EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103004/B/2020/36114651172394244

IT pre analýzu zákazníckeho sentimentu

Bakalárska práca

2020 **Ondrej Izakovič**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

IT pre analýzu zákazníckeho sentimentu

Bakalárska práca

Študijný program: Hospodárska informatika

Študijný odbor: Hospodárska informatika

Školiace pracovisko: Katedra aplikovanej informatiky FHI

Vedúci záverečnej práce: Eva Rakovská RNDr. PhD.

Bratislava, 2020 **Ondrej Izakovič**

Čestné vyhlásenie

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

Dátum: 20. 05. 2020 .....

Ondrej Izakovič

ABSTRAKT

IZAKOVIČ, Ondrej: *IT pre analýzu zákazníckeho sentimentu* – Ekonomická univerzita v Bratislave. Fakulta Hospodárskej informatiky; Katedra aplikovanej informatiky FHI. – Vedúci záverečnej práce: Eva Rakovská RNDr. PhD. – Bratislava: FHI EU, 2020, počet strán 38.

Cieľom záverečnej práce je: mapovanie dostupných IT nástrojov pre analýzu zákazníckeho sentimentu a ich porovnanie. Dáta pochádzajúce z neštruktúrovaných údajov akým je zákaznícky sentiment sú podstatné pre podniky, avšak bez IT nástrojov je nemožné ich sledovať. Práca je rozdelená do 3 kapitol. Prvá kapitola je venovaná súčasnému stavu riešenia problematiky doma a v zahraničí. Zaoberá sa teoretickým vymedzením zákazníckeho sentimentu a jeho analýzy, úlohou analýzy sentimentu a prekážkami, ktoré sú spojené so skúmaním ľudského sentimentu.

Ďalšia časť sa zaoberá analýzou sentiment na rôznych úrovniach a opisuje prístupy k meraniu zákazníckeho sentimentu.

Tretia kapitola charakterizuje cieľ bakalárskej práce, ktorým je mapovanie a porovnanie algoritmov a postupov, ktoré sa využívajú pri analýze zákazníckeho sentimentu a vychádzajú z princípov strojového učenia sa (neurónové siete, heuristiky, spracovanie jazyka a podobne).

Záverečná časť sa zaoberá riešeniami danej problematiky dosiahnutými za použitia algoritmov a postupov pre analýzu zákazníckeho sentimentu.

Kľúčové slová:

zákaznícky sentiment, analýza zákazníckeho sentimentu, dolovanie textu, predspracovanie údajov, strojové učenie

ABSTRACT

IZAKOVIČ, Ondrej: *IT for customer sentiment analysis* – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Applied Informatics. – The supervisor of final work: Eva Rakovská RNDr. PhD. – Bratislava: FHI EU, 2020, number of pages 38.

The aim of the final work is: The aim of the thesis is to map the available IT tools for the analysis of customer sentiment and their comparison. Data acquisition based on non-measurable indicators such as customer sentiment becomes part of the business and cannot be tracked without IT. The thesis is divided to 3 chapters. The first chapter is devoted to the current state of the issue at home and abroad, which deals with the theoretical definition of customer sentiment and its analysis, the goal of sentiment analysis, and the challenges that are connected to extracting data from human sentiment.

The next part deals with sentiment analysis on different levels and describes different approaches to quantification of the customers sentiment.

The third chapter describes the goal of the bachelor thesis which is mapping and comparison of algorithms and procedures that are used for analysis of customer sentiment and are based on the principles of machine learning (neural networks, heuristics, language processing, etc.).

The final part deals with result of the solution of this issue achieved with algorithms and procedures that are used for analysis of customer sentiment.

Keywords:

customer sentiment, customer sentiment analysis, text mining, text pre-processing, machine learning

O B S A H str.

Úvod

V dnešnej dobe nie sú spotrebitelia izolovaní pri nákupnom rozhodovaní. Sociálne siete a rôzne iné internetové média umožňujú kupujúcim prístup do sveta veľkého množstva recenzií iných spotrebiteľov. Ich kritika, spätná väzba, odporúčania a varovania sú nápomocné pri rozhodovaní o kúpe rôznych produktov a služieb. Tieto subjektívne a na skúsenostiach založené informácie môžu spotrebitelia uverejniť prostredníctvom mnohých sociálnych médií a platforiem. Najbežnejšími sú fóra, diskusné skupiny, recenzné portály, blogy a stránky sociálnych sietí. Nemôžme opomenúť vizuálne platformy ako YouTube, Flickr, Instagram a Pinterest. Takýto obsah na sociálnych sieťach je známy ako user-generated content, teda užívateľmi vytvorený obsah. Tento obsah je nesmierne cenný a to nie len pre samotných spotrebiteľov. Poskytuje totiž okrem samotných hodnotení aj informácie o trendoch na trhoch, zmenách v reputácii spoločností a produktoch, správaní zákazníkov, procese rozhodovania a o tom, do akej mieru sú konzumenti spokojní so službami a výrobkami.

Zatiaľ čo istá časť obsahu vytvoreného užívateľmi je v podobe obrázkov, videozáznamu a číselných hodnôt, akými sú napríklad hodnotiace stupnice, drvivá väčšina informácií je zverejňovaná formou textu na sociálnych sieťach. Takýto text obsahuje okrem faktických údajov aj osobné ohodnotenie, ktoré sa nazýva sentiment. Zákaznícky sentiment je hodnotný pre marketérov, ktorí analyzujú trh, ale aj pre samotných spotrebiteľov, ktorým umožňuje rýchlo a ľahko získať prehľad o službe, výrobku alebo určitej destinácii (Jing et al., 2018).

Analýza sentimentu sa osvedčila pri získavaní podstatných informácií tak ako o výrobkoch, tak aj o zákazníkoch. Na jednej strane podnik získa prehľad o názoroch svojich zákazníkov. Ak je reakcia na produkt z veľkej časti negatívna, je jasné, že sa výrobok neteší veľkej obľube. V takom prípade je na mieste zvážiť, čo bude ďalej s daným produktom. Na druhej strane pozitívne ohlasy nám môžu pomôcť identifikovať tie aspekty produktu, ktoré sú pre zákazníkov dôležité. Zameranie sa na tieto aspekty má potenciál prilákať nových zákazníkov. V konečnom dôsledku nám analýza zákazníckeho sentimentu pomáha vytvoriť produkt, ktorý je prispôsobený určitej časti trhu (Jing et al., 2018).

1 Súčasný stav riešenej problematiky doma a v zahraničí

Analýza sentimentu sa osvedčila pri získavaní podstatných informácií tak ako o výrobkoch, tak aj o zákazníkoch. Na jednej strane podnik získa prehľad o názoroch svojich zákazníkov. Ak je reakcia na produkt z veľkej časti negatívna, je jasné, že sa výrobok neteší veľkej obľube. V takom prípade je na mieste zvážiť, čo bude ďalej s daným produktom. Na druhej strane pozitívne ohlasy nám môžu pomôcť identifikovať tie aspekty produktu, ktoré sú pre zákazníkov dôležité. Zameranie sa na tieto aspekty má potenciál prilákať nových zákazníkov. V konečnom dôsledku nám analýza zákazníckeho sentimentu pomáha vytvoriť produkt, ktorý je prispôsobený určitej časti trhu (Jing et al., 2018).

1.1 Sentiment

Vo všeobecnosti sentiment môžeme definovať ako určitý osobný názor, alebo hodnotenie na základe emócií. V informatike pod pojmom sentiment rozumieme názor ľudí na určitý predmet. Názor na predmet je vyjadrený prostredníctvom orientácie – polarity, ktorá môže byť pozitívna alebo negatívna (Yi et al., 2003).

Vo všeobecnosti sa za zložky sentimentu považujú názor a cieľ, na ktoré sa názor vzťahuje. V informatike sa používajú aj viaceré iné definície, ktoré sentiment rozkladajú na viaceré zložky.

1.1.1 Definícia sentimentu ako pätice

Najkomplexnejšou a zároveň pre analýzu sentimentu optimálnou definíciou je definícia sentimentu ako pätice. Podľa nej sentiment obsahuje tieto komponenty (Bing, 2012):

- cieľovú entitu,
- hodnotené aspekty cieľovej entity,
- sentiment voči hodnoteným aspektom,
- vlastníka názoru,
- čas vyjadrenia názoru jeho vlastníkom.

Je nutné poznamenať, že každý komponent tejto pätice sa musí vzťahovať na ostatné komponenty. Každý jeden komponent je podstatný a vynechanie ktoréhokoľvek z nich môže viesť k neistým výsledkom. Ak napríklad nepoznáme časový komponent, nedokážeme ani analyzovať entitu v čase. Analýza v čase je v praxi veľmi dôležitá. Názor starý dva roky má totiž rozdielny význam od názoru vyjadreného dnes.

Z hľadiska polarity môže byť názor pozitívny, negatívny, alebo neutrálny. Sentiment môže byť taktiež vyjadrený hodnotou na stupnici podľa intenzity. Túto metódu aplikujú mnohé webových stránky, ktoré používajú stupnice, napríklad od 1 do 5 hviezdičiek. Pokiaľ sa názor vzťahuje na entitu ako celok, hovoríme o všeobecnom názore. Vtedy sa cieľová entita a jej aspekty zhodujú.

Sentiment vyjadrený ako pätica predstavuje nástroj, pomocou ktorého je možné premeniť neštruktúrovaný text do štruktúrovaných údajov. Vyššie uvedená pätica definuje optimálnu schému, na základe ktorej je možné získané údaje vkladať do databázy (Bing, 2012).

1.2 Analýza sentimentu

Analýza sentimentu, známa aj pod pojmom opinion mining, skúma sentiment, čiže názory na produkty a služby, firmy, jednotlivcov, destinácie, ale aj udalosti a témy. Jej podstata tkvie v dolovaní informácií z textu, na základe ktorých sa identifikuje subjektívny názor a s ním spojená konkrétna emócia. Väčšinu skúmaného textu tvorí užívateľmi generovaný obsahom na internete.

Podstatou analýzy sentimentu je získať prehľad o celkovom dojme, ktorý zanechal určitý produkt na zákazníkoch. V praxi sa táto metóda spolieha na zjednodušený, binárny systém pozitívnych a negatívnych reakcií. Pozeráme sa teda na skúsenosť, ktorú mal zákazník s daným produktom. Výsledok tejto zjednodušenej analýzy je ľahko spracovateľný a hodnotiteľný (Allouch, 2018).

Najväčší rozmach zaznamenala analýza sentimentu po roku 2000, kedy začali ľudia po prvý krát vo veľkom zverejňovať svoje názory na internete. To podnietilo záujem podnikov o spracovanie a analyzovanie týchto informácií pre komerčné použitie. Rastie taktiež aj počet firiem, ktoré poskytujú nástroje na meranie zákazníckeho sentimentu. Je však nutné poukázať aj na negatívne aspekty, ktorými sú vysoké náklady spojené s použitím nástrojov na dolovanie dát (z angl. data mining) a neistota návratnosti použitia týchto nástrojov. Jeho podstata tkvie v dolovaní informácií z textu so zameraním na určenie subjektívneho názoru v určitom vyjadrení a identifikácii konkrétnej emócie, predovšetkým v užívateľmi vytvorenom obsahu na internete (Jing et al., 2018).

Subjektívne, na skúsenostiach založené informácie uverejňujú spotrebitelia prostredníctvom mnohých sociálnych médií a platforiem. Najbežnejšími sú fóra, diskusné skupiny, recenzné portály, blogy a stránky sociálnych sietí. Nemôžme opomenúť vizuálne

platformy ako YouTube, Flickr, Instagram a Pinterest. Používatelia často uvádzajú svoje recenzie produktov aj na stránkach internetových predajcov alebo rôznych internetových poradcov. Príkladom takéhoto poradcu je portál heureka.sk.

Recenzia mobilného telefónu na portáli heureka.sk (Zdroj: heureka.sk)

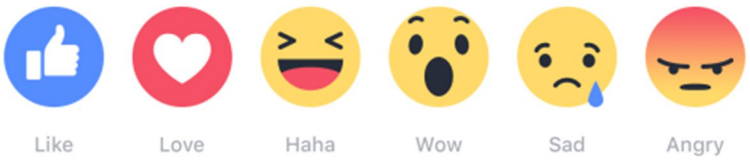
1.2.1 Prekážky analýzy sentimentu

Rastúce množstvo subjektívneho obsahu v sociálnych médiách, ako odlišnosti jednotlivých médií, pomocou ktorých je subjektívny názor publikovaný na internete, so sebou prinieslo určité výzvy. Používatelia okrem textu zverejnenú s úmyslom vyjadriť sa aj obrázky, videá a emotikony. Aj keď je video vo všeobecnosti chápané ako jedno médiu, pri analýze sentimentu musí byť rozdelené na dve časti, a to zvukovú a vizuálnu. Zvuková časť okrem hovorenej reči obsahuje aj iné zvukové vyjadrenia, akými sú prozódia a smiech. Pri analýze obrazovej časti sú podstatné gestá a mimika. Taktiež treba brať do úvahy, že spôsobom akým sa ľudia vyjadrujú tvárou v tvár sa rôzni. Niektorí ľudia pracujú viac s hlasom, zatiaľ čo iní vyjadrujú svoje emócie pomocou gestikulácie. Tieto aspekty pridávajú na zložitosti analýzy. Pre správnu identifikáciu sentimentu je nutné dekódovať

a podrobne rozobrať obe dimenzie. Podobný prístup si vyžaduje aj analýza statických obrazov.

Ďalšiu výzvu pri identifikácii sentimentu v komunikácii predstavujú emotikony a iné grafické ikony. Okrem emotikonov, predstavujúcich rôzne výrazy tváre, majú používatelia možnosť vyjadriť svoj sentiment prostredníctvom veľkého množstvo emotikonov predstavujúcich gestá a ikon znázorňujúcich symboly a objekty. Podobne ako pri sentimentových slovách, musia analytici priradiť týmto symbolom význam a vytvoriť z nich pozostávajúci lexikón. Ďalej je nutné vytvoriť algoritmus schopný chápať tieto ikony a správne im priradiť orientáciu a frekvenciu. Okrem očividného významu, môžu mať emotikony iný význam, viažuci sa na špecifickú kultúru (Jing et al., 2018).

Emotikony však poskytujú možnosť zachytiť zákazníkove nálady formou tlačidiel reakcie. Skvelým príkladom sú tlačidlá, aké používa sociálna sieť Facebook. Tie formou emotikonu reprezentujú 6 emócii: Páči sa mi to (Like) , Super (Love), Haha (Haha), Žasnem (Wow), Je mi to ľúto (Sad), Štve ma to (Angry). Výhodou tohto systému je reakcia zákazníka v reálnom čase a relatívna jednoduchosť jej interpretácie (Allouch, 2018).



Emotikony ako tlačidlá reakcie (Allouch, 2018)

1.2.2 Použitie analýzy sentimentu

Človek je do veľkej miery ovplyvniteľný názorom okolia. Tento fenomén sa týka aj nákupného správania. Kupujúci sa pri výbere produktu prirodzene obracia na ľudí, ktorí už majú s daným produktom skúsenosti. Firmy a spoločnosti chcú mať prehľad o názore spotrebiteľov, ako aj ostatných spoločností, na produkty a služby, ktoré ponúkajú. V minulosti sa kupujúci pýtali na názor svojej rodiny a okolia. Ak chceli firmy poznať verejnú mienku, spravili prieskum trhu.

V dnešnej dobe ak konzument uvažuje nad kúpou určitého produktu, nepýta sa na názor svojich priateľov, pretože na internete je obrovské množstvo recenzií na daný produkt.

Pre firmy už nie je nutné vytvárať ankety a dotazníky, aby získali informácie o verejnej mienke, pretože skrz internet sú tieto informácie verejne dostupné. Avšak získavanie konkrétnych informácií z veľkého množstva textu na internete je časovo náročné. Preto je potrebná automatizovaná analýza zákazníckeho sentimentu.

Analýza sentimentu sa v posledných rokoch rozšírila takmer do všetkých možných odvetví, od spotrebiteľských produktov, služieb, zdravotníctva, finančníctva až po politiku. Praktické použitie analýzy sentimentu a záujem podnikovej praxe oň podnietil rozmach výskumu v tomto obore. Len za posledné roky sa v oblasti analýzy zákazníckeho sentimentu angažovalo minimálne 60 startupov len v USA. Mnohé veľké spoločnosti vytvorili svoje vlastné systémy na analýzu sentimentu, ako napríklad Microsoft, Google, Hewlett-Packard, SAP a SAS.

Ďalšími príkladmi využitia analýzy sentimentu je napríklad predpovedanie výsledkov volieb skúmaním blogov na sociálnych sieťach alebo stanovenie prognózy vývoja na finančných trhoch (Bing, 2012).

Z pohľadu marketingu sú dôležité práve pocity zákazníkov a to ako reagujú na službu alebo tovar. V praxi však nie je jednoduché tieto nálady kvantifikovať. Emočná analýza je základnou metódou na meranie spokojnosti zákazníkov.

1.3 Emočná analýza

V kontraste s analýzou sentimentu, podstata emočnej analýzy tkvie v komplexnejších a sofistikovanejších systémoch. Zatiaľ čo prvá používa jednoduchú binárnu kategorizáciu, emočná analýza sa zameriava na dôkladné skúmanie ľudských pocitov. Skúma nuansy v odpovediach respondentov. Na základe týchto odtienkov je možné vyjadriť odchýlku rôznych emócií. Emočná analýza zohľadňuje ľudskú mentálnu subjektivitu, ktorá pozostáva zo širokého spektra nálad, nie len z niekoľkých nemenných kategórií. Pozitívna reakcia, môže byť výsledkom rôznych emócií, ako napríklad radosť, uspokojenie, alebo vzrušenie. Emočná analýza skúma motivácie zákazníkov a poskytuje informácie o tom, ako podnietiť zákazníkov k želanému správaniu.

Vo všeobecnosti je možné považovať emočnú analýzu za hlbšiu vrstvu analýzy sentimentu. Každá z týchto metód ponúka odlišný pohľad na nálady zákazníka.

1.3.1 Špecifiká emočnej analýzy

Emočná analýza ponúka holistický obraz, ktorý je pre marketing nesmierne dôležitý. Ľudské emócie sú komplexné a pokiaľ je spätná väzba zákazníkov zjednodušená len na pozitívne a negatívne reakcie, môže to viesť k nepresnej predstave o imidži podniku.

Na rozdiel od analýzy sentimentu, emočná analýza poskytuje mnohonásobne hlbšie pochopenie ľudského správania. Pre pochopenie toho, prečo zákazníci kupujú, alebo nekupujú určitý produkt, nestačí percentuálne ohodnotenie produktu. Úspešnosť závisí aj od toho, aké emócie v kupujúcom vyvoláva.

Na základe percentuálneho ohodnotenia je možné povedať, ako sa danému produktu darí na trhu, avšak emočná analýza umožňuje pochopiť, prečo tomu tak je.

Zlepšiť produkt je takmer nemožné len na základe polaritných reakcií. Vedeniu podniku nestačí vedieť, že spotrebitelia považujú produktu jednoducho za zlý. Len na základe takejto informácie je náročné produkt vylepšiť. Na druhej strane, ak sú známe konkrétne emócie, ktoré zákazníci voči produktu pociťujú, je jednoduché určiť konkrétny aspekt produktu, na ktorom treba pracovať (Allouch, 2018).

1.3.2 Využitie emočnej analýzy v praxi

Na trhu je možné nájsť viacero nástrojov emočnej analýzy. Príkladom je softvér spoločnosti EMOTION RESEARCH LAB, ktorý identifikuje tváre zákazníkov, ktorých produkt fyzicky zaujal. Ľudia sú citlivý na podnety akými sú vôňa, chuť, ale aj na to, ako produkt vyzerá. Tieto podnety zanechávajú na zákazníkoch veľký dojem. Emócie ktoré zákazník pociťuje sa podvedome odrážajú na tvári. Za použitia metód strojového učenia dokáže softvér z tváre vyčítať emócie testovaného subjektu. Ide teda o skúmanie nevyslovených, podvedomých reakcií testovaného subjektu. Takto je možné zistiť, akú reakciu vyvolávajú jednotlivé aspekty produktu, aký prvý dojem produkt zanechal, ale aj vzťah medzi vyslovenými a nevyslovenými emóciami.

Takýto softvér sa dá použiť pri testovaní produktu pred jeho uvedením na trh, a to najmä v oblasti gastronómie, kozmetiky, alebo pri dizajne obalu produktu. Ďalšie možné použitie takýchto nástrojov je priamo pri predaji výrobku, kedy sa z čŕt tváre s vysokou presnosťou zisťujú okrem emócií aj pohlavie, rasa a vek. To umožňuje k emočnej reakcii priradiť profil zákazníka. Výsledkom takejto analýzy je obraz záujmu zákazníka o tovar v konkrétnej lokalite, podľa veku, pohlaviu a etnicity.

2 Cieľ práce, metodika práce a metódy skúmania

Hlavným cieľom práce je urobiť mapovanie dostupných IT nástrojov pre analýzu zákazníckeho sentimentu a ich porovnanie. Pre dosiahnutie cieľa práce je nutné pochopiť, čo je to zákaznícke sentiment a jeho analýza a aký rozdiel je medzi analýzou sentimentu a emočnou analýzou. Tým sa do hĺbky zaoberá prvá kapitola. Informácie, ktoré sa nachádzajú v tejto kapitole pochádzajú zo zahraničnej literatúry a vo väčšine sa zaoberajú zberom a spracovaním neštruktúrovaných údajov a ich použitím pri podnikaní.

Ďalej sa práca zaoberá stupňami, na ktorých je možné podniknúť analýzu sentimentu, ako aj prístupmi k nej. Prístupy k analýze sentimentu predstavujú použitie určitej technológie pri jej realizácii. Zvyčajne ide o syntézu poznatkov z viacerých vedných oblastí, akými strojového učenia a spracovanie prirodzeného jazyka. Jednotlivé prístupy definuje, ku ktorej oblasti sa viac prikláňa. Pri skúmaní sentimentu sa do značnej mieri využíva metóda abstrakcie. Údaje o sentimente zákazníka sa získavajú v prevažnej miere z neštruktúrovaného textu, sentimentové slová sú vyhľadávané a spracované a hodnotené na základe spoločných znakov v podobe tokenov, príkladom je takéhoto prístupu je zhlukovanie sentimentových slov. Subjektivita sentimentu je predpokladom použitia dedukcie pri vyvodzovaní záverov, týkajúcich sa efektivity jednotlivých prístupov a metód jeho analýzy.

Ako už bolo vyššie spomenuté, užívateľmi tvorený obsah je cenný pri analýze sentimentu a má veľký potenciál pre využitie v podnikateľskej sfére. Väčšina týchto dát sa na internete nachádza vo forme neštruktúrovaných dát, predovšetkým v textovej forme. V odbornej literatúre sa uvádza, že neštruktúrované dáta tvoria 80 % pre podnikateľskú sféru relevantného obsahu. Vzhľadom na obrovské a neustále rastúce množstvo údajov je nepraktické získavať údaje potrebné pre analýzu sentimentu manuálne. To podnietilo vznik analýzy sentimentu a zavádzania mechanizmov a algoritmov na jej automatizáciu. Automatizácia analýzy sentimentu v neštruktúrovaného obsahu má niekoľko nezanedbateľných výhod:

- analýza sentimentu umožňuje spracovanie veľkého množstva údajov efektívne, bez ľudskej práce a pri nízkych nákladoch,
- informácie o sentimente zákazníkov je možné získavať v reálnom čase, ak si to situácia žiada,
- sentiment je automaticky spracovaný do podoby využiteľnej manažmentom podniku,

- výsledky analýzy sú presnejšie, než pri manuálnom spracovaní, keďže je sentiment pozorovaný na základe konzistentných kritérií (Hussein, 2019).

2.1 Stupne analýzy sentimentu

Analýza sentimentu môže byť vykonaná podľa požadovanej diskrétnosti na troch úrovniach.

2.1.1 Analýza sentimentu na úrovni dokumentu

Ide o najjednoduchšiu formu analýzy, kedy je celý text obsahujúci názor považovaný za základnú jednotku informácie. Predpokladom je, že dokument poskytuje názor tákajúci sa len jedného objektu. Analýzu na tejto úrovni nie je vhodné vykonať, pokiaľ dokument obsahuje aj názory týkajúce sa viacerých objektov. Výsledkom je dokument klasifikovaný na pozitívny, alebo negatívny. Irelevantné vety musia byť pred spracovaním dokumentu odstránené (Kolkur et al., 2015).

2.1.2 Analýza sentimentu na úrovni viet

Na tejto úrovni sa určuje sentiment obsiahnutý v každej jednej vete zvlášť. Sentiment môže byť pozitívne, negatívne alebo neutrálne orientovaný. Neutrálna emócia je všeobecne považovaná za nevyjadrenie žiadneho názoru. Táto úroveň analýzy je úzko spojená s klasifikáciou subjektivity. Na základe klasifikácie subjektivity členíme vety na objektívne, ktoré sú vyjadrením faktických informácii a subjektívne, ktoré vyjadrujú subjektívny pohľad a názor. Je nutné poznamenať, že subjektivita nie je ekvivalentom sentimentu, keďže aj objektívne vety môžu zahrňovať sentiment (Bing, 2012).

2.1.3 Analýza na úrovni entity a aspektu

Analýza na úrovni viet, ani analýza na úrovni dokumentu nie sú schopné presne určiť, čo sa ľuďom páčilo, a čo nie. Na úrovni aspektu sa nehľadí na jazykové štruktúry, akými je text, paragraf, alebo fráza, ale skúma sa samotný názor. Táto analýza je založená na myšlienke, že názor sa skladá zo sentimentu a cieľa. Názor, bez jeho cieľa má len obmedzené použitie. Pri skúmaní sentimentu je teda podstatné identifikovať aj cieľ. Ak veta znie pozitívne, nemusí automaticky vyjadrovať pozitívnu emóciu. V praxi sú názory na určitý cieľ vyjadrené názorom, ktorý sa vzťahuje len na niektorý z aspektov cieľa, poprípade inej entity. Preto je potrebné identifikovať všetky aspekty a entity pri meraní sentimentu. Výsledkom analýzy na tejto úrovni je sumár, ktorý zlučuje jednotlivé názory na entity do jednej štruktúry. Takto získané dáta sú ďalej ľahko spracovateľné (Bing, 2012).

2.2 Prístupy pri analýze sentimentu

K analýze sentimentu môžeme pristupovať tromi spôsobmi. Prvým je prístup založený na skúmaní lexiky obsiahnutej v lexikónoch sentimentu, druhý skúma sentiment za použitia jazykového korpusu a tretí je založený na strojovom učení.

2.2.1 Lexikálny prístup

Takzvané sentimentové slová (z angl. sentiment words) zohrávajú dôležitú úlohu pri identifikácii sentimentu, jeho orientácie a intenzity. Tieto kľúčové slová sa delia na dve kategórie. Prvá zahŕňa základné slová, ktoré vyjadrujú určitý sentiment, druhá porovnávacie výrazy, ktoré slúžia na identifikáciu sémanticky zložitejšieho sentimentu. Základnými slovami môžu byť prídavné mená, príslovky, podstatné mená a slovesá, pričom najbežnejšími sú práve prídavné mená v podobe superlatívov. Napríklad slová ako dobrý, výborný a vynikajúci vyjadrujú pozitívny sentiment, zatiaľ čo slová ako zlý, hrozný a otrasný vyjadrujú negatívny sentiment. Vo vete „Tento fotoaparát *fotí perfektne.*“ príslovka určuje sentiment vo vzťahu k slovesu, na ktoré sa viaže. Slovesá je možné zaradiť do oboch kategórií, dajú sa použiť na vyjadrenie emócií priamo, aj nepriamo. Za priame vyjadrenie názoru slovesom považujeme „Jedlo nám *nechutilo*“. „Elektrické autá stále *zaostávajú* za autami so spaľovacím motorom.“ je veta, v ktorej je slovesom vyjadrený sentiment nepriamo, vzhľadom na porovnávané objekty.

Okrem jednotlivých slov, pri vyjadrení názoru zohrávajú dôležitú rolu aj frázy a frazeologizmy, napríklad „*šľape ako hodinky*“. Sentiment môže byť zdôraznený aj pravopisne: „*celeee* zle“, alebo za použitia diakritiky: „*fantastické!!!*“. Problém pri identifikácii sentimentu môže spôsobiť negácia slova vyjadrujúca určitý sentiment, ako vo vete „*Nie som veľmi spokojný*“, kde spokojný je slovo bežne chápané ako pozitívne. Je nutné poznamenať, že v rôznych jazykoch môže byť sentiment vyjadrený rôznym spôsobom. Automatizácia hodnotenia sentimentu na základe kľúčových slov je obtiažna v jazykoch, ktoré slová spájajú alebo nerozdeľujú. Medzi takéto jazyky patrí napríklad Čínština.

Spoločne tieto slová a výrazy tvoria lexikóny sentimentu. K tvorbe týchto lexikónov je možné pristupovať dvomi spôsobmi.

Manuálny prístup

Jedným z prístupov tvorby lexikónov sentimentu je manuálny, kedy programátori zostavujú tieto slovníky ručne. Takýto prístup je náročný na ľudskú prácu a zaberá veľké množstvo času. Na druhej strane ľudia dokážu do lexikónov zahrnúť aj jemné nuansy, teda odtienky ľudskej reči a pojmy, ktoré majú v rozličných situáciách rozličný význam.

Lexikálny prístup

Ďalším spôsobom tvorby slovníkov sentimentu je takzvaný lexikálny prístup. Pod týmto prístupom chápeme proces využívania znalostí z už vytvorených slovníkov pri identifikácii nových sentimentových slov. Pri zostavovaní slovníkov sa často používa General Inquerer (Jing, 2018).

General Inquirer je súbor procedúr slúžiacich na identifikáciu opakujúcich sa vzorov v ľudskej reči a písomnej komunikácii. Tento systém je naprogramovaný tak, aby slová a frázy z textu vyhľadával v slovníkoch, určoval kľúčové slová texte, porovnával vzory použitia týchto kľúčových slov, počítal, koľko krát sa v texte objavia a vracal vety so špecifickými charakteristikami. Problém však môže nastať, keď vyhľadávané slová majú v rôznych situáciách rozličný význam (Stone et al., 1966) .

Takéto slovníky obsahujú nielen slová a ich význam ale často aj ich synonymá a antonymá. Tento prístup umožňuje rýchlo vyhľadať veľké množstvo sentimentových slov aj s ich orientáciou.

Využívanie lexikónov sentimentu je podstatné pri meraní sentimentu, nie je však samé o sebe dostačujúce. Vo väčšine prípadov je sentiment obsiahnutý vo viac než jednom kľúčovom slove a tak môžu nastať viaceré problémy:

1. Jedno slovo môže byť použité na vyjadrenie opačného sentimentu. Ako príklad uvediem slovo *dobrý*, ktoré sa zvyčajne používa na vyjadrenie pozitívneho sentiment, avšak vo vete „Toto auto je *dobrý* šrot.“ vyjadruje presný opak.
2. Vety obsahujúce „sentimentové“ slovo nemusia vyjadrovať sentiment. Tento fenomén nastáva v otázkach, alebo v kondicionálnych vetách. Príkladom sú vety: „Môžete mi poradiť nejaký *dobrý* fotoaparát?“ a „Keby sa mi podarí nájsť *dobrý* fotoaparát, tak si ho kúpim“. Nie všetky podmienkové vety však nevyjadrujú žiadny sentiment. Napríklad veta „Keby bol tento fotoaparát *dobrý*, tak ho nereklamujem“ je negatívna.

3. Je ťažké správne identifikovať sentiment vo vetách obsahujúcich sarkazmus. Sarkazmus sa najčastejšie vyskytuje v politických diskusiách, nie sú ale nezvyčajné ani v spotrebiteľských recenziách produktov a služieb. Príkladom môže byť veta: „Skvelý fotoaparát, vydržal celé dva dni.“
4. Vety môžu obsahovať sentiment, aj keď neobsahujú žiadne „sentimentové“ slová. Často ide o objektívne vety, ktoré obsahujú fakty. Napríklad veta „Batéria v tomto fotoaparáte vydrží len niekoľko hodín“ je vyjadrením negatívneho sentimentu aj bez konkrétneho slova nesúceho sentiment (Bing, 2012).

2.2.2 Prístup založený korpuse

Korpus textov predstavuje špecifický súbor jazykových dát, ktorý sa buduje v elektronickej podobe. Jeho základom sú texty zvyčajne rôznych štýlov a žánrov, ku ktorým sa pridávajú lingvistické informácie na úrovni slova (textovej jednotky), vety aj celého textu. Výkonné vyhľadávacie nástroje umožňujú vyhľadávanie a triedenie skúmaných jazykových prostriedkov a informácií. Lingvisti na základe autentického jazykového materiálu opisujú významy a funkcie slov i ďalších jazykových javov, ich štatistiky, spájateľnosti a pod. Bežným používateľom jazyka môže korpus poslúžiť ako zdroj praktického poznania systému jazyka a overenia či doplnenia jednotlivých poznatkov o reálnom fungovaní jazykových prostriedkov v praxi. Korpus nie je elektronickou knižnicou (texty v ňom sa nedajú čítať ako jeden celok), ani nenahrádza kodifikačné či gramatické príručky (Šimková, 2019).

Prístup založený na korpuse tkvie v tvorbe slovníka pre určitú oblasť. Existujú dve metódy aplikácie tohto prístupu. Pri prvej sú vyhľadávané sentimentové slová špecifické pre určitú oblasť na základe primárneho zoznamu sentimentových slov. Pri druhej sú základné sentimentové slová adaptované pre určitú oblasť za pomoci lingvistického korpusu špecifického pre danú oblasť.

Pointa tvorby lexikónov špecifických pre určitú oblasť vychádza z toho, že sentimentové slová môžu vyjadrovať v rôznych oblastiach rôzny sentiment. Samozrejme je možné využiť tento prístup pri tvorbe všeobecných slovníkov, ktoré sa neviažu na žiadnu oblasť, avšak v takom prípade je považované za efektívnejšie použiť lexikálny prístup.

Pri použití korpusu a jeho skúmaní sa prišlo na niekoľko pravidiel týkajúcich sa kľúčových slov, ktoré sú zvlášť užitočné pri analýze sentimentu. Pre spojku *a* napríklad platí,

že dve prídavné mená spojené touto spojkou majú rovnakú polaritu. Toto pravidlo vychádza z poznatku že, ľudia zvyknú vyjadriť rovnaký názor pred aj za touto spojkou, ako napríklad vo vete „Interiér tohto auta je pohodlný *a* priestranný,“. Podobné pravidlá boli nájdené aj pre iné spojky, akými sú *alebo*, *ale*, *buď-alebo* atď. Táto koncepcia sa nazýva konzistencia sentimentu. Obdobne, v rámci viet, ale aj medzi susednými vetami ľudia často vyjadrujú ten istý sentiment. Tento úkaz sa nazýva koherencia.

Určenie sentimentu slov v rámci určitej oblasti použitia slov je užitočné, avšak ani to nemusí byť dostatočné pri analýze sentimentu v praxi. Napríklad v oblasti techniky slovo *dlhú* môže mať rôznu polaritu. Príkladom je dvojica vie „Batéria v tomto telefóne ma *dlhú* životnosť“ a „Dotykový displej tohto telefónu má *dlhú* odozvu“, kedy má slovo *dlhú* v prvej vete pozitívnu polaritu, zatiaľ čo v druhej negatívnu. Táto situácia nastáva aj pri iných kvantifikátoroch, akými sú krátky, veľký, malý a pod. Dôležitý je tak aspekt, ako aj slovo vyjadrujúce sentiment. Preto sa začal používať pár *(aspekt, sentimentové_slovo)* ako kontext názorov (z angl. opinion context). Príkladom je dvojica („výdrž batérie“, „dlhá“). Podľa tejto metódy je sentiment kľúčových slov stanovený na základe vzťahu kľúčového slova a aspektu, ktorý ho bližšie určuje (Bing, 2012).

2.2.3 Prístup založený na strojovom učení

Strojové učenie vo všeobecnosti rozumieme podoblasť umelej inteligencie, zaoberajúcu sa programami, ktoré sa dokážu učiť a na základe nadobudnutých vedomostí vykonávať procesy, ktoré nemali pôvodne naprogramované. Algoritmy strojového učenia umožňujú programom učiť sa na trénovacích datasetoch. Dataset je súbor cvičných dát, z ktorých sa funkcia učí nájsť vzory a súvislosti medzi údajmi. Trénovacie datasety tvoria vždy dve skupiny prvkov – features a labels. Features sú skúmané vlastnosti objektu - príznaky, zatiaľ čo label je atribút, ktorý označuje vlastnosti objektu. Na týchto datasetoch sa programy učia odhadovať výsledky a generovať modely.

Algoritmy strojového učenia využívané pri analýze sentimentu spadajú do dvoch kategórii, učenie s učiteľom, ktoré je známe aj ako kontrolované učene (z angl. supervised learning) a učenie bez učiteľa (z angl. unsupervised learning). Metódy kontrolovaného učenia je pred použitím na ostrých dátach nutné trénovať na tzv. trénovacej množine. Trénovacia množina predstavuje vstupné údajmi s už priradenými atribútmi.

Pri nekontrolovanom učení tieto atribúty chýbajú. Z textu sú extrahované vety, ktoré obsahujú sentimentové slová ako prídavné mená a príslovky. Algoritmus potom musí nájsť štruktúru v týchto údajoch. Keďže pri týchto sentimentových slovách je dôležitý kontext,

algoritmus vyhľadáva aspekt, na ktorý sa viažu. Celková orientácia vlastnosti objektu je potom vypočítaná ako priemer jednotlivých extrahovaných fráz (Kamble – Itkikar, 2018).

3 Výsledky práce a diskusia

V predchádzajúcej kapitole sú opísané prístupy k analýze zákazníckeho sentimentu. V tejto kapitole sú opísane techniky a algoritmy na jej realizáciu.

3.1 Analýza sentimentu v neštruktúrovaných údajoch

Ako už bolo v úvode spomenuté, užívateľmi tvorený obsah je cenný pri analýze sentiment a má využitie v podnikateľskej sfére. Až 80 % užívateľmi generovaného obsahu, ktorý je relevantný pre firmy, tvoria neštruktúrované údaje

Myšlienka hľadania vzorov v ľudskej reči pochádza už zo 60. rokov 20. storočia. Pôvodne mala byť táto technológia použitá pri napodobovaní rozhovoru terapeuta s pacientom. Umelá inteligencia sa začala používať na analýzu prirodzeného jazyka už začiatkom 90. rokov. Trvalo však ešte mnoho rokov, kým sa spracovanie prirodzeného jazyka stalo bežným javom aj v komerčnej sfére.

Pri zavádzaní business intelligence sa veľmi skoro zistilo, že väčšina údajov v podniku je vo forme neštruktúrovaného textu. To naštartovalo rozvoj v analýze neštruktúrovaných údajov. Do relačných databáz boli zavádzané systémy riadenia bázy dát, ktoré, okrem iného, boli schopné spracovať, indexovať a ukladať neštruktúrovaný text. S nástupom internetu však tieto systémy zastarali a boli nahradené novými technikami (Grimes, 2008).

3.2 Text Mining

Text mining, alebo aj dolovanie z textu, je disciplína, ktorá vychádza z prediktívnej analýzy a zaoberá sa dolovaním – získavaním údajov z textu. Analýza textu vychádza z potreby spracovania ľudského jazyka, ktorý prakticky neexistuje v štruktúrovanej forme. Pod štruktúrovanými údajmi rozumieme dáta zotriedené do riadkov, tvoriacich záznamy a stĺpcov tvoriacich atribúty. Text mining je teda neštruktúrované dolovanie údajov. Text mining bol prvý krát aplikovaný v 90. rokoch pri organizovaní textových dokumentov, za použitia výpočtovo riadenej a užívateľsky vedenej analýzy.

Základom text mining-u je predspracovanie neštruktúrovaných údajov na polo štruktúrované údaje. Takto predspracované údaje je možné ďalej spracovať podľa štandardných analytických techník, akými je klasifikácia alebo zhlukovanie. Predspracované údaje sú taktiež vhodné na hľadanie vzorov medzi týmito údajmi. Hľadanie vzorov sa dá ďalej automatizovať pomocou modelov (Kotu et al., 2015).

3.3 Predspracovanie údajov

Užívateľsky generovaný obsah nesúci sentiment pozostáva z obrovského množstva údajov. Tento obsah je heterogénny, nekonzistentný a obsahuje veľké množstvo „hluku“. Analýza nekonzistentných údajov môže viesť k nepresným výsledkom. Aby bola zabezpečená konzistencia údajov, údaje musia byť predspracované. Dáta totiž môžu obsahovať špeciálne formátovanie, ako napríklad neštandardné číselné formáty alebo formáty pre zápis dátumov. Pomocou procesov, akými sú napríklad tokenizácie, odstraňovania stop slov a normalizácie slov sa získava normalizovaný typ – term.

Vhodné je zbaviť sa aj slov, ktoré nemajú pri text mining-u žiadnu, alebo len veľmi malú hodnotu, akými sú napríklad zámená a predložky. Predspracovaním sa, okrem iného, zvyšuje kvalita údajov a zjednodušuje sa celý proces ich analýzy.

Ďalej sa budeme zaoberať metódami pre efektívne predspracovanie údajov.

3.3.1 Tokenizácia

Prvým krokom morfologickej analýzy je tokenizácia. Úlohou tokenizátora je zabezpečiť konzistenciu textu. Pod pojmom token chápeme arbitrárnu jednotku textu, ktorá rozširuje lingvistický význam pojmu slovo. Za token sa v automatickej segmentácii textu považuje akýkoľvek reťazec znakov medzi dvoma medzerami (bielymi znakmi), aj jednotlivé znaky interpunkcie, ktoré nemusia byť oddelené medzerou od predchádzajúceho alebo nasledujúceho tokenu. Text sa teda z formálneho hľadiska skladá z tokenov a medzier (Garabík, 2004).

Prioritou tokenizácie pri dolovaní z textu je identifikácia podstatných slov. Tokenizácia sa zameriava na text ako celok. Syntaktický analyzátor získava z viet a jednotlivé slová a ukladá ich datasetu. Aj keď sa tento proces môže zdať triviálny, keďže sa predpokladá, že text už je vo forme, ktorú stroj dokáže prečítať. Tokenizácia sa však zameriava na spracovanie interpunkcie a iných špeciálnych znakov, ako sú napríklad zátvorky a spojovníky. Okrem iného sa tokenizáciou skratky a skratkové slová upravujú na základný tvar.

3.3.2 Eliminácia stop slov

Väčšinu textu tvoria slová, ktoré nereprezentujú akýkoľvek sémantický obsah. Spravidla majú len syntaktický význam. Ide o takzvané stop slová. V slovenčine sa medzi ne zaraďujú napríklad spojky, predložky, zámená, niektoré slovesá (byť), častice a pod.

V závislosti od typu textu, môžu byť za stop slová označené iné skupiny slov. Keďže tieto slová sú pri dolovaní z textu nepodstatné, vyraďujú sa. Elimináciou nepotrebných slov sa redukuje množstvo vstupných údajov a zvyšuje sa výkonnosť celej analýzy (Kumar, 2012).

3.3.3 Normalizácia slov

Cieľom normalizácie tokenov je redukcia rôznych slovných tvarov na základnú, spoločnú reprezentáciu. Existujú dva hlavné prístupy normalizácie slov, a to stemovanie a lematizácia.

Stemovanie predstavuje heuristické odstraňovanie sufixov a prefixov, najčastejšie na základe sady pravidiel definovanej pre daný jazyk (Vincúr, 2015).

Lematizácia je technika rozkladania alebo úpravy slov na ich základný tvar – lemu. Lema sa zvyčajne definuje ako „slovníkový" tvar tokenu. Lemy bývajú obsiahnuté v korpusoch, v prípade slovenčiny ide o Slovenský národný korpus (Garabík, 2004).

Výhodou takto získanej lemy je, že všetky jej vyčasované a vyskloňované tvary sú zhodné. Proces stemovania môže viesť k získaniu nesprávneho koreňa slova, čo však nie je pri analýze sentimentu veľkým problémom. Problém môže nastať, ak je systém case-sensitive, čiže veľké a malé písmená považuje za rôzne znaky. V takom prípade sú výsledkom stemovania dve koreňové slová s rovnakým významom (Kumar, 2012).

3.3.4 Vektorový model

Vektorový model (z angl. Vector Space Model) je najčastejšie používanou formou reprezentácie textových dokumentov. Vo všeobecnosti ide o matematický model, ktorý je definovaný ako pevná matematická štruktúra opisujúca určité fyzikálne, biologické, sociálne, psychologické alebo iné abstraktné entity. Pri text miningu sa využíva dátovo centrická úroveň abstrakcie modelovania. Príznakový priestor pre túto reprezentáciu je konštruovaný na základe množiny slov, pričom každý príznak korešponduje s textovou jednotkou v dokumente. Za textovú jednotku je považované slovo, fráza, resp. iná lexikálna jednotka. Daný model vychádza len zo štatistických charakteristík dokumentu, a to najmä z distribúcie pravdepodobnosti jednotlivých slov extrahovaných z dokumentov.

Jej riadky predstavujú jednotlivé termy dokumentu a stĺpce predstavujú dokument. Váhy v jednotlivých bunkách indikujú výskyt týchto elementov v dokumente. Jednotlivé prípady sú modelované vo vektorom priestore pre uľahčenie ich analyzovania (Dubin, 2004).

Vektorový model odstraňuje nedostatky booleovského modelu v tom, že jednotlivé váhy reflektujú nielen výskyt príslušného termu (booleovsky „0“, „1“), ale opisujú aj počet výskytov príslušného termu v dokumente, resp. jeho váhu, teda jeho dôležitosť či významnosť (Halčinová, 2009).

So zväčšujúcim sa počtom dokumentov rastie aj počet dimenzií, čo vedie k mnohodimenzionalite.

3.3.5 Redukcia priestoru

S neustále narastajúcim množstvom údajov, ktoré je potrebné analyzovať, je spojený problém rastúceho množstva rozmerov. S rastúcim množstvom rozmerov sa zvyšujú náklady spojené s uchovávaním údajov. Tradičné štatistické metódy sa môžu pri veľkom počte dimenzií ukázať ako neefektívne. Čím väčší je počet rozmerov, tým je zložitejšie daný algoritmus overiť a vzrastajú nároky na spôsob zobrazenia výsledkov zhlukovania.

Pod pojmom dimenzia údajov rozumieme počet premenných, ktoré sú merané pri každom pozorovaní. Mnohodimenzionálne datasety so sebou nesú množstvo problémov. Jedným z nich je, že nie všetky merané veličiny sú podstatné pre výsledky analýzy. Skúmanie mnohodimenzionálnych datasetov nie je síce nemožné, pred aplikáciou akýchkoľvek modelovacích metód je omnoho výhodnejšie redukovať počet rozmerov.

Výhodou redukcie mnohodimenzionality je, že znižuje výpočtovú náročnosť, zefektívňuje celý proces identifikovania vzťahov medzi údajmi v texte. Okrem iného prichádza aj k znižovania času potrebného na analýzu.

3.3.6 Singulárny rozklad

Podstatou singulárneho rozkladu (z angl. Singular value decomposition - SVD) je zjednodušiť dataset obsahujúci veľké množstvo hodnôt na podstatne jednoduchší, teda dataset obsahujúci mnohonásobne menší počet hodnôt. Singulárny rozklad vychádza z pravidla lineárnej algebry, ktoré hovorí, že regulárna matica A je rozložiteľná na súčin troch matíc. Singulárny rozklad je definovaný takto:

$$A_{m \times n} = U_{n \times r} D_{r \times r} V_{r \times n}^{T}$$

V oblasti text mining-u chápeme maticu A rozmeru m x n, kde m reprezentuje výrazy a n dokumenty. U je matica rozmeru n x r, pričom n reprezentuje dokumenty a r koncepty. D je diagonálna matica rozmeru r x r (Kumar, 2012).

Nenulové prvky matice D predstavujú singulárne čísla matice A, a teda predstavujú odmocniny vlastných hodnôt (z angl. eigenvalues) matice A^TA alebo AA^T . Stĺpce matice U, ľavé singulárne vektory, získame horizontálnym spojením vlastných vektorov (z angl. eigenvectors) matice AA^T. Stĺpce matice V, pravé singulárne vektory, získame horizontálnym spojením vlastných vektorov matice A^TA. Nakoľko vlastné vektory symetrickej matice sú ortogonálne a lineárne nezávislé, môžu byť použité ako bázické vektory, a teda ako reprezentácia multidimenzionálneho priestoru. Takýto rozklad nám umožňuje aproximovať pôvodnú maticu, a teda mapovať stĺpce a riadky pôvodnej matice do k-dimenzionálneho priestoru. Túto k-aproximáciu pôvodnej matice získame tak, že berieme do úvahy len k stĺpcov matice U a k riadkov matice V známe tiež ako redukovaná SVD (Vincúr, 2015).

3.4 Podobnosť dokumentov

V mnohých algoritmoch je zhlukovanie založené na zisťovaní podobnosti, na základe ich vzájomnej vzdialenosti. Dokumenty kolekcie sú najčastejšie reprezentované ako body alebo ako vektory v n-rozmernom priestore. V prípade bodov dokážeme určiť mieru podobnosti na základe ich vzájomnej vzdialenosti, korelácie, alebo asociácie. Miery podobnosti v ideálnom prípade nadobúdajú hodnoty z intervalu <0,1>, pričom 0 vyjadruje maximálnu rozdielnosť objektov a hodnota 1 maximálnu totožnosť.

