UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

Sentimentalna analiza tweetov

(Sentiment analysis of tweets)

Ime in priimek: Matic Adamič
Študijski program: Računalništvo in informatika
Mentor: doc. dr. Jernej Vičič

Koper, september 2019

Ključna dokumentacijska informacija

Ime in PRIIMEK: Matic ADAMIČ

Naslov zaključne naloