jadra n-rádu je vážený súčet všetkých možných sekvencií n-dĺžky slova, ktoré sa vyskytujú v oboch porovnávaných reťazcoch.[7]

Matematicky slovo sekvencie jadra je zapísané:

6

7

$$K_n(s,t) = \sum_{u \in \Sigma^\star} \sum_{\mathbf{i}:s[\mathbf{i}]=u} \sum_{\mathbf{j}:t[\mathbf{j}]=u} \lambda^{(i[n]-i[1]+1)+(j[n]-j[1]+1)}$$

kde λ je parameter jadra.

Dĺžka n, ktoré sa skladá z indexových reťazcov, ktoré zodpovedajú subsekvenci u.

Hodnota aj [n] - i [1] + 1 môže byť považovaná za celkovú dĺžku rozpätia

ktorý predstavuje výskyt subsekvencie u.

Kombináciou sekvenčného jadrá v rôznom poradí:

$$K(s,t) = \sum_{i=1}^{\infty} \mu^{1-i} K_i(s,t)$$

1.10 Časti reči

Informácie o časti reči sú najčastejšie využívane vo všetkých úlohách. Jedným z hlavných dôvodoch je , že poskytujú hrubú formu slova a jeho adjektíva. Prídavné mená sa používajú najčastejšie zo všetkých častí reči. To znamená že autor používa viacej prídavných mien ako iných slovných druhov. Bola zistená korelácia medzi adjektívami a subjektívami. Aj keď všetky časti ľudskej reči sú dôležité najviac sú využíva adjektíva. Zistilo sa že pomocou prídavných mien človek najlepšie vyjadrí svoju emóciu. [8]

1.11 Adjektíva, príslovky a ich kombinácia

Väčšina prísloviek nemá žiadnu predchádzajúcu polaritu. Ale ak dôjde ku kombinácií príslovky s adjektívom, môže hrať príslovka významnú úlohu pri určovaní sentimentu vety. Bolo ukázané, že príslovky ktoré sa používajú menia hodnotu emócií. Príslovky ktoré menia emócie vo vete môžeme klasifikovať takto:

- Príslovka potvrdenia: absolútne, iste
- Príslovka pochybností: možno, pravdepodobne
- Silne zintenzívnené príslovky : nesmierne
- Slabo zintenzívnené príslovky: ťažko, mierne

Kombinácie sú definované do dvoch typov:

1.Unárne kombinácie: Obsahujú jednu príslovku a jedno prídavné meno. Sentimentálne skóre ktoré prídavného mena je upravené pomocou príslovky, ktorá s ním susedí.

8

2.Binárne kombinácie: Obsahujú viac ako jednu príslovku a prídavné meno. Sentimentálne skóre tejto kombinácie sa vypočíta opakovanou úpravou skóre každej príslovky ktorá sa pridá k prídavnému menu.

Toto je ekvivalent k definovaniu binárnych kombinácií, pokiaľ ide o dva unárne kombinácie ktoré sú opakované. Niektoré axiomatické pravidlá majú za úlohu špecifikovať spôsob, ako príslovka zmení sentiment adjektíva. Jeden takýto axióm je možné charakterizovať ako:

"Každá slabo zintenzívnená príslovka a každá príslovka pochybností má skóre menšie alebo rovné ako každá silne zintenzívnená príslovka a príslovka potvrdenia."

Potom, niektoré funkcie sú popísané za účelom kvantifikácie axiómov. Funkcia f vyhodnotí

adjektívum a jeho kombináciu s príslovkou a následné vráti jeho výsledné skóre.

1.11.1 Kladná a silne zintenzívnená príslovka

AAC-1. If sc(adj) > 0 and adv ∈ AFF U STRONG

then f (adv, adj) ≥ sc (adj).

AAC-2. If sc (adj) < 0 and adv ∈ AFF U STRONG,

then f (adv, adj) ≤ sc (adj).

Napríklad, f hodnota "nesmierne dobré" je pozitívnejšie než skóre pozitívneho adjektíva "dobré".

1.11.2 Slabo zintenzívnená príslovka

AAC-3. If sc(adj) > 0 and adv ∈ WEAK, then

f (adv, adj) ≤sc (adj).

AAC-4. If sc (adj) < 0 and adv ∈ WEAK, then

f (adv, adh) ≥ sc (adj).

Napríklad, f hodnota "málo dobré" je negatívnejšia než skóre pozitívneho adjektíva "dobré". Je to účinok slabej príslovky ktorá s adjektívom susedí.

1.11.3 Príslovka pochybností

AAC-5 If sc(adj) > 0, adv ∈ DOUBT, and adv' ∈

AFF U STRONG, then f (adv, adj) ≤ f (adv', adj').

AAC-6 If sc (adj) < 0 is negative, adv ∈ DOUBT, and

adv' ∈ AFF U Strong, then f (adv, adj) ≥ f(adv', adj).

Napríklad, f hodnota "pravdepodobne dobrý" je menšia ako "nesmierne dobrý".

1.11.4 Minimizers

AAC-7 If sc(adj) > 0 and adv ∈ MIN, then

f(adv,adj) ≤ sc(adj).

AAC-8 If sc(adj) < 0 and adv ∈ MIN, then

f(adv, adj) ≥ sc(adj).

Napríklad, "ťažko dobré" je menej pozitívne než pozitívne adjektívum "dobré"

1.12 Bodové algoritmy

1.12.1 Variabilný bodový algoritmus

Algoritmus modifikuje skóre kombinácie použitím funkcie f , ktorá je definovaná nasledujúcim spôsobom:

If adv ∈ AFF U strong, then

fvs (adv, adj) = sc(adj) + (1-sc(adj)) * sc(adv)

if sc(adj) > 0. If sc(adj) < 0,

fvs(adv, adj) = sv(adj) - (1-sc(adj))*sc(adv)

IF adv ∈ WEAK U DOUBT, VS reverses the above and returns

fvs (adv, adj) = sv(adj) - (1 - sv(adj)) * sv(adv)

if sc(adj) > 0. If sv(adj) < 0, it returns

fvs(adv, adj) = sv(adj) + (1 - sv(adj)) * sc (adv).

To znamená, že výsledný počet bodov kombinácie je skóre adjektíva, ktoré je vhodne upravené účinnou príslovkou.

1.12.2 Adjektívne prioritný bodový algoritmus

If adv ∈ AFF U STRONG, then

f_{APSr} (adv, adj) = min (1,sc(adj) + r * sv(adv)).

if sc(adj) >0. If sc(adj) > 0,

f_{APSr} (adv, adj) = min (1,sc(adj) - r * sv(adv)).

If adv ∈ WEAK U DOUBT, then APSr reverses the above

and sets f_{APSr} (adv, adj) = max (0, sc(adj) - r* sc(adv))

If sc(adj) >0. If sc(adj) < , then fAPSr (adv, adj) =

max(0, sc(adj) + r * sc (adv)).

Dáva sa prednosť adjektívu pre príslovkou. Skóre adjektíva určuje hmotnosť r. Táto hmotnosť r rozhodne, do akej miery príslovka ovplyvňuje skóre adjektíva.

1.12.3 Príslovková priorita bodového algoritmu

If adv ∈ AFF U STRONG, then

f_{advFSr} (adv, adj) = min (1, sc(adv) + r * sc(adj))

if sc(adj) > 0. If sc(adj) <0,

f_{advFSr} (adv, adj) = max (0, sc(adv) - r * sc (adj)).

Témovo orientované črty

"Bag-of-word" a frázy sú široko používané ako rysy. Avšak v mnohých oblastiach, idividuálne frázy a ich hodnota ma celkovo malý vzťah s celkovým sentimentom textu. Výzvou v analýze sentimetu je využitie takých aspektov textu, ktoré sú nejakým spôsobom reprezentatívnym tónom celého textu. Často zavádzajúce vety sa používajú na posilnenie sentimetu v texte. Ak použijeme "bag-of- word" alebo frázy, algoritmus nebude schopný rozlíšiť medzi tým čo sa hovorí v lokálnych frázach, a čo je myslené v globálnom texte. Ako príklad si môžeme uviesť že autor niekedy odbočuje od témy, tak isto môže používať sarkazmus a algoritmus nie je schopný zhodnotiť túto situáciu. Preto boli vyvinuté nasledujúce postupy:[9]

1.13.1 Sémantické orientácie PMI

Sémantická orientácia (SO) odkazuje k reálnemu číslu miery pozitívneho alebo negatívneho sentimentu ktorý je vyjadrený pomocou slov alebo fráz. Hodnotou frázy sú frázy, ktoré sú zdrojom SO hodnoty. Akonáhle boli požadované frázy obsiahnuté v texte, každej z nich sa priradí hodnota SO.

Vzájomná výmena informácií (*pointwise mutual information* (PMI) so slovami "vynikajúci" a "chudobný". PMI je definovaný nasledovne:

PMI (w1, w2) = log2 (p (w1 a w2) / p (w1) p (w2)).

SO fráza je rozdiel medzi PMI so slovom "vynikajúci" a jeho PMI so slovom "chudobný". tj

SO (výraz) = PMI (fráza, "vynikajúci") - PMI (fráza, "chudobný")

9

Intuitívne, to sú hodnoty nad nulou pre frázu s väčšou PMI a to je slovo "vynikajúci" a pod nulou pre väčšie PMI so slovom "chudobný". SO hodnota nula naznačuje úplne neutrálnu sémantickú orientáciu.

1.13 Sémantická diferenciácia WordNetom

WordNet vzťahy sú použité na odvodenie troch hodnôt vzťahujúcich sa k emocionálnemu zmyslu adjektíva. Tieto tri hodnoty môžeme rozdeliť takto. Podľa ich potencie (silné a slabé adjektívum), podľa aktivity (aktívne alebo pasívne adjektívum) a podľa hodnotenia (dobré alebo zlé adjektívum). Tieto hodnoty sú odvodené meraním relatívnej *minimal path length* (MPL) s WordNetom. Je to minimálna dĺžka dráhy medzi adjektívom a dvojicou slov vhodných pre daný faktor. *Evaluative factor* (EVA) je pomerový faktor na porovnanie MPL medzi adjektívom z kategórie "dobré" a medzi MPL adjektíva z kategórie "zlé".

Sentiment vyjadrený vzhľadom na konkrétny predmet je najlepšie identifikovaný s odkazom na subjekt samotný. V niektorých aplikačných oblastiach je známe dopredu, aká je téma a ku ktorému objektu je potrebné vyhodnotiť sentiment.

To sa dá využiť vytvorením niekoľkých tried funkcií, ktoré vychádzajú z SO hodnoty viet, vzhľadom na ich postavenie vo vzťahu k téme textu. V textoch je všeobecne jeden primárny subjekt, a stanovisko k danému predmetu je buď pozitívne alebo negatívne. Avšak sekundárne témy sú tiež do určitej miery použiteľné.

Príklad: Stanovisko (referencie) autorova na recenziu knihy môže byť užitočné pri recenzii knihy.

Príklad: V prehľade výrobkov, postoj voči spoločnosti, ktorá vyrába produkt môže byť relevantný.

1.14 Strojové prístupy učenia

Vo svojej práci, Pang Lee et al (2002,2004) porovnal výkonnosti Naive Bayes, Maximum Entropy a Support Vector Machines v analýze senzitivity v rôznych funkciách. Pri porovnávaní výkonnosti bral ohľad na unigramy, bigramy a ich kombináciu. Tieto unigramy a bigramy obsahovali len informácie o adjektívach.

Výsledky tejto práce sú zahrnuté v tabuľke

Tabuľka č.1 : Presnosť porovnania rôznych klasifikátorov v AS na súbore filmových recenzií.

Funkcia	**Početnosť funkcie**	**Frekvencia a prítomnosť**	**NB**	**ME**	**SVM**
	16165	frekvencia	**78.7**	N/A	72.8
Unigramy					
Unigramy	16165	prítomnosť	81.0	80.4	**82.9**
Unigramy+bigramy	32330	prítomnosť	80.6	80.8	**82.7**
Bigramy	16165	prítomnosť	77.3	**77.4**	77.1
Unigramy+POS	16695	prítomnosť	81.5	80.4	**81.9**
Prídavné mená	2633	prítomnosť	77.0	**77.7**	75.1
Top 2633 unigramov	2633	prítomnosť	80.3	81.0	**81.4**
Unigramy+pozícia	22430	frekvencia	81.0	80.1	**81.6**

Z týchto pozorovaných výsledkov vyplýva, že:

a. Prítomnosť funkcie je dôležitejšie ako frekvencia.

b. Pomocou Bigramov presnosť skutočne klesne.

c. Presnosť sa zlepší, keď sú všetky často sa vyskytujúce slová zo všetkých slovných druhov prijaté, nie len prídavné mená.

d. Zahrnutie informácií o polohe zvyšuje presnosť.

e. Ak je funkcionálny priestor malý, Naive Bayes je lepšie ako SVM. Ale SVM dosahuje lepšie výsledky ak sa funkionálny priestor zväčšuje.

Ak sa funkcionálna plocha zväčšuje, môže Maximum Entropy dosahovať lepšie výsledky ako Naive Bayes.

4. Kognitívy prístup [10]

1.15 Čo je to subjektívna analýza, rozprávanie

Vstupom do klasifikatóra sentimentu je vždy takzvaný "tvrdohlavý" text. Pod pojmom "tvrdohlavý" text rozumieme text, ktorý obsahuje negatívny aj pozitívny

[10]

sentiment. Po vstupe textu je potrebné odfiltrovať objektívne fakty z textu. Táto práca extrakcie alebo filtrovania objektívnych faktov od subjektívnych faktov sa na nazýva analýza subjektivity. Časť textu často obsahuje pohľad druhej osoby alebo dokonca viacerých osôb. Text obsahuje širokú škálu emócií, názorov a perspektív autora. Preto je potrebné identifikovať znak, ktorým sa bude odlišovať. Cieľom je nielen detekovať časť textu, ale aj čomu daný text zodpovedá.

Rozprávanie je príbeh, ktorý je vytvorený v konštruktívnom formáte. Opisuje frekvenciu vymyslených a pravdivých udalostí. Rozprávanie môže byt tiež celkom fiktívne, kde autor priamo komunikuje s čitateľom. Perspektíva je úloh pohľadu. Príbeh je rozprávaný z pohľadu jedného alebo viacerých znakov. Môže tiež obsahovať znaky prechody ktoré nesúvisia so žiadnym charakterom.

1.16 Analýza úrovne

Aby bolo možné identifikovať charakter a jeho perspektívu, nebudeme robiť analýzu vety. Vyžaduje si to analýzu úrovne, pretože vety nie sú vždy výslovne označené subjektívnymi prvkami a subjektívne vety sa nedajú priamo označiť ako subjektívne.
Príklad 1 : Chcel sa rozprávať s Michalom.
Príklad 2 : Keď vyjdeme von, môžeme s úžasom sledovať, ako sa príroda po viacmesačnom oddychu pripravuje na znovuzrodenie.
V prvom príklade je zastúpená myšlienka priamo. V druhom príklade je zastúpene vnímanie, predstavuje čo charakter vidí, ako vidí situáciu ale neopisuje situáciu priamo ako takú. Taktiež označuje subjektívny charakter. Subjektívne vety, neobsahujú žiadne subjektívne prvky alebo subjektivitu.[11]

11

2 Cieľ, metodika práce a metódy skúmania

Cieľom Diplomovej práce je charakterizovať pojem sentiment, analýza sentimentu a nástroje ktoré sa využívajú pre analýzu sentimentu.

Aby sme neostali len pri teoretických poznatkoch, stanovíme si dva hlavné ciele. Prvým je využitie nástrojov pre analýzu sentimentu a ich aplikovanie na sociálnej sieti Twiiter.

Výsledkom prvého cieľa bude analýza ktorá ukáže ako jednotlivé nástroje pre analýzu zákazníckeho sentimentu fungujú a aké ponúkajú možnosti. K dosiahnutiu tohto cieľa budeme musieť splniť nasledujúce čiastkové ciele:

- Vyhľadať správne analytické nástroje pre analýzu
- Vhodne nastaviť analytický systém
- Vykonať analýzu pomocou analytických nástrojov

V prípade úspešného splnenia čiastkových cieľov prvého cieľa sa budeme snažiť splniť aj druhý cieľ, ktorým je porovnanie analytických nástrojov ktoré sme využili.

Tvorbou tejto práce sme sa snažili priniesť objektívne informácie a výsledky, súvisiace s témou tejto práce. Dôležité bolo podrobne oboznámiť so súčasným stavom danej problematiky. Vyhľadali sme dostupné knižné zdroje ktoré sa aspoň časťou zaoberajú danou problematikou a tieto informácie sme doplnili informáciami z internetu.

Keďže sa táto práca nezaoberá len samotnou teóriou, jedných z hlavných problémov bolo využiť nástroje ktoré sú zamerané na analýzu zákazníckeho sentimentu a uskutočniť analýzu na sociálnej sieti .Prvou úlohou bolo vybrať sociálnu sieť na ktorej sa daná analýza bude robiť. Vybrali sme sociálnu sieť Twitter keďže je to jedna z najväčších sociálnych sietí a užívatelia na nej často prejavujú svoje názory o produktoch a firmách.
Najzložitejšou úlohou bolo vybrať nástroje ktoré sme využili pre analýzu zákazníckeho sentimnetu. Bolo potrebné naštudovať z internetových zdrojov, ktoré nástroje budú najlepšie pre analýzu zákazníckeho sentimentu na sociálnej sieti Twiiter.

Ako druhý cieľ sme si stanovili porovnanie jednotlivých nástrojov pre analýzu zákazníckeho sentimnetu.

Metódy ktoré boli využité pri tvorbe tejto práce boli:

- Analýza nových poznatkov
- Pokus a omyl
- Intuícia
- Subjektívnosť
- Meranie a komparácia

3 Výsledky práce a diskusia

Výsledkom tejto práce je bližšie oboznámenie čo je to analýza sentimetu a následná analýza zákazníckeho sentimentu na sociálnej sieti Twitter za pomoci nástrojov ktoré sú na to určené.

3.1 Sociálna sieť Twitter

Twitter je jednou z najväčších sociálnych sietí. Využíva ju viac ako 550 miliónov užívateľov po celom svete. Títo užívatelia generujú obrovský počet dát. Jedná sa o mikro blogovací systém ktorý sa nazýva aj "databáza názorov". Užívatelia tu môžu komunikovať pomocou textových správ s maximálnou možnou dĺžkou 140 znakov (tzv.tweet). Toto obmedzenie má vplyv na výpovednú hodnotu každého príspevku. Preto sociálna sieť Twitter poskytuje špeciálny formát pre kľúčové slová a označenie užívateľov. Kľúčové slová , ktoré sú nazývané sémantickými značkami (hashtags) sú zapísané vo forme #slovo a užívateľov označujeme pridaním znakov " @ ", takže @užívateľ. Jednotlivé príspevky je možno zaradiť medzi obľúbené alebo ich je možné ďalej preposielať (retweet), čoho sa dosiahne použitím špeciálnej funkcie alebo pridaním reťazca " RT " pred posielanou správou.[12] Tento princíp preposielania je podstatou rýchleho šírenia správy naprieč celou sieťou. Každý užívateľ môže sledovať (follow) príspevky iných užívateľov, čím sa stáva sledujúcim (follower). Pre zhromažďovanie príspevkov určitých užívateľov slúžia zoznamy (lists), ktoré odoberateľov (subscribers) dovoľujú odoberať príspevky od členov (members). Zoznamy môžu byť súkromné alebo verejné. Podľa toho, koho sledujete a kde sa práve nachádzate, vám sociálna sieť Twitter ponúka trendy - kľúčové slová vybrané pre každého užívateľa na mieru. Okrem príspevkov môžeme posielať súkromné správy (direct messages), ktoré sú viditeľné len medzi komunikujúcimi užívateľmi. Twitter umožňuje okrem textu pridávať k príspevkom aj obrázky, ktré sú najprv nahrané na server a potom vo forme URL adresy pridané k príspevku. Pretože URL adresy môžu obsahovať veľké množstvo znakov, často využívame služby skracovania URL adresy systémom Twitteru alebo tretích strán.

3.2 Sentiment na sociálnej sieti Twitter

12

V porovnaní s ostatnými sociálnymi sieťami sa na Twitteri objavuje menšie percento neutrálnych príspevkov. K tomuto javu môžu dopomôcť média často vyzývajúce obecenstvo ku aktivite pomocou zverejňovania sémantických značiek alebo reakcií na zadanú otázku. Slovenský Twitter má viac ako 130 tis. užívateľov a ich počet stále narastá. [13]

3.3 Funkcia ohodnotenia n-gramu

Táto funkcia je využívaná všetkými metódami pre výpočet sentimentu. Jej úlohou je vybrať hodnotenie pre konkrétny n-gram, ktoré bude ďaľej používané v metódach pre vyhodnotenie sentimentu. Funkcie sa rozlišujú číselnou hodnotou sentimentu a špeciálnymi značkami. Pokiaľ je väčšinové hodnotenie n-gramu špeciálna značka, bude tomuto n-gramu priradená ako hodnota práve táto špeciálna značka. Naopak , ak je väčšinové hodnotenie číselné, potom sa vypočíta priemer a tento priemer bude použitý ako hondota n-gramu.

3.3.1 Triviálna metóda

Jedná sa o základnú metódu pre vyhodnocovanie sentimentu príspevku. Je vypočítaná podľa označkovaných unigramov. Funguje na vyhľadávaní všetkých unigramov v korpuse a získaní ich ohodnotení. Neutrálne unigramy algoritmus ignoruje a nezapočítava do hodnotenia. Pokiaľ algoritmus objaví zosilujúce značky, tak ich nasledovník bude ohodnotený dvojnásobnou mierou, naopak ak sa jedná o značku zoslabenia potom hodnota unigramu bude vydelená dvoma. Cudzojazyčné a nevyžiadané slová algoritmus ignoruje, ale zároveň ich ukladá do korpusu, aby sa ďalej nezobrazovali pri značkovaní ostatných užívateľov. Pri nájdení značky negácie je negovaná hodnota nasledujúceho unigramu. Výsledná hodnota sentimentu je priemer všetkých nenulových hodnôt (vrátane tých zosílených / zoslabených) pre každý príspevok. Túto metódu je tiež možné modifikovať tak, že sa robí na bigramoch alebo trigramoch. Pri tejto metóde dochádza k najlepším výsledkom pri silne negatívnych alebo silne pozitívnych príspevkoch. Pri malom počte neutrálnych n-gramov platí, že pomocou bigramov a trigramov sú výsledky presnejšie. Najväčšou slabinou tejto metódy je detekcia neutrálnych príspevkov, pokiaľ sa jedná o príspevok obsahujúci ojedinelo sentimentálne zafarbené

[13]

slová. Tieto väčšinou neutrálne príspevky sú často hodnotené ako pozitívne alebo negatívne.

Triviálna metóda zapísaná pomocou pseudokódu

```
public function trivialRanking (grams)  {
grams = map(toLower(grams));
rankedGrams = getRankedGrams(grams); // vráti z databázy pole gramov
sentimentScore = 0;
count = 0;
modifier = 1;
foreach (grams as gram) {
 if (existuje gram medzi označkovanými gramami z databázy)
score = getScore(grams); // vráti ohodnotenie gramu z databázy
else
continue;
if (score > -1 AND score < 1) {
modifier = 1;
continue;
}
else if (isNumeric(score)) {
sentimentScore += modifier * score;
modifier = 1;
}
else if (score == "+")
modifier *= 2;
else if (score == "-")
modifier *= 0.5;
else if (score == "!")
modifier = -1;
else if (score == "?") {
continue;
modifier = 1;
}
count++;
```

```
}
if (count == 0)
return 0;
return sentimentScore / count;
}
```[14]

3.3.2 Priemerová metóda

Táto metóda vychádza z predchádzajúcej triviálnej metódy a obohacuje ju o koeficient počtu neutrálnych n-gramov, ktorý prináša " zjemnenie " pre detekciu neutrálnych príspevkov. Algoritmus vypočíta podiel neutrálnych n-gramov (-1 < N > 1), ktorý sa nazýva koeficient počtu núl. Týmto koeficientom sa ďalej násobí aritmetický priemer nenulových n-gramov. Toto roznásobenie berie do úvahy veľký počet neutrálnych slov a " zjemňuje " význam ojedinelých sentimentálne zafarbených slov. Tento algoritmus je taktiež modifikovaný pre prácu na bigramoch a trigramoch. Ojedinelé mierne pozitívne alebo negatívne slová automaticky nezaznamenajú príspevok s takýmto sentimentom, ale väčšinou sú vyhodnotené ako neutrálne. Toto je hlavnou prednosťou tohto algoritmu. Avšak na druhú stranu, tento algoritmus nedokáže detekovať veľmi slabý sentiment či je pozitívny alebo negatívny.