Miery vzdialenosti predstavujú najpoužívanejšie miery založené na prezentácii objektov v priestore, ktorého súradnice tvoria jednotlivé premenné a využívajú sa v štatistických programoch. Na výpočet vzdialenosti, resp. podobnosti sa slúžia metriky. K najznámejším typom mier vzdialenosti patria Minkowského metriky L1 a L2, známe tiež ako Manhattanova a Euklidova vzdialenosť. Čím väčšia je táto vzdialenosť, tým menšia je podobnosť dokumentov.

Manhattanova vzdialenosť, metrika L1, alebo aj cityblock metrika je metrika, ktorá vracia sumu absolútnych hodnôt rozdielov každej dimenzie dvoch bodov v n-rozmernom priestore. Majme body $U = [x_1, x_2, \dots, x_n]$ a $V = [y_1, y_2, \dots, y_n]$. Manhattanova vzdialenosť – MD, týchto dvoch bodov je určená nasledovne:

$$MD(U,V) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n| \quad (1.)$$

$$MD(U,V) = \sum_{i=1}^{n} |x_i - y_i| \quad (2.)$$

Euklidovská vzdialenosť, metrika L2 je pre tú istú dvojicu bodov U, V je Euklidovská vzdialenosť určená vzťahom:

(1.) $$ED(U,V) = \sqrt{(x_1 - y_1)^2 + (x_2 y_2)^2 + \cdots + (x_n - y_n)^2}$$

(2.) $$ED(U,V) = \sqrt{\sum_{i=1}^{n} \left(x_i - y_j\right)^2}$$

3.5 Zhlukovanie

Zhluková analýza, alebo aj zhlukovanie, je technika na spracovanie polo štruktúrovaných údajov. Je to proces zaraďovania konkrétnych, či abstraktných entít do tried podľa rozličných vlastností. Každá trieda, zhluk (z angl. cluster), je množina dátových objektov, ktoré sú si navzájom podobné a zároveň sú odlišné od objektov v iných zhlukoch. Hlavným cieľom zhlukovania je zjednodušenie veľkého množstva údajov na jednoduchšie zhluky. Zhluková analýza je problémom nekontrolovaného učenia, pretože údaje sú zoskupované len na základe podobnosti, nie konkrétnych stanovených vlastností. Zhlukovanie je teda učenie pozorovaním (Jiawei et al., 2015).

3.5.1 Nehierarchické metódy zhlukovej analýzy

Pri nehierarchických metódach zhlukovej analýzy ide o zaradenie objektov do vopred stanoveného počtu disjunktných zhlukov. Priradenie ku zhlukom je buď jednoznačné, alebo sa počíta miera príslušnosti jednotlivých objektov ku zhlukom. Hlavným kritériom posudzovania príslušnosti ku skupine je vzdialenosť.

Metóda k-priemerov (k-means)

Metóda k-priemerov sa označuje aj ako metóda najbližších ťažísk. Postup je založený na určení najbližšieho ťažiska a objekt s najmenšou vzdialenosťou k ťažisku, nazývaného aj centrum gravitácie, je zaradený do zhluku. Podobnosť zhlukov sa teda meria na základe vzdialenosti na priemernú hodnotu zhluku. Princíp metódy k-priemeru je založený na rozdelení *n* objektov s *m* znakmi do *k* zhlukov tak, aby suma štvorcov medzi zhlukmi bola minimalizovaná. Algoritmus zhlukovania uvedenou metódou pracuje iteratívne a vychádza vždy z iného počiatočného usporiadania zhlukov. Nakoniec sa z možných usporiadaní zhlukov vyberie vhodné riešenie (Halčinová, 2009).

Ak je vopred známy a zadaný počet zhlukov použije sa metóda klasifikácie so známymi ťažiskami zhlukov. Najprv sa na náhodný výber objektov aplikuje hierarchická

analýza a na základe toho sa určia počiatočné ťažiska zhlukov pre následné použitie metódy *k*-priemeru. Na začiatku procesu zhlukovania užívateľ zadá počet zhlukov, ktoré majú byť nájdené a potom sú vytvorené priestorové zhluky nájdením súboru ťažísk zhluku takým spôsobom, že každý objekt je priradený do jedného zhluku.

Existujú prípady, keď nie je vopred známe ťažisko zhluku, vtedy sa pre zhlukovanie použijú metódy klasifikácie objektov s neznámymi ťažiskami zhlukov. Tieto metódy obsahujú rôzne postupy pre výpočet ťažísk. Väčšina z nich vyšetruje údaje niekoľkokrát. Pri správnom ťažisku sú zhluky jednoducho separovateľné. Na začiatku aplikácie tejto metódy sa vyberú objekty, ktoré majú veľkú vzdialenosť medzi sebou a tieto hodnoty použijeme ako počiatočné odhady ťažísk zhlukov. Počet vybraných objektov musí byť rovný počtu zadaných zhlukov. Zhlukovanie pracuje tak, že prvých *k* objektov v údajoch sú vybrané ako dočasné ťažiská. V nasledujúcich krokoch analýzy objekt nahradí ťažisko, keď jeho najmenšie vzdialenosť k ťažisku bude väčšia ako vzdialenosť medzi dvoma najbližšími ťažiskami. Ťažisko, ktoré je bližšie k objektu je preto vymenené. Objekt sa dosadí na miesto ťažiska, keď vzdialenosť z objektu k ťažisku je väčšia než najmenšia vzdialenosť medzi ťažiskom a ostatnými ťažiskami. Postup sa opakuje a znovu sa vymení ťažisko najbližšie k nemu. Počiatočné priblíženie teda ovplyvňuje konečné usporiadanie zhlukov, preto algoritmus zhlukovania náhodne priraďuje pri každom pokuse každý objekt jednému zhluku. Toto usporiadanie je následne optimalizované (Halčinová, 2009).

Metóda k-medoidov (k-medoids)

Medoid, označovaný tiež ako optimálny stred zhluku, je stredný objekt, pre ktorý platí, že priemerná vzdialenosť k ostatným objektom v tomto zhluku je minimálna. Ak sa požaduje počet zhlukov *k*, existuje tiež počet medoidov *k*. Pre nájdenie medoidu sa údaje klasifikujú do zhlukov vždy okolo najbližšieho medoidu. Medoidy a následne zhluky sa vytvárajú na základe nepodobnosti. Výhodou metódy k-medoidov oproti metóde k-priemerov je jej menšia citlivosť na rušivé a nevhodné údaje. Pri aplikácií metódy k-medoidov sa využívajú ďalšie metódy.

Späthova metóda minimalizuje účelovú funkciu premiestňovaním objektov z jedného zhluku do druhého. Postup zhlukovania sa začína počiatočným usporiadaním zhlukov a nájdením lokálneho minima presúvaním objektov zo zhluku do zhluku. Metóda končí, ak sa nepremiestni už žiaden objekt.

Metóda PAM minimalizuje celkovú vzdialenosť *D*. Proces zhlukovania začína nájdením reprezentatívneho súboru *k* objektov. Prvý objekt má najkratšiu vzdialenosť k ostatným objektom, tzn. predstavuje stred. Možné alternatívy polohy *k* objektov sú vyberané iteračným spôsobom. Algoritmus vyhľadáva nezaradené objekty a premiestňuje ich tak, aby sa hodnota *D* znižovala (Halčinová, 2009).

Silueta je proces validácie konzistencie údajov v zhluku. Silueta sa vyčísli vzdialenosť objektu v zhluku k objektom v susediacich, čím umožňuje vizualizovať parametre zhlukov. Vzdialenosti sa merajú v intervale [-1, 1]. Silueta je mierou úspešnej klasifikácie objektov do zhlukov.

3.5.2 Hierarchické metódy zhlukovej analýzy

Myšlienkou hierarchickej zhlukovacej analýzy je vytvorenie binárneho stromu dát, ktorý spája podobné skupiny objektov. Hierarchický systém zhlukov je systém navzájom rôznych neprázdnych podmnožín pôvodnej množiny, v ktorom prienikom každých dvoch podmnožín je buď jedna z nich, alebo prázdna množina. Hierarchický systém zhlukov je charakteristický tým, že vytvára taký rozklad pôvodnej množiny objektov, v ktorom každý z čiastkových rozkladov je zjemnením nasledujúceho alebo predchádzajúceho rozkladu. Existujú dva hlavné typy algoritmov:

Aglomeratívne zhlukovanie predstavuje prístup zdola-nahor, teda jednotlivé zhluky sú iteratívne spájané do väčších celkov. Na začiatku zhlukovania každý dokument predstavuje jeden zhluk, triedu. V nasledujúcich fázach sú spájané najpodobnejšie zhluky, až pokým nie sú všetky objekty zlúčené do jedného zhluku. Každá úroveň výsledného stromu predstavuje rozdelenie objektov. Existuje množstvo kritérií na určovanie podobnosti zhlukov, ako príklad použijeme algoritmus SLCA (Single-linkage Clustering Algorithm), ktorý ako mieru podobnosti dvoch zhlukov používa minimálnu vzdialenosť medzi ich objektmi.

Divízne zhlukovanie predstavuje opačný prístup ako aglomeratívne zhlukovanie, čiže prístup zhora-nadol. Všetky objekty na začiatku predstavujú jeden veľký zhluk. Objekty zhlukov sú následne rozdeľované do dvoch tried na základe určitého kritéria, až kým nie sú zhluky jednoprvkové, čiže obsahujú len jeden objekt. Príkladom tejto skupiny je algoritmus BK-means. Tento algoritmus využíva algoritmus k-means pri rozdeľovaní počiatočného zhluku. Kritický je v tomto prípade výber zhluku, ktorý sa má rozdeliť. Jedným z možným kritérií je výber zhluku s najmenšou kvalitou alebo toho najväčšieho. (Vincúr, 2015).

Výsledkom predspracovania textu normalizovaných dokument. Dokument je konzistentný a neobsahuje hluk, teda žiadne stop slová, neštandardné formáty. Termy sú štruktúrované do zhlukov, na základe ich orientácie, poprípade frekvencie.

3.6 Algoritmy analýzy sentimentu

Pri analýze sentimentu sú tri možné prístupy tvorby a implementácie algoritmov a tie sú:

- prístup založený na pravidlách,
- automatický prístup,
- hybridný prístup.

3.6.1 Prístup založený na pravidlách

Prístup založený na pravidlách využíva systémy analýzy, ktoré vyhodnocujú sentiment a klasifikujú termy na základe vopred stanovených pravidiel, čiže logickej konjunkcie. Tieto pravidlá sa vytvárajú ručne a zväčša majú podobu skriptovacieho jazyka. Pravidlá sú zamerané na identifikáciu subjektivity vyjadreného názoru. Okrem pravidiel na identifikáciu sentimentu, obsahujú tieto systémy aj rôzne algoritmy na predspracovania prirodzeného jazyka, ako vyššie spomenutú lematizáciu, tokenizáciu alebo zhlukovanie. Prístup založený na pravidlách vychádza z lexikálneho prístupu. Dôležitou súčasťou sú lexikóny obsahujúce sentimentové slová, ich význam a k nim patriace synonymá a antonymá (Hussein, 2019).

Implementácia prístupu záložného na pravidlách môže mať takýto postup:

1. Vytvorenie zoznamu sentimentových slov s ich polaritou,
2. Tvorba pravidla na počítanie slov v texte s pozitívnou polaritou,
3. Tvorba pravidla na počítanie slov v texte s negatívnou polaritou,
4. Tvorba pravidla na hodnotenie sentimentu na základe počtu sentimentových slov. Ak skúmaný dokument obsahuje viac slov s kladnou polaritou, návratová hodnota predstavuje pozitívny sentiment, naopak, ak obsahuje text väčší počet negatívnych slov, návratová hodnota je negatívny sentiment. Ak sú počty slov zhodné, výsledkom je neutrálny sentiment.

Na zefektívnenie tohto prístupu a automatizáciu sa používajú algoritmy generujúce logické konjunkcie. Výsledná znalosť má formu klasifikačného pravidla, ktorého

podmienková časť je konjunkciou podmienok a záverová časť je zaradenie do triedy (Halčinová, 2009).

3.6.2 Automatické prístupy

Automatické metódy analýzy sentimentu predstavujú protiklad k prístupom založeným na pravidlách. Tie sa nespoliehajú na ručne vytváraných pravidlách, ale na strojovom učení.

Model vytvorený na základe automatických prístupov analýzy sentimentu zvyčajne predstavuje klasifikačný problém. Klasifikácia je problém, pri ktorom algoritmus podľa vstupný údajov musí zaradiť objekt do triedy, pričom prvky jednej triedy sú si podobné, ale od prvkov ostatných tried sa odlišujú. Objekty sú klasifikované na základe ich polarity (Hussein, 2019).

Na riešenie tohto problému sa používajú algoritmy kontrolovaného učenia. Trénovacie datasety sú poskytované klasifikátoru. Trénovacie údaje obsahujú aj požadovaný výstup – label, na základe ktorého sa klasifikátor naučí predpovedať požadovaný výstup z vstupných dát, ktoré neobsahujú label. Výsledkom je klasifikátor, ktorý priraďuje sentimentovým slovám numerickú hodnotu. Táto numerická hodnota reprezentuje frekvenciu polarity objektu, čiže nie len to, či je term pozitívny, negatívny alebo neutrálny, ale aj do akej miery (Kamble – Itkikar, 2018).

V analýze sentimentu patria medzi najpoužívanejšie algoritmy strojového učenia Naivný Bayesov klasifikátor, Klasifikátor NBCI, Support vector machines a Klasifikátor maximálnej entropie.

Naivný Bayesov klasifikátor

Naivný Bayesov algoritmus je založený na teórii pravdepodobnosti. Využíva Bayesov teorém na trénovanie klasifikátora. Výhodou tohto klasifikátor je použitie jednoduchej pravdepodobnostnej definície triedy. Naopak nevýhodou je predpoklad nezávislosti atribútov opisujúcich klasifikovaný objekt, čo vo väčšine reálnych domén nie je splnené. Aj napriek tejto nevýhode však v doméne klasifikácie textových dokumentov dosahuje naivný Bayesov klasifikátor výsledky kvalitatívne porovnateľné s principiálne zložitejšími algoritmami (Halčinová, 2009).

Naivný Bayesov klasifikátor vypočíta pravdepodobnosť, s akou term patrí do určitej triedy na základe toho, koľko krát sa daný term nachádza v dokumente. Klasifikačný model

využíva extrakciu modelu Bag-of-Words, ktorá ignoruje pozíciu slov v rámci dokumentu (Kamble – Itkikar, 2018).

Klasifikátor NBCI

Klasifikátor NBCI (Naive Bayes Classifier using Itemsets) je metóda induktívneho strojového učenia bola vyvinutá špeciálne na klasifikáciu textových dokumentov. Používa kombináciu Naivného Bayesovho klasifikátora a klasifikátora Itemset. Princíp spočíva vo vyhľadávaní frekventovaných množín termov charakterizujúcich textový dokument, s ktorými sa ďalej pracuje použitím Naivného Bayesovho klasifikátora. Klasifikátor NBCI dosahuje veľmi dobré výsledky pri klasifikácii krátkych textových (Halčinová, 2009).

Support vector machines

Support vector machines (SVM), čiže mechanizmy podporných vektorov využívajú výhody poskytované efektívnymi algoritmami pre nájdenie lineárnej hranice a zároveň sú schopné reprezentovať vysoko zložité nelineárne funkcie. Neurónové siete a SVM sú techniky kontrolovaného učenia na modeli alebo na vzore založenom na cvičných dátach, pri analýze sentimentu sa tento model využívajú na klasifikáciu dát. Údaje v textovej podobe sú ako stvorené pre klasifikáciu za použitia SVM. V texte je mnoho nepodstatných údajov, ktoré však navzájom korelujú a je možné ich prirodzene separovať do tried.

SVM je typický algoritmus strojového učenia, hľadajúci nadrovinu, ktorá v priestore príznakov optimálne rozdeľuje cvičné dáta. Algoritmus slúži k lineárnej separácii dát, aj takých, ktorých separácia nie je možná. Požiadavkou pri hľadaní nadroviny je vzdialenosť medzi nadrovinou a najbližším prvkom jednotlivých tried. Cieľom SVM je nájdenie iba jedného optimálneho lineárneho oddeľovača. Optimálny lineárny oddeľovač poskytuje čo najširšie pásmo medzi ním a pozitívnymi príkladmi na jednej strane a negatívnymi na druhej. V okolí nadroviny je po oboch stranách čo najširší pruh, v ktorom sa nenachádza žiaden iný bod. K popisu tejto nadroviny slúžia najbližšie body, ktorých býva veľmi málo. Tieto body sa nazývajú podporné vektory. Svoje meno získala metóda podľa týchto vektorov. Metóda SVM je binárna, čo znamená, že dáta rozdeľuje do dvoch tried.

V prípade, že dáta v priestore nie sú lineárne separovateľné, transformuje sa pôvodný vstupný priestor pomocou nelineárnej kernelovej funkcie do priestoru s vyššou dimenzionalitou, kde je možné od seba oddeliť triedy lineárne (Korčuška, 2015).

Klasifikátor maximálnej entropie

Klasifikátor maximálnej entropie, známy aj ako klasifikátor MaxEnt je založený na konvertovaní term a ich frekvencie polarity na vektory. Takto zakódované vektory sú použité na výpočet váh každého príznaku objektov, ktoré sa použijú pri klasifikácii budúcich objektov. Klasifikátor maximálnej entropie je založený na podobnom princípe ako Naivný Bayesov algoritmus, na rozdiel od neho však nepredpokladá žiaden vzťah medzi príznakmi (Kamble – Itkikar, 2018).

3.6.3 Hybridné prístupy

Ako už názov napovedá, jedná sa o metódy, ktoré v sebe kombinujú vlastnosti automatických prístupov a prístupov založených na pravidlách. Kvôli odlišnosti oboch typov modelov je obtiažne ich zlúčiť do jedného. Zlučované sú až výsledky týchto modelov.

3.7 Využitie neurónových sietí pri analýze sentimentu

Neurónové siete patria do odboru strojového učenia nazývaného deep learning. Ide o model konštruovaný na základe abstrakcie vlastností biologických nervových systémov. Neurónové siete si našli využitie pri analýze sentimentu najmä kvôli schopnosti nájsť vzory v neštruktúrovaných dátach, a zo zdrojov akými sú obrázky a videá.

Základnou súčasťou umelej neurónovej siete je neurón, ktorý imituje funkciu ľudského neurónu. Neurón pozostáva z N vstupov a M výstupov a dráh W. Dráhy spájajú neuróny v jednotlivých vrstvách. Signály vstupujú do neurónu dráhami. Ak je signál vstupujúci do neurónu, tzv. aktivácia neurónu, vyšší ako prah citlivosti neurónu, neurón sa aktivuje, teda vyšle signál.

Trénovanie neurónovej siete spočíva v predkladaní vzorov s N vstupmi a M výstupmi, na základe ktorých sa sieť snaží prispôsobiť veľkosti váh a prahy citlivosti tak, aby neurón na N-ticu vstupov reagoval M-ticou výstupov. Váhy určujú dráhu, po ktorej sa po aktivácii neurónu signál vydá (Gurney, 1997).

3.7.1 Konvolučné neurónové siete

Konvolučné neurónové siete (z angl. Convolutional Neural Networks – CNN) sú podmnožinou neurónových sietí, ktorá sa ukázala byť zvlášť efektívna pri spracovaní obsahu na obrázkoch a ich klasifikácii. CNN sú schopné rozpoznať na obrázkoch nielen bežné objekty, ľudí a ľudské tváre, ale opísať situáciu, ktorá je na obrázku zachytená. CNN majú svoje využitie aj pri spracovaní prirodzeného jazyka.