Priemerová metóda zapísaná v pseudokóde

```
private function averageRanking(grams) {
grams = map(toLower(grams));
rankedGrams = getRankedGrams(grams); // vráti z databázy pole gramov
sentimentScore = 0;
count = 0;
countZeros = 0;
modifier = 1;
foreach (grams as gram) {
if (existuje gram medzi označkovanými gramami z databázy)
score = getScore(grams); // vráti ohodnotenie gramu z databázy
else
continue;
if (score > -1 AND score < 1) {
countZeros++;
```

[14]

```
modifier = 1;
continue;
}
else if (isNumeric(score)) {
sentimentScore += modifier * score;
modifier = 1;
}
else if (score == "+")
modifier *= 2;
else if ($score === "-")
modifier *= 0.5;
else if (score == "!")
modifier = -1;
else if ($score == "?") {
continue;
 modifier = 1;
}
count++;
 }
if (count == 0)
return 0;
numbersRatio = (count) / (count + countZeros);
return (sentimentScore / count) * numbersRatio;
}
```
[15]

3.3.3 Rozdielová metóda

Táto metóda pracuje s n-gramami podľa priority, kde trigram má najväčšiu váhu a unigram najmenšiu. Algoritmus prehľadáva v korpuse najprv trigrami, ktorých sentiment je kladný alebo záporný a až potom hľadá bigramy a unigrami. Pokiaľ nenájde žiadny sentimentálne zafarbený n-gram, dosadí najmenšiu neutrálnu jednotku - unigram. Po rozdelení nasleduje výpočet pozitívneho a negatívneho koeficientu. Algoritmus spočíta podiel pozitívnych a negatívnych n-gramov a ten vynásobí ich priemernou hodnotou. Nakoniec spočíta rozdiel medzi týmito koeficientmi, ktorý sa považuje za výslednú

[15]

hodnotu sentimentu. Rozdielová metóda bola navrhnutá, aby dokázala brať v úvah nejednoznačné príspevky, v ktorých sa objavujú pozitívne ale aj negatívne n-gramy. Jej hlavnou nevýhodou sú príspevky, v ktorých sa nachádza rovnomerný počet pozitívnych a negatívnych n-gramov. Túto situáciu môže algoritmus chybne považovať ako neutrálnu, pretože rozdielový koeficient sa bude blížiť k nule.

Rozdielová metóda zapísaná pomocou pseudokódu

```
private function differenceNgramsRanking(text) {
tweet = new Tweet(text);
tweetGrams = tweet->getGrams();
// funkce merge zlúči  pole do jedného
grams = merge(tweet->getGrams(), tweet->getBigrams(), tweet->getTrigrams());
grams = map(toLower(grams));
rankedGrams = getRankedGrams(grams); // vráti z databázy pole gramov
positiveScore = 0;
negativeScore = 0;
count = 0;
positiveCount = 0;
negativeCount = 0;
modifier = 1;
move = 0;
for (i = 0; i < count(tweetGrams); i++) {
count++;
score = 0;
trigram = null;
bigram = null;
gram = tweetGrams[i];
if ((i+2) < count(tweetGrams))
// funkce + zretazí
trigram = tweetGrams[i]  +  "  "  +  $tweetGrams[i+1]  +  "  "  +
$tweetGrams[i+2];
if ((i+1) < count(tweetGrams))
bigram = $tweetGrams[i]  +  "  "  +  $tweetGrams[i+1];
scoreGram = 0;
scoreBigram = 0;
```

```
scoreTrigram = 0;
if (existuje unigram medzi označkovanými gramami z databáze)
scoreGram = getScore(grams); // vráti ohodnotenie gramov  z databázy
if (existuje bigram medzi označkovanými gramami z databáze)
scoreBigram = getScore(grams); // vráti ohodnotenie gramov  z databázy
if (existuje trigram medzi označkovanými gramami z databáze)
scoreTrigram = getScore(grams); // vráti ohodnotenie gramov z databázy
if (scoreTrigram != 0) {
score = scoreTrigram;
i += 2;
} else if (scoreBigram != 0) {
score = scoreBigram;
i += 1;
} else {
score = scoreGram;
}
if (isNumeric(score) AND score > -1 AND score < 1) {
modifier = 1;
continue;
}
else if (isNumeric(score) AND score >= 1) {
positiveScore += modifier * score;
positiveCount++;
modifier = 1;
}
else if (isNumeric(score) AND score <= -1) {
negativeScore += modifier * score;
negativeCount++;
modifier = 1;
}
else if (score == "+")
modifier *= 2;
else if (score == "-")
modifier *= 0.5;
```

```
else if (score == "!")
modifier = -1;
else if (score == "?") {
modifier = 1;
count--;
continue;
}
}
if (count == 0)
return 0;

positiveWeight = positiveCount / count;
negativeWeight = negativeCount / count;
averagePositive = positiveCount > 0 ? positiveScore / positiveCount : 0;
averageNegative = negativeCount > 0 ? negativeScore / negativeCount : 0;
return positiveWeight * averagePositive + negativeWeight * averageNegative;
}
```
[16]

3.4 Problémy analýzy sentimentu

Väčšina nástrojov na monitorovanie sociálnych sietí poskytuje strojovú analýzu sentimentu. Problém nastáva pri overovaní presnosti jej výsledku. Bohužiaľ neexistuje žiadna množina dát, ktorá by bola autorizovaná a ku ktorej by bolo možné porovnávať konkrétne postupy. Jednou z príčin, prečo takýto súbor dát neexistuje je to, že s vyhodnocovaním sentimentu majú problém nie len počítače, ale aj ľudia. Predstavme si vetu " Vaša izba sa nachádza v prízemí hneď vedľa recepcie". Pokiaľ autor chcel izbu na pokojnom mieste s výhľadom na more, jeho postoj k tejto vete bude negatívny. Na druhú stranu, ak sa jedná o osobu s invalidným vozíkom, bude spokojná, že má izbu na ľahko prístupnom mieste. Pre všetkých ostatných bude táto veta iba obyčajný neutrálny fakt. Bolo dokázané, že ľudia sa nie sú schopný zhodnúť na tom, čo je pozitívne, neutrálne a čo je negatívne. Ďalším faktom je to, že lepšie sa detekuje extrémna negativita ako drobná pozitivita. Okrem toho, že vyhodnotenie sentimentu je zložité aj pre ľudí, existujú aj ďalšie problémy ako sú kontext, negácia, irónia, ktoré robia vyhodnotenie sentimentu textu

16

extrémne zložité. Čo sa týka sociálnych sietí a Twitteru platí jednoduchá rovnica - čím kratší text, tím zložitejšia analýza.

3.5 Twitter

Analýza sentimentu na Twitteri sa značne odlišuje od iných štúdií sentimentu. Príspevky na sociálnej sieti Twitter môžu mať maximálne 140 znakom, ale často sú oveľa kratšie a obsahujú na viac hypertextové odkazy, ktoré sú pre detekcie sentimentu irelevantné. Z dôvodu limitu znakov sú často používané skratky alebo slangový jazyk. Navyše jazyk, ktorý sa používa na sociálnych sieťach sa odlišuje od tradičných textov. Príkladom môžu byť slová ako "lol", "rofl", "wtf", emotikony alebo špeciálne značky pre Twiiter ako "RT", "#" či "@". Samotné slová môžu byť hodnotené ako negatívne alebo pozitívne, ale kontext ktorý vytvárajú môže byť úplne iný.[17]

3.5.1 Kontext a sarkazmus na Twitteri

Ďalším častým problémom je kontext, ktorý plní dôležitú úlohu a môže často meniť význam celého textu. Súčasné metódy analýzy sentimentu sú statické a neberú do úvahy kontext, ktorý je pre ľudský intelekt úplne prirodzený ale počítaču spôsobuje obrovské problémy. Ako príklad si môžeme uviesť vetu " Moji chudáci z Realu znova prehrali s Barcou 3:5. Do toho Real!" Detekované boli dve silno negatívne slová "chudáci" a "prehrali". Tento príspevok je však používaní v inom kontexte. Autor ľutuje Real Madrid a v závere im fandí. Kontext tiež spôsobuje problémy u fráz, ktoré majú špecifický význam a súčasne synonymne popisujú nejakú vec. Príkladom je napríklad veta: "To je bomba!" Autor mohol mať na mysli, že je niečo naozaj skvelé - pozitívny sentiment. Na druhú stranu mohol tiež popisovať nejaký druh bomby, napr. vodíkovú bombu. Čo sa týka sarkazmu, ako u každého typu spracovania prirodzeného jazyka záleží na kontexte. Analyzovanie prirodzeného jazyka je veľmi komplikovaný problém a sarkazmus a iné druhy ironického jazyka sú neodmysliteľne problematické pri detekcii počítačom, pokiaľ k nim pristupujeme pomocou izolovaných slov. Príkladom môže byť nasledujúca veta: " To je teda naozaj úžasné. " Slovo "úžasný" je veľmi pozitívne a na viac je zosílené slovom "naozaj", ale pokiaľ by sa jednalo o sarkasticky myslenú vetu, tak by mala byť táto veta myslená ako negatívna. Aby sme mohli presnejšie detekovať podobné javy, potrebovali by

17

sme poznať profil autora a vedieť, ako často používa sarkazmus alebo iróniu. Posledným z kontextuálnych problémov sú viacznačné slová. Niektoré slová môžu mať dva a viac významov, pričom jeden význam môže znamenať pri značkovaní určitú mieru sentimentu a ďalší môže označovať špeciálnu značku. Príkladom na takéto slovo môže byť slovo "strašne". Toto slovo môže byť považované pri značkovaní ako zosílené alebo ako slovo s negatívnym sentimentom. Problém potom nastáva pri vyhodnocovaní algoritmov, ktoré do svojho výpočtu dosadzujú len tú hodnotu, ktorá je väčšinová.

3.5.2 Cudzojazyčné slová

Ako už bolo zmienené , na "slovenskom Twitteri" užívatelia často používajú anglické názvy. Väčšinou sa jedná o klasický počítačový žargón alebo slová bežne používané na internete. Vo väčšine prípadov ich môžeme považovať za neutrálne, ale objavujú sa aj výnimky, ktoré majú veľmi silný sentiment. Príkladom môže byť najpopulárnejšia sémantická značka "#fail", ktorá je veľmi negatívna a preto ju nemôžeme ignorovať ako ostatné cudzojazyčné slová.

3.5.3 Negácia

Používanie negácie sa v slovenskom jazyku riadi určitými pravidlami a pri analýze sentimentu prináša nemalé problémy. Zápory rozlišujeme podľa toho, či chceme negovať obsah celej vety, jej časti alebo len slova. Rozlišujeme tri kategórie : 1. vetný zápor - popiera obsah celej vety " Čakal som tam, ale zase nikto neprišiel". 2. členský zápor - popiera platnosť jedného vetného člena "Nie mne, ale mame to hovor. " 3. slovný zápor - popiera význam samotného slova " Dnešný obed bol naozaj nechutný." Niektoré druhy negácií dokážeme úspešne vyhodnotiť, ale problém môže nastať pri vete : " To naozaj nie je zlé. " Veta obsahuje dve negatívne slová, ktoré sú v skutočnosti súčasťou pozitívnej vety.

3.6 Nástroje pre analýzu sentimentu

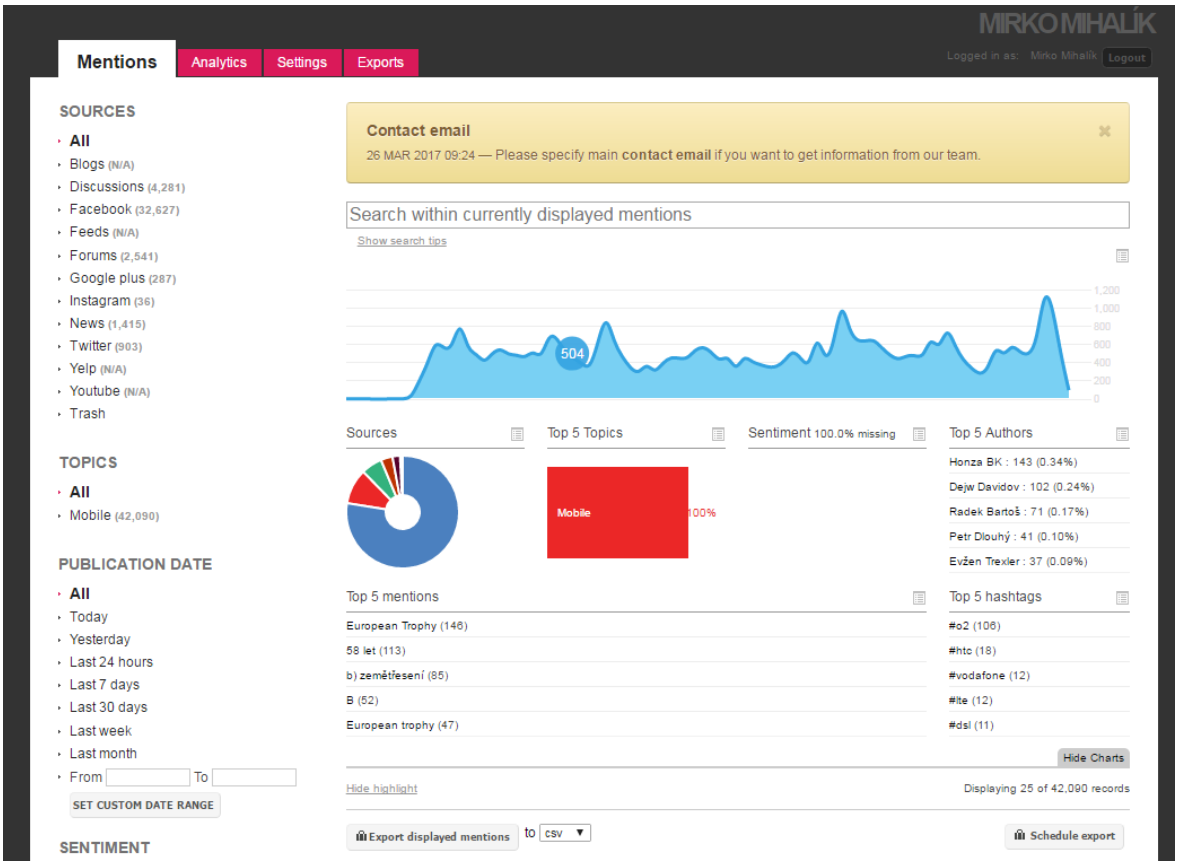
Veľmi jednoduchý systém pre vyhľadávanie sentimentu má v sebe zbudovaná sociálna sieť Twitter a konkrétnejšie jeho model vyhľadávanie. V rozšírenom vyhľadávaní priamo ponúka funkcia vyhľadávanie pozitívnych alebo negatívnych príspevkov. Táto funkcia je založená na tom, či príspevok obsahuje emotikon, čo sa nedá nazvať ako nástroj pre analýzu sentimentu. Jedným známym nástrojom, ktorý dokáže analyzovať sentiment v

slovenčine je produkt Ataxo Social Insider. V ďalšej časti tejto kapitoly sú popísané nástroje pre analýzu sentimentu, ktoré sú v súčasnosti považované za najefektívnejšie.

3.6.1 Ataxo Social Insider

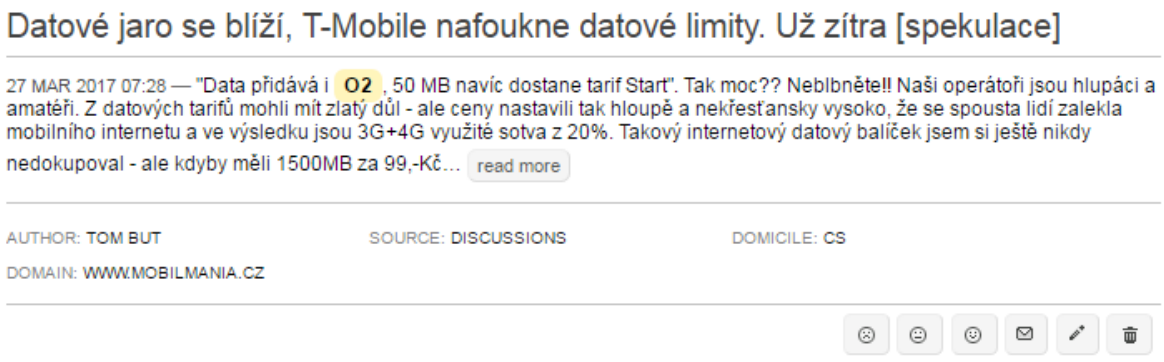
Na slovenskom webe sú dve väčšie firmy s nástrojmi pre monitoring a analýzu sociáoknych sietí zameriavajúce sa na slovenské prostredie. Tím prvým je nástroj Social Insider od spoločnosti Ataxo Interative (Social Insider, 2014), ktorý je plne lokalizovaný pre Slovenskú a Českú republiku.[18] Ide o komerčný nástroj ale užívateľom môže byť poskytnutá trial verzia na podobu 14 dní. Je potrebné sa prihlásiť pomocou Facebooku alebo Twitteru a odoslať žiadosť spoločnosti na sprístupnenie trial verzie. Ataxo Social Insider (ďalej len ASI) dokáže prehľadávať v blogoch, diskusných fórach, sociálnych sieťach a aj na spravodajských serveroch. ASI pre vyhľadávanie zadaných dotazov používa vyhľadávač, RSS aj crawler. Vyhľadávanie prebieha na základe zvolených kľúčových slov (max.500) alebo podľa zvolených tém. Ďalší porovnateľný komerčný nástroj Monitoring sociálnych sietí od spoločnosti Newton Media, a.s nemá v skúšobnej verzii sprístupnené vyhľadávanie podľa ľubovoľných kľúčových slov, ale sú len k dispozícií iba vybrané témy týkajúce sa finančného sektora (produkty, banky, poisťovne), tak sú aj v ASI vyhľadávané podobne témy podľa kľúčových slov : banka, pôžička, úver. V pokročilom vyhľadávaní je možné pri kľúčových slovách využiť logické operátori. ASI má pomerne prepracované filtrovanie a kategorizáciu výsledkov vyhľadávania. Vyhľadávané výsledky je možné členiť podľa zdroja (blogy, diskusie, Facebook, fóra, Google +, Instagram, Twitter, Yelp, Youtube a spravodajské servery), podľa kľúčových slov (aj v rámci témy), dát publikácie (v závislosti na zdroji), sentimentu (pozitívny, negatívny, neutrálny a neoznačený), a v neposlednej rade podľa "bydliska" (či je príspevok z Českej alebo Slovenskej republiky).

18



Obrázok 1: Hlavná stránka nástroja ASI

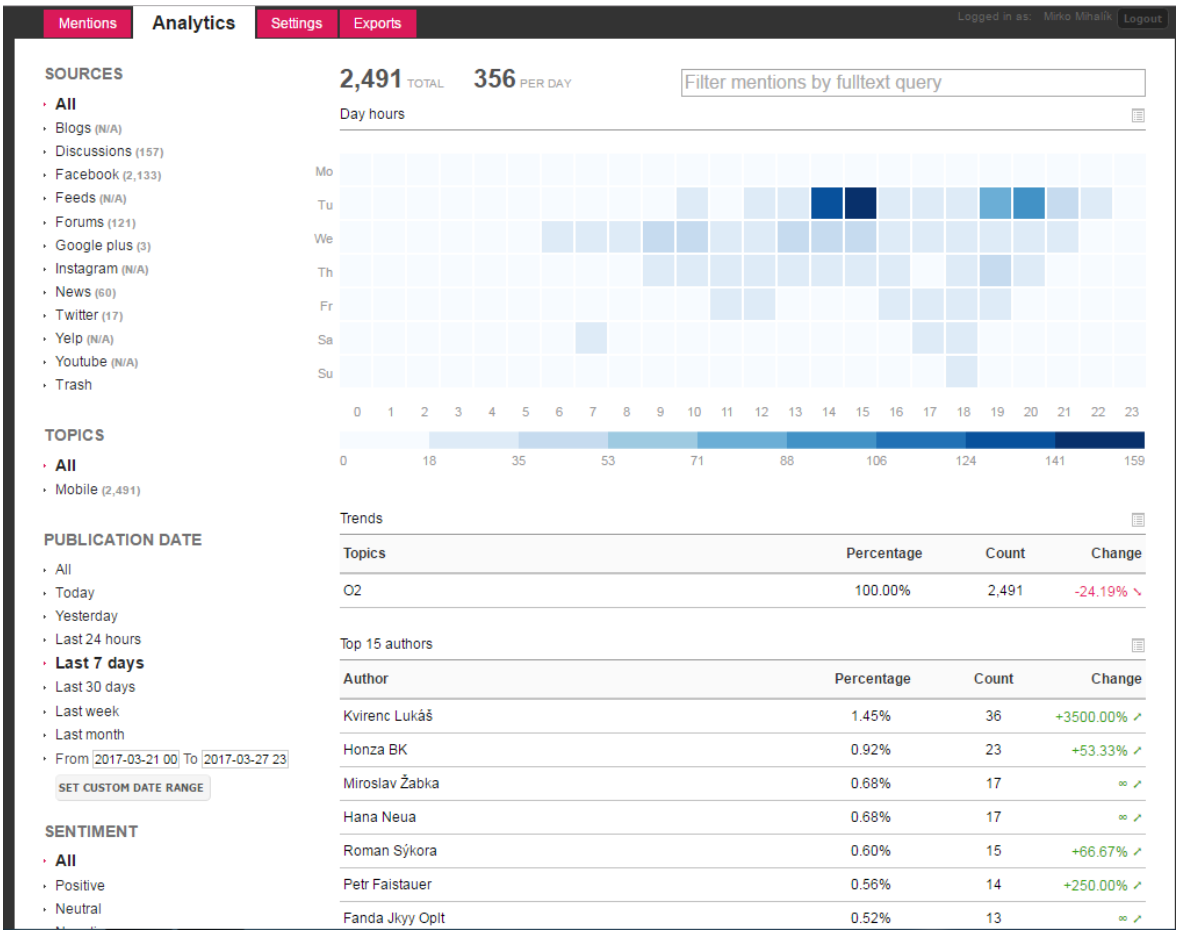
Vyhodnocovanie sentimentu nie je automatické. Užívateľ musí zvolený príspevok ručne ohodnotiť. Množstvo ohodnotených príspevkov sa potom sumarizuje a podľa príslušnej kategórie a je zobrazené v prstencovom grafe.



Obrázok 2 : Vyhľadaný príspevok v prostredí ASI

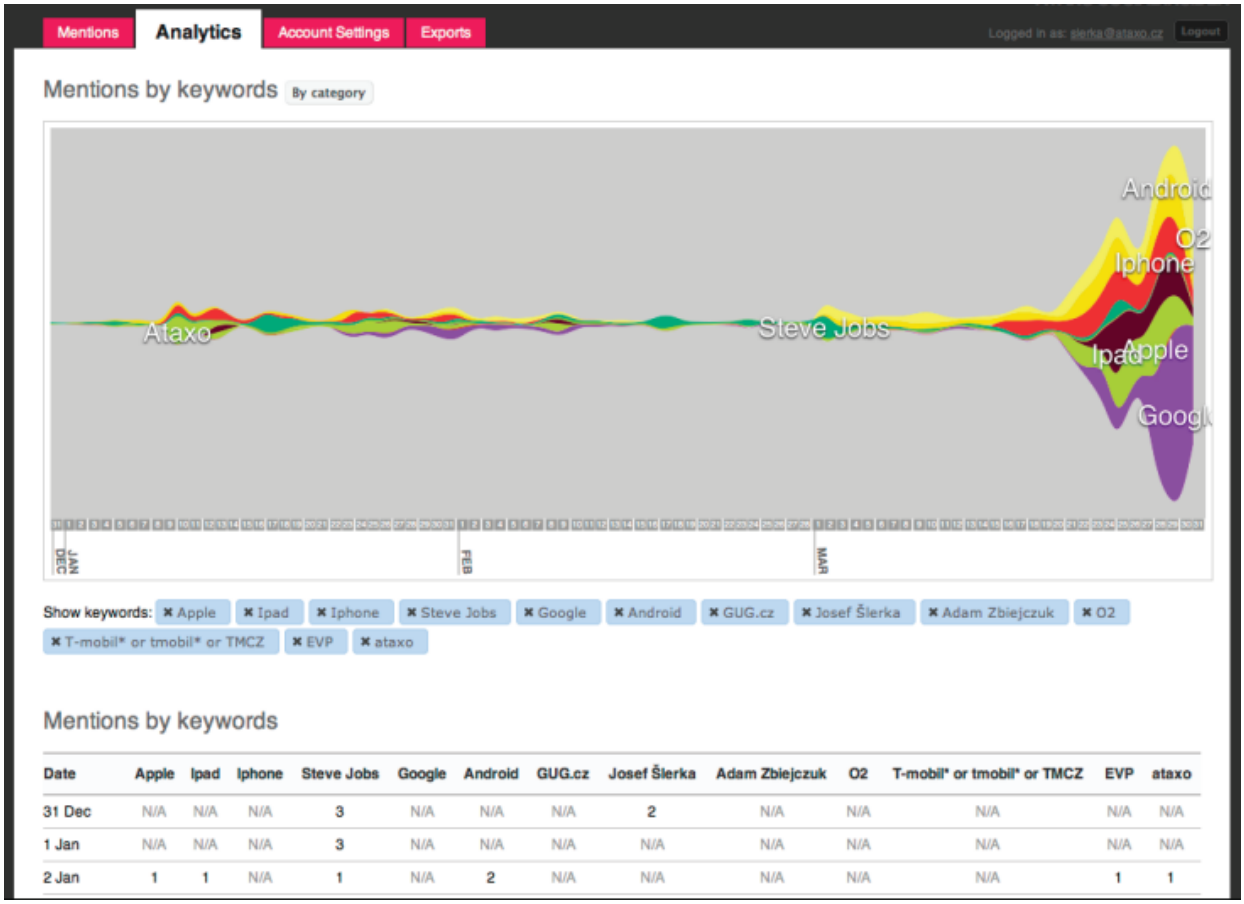
Pre ilustráciu, ako sa dá s vyhľadaným príspevkom v aplikácií pracovať, je na obrázku 2 zobrazený príspevok vyhľadávaný podľa kľúčového slova " O2", slovo je zvýraznené. Každý príspevok obsahuje dátum, kedy bol publikovaný, vlastný text, meno autora, zdroj, bydlisko a v prípade Twitteru ešte počet "followers", teda tých ktorý užívateľa sledujú a počet "friends", tých, ktorých užívateľ sleduje. Ďalej je tam zobrazená ešte hodnota "klout", ktorá symbolizuje jeho vplyv na sociálnych sieťach. V pravom dolnom rohu sú tlačidla na označenie sentimentu, na odoslanie príspevku e-mailom, na zapísanie poznámky k príspevku a nakoniec sa tam nachádza tlačidlo na odstránenie príspevku z výberu. Export dát je možný hneď do štyroch formátov (CSV,XLXS, HTML a DOCX). Exportuje sa všetko, čo je vyhľadané podľa navolených kritérií. Data potom obsahujú 12 kategórií, ako napríklad obsah príspevku, autora, dátum, odkaz na zdroj, označený sentiment, kľúčové slovo, a ďalšie údaje z Twitteru. Export je možný naplánovať denne, týždenne alebo mesačne.