CNN pracujú na princípe konvolúcie. Konvolucia je výpočtová operácia, ktorá sa používa na spracovanie dvojrozmerného diskrétneho obrazu. Na obraz sú opakovane aplikované rôzne filtre, tzv. aplikačné masky, ktoré vytvárajú vrstvy neurónovej siete. Aplikačná maska je reprezentovaná tabuľkou hodnôt. Pôvodný obraz je upravený za pomoci konvolučnej masky a to tak, že každý pixel pôvodného obrazu je vynásobený koeficientom hodnôt tabuľky. Pri trénovaní sa CNN učí jednotlivé hodnoty filtrov. Príkladom aplikácie CNN na obraz môžu byť štyri vrstvy. Prvá vrstva sa naučí v obrazci detekovať hrany, druhá sa naučí medzi hranami hľadať tvary, tretia dokáže medzi týmito tvarmi nájsť tváre a štvrtá dokáže skúmať tieto tváre na pôvodnom obraze.

CNN majú svoje využitie aj pri spracovaní prirodzeného jazyka a teda aj extrahovaní sentimentu z textu. CNN sa aplikuje na predspracovaný text, napríklad vektorovým modelom do tvaru matice, kde každý riadok reprezentuje token. Na každý riadok sú opakovane aplikované filtre, podobne ako na jednotlivé pixely obrazu. Úlohou jednotlivých filtrov je identifikovať entity a hľadať vzťahy medzi nimi. Výhodou použitia CNN pri spracovaní textu je rýchlosť, s akou dokážu spracovať text. Za výpočty na hardvérovej úrovni je zodpovedná GPU, keďže konvolucia je používaná najmä v oblasti počítačovej grafiky (Britz, 2015).

3.8 Vyhodnocovanie efektívnosti analýzy sentimentu

Výsledky analýzy sentimentu vo veľkej miere ovplyvňujú rozhodovanie. Chybné, alebo zavádzajúce výsledky vedú k zlým rozhodnutiam. Preto je dôležité sledovať metriky ako precíznosť, priepustnosť, a presnosť týchto nástrojov. Precíznosť predstavuje podiel správne klasifikovaných textov do určitej triedy. Priepustnosť porovnáva počet správne kategorizovaných textov do určitej triedy s počtom textov, ktoré mali do tejto triedy byť kategorizované. Presnosť meria, koľko bolo správne kategorizovaných textov v rámci celého korpusu.

Jedným z najpoužívanejších spôsobov na meranie presnosti modelu je krížová validácia. Pokiaľ sa model trénuje na tej istej trénovacej množine, môže to viesť k prispôsobeniu sa modelu práve tejto jednej množine. Takýto model je náchylný k chybnému analyzovaniu akéhokoľvek iného datasetu. Riešením je práve krížová validácia. Pri krížovej validácií sa vstupné údaje rozdelia na trénovaciu podmnožinu, ktorá tvorí 75 % trénovacieho datasetu a testovaciu podmnožinu, ktorá predstavuje zvyšných 25 %. Klasifikátor je potom trénovaný na trénovacej podmnožine, zatiaľ čo testovacia podmnožina

testuje výkonnosť modelu. Tento proces sa niekoľkokrát opakuje. Výsledkom je funkcia, ktorá je aproximáciou výkonnosti modelu každej iterácie.

Pri tak náročnej úlohe, ako je analýza sentimentu, zvyknú byť miery precíznosti a priepustnosti zo začiatku nízke. Avšak s rastúcim počtom dát spracovaným modelom, stúpa aj efektívnosť modelu (Hussein, 2019).

Záver

Odkedy sa internet stal bežnou súčasťou našich životov, zdieľame na ňom obrovské množstvo informácií. Veľká časť týchto informácií sa týka produktov, ktoré konzumujeme. Dôvodov na zdieľanie osobného názoru je mnoho. Niektorí ľudia chcú prejaviť svoju vernosť značke, iní upozorňujú na chybné alebo inak závadné produkty. Na sociálnych platformách ako je YouTube je množstvo používateľov, ktorých prácou je hodnotiť produkty z určitej oblasti. Netrvalo dlho kým firmy objavili potenciál, ktorý má tento obsah. Zákaznícky sentiment je totiž zdrojom pre podnik nevyhnutných informácií. To podnietilo vznik analýzy sentimentu. Analýza zákazníckeho sentimentu pozostáva z extrakcie názoru z užívateľmi generovaného obsahu, jeho spracovania a poskytovania výsledkov.

Vo svojej práci opisujem veľký vplyv analýzy sentimentu na podniky a jednotlivé informačné technológie používané na jej realizáciu. Po zadefinovaní analýzy sentimentu rozoberám prístupy k nej a metódy na jej realizáciu. Sentiment je totiž možné skúmať na úrovni celého dokumentu, vety, alebo len jednej entity. Keďže je väčšina zákazníckeho sentimentu vo forme neštruktúrovaného textu, venoval som sa jednotlivým technológiám skúmania práve takéhoto obsahu. Na predspracovanie textu sa používa zvyčajne tokenizácia, eliminácia stop slov, normalizácia, redukcia dimenzií, singulárny rozklad, vektorové modelovanie a zhlukovanie. Výsledkom predspracovania takéto textu je normalizovaných dokument. Normalizovaný dokument tvoria normalizované termy a je konzistentný. V takejto podobe je vhodný na použitie metód dolovania textu a aplikáciu analýzy sentimentu na dokument ako celok. V práci rozoberám jednotlivé prístupy analýzy textu, na základe stupňa automatizácie. Sentiment okrem textu zvykne byť vyjadrený pomocou zvukového alebo obrazového záznamu, alebo oboch súčasne. Na spracovanie takéhoto obsahu sa využívajú neurónové siete.

Po aplikácii analýzy sentimentu je možné posúdiť orientáciu vyjadreného názoru, určiť frekvenciu jeho polarity, identifikovať vyjadrené pozitíva a negatíva hodnoteného produktu.

Zoznam použitej literatúry

ALLOUCH, Nada. *Sentiment and Emotional Analysis: The Absolute Difference* [online]. Emojics, 21. 05. 2018. Dostupné na https://www.emojics.com/blog/emotional-analysis-vs-sentiment-analysis/

BING, Liu. *Sentiment Analysis and Opinion Mining*. 1. vyd. Morgan & Claypool Publishers, May 2012. 167 s. ISBN 9781608458851.

BRITZ, Denny. *Understanding Convolutional Neural Networks for NLP*. In: WILDML, 07. 11. 2015. Dostupné na: http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

DUBIN, David. *The Most Influential Paper Gerard Salton Never Wrote*. In: Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign. jar 2004. s 748 – 764.

GARABÍK, Radovan et al. *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu*. In: Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied, 04. 05. 2004. Dostupné na: https://korpus.sk/attachments/publications/2004-garabik-gianitsova-horak-simkova-tokenizacia.pdf

GRIMES, Seth. *Unstructured data and the 80 percent rule,* In: Clarabridge Bridgepoints newsletter 23, 01. 08. 2018. Dostupné na: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

GURNEY, Kevin. *An Introduction to Neural Networks*. Londýn: UCL Press, august 1997. 234 s. ISBN-13: 978-1857285031

HALČINOVÁ, Jana – TREBUŇA Peter. *Podstata nehierarchických metód zhlukovej analýzy.* In Transfer inovácií 13/2009 [príspevok]. Košice, 2009. Dostupné na https://www.sjf.tuke.sk/transferinovacii/pages/archiv/transfer/23-2012/pdf/138-140.pdf

JIAWEI, Han – PEI, Jian – KAMBER, Micheline. *Data Mining: Concepts and Techniques*. 3. vyd. Waltham: Morgan Kaufman, 2012. 701 s. ISBN 978-0-12-381479-1

JING, Ge – MARISOL, Alonso – GRETZEL, Ulrike. *Sentiment Analysis: A Rewiev*. In *Advances in Social Media for Travel, Tourism, and Hospitality* [elektronický zdroj]. Singala, 2018, s. 243 – 261. Dostupné na https://www.researchgate.net/publication/319928524_Sentiment_Analysis_A_Review

KAMBLE, Sangharsjit – ITKIKAR, A. R. *Study of supervised machine learning approaches for sentiment analysis.* In International Research Journal of Engineering and Technology (IRJET). apríl 2018.

KOLKUR, Seema – DANTAL, Gayatri – MAHE, Rena. *Study of Different Levels for Sentiment Analysis*. 2. vyd. Mahashrta: International Journal of Current Engineering and Technology, 2015. 3 s. E-ISSN 2277 – 4106. Dostupné na https://pdfs.semanticscholar.org/01b6/99a840217e3ec5551b692def1e1d25f0ca12.pdf

KORČUŠKA, Robert. *Segmentace tomografických dat v prostředí 3D Slicer* [diplomová práca]. In Vysoké učení technické v Brně, Brno 26. 05. 2015

KOTU, Vijay – DESPANDE. Bala. *Predictive Analytics and Data Mining: Concepts and Practice with Rapidminer*. Elsevier Inc., 2015. 446 s. ISBN 978-0-12-801460-8

KUMAR, Anil. *Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering*. In: International Journal of Engineering Research & Technology (IJERT). Srikakulam, 05. 07. 2012. ISSN: 2278-0181

MACHOVÁ, Kristína. *Strojové učenie v systémoch spracovania informácií*. In Edícia vedeckých spisov Fakulty elektrotechniky a informatiky TU Košice. Košice 2010. Dostupné na: http://people.tuke.sk/kristina.machova/pdf/monoSUvSSI.pdf

HUSSEIN, Sajid. *Deep Learning for Sentiment Analysis*. In: *Medium*, 06. 06. 2019. Dostupné na: https://medium.com/@hussein.sajid7/deep-learning-for-sentiment-analysis-7da8006bf6c1

ŠIMKOVÁ, Mária. *Čo je korpus?*. Bratislava: Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied, 2019. Dostupné na: https://korpus.sk/what.html

STONE, Philip J. et al. *The General Inquirer: A computer Approach to Content Analysis*. 1. vyd. Cambridge: The MIT Press, 1966. 661 s. ISBN: 9780262190305

VINCÚR, Juraj. Identifikácia tém zdrojového kódu [diplomová práca]. In: Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava. máj 2015

YI, Jeonghee et al. *Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques*. In Third IEEE International Conference on Data Mining [konferenčný príspevok]. Melbourne, 2003, 8 s. Dostupné na https://www.researchgate.net/publication/4047551_Sentiment_Analyzer_Extracting_sentiments_about_a_given_topic_using_natural_language_processin