http://klout.com/home



Obrázok 3 : Analytika v prostredí ASI (trial verzia)

V trial verzii je analytický nastroj avšak nie je nejako bohatá. Na obrázku 4 je porovnaný graf z platenej verzie, ktorá už ponúka vizuálne pútavejšie výstupy. Na hlavnej záložke Mentions je zobrazený iba vývoj počtu príspevkov v čase a dve prstencové grafy znázorňujú podiel vyhľadávaných zdrojov a podiel označeného sentimentu (obe v %). Ďalšia záložka Analytics už je niečo zaujímavejšia. Na obrázku 3 je zachytený počet príspevkov na Twitteri za posledných 7 dní podľa kľúčového slova "O2" . Čím tmavšie políčko, tím viac príspevkov, ako je vidno na modrej škále pod grafom. Z grafu je možné vyčítať, že utorok medzi 14. a 15. hodinou bolo napísaných najviac príspevkov týždeň tj. 159. Pod grafom sú ďalšie prehľady ale tie už nie sú grafické. Ako prvé sú trendy, teda aký bol vývoj príspevkov s kľúčovým slovom v danom týždni oproti minulému. Ďalej sú tam jednoduché štatistiky najviac čítaných príspevkov, ako napríklad rebríček 15 autorov, ktorý najviac prispievali, 15 najčastejších "hastagov" vo vybraných príspevkoch, najčastejšie písane emotikony a najčastejšie spomínané odkazy a domény v príspevkoch. Ako už bolo zmienené, ASI je lokalizovaný pre Českú a Slovenskú republiku. Pri hľadaní kladie dôraz aj na skloňovanie slov. Avšak so slovenskými slovami má značný problém preto je lepšie využívať české slová.



Mentions by keywords

Date	Apple	Ipad	Iphone	Steve Jobs	Google	Android	GUG.cz	Josef Šlerka	Adam Zbiejczuk	O2	T-mobil* or tmobil* or TMCZ	EVP	ataxo
31 Dec	N/A	N/A	N/A	3	N/A	N/A	N/A	2	N/A	N/A	N/A	N/A	N/A
1 Jan	N/A	N/A	N/A	3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2 Jan	1	1	N/A	1	N/A	2	N/A	N/A	N/A	N/A	N/A	1	1

Obrázok 5: Analytika v prostredí ASI (plná verzia)

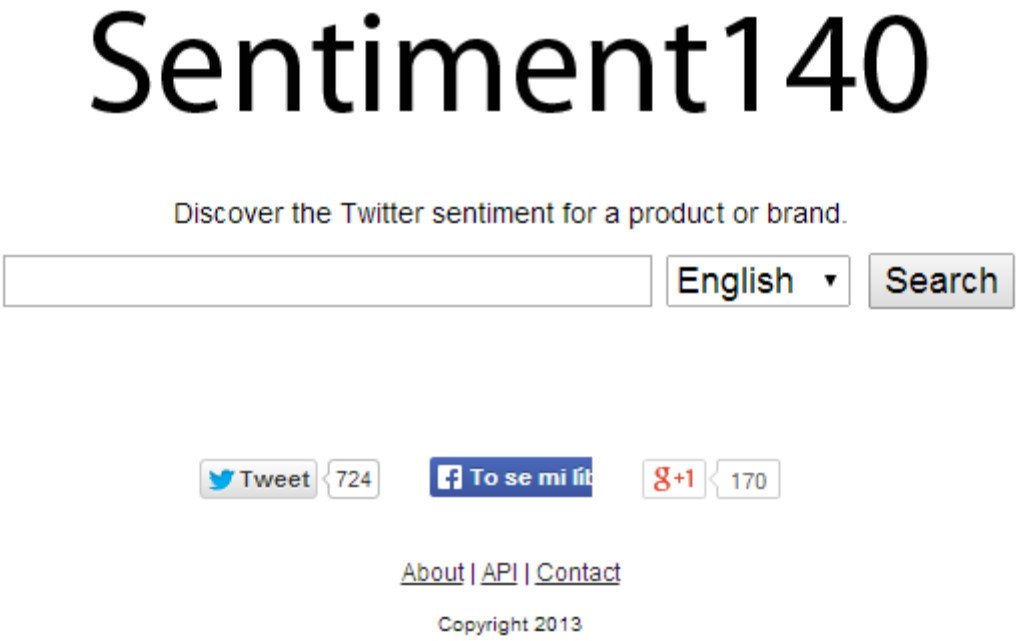
3.6.2 Sentiment 140

Sentiment140 je na webe voľne prístupný nástroj na analýzu sentimentu. Ide o akademický projekt troch študentov počítačových vied zo Standfordskej univerzity z roku 2009. Pomocou nástroja Sentiment140, predtým známeho ako Twitter Sentiment, je možné objaviť sentiment značiek , produktov a tém iba na sociálnej sieti Twitter. Klasifikátor funguje na princípoch strojového učenia, konkrétne využíva metódu klasifikácie maximálnej entropie. Klasifikátor bol naučený na dátach, ktoré boli stiahnuté z Twitter API na základe emotikonov, za predpokladu, že príspevky obsahujúce :) sú pozitívne a príspevky obsahujúce :(sú naopak negatívne. Klasifikátor dosahuje viac ako 80 % úspešnosť. [19]

Nástroj klasifikuje príspevky písane iba v anglickom a španielskom jazyku, takže pre slovenčinu nie je vhodný. Avšak podnik ho môže využiť pre vlastnú analýzu, pokiaľ tweetuje v anglickom jazyku. Tento nástroj je zameraný na zahraničný trh alebo aj ako monitoring konkurencie v zahraničí.

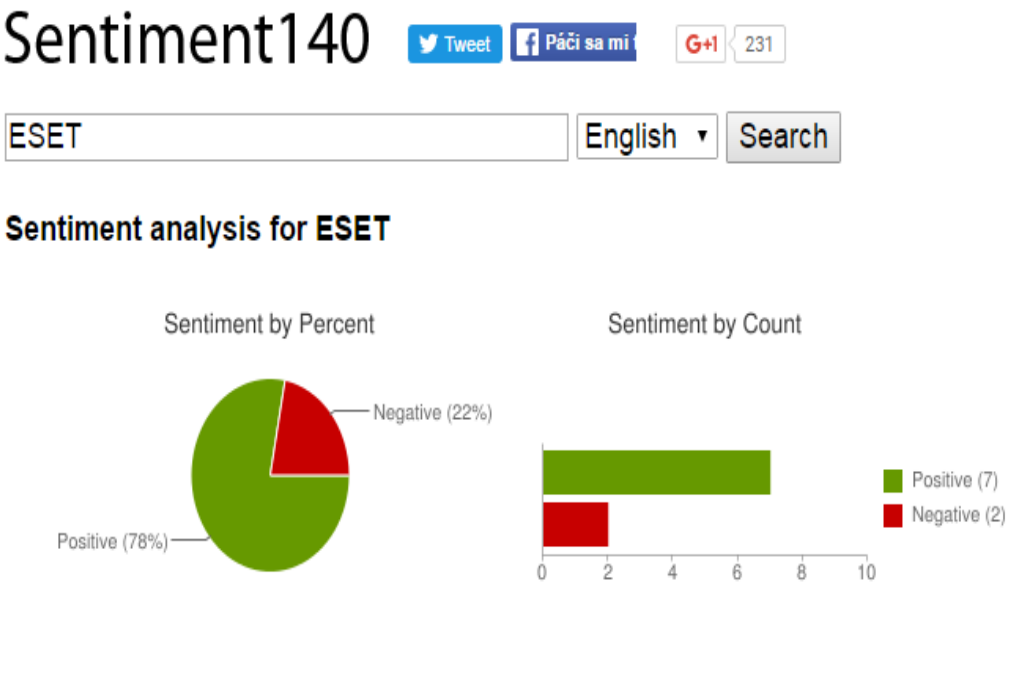
Užívateľské rozhranie webovej aplikácie je veľmi jednoduché. Na obrázku 6 je hlavná stránka, ktorá sa načíta pri autorizovaní aplikácie cez vlastný účet na Twitteri. Stačí zadať ľubovoľné slovo čí slovné spojenie, vybrať jazyk (angličtina alebo španielčina) a vyhľadať.

19



Obrázok 6:Úvodná stránka Sentiment140

Výsledky vyhľadávaní sú tak isto jednoduché a prehľadné. Je otázkou, či je to skôr výhodou (rýchla a jednoduchá orientácia), alebo nevýhodou (napr. obmedzenia výsledkov a manipulácia s nimi). Pri vyhľadávaní slovenskej firmy ESET sa zobrazia pod vyhľadávacím poľom najprv dve jednoduché grafy ktoré porovnávajú pozitívny a negatívny sentiment. Obrázok číslo 7. Pod grafmi sú už vyhľadané a klasifikované tweety.(Obrázok číslo 8).Zelené pozadie tweetu znázorňuje pozitívny sentiment, červené pozadie znázorňuje negatívny sentiment a biely negatívny sentiment. Iné výsledky, alebo aspoň manipulácia s tými ktoré sme už získali nie sú k dispozícií. Výsledných tweetov nástroj negeneruje veľa. Rádovo sú ich maximálne desiatky, takže pre užívateľa je nutné kontrolovať výsledky opakovane. Sentiment140 tiež nezobrazuje historické výsledky, ktoré siahajú viac do minulosti, maximálne niekoľko dní späť. Je to zásluhou obmedzenia Twitter API. Aj napriek svojím obmedzeným výsledkom a manipuláciou s nimi , Sentiment140 je rýchly a prehľadný nástroj pre analýzu sentimentu, ktorý je bezplatný a úspešný. Na viac autori na stránkach priznávajú, že oblasť analýzy sentimentu má ešte veľa nevyriešených problémov a že sa budú niektorými záležitosťami zaoberať a nástroje tak vylepšovať. Napríklad vyriešenie problémov s negáciou, vylepšenie parseru a aby si lepšie poradili s neformálnym štýlom jazyka.

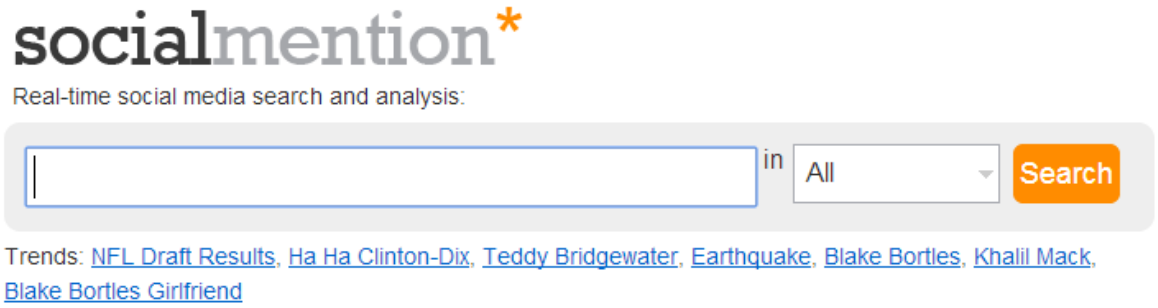


Obrázok 7: Jednoduchá analytika Sentimentu140

Obrázok 8: Vygenerované a klasifikované tweety v Sentiment140

3.6.3 Social Mention

Social Mention je ďalšia jednoduchá a bezplatná webová aplikácia pre monitorovanie a analýzu sociálnych médií, zameraná na blogy, komentáre, udalosti, správy, videa a sociálne siete. Je vhodnou a nenáročnou alternatívou pre monitorovanie značky a analýzu sentimentu v reálnom čase.[20]



Obrázok 9: Úvodná stránka Social Mention

Pri hľadaní je možné si vybrať v akých zdrojoch vyhľadávať, ale nie je možné vybrať konkrétny zdroj, ale vždy len celú skupinu zdrojov, akže v skupine microblogs je napríklad Twitter aj FriendFeed. Konkretizovať dotaz nielen podľa zdroja je možné až na hlavnej stránke výsledku oproti Sentimentu140. Na obrázku 10 v strede je súvislé vlákno nájdených príspevkov zoradených podľa dátumu publikácie od najnovšieho. Rozkliknutím daného príspevku sa je možné dostať k zdroju na stránke Twitter.

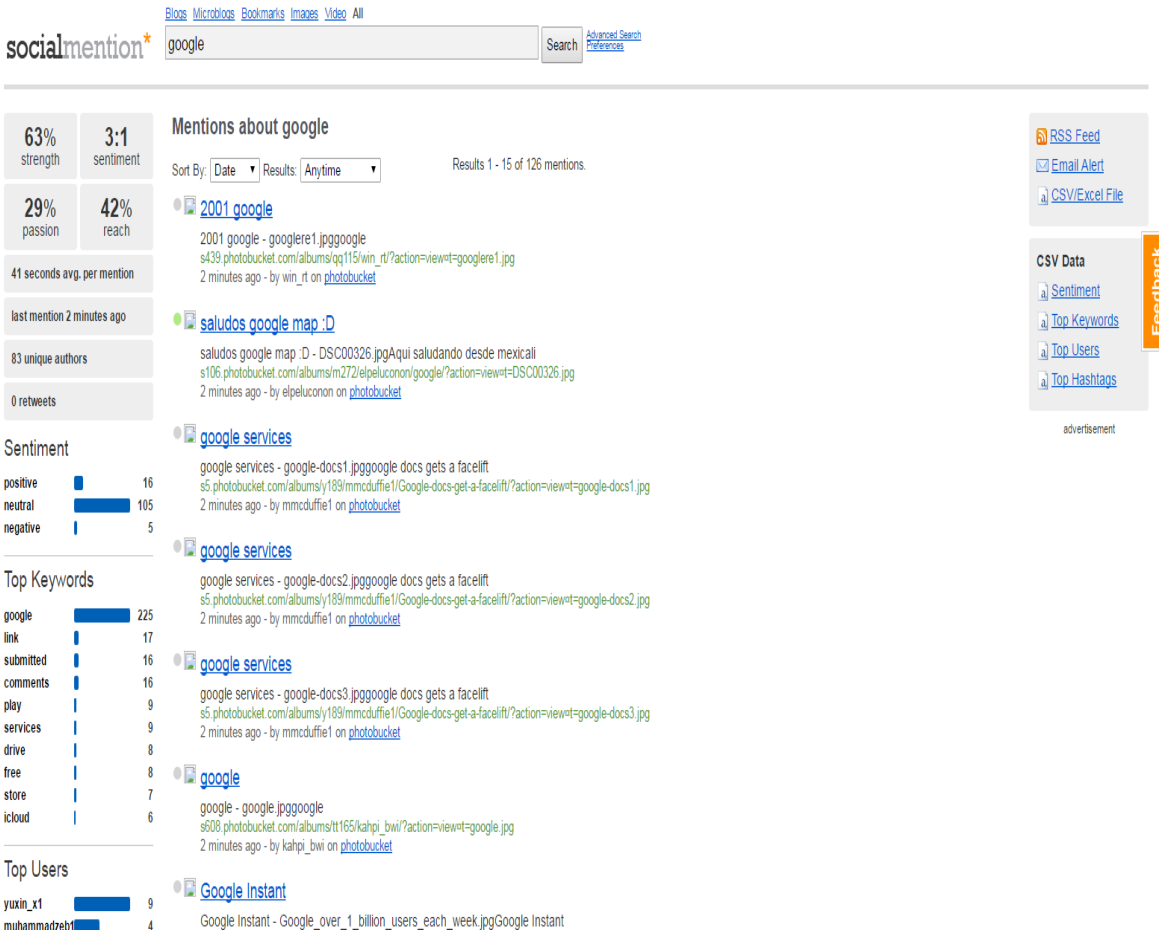
Social Mention ponúka tieto štyri hlavné metriky:

- Strength (pravedpodobnosť, že ľudia diskutujú o hľadanom termíne)
- Sentiment (pomer pozitívnych príspevkov k negatívnym)
- Passion (ukazovateľ toho, ako často rovnaký ľudia diskutujú o hľadanom termíne, naopak čím viac jedinečných autorov ktorý spomínajú hľadaný termín, tím nižšie %)
- Reach (pomer počtu jedinečných autorov voči celkovému počtu príspevkov)

Ďalej Social Mention ponúka prehľad o priemernom čase pridávania príspevkov, počet jedinečných autorov a reetweetov. Mimo iné nástroj zobrazuje stĺpcové grafy

20

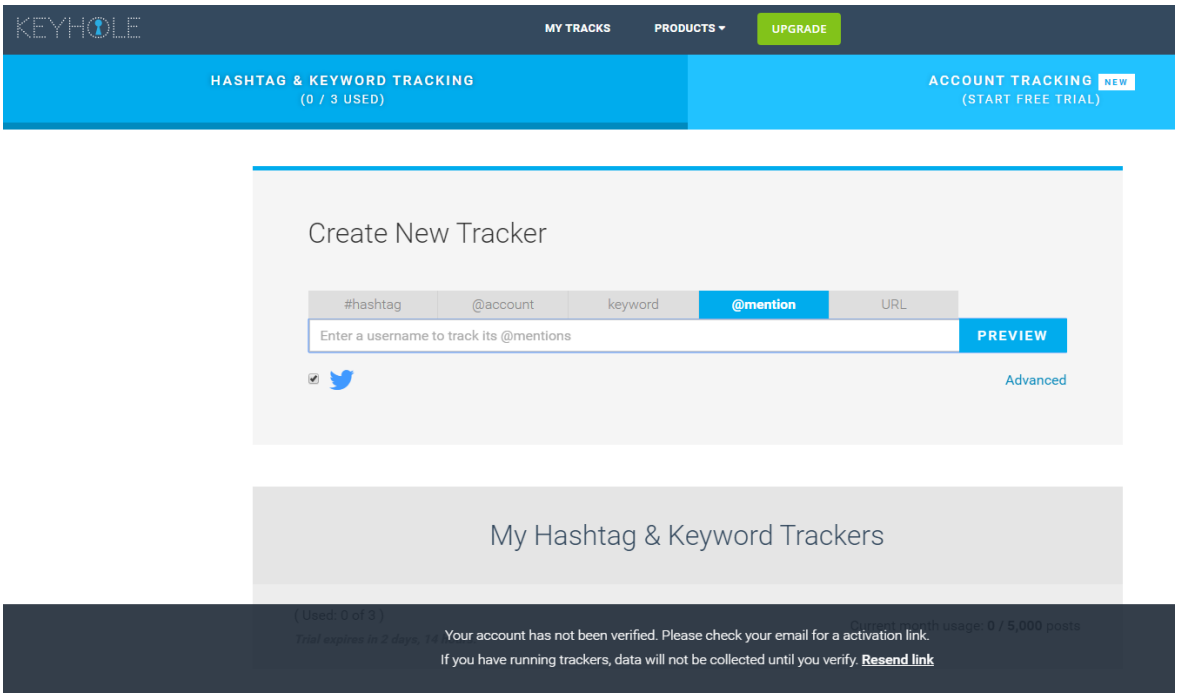
sumarizujúce sentiment príspevkov a "Top Keywords", čo je pomerne zaujímavý ukazovateľ slov ktoré sú najčastejšie používané vo vzťahu k hľadanému slovu. Nechýba ani prehľad užívateľov, ktorý najčastejšie píšu o danej téme , a ani prehľad najviac používaných "hashtagov" ktoré sa vzťahujú k vyhľadávanému slovu. Social Mention ponúka pokročilé vyhľadávanie, teda špecifikáciu vyhľadávaného termínu. Užívateľ aplikácie môže vyhľadať nie len , ktoré obsahujú všetky dané slová alebo presné slovné frázy, ale môže aj filtrovať slová ktoré sú vo výsledku nežiadané. Vyčleniť z výsledkov je možné dokonca aj autorov. Zaujímavosťou je aj vyhľadávanie podľa jazyka, Social Mention ich podporuje hneď 43, vrátane slovenčiny. Avšak nejde o veľmi podarenú pomôcku, nakoľko vyhľadávanie podľa jazyka slova "Ukrajina" je krkolomné a konkrétne v slovenčine aplikácia motá niekoľko jazykov dokopy (napríklad ruštinu, poľštinu, češtinu). Z toho je možné vyvodiť, že nástroj filtruje príspevky prevažne podľa zadaných termínov, ktoré sa objavujú v príspevkoch. Ale aj pri samotnom pokročilom vyhľadávaní nie je ojedinelé opakovanie všetkých úkonov, nakoľko aplikácia čas od času užívateľa po nastavení kritérií vráti spať na úvodnú stránku a je potrebné všetko robiť od začiatku.



Obrázok 10: Výsledky hľadania Social Mention

3.6.4 Keyhole

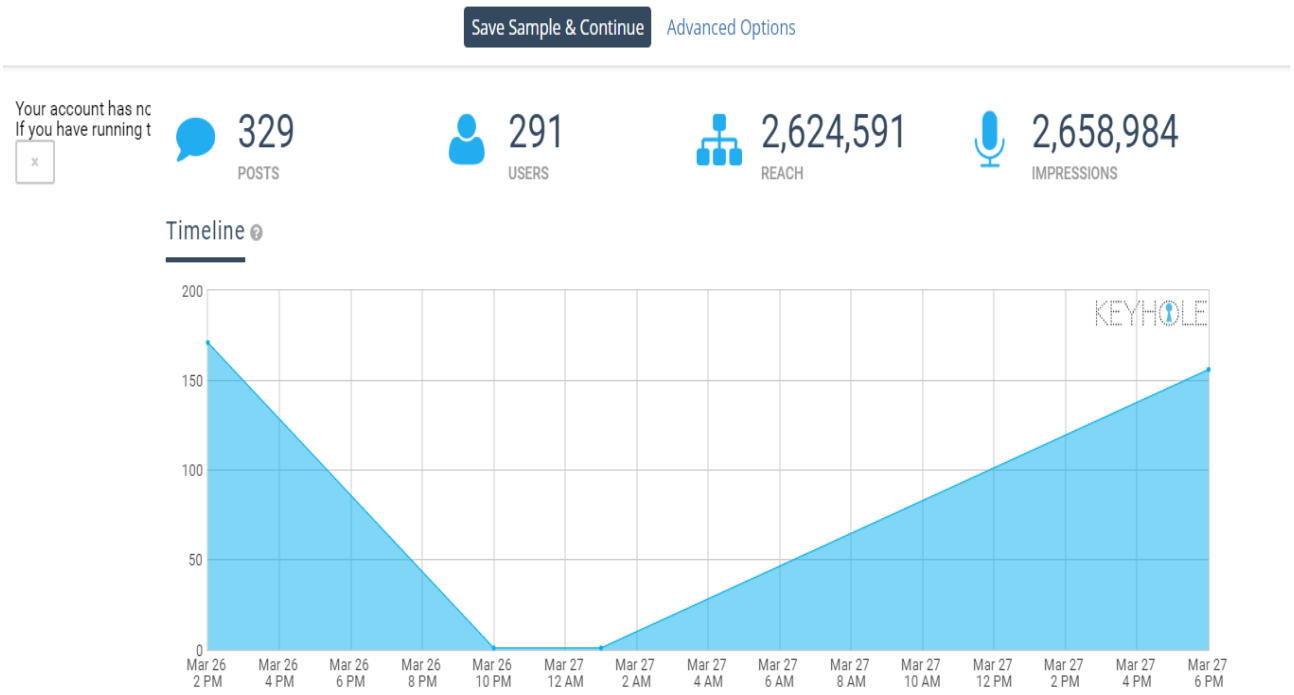
Posledným nástrojom pre analýzu sentimentu je aplikácia KeyHole. Demo verzia ponúka širokú škálu možnosti pre analýzu sentimentu. Na úvodnej stránke sa zobrazí menu pomocou ktorého je možné vykonávať analýzu pomocou hashtagov, účtov, kľúčových slov, zmienky alebo podľa URL adresy.[21]



Obrázok 11: Úvodné menu aplikácie Keyhole

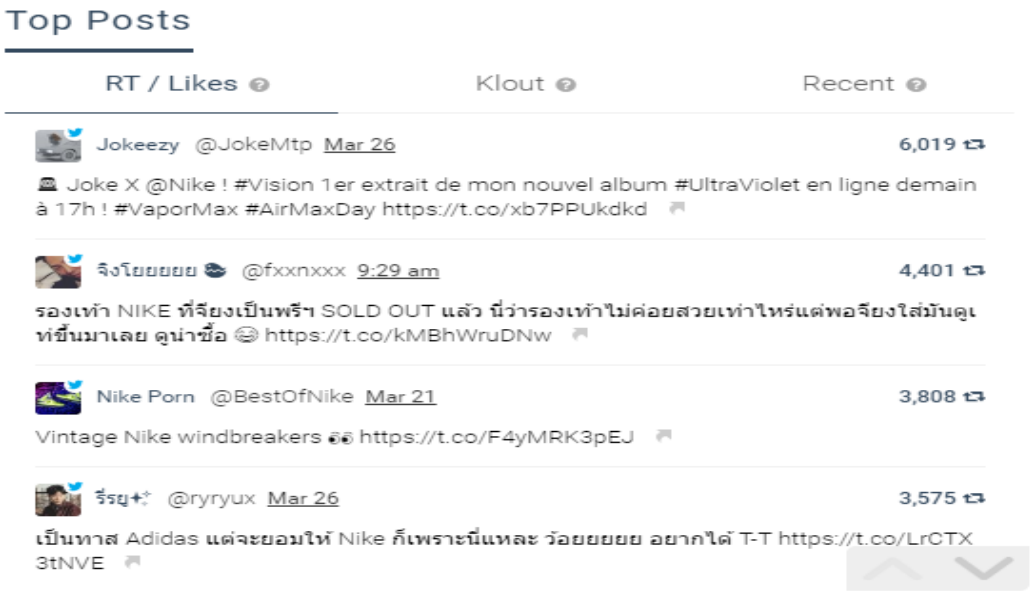
Po zadaní kľúčového slova "Nike" prebehne analýza sentimentu na sociálnej sieti Twitter. Výsledná obrazovka nám ponúka široké množstvo grafov. Na obrázku číslo 12 je graf ktorý nám hovorí o tom, o ktorej hodine bolo najviac príspevkov kde bolo kľúčové slovo "Nike". Podľa toho je zrejmé že najviac príspevkov bolo 26 marca o 14 hodine. Tak isto môžeme vidieť koľko príspevkov a používateľov použilo nami vybrané kľúčové slovo. V demo verzii nám aplikácia ponúka analýzu len za posledných 24 hodín. A za posledných 24 hodín je zrejme že až 291 používateľov sociálnej siete Twitter použilo nami vybrané kľúčové slovo.

21



Obrázok č.12: Časové rozhranie poslaných príspevkov

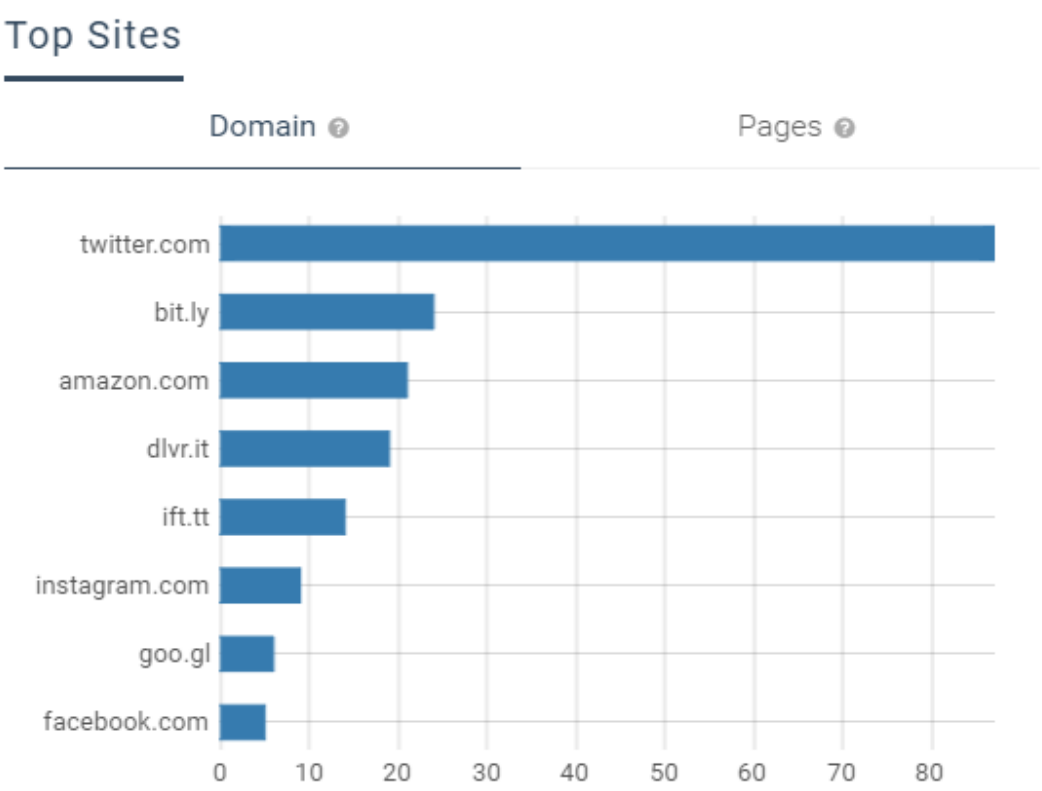
Ako ďalší výstup nám Keyhole na obrázku číslo 13 ponúka výstup kde sú znázornené príspevky s najväčším ohodnotením. To znamená príspevky užívateľov, ktoré boli najviac krát videné a "reetweetnuté" Ďalším užívateľom. Ako ďalšiu funkciu nám ponúka znázornenie tém, ktoré s daným kľúčovým slovom súvisia. Tieto súvisiace témy sú znázornené na obrázku číslo 14.



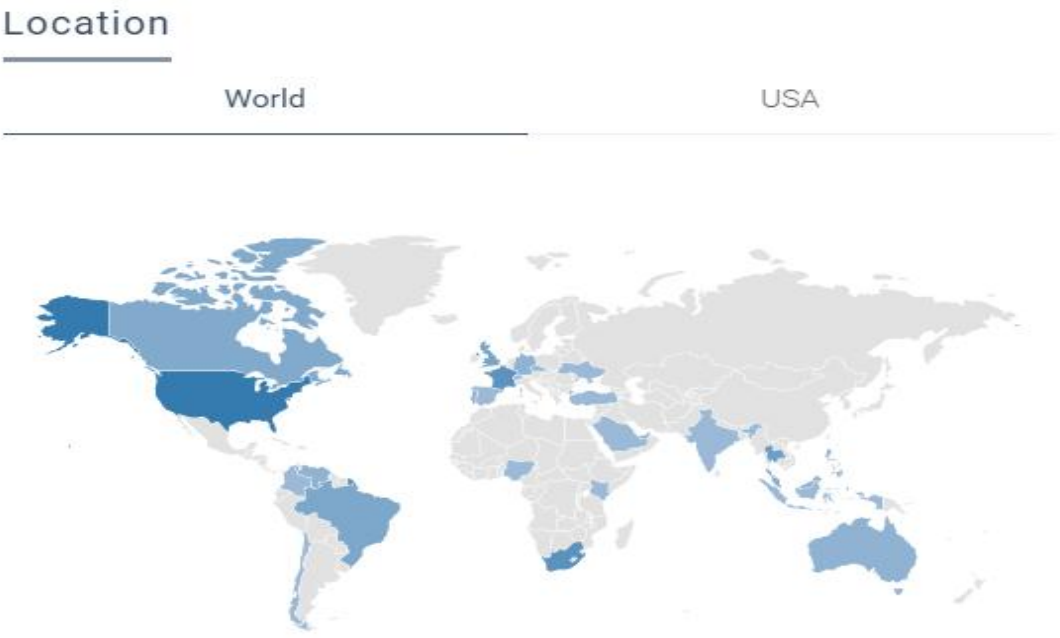
Obrázok č.13: Najlepšie komentáre podľa ohodnotenia

Obrázok č.14: Znázornenie tém súvisiacich s kľúčovým slovom

Nasledujúci výstup na obrázku číslo 15, ktorý nám aplikácia ponúka, znázorňuje sociálne siete kde bolo najviac príspevkov súvisiacich s kľúčovým slovom "Nike". Z grafu je zrejmé, že najviac príspevkov pochádza zo sociálnej siete Twitter. Ďalej ho nasledujú bit.ly, amazon.com, dlvr.it, ift.tt, instagram.com, goo.gl a facebook.com. Na sociálnej sieti Twitter bolo viac ako 80 príspevkov zo všetkých 329 za posledných 24 hodín. Na obrázku číslo 16 je znázornené, z ktorých krajín najviac príspevkov pochádza. Podľa danej mapy je jasné, že najviac príspevkov pochádza zo Spojených štátov amerických.

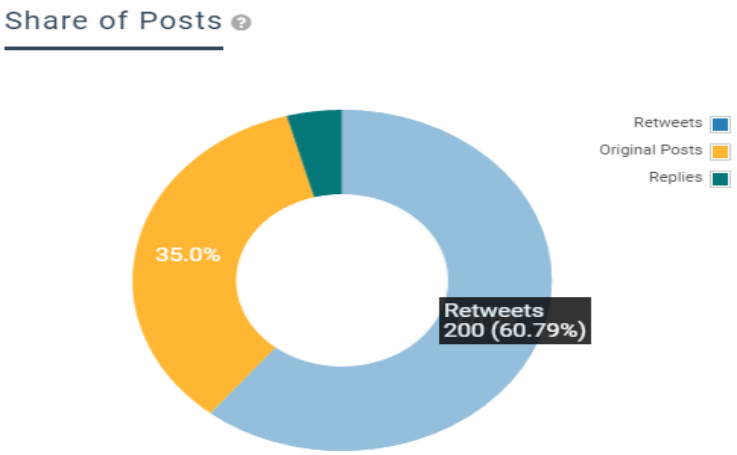


Obrázok č.15 : Zoznam sociálnych sietí s najväčším počtom príspevkov

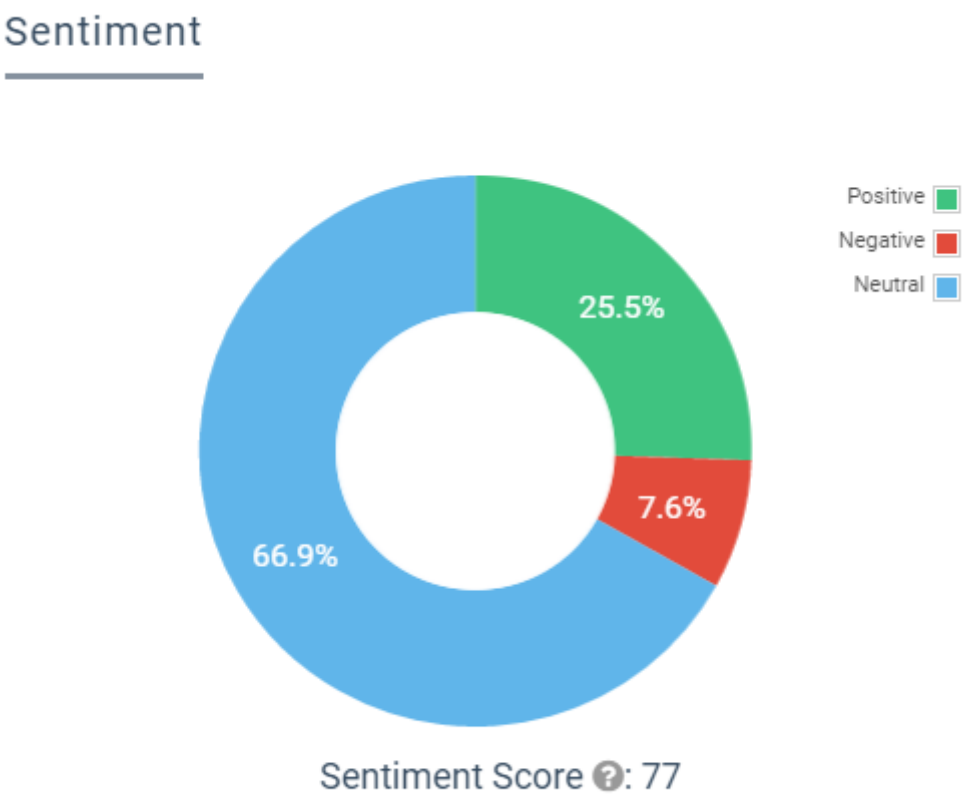


Obrázok č16 : Mapa s najväčším počtom príspevkov

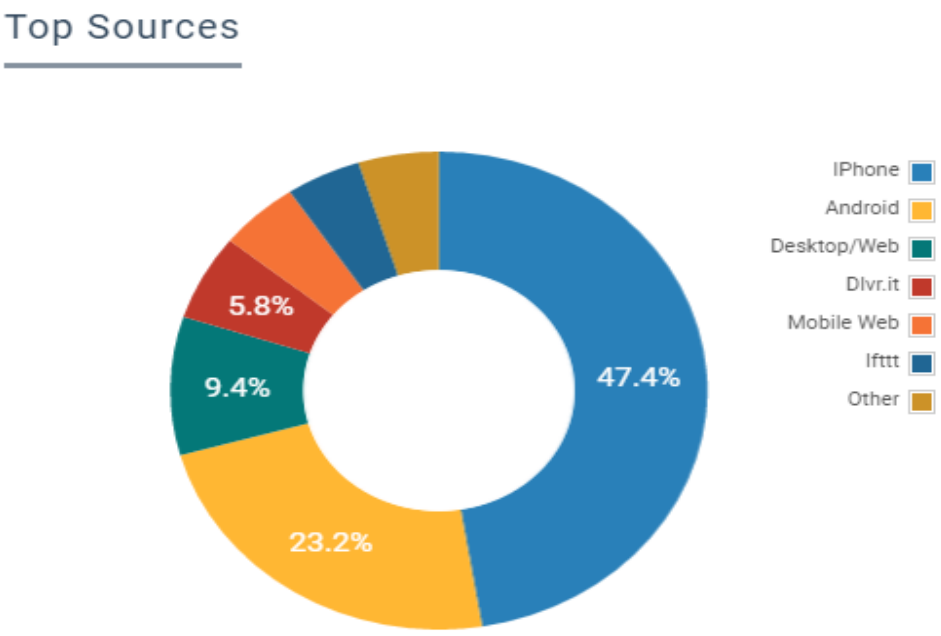
V ďalšom výstupe nám aplikácia ponúka nasledujúce štyri grafy. V prvom grafe na obrázku číslo 17 s názvom " Share of Posts" je vidieť, percentuálny podiel koľko tzv. "tweetov" bolo originálnych, koľko ich bolo "reetwetnutých" a koľko bolo odpovedí na tieto príspevky. Z toho grafu je zrejmé, že originálne príspevky zo všetkých príspevkov tvoria 35 % zo všetkých analyzovaných príspevkov. "Reetwetnutých" príspevkov, čiže príspevkov ktoré boli preposlané znova iným užívateľom bolo 60,79 %. Odpovedí na príspevky s kľúčovým slovom bolo len 4,21%. V druhom grafe je znázornený sentiment všetkých vybraných príspevkov. Z tohto grafu je zrejmé, že sentiment používateľov siete "Twitter" k danému kľúčovému slovu je rozmanitý. Podľa grafu na obrázku číslo 18 je zrejmé, že 25,5 % všetkých príspevkov je pozitívne, 7,6 % príspevkov je negatívnych a 66,9 % príspevkov je neutrálnych. Ak neberieme do úvahy neutrálne príspevky tak môžeme povedať, že sentiment používateľov ku vybranému kľúčovému slovu je pozitívny. Na obrázku číslo 19 je výstupný graf, ktorý nám ukazuje, aký prístroj bol použitý na reakciu na kľúčové slovo. Z tohto grafu je možné vyčítať, že najväčší podiel používateľov, ktorý reagujú na "tweeteri" používa zariadenie iphone. Z celkového počtu je to až 47,4 % zo všetkých prístrojov ktoré sa používajú na reakciu na sociálnej sieti "twitter". Druhým prístrojom boli zariadenia ktoré používajú software android. Je ich 23,2 % zo všetkých príspevkov. Ďalej ich nasledujú desktop/web, Dlvt.it, Mobile Web,Iftt a Other čiže ostatné. Na poslednom grafe ktorý je na obrázku číslo 20 je znázornené demografické rozloženie príspevkov na sociálnej sieti Twitter. Z neho je jasné že dané kľúčové slovo zmieňuje vo svojich príspevkoch viac mužov ako žien. Je to v pomere 74% ku 24%. Z toho by sme mohli povedať, že značka "Nike" ktorú sme použili ako kľúčové slovo je viac mužskou značkou ako ženskou.



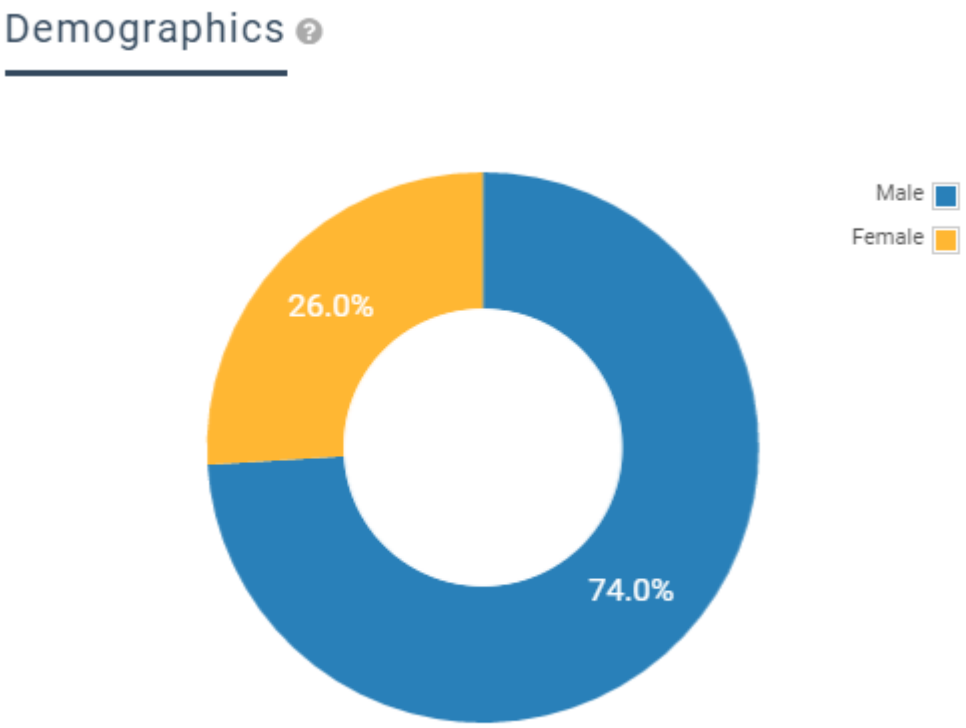
Obrázok č.17: percentuálny podiel príspevkov



Obrázok č.18: Sentimentálne ohodnotený príspevok



Obrázok č. 19: Hlavné zdroje príspevkov



Obrázok č.20: Demografické rozdelenie príspevkov

3.6.5 Porovnanie nástrojov

Nástroj	Jazyk	Zameranie na Twitter	API	Grafy	Licencia
Social Insider	SK,CS,EN	Nie	Áno	Áno	Komerčný
Sentiment140	EN, ES	Áno	Áno	Áno	Komerčný
SocialMention	EN	Nie	Áno	Áno	Komerčný
KeyHole	EN	Áno	Áno	Áno	Komerčný

Tabuľka š.2: porovnania nástrojov analýzy sentimentu

Záver

V porovnaní so súčasnými nástrojmi, je veľmi pravdepodobné, že v budúcnosti by analýza sentimentu mohla spĺňať oveľa viacej kritérií. Ideálny nástroj by mohol poskytovať nasledujúce funkcie:

- Vyhľadávanie ľubovoľného slova, skratky, symbolu či emotikony, ktoré sa objavia v ľubovoľnom príspevku.
- Pokročilé vyhľadávanie obsahujúce vyhľadávanie podľa dát, užívateľa alebo značky.
- Rozlišovanie medzi príspevkami od určitého užívateľa, príspevky pre určitého užívateľa a príspevky ktoré sa zmieňujú o užívateľovi.
- Generovanie farebných grafov znázorňujúcich odlišné nálady
- Možnosť vyhľadávania za dlhé časové obdobie
- Trvalé grafy znázorňujúce trendy a relevantné príspevky
- Minimálne 90 percentnú presnosť vyhodnocovania príspevku
- Možnosť aktualizácie v reálnom čase

Žiadny so súčasných nástrojov nedokáže splniť všetky podmienky a dokonca sa im niektoré len z ďaleka približujú. Niektoré sú na dobrej ceste a ich výsledky sa každou novou verziou zlepšujú. Pravdepodobným smerom, ktorým sa vydajú nástroje analýzy sentimentu v ďalších rokoch, bude zlepšovanie súčasných metód a väčšie zapojenie umelej inteligencie. Nástroje by sa mohli učiť zo svojich chýb, čo bude vyžadovať, aby im niekto povedal, že urobili chybu. Tieto algoritmy budú veľmi pravdepodobne podporené štatistickými a pravdepodobnostnými algoritmami, ktoré budú využívať pri svojom následnom rozhodovaní. V neposlednej rade by mali využiť podobný princíp, ako je v projekte sentiment140. Vedci zo Standfordskej univerzity využili ľudí, aby vytvorili tabuľky kľúčových slov, s ktorými neskôr pracujú algoritmy. Ľudský faktor bude treba využiť pre maximálnu presnosť vyhodnocovania. Mohlo by to vyzerať tak, že by človek robil finálnu kontrolu nad vyhodnocovacím algoritmom. Kombinácia umelej inteligencie, štatistiky, pravdepodobnosti a ľudskej manuálnej kontroly sa javí ako riešenie budúcnosti.

Hlavným cieľom práce bolo poukázať na nový spôsob analyzovania sentimentu a pripraviť predpoklady pre ďalší možný výskum detekcie sentimentu pomocou metód pracujúcich s jazykovým korpusom.

V prvej kapitole sme sa zamerali na definíciu pojmov sentimet a analýza sentimentu. Vysvetlili sme pojmy ktoré súvisia s analýzou sentimentu a sentimentom samotným.

V druhej kapitole sme opísali ciele a metodiky, ktoré boli použité.

V tretej kapitole sme sa zamerali na sociálnu sieť Twitter. Definovali sme čo je to sociálna sieť Twiiter a tak isto aj jej súčasti. Opísali sme metódy, ktoré sa v dnešnej dobe využívajú pre analýzu a tak isto aj problémy s ktorými sa analýza stretáva. V druhej časti tejto kapitoly sme spravili analýzu sentimentu na sociálnej sieti Twiiter pomocou vopred vybraných nástrojov . Pomocou výstupov z týchto nástrojov sme opísali ich základné funkcie a poukázali na ich silné a slabé stránky. Na záver sme opísali budúcnosť, ktorá muí čeliť veľkému množstvu problémov. Pretože v dnešnej dobe neexistuje nástroj ktorý by bol dokonalý, je pred analýzou sentimentu ešte dlhá cesta. Dnešné algoritmy majú závažné problémy s detekciou sentimentu v písaných prejavoch pretože človekom napísaný prejav je pre detekciu často náročný a sentiment nemusí byť zjavný. Avšak ak sa nástroje ktoré sú využívané pre analýzu sentimentu zdokonalia, stanú sa tak jedným s najsilnejším marketingovým nástrojom pre firmy ktoré sa prezentujú na sociálnych sieťach.

Zoznam použitej literatúry

[1] KOUDELA, Michal. Emoce a jejich fyziologický význam, Masarykova univerzita, 2012

[2] SOCHMAN, Jan, MATAS, Jiří. AdaBoost [online]. Dostupné z WWW: http://cmp.felk.cvut.cz/~sochmj1/adaboost_talk.pdf.

[3] Dey, Lipika and Haque, Sk. Opinion Mining from Noisy Text Data, International Journal on Document Analysis and Recognition 12(3). pp 205-226, 2009 Dostupné z WWW: https://link.springer.com/article/10.1007/s10032-009-0090-z

[4] Pang, Bo and Lee, Lillian, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval Dostupné z WWW: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[5] Bing LiuSentiment Analysis: Mining Opinions, Sentiments, and Emotions 1st Edition ISBN 978-1107017894

[6] Alekh Agarwal and Pushpak Bhattacharyya, Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified, International Conference on Natural Language Processing (ICON 05), IIT Kanpur, India, December, 2005 Dostupné na WWW: https://www.ijircce.com/upload/2014/august/6_AnEfficient.pdf

[7] Asher, Nicholas and Benamara, Farah and Mathieu, Yvette Yannick. Distilling opinion in discourse: A preliminary study, In Proceedings of Computational Linguistics (CoLing), 2008 Dostupné na WWW: https://www.aclweb.org/anthology/J/J11/J11-2001.pdf

[8] Taboada, Maite and Brooke, Julian and Tofiloski, Milan and Voll, Kimberly and Stede, Manfred, Lexicon-based methods for sentiment analysis, Computational Linguistics, 2011 Dostupné na WWW: https://www.aclweb.org/anthology/J/J11/J11-2001.pdf

[9] PECINOVSKÝ, Rudolf. Návrhové vzory. Computer Press, 2007. ISBN: 9788025115824.

[10] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar, Thumbs up?: sentiment classification using machine learning techniques, In Proceedings of the ACL-02 conference on Empirical Methods in Natural Language, ISBN 978-1-60198-150-9

[11] McCarthy, Diana and Koeling, Rob and Weeds, Julie and Carroll, John, Finding Predominant Word Senses in Untagged Text, Proceedings of the 42nd Meeting of the

Association for Computational Linguistics (ACL'04), 2004 Dostupné na WWW: http://dl.acm.org/citation.cfm?id=1218991
[12] Twitter Statistics | Statistic Brain [online] Dostupné na WWW: http://www.statisticbrain.com/twitter-statistics.

[13] Twitter Blog [online] Dostupné na WWW: http://blog.twitter.com/ 2010/06/links-and-twitter-length-shouldnt.html.

[14] CHEN, Zhe. Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond [online]. Dostupné na WWW: http://library.utia.cas.cz/separaty/2014/AS/pavelkova-0422958.pdf

[15] SOCHMAN, Jan, MATAS, Jiří. AdaBoost [online]. Dostupné na WWW: http://www.robots.ox.ac.uk/~az/lectures/cv/adaboost_matas.pdf

[16] HAN, Jiawei, KAMBER, Micheline. Data mining :concepts and techniques. San Francisco, 2006. ISBN: 1-55860-901-6.

[17] Twitter Libraries [online]. Dostupné na WWW: https://dev. twitter.com/docs/twitter-libraries.
[18] Ataxo - Social Insider [online]. Dostupné na WWW: http:// ataxosocialinsider.cz/.

[19] Sentiment140 [online] Dostupné na WWW: http://www. sentiment140.com.

[20] Real Time Search - Social Mention [online]. Dostupné na WWW: http://www.socialmention.com.

[21] KeyHole[online]. Dostupné na http://keyhole.co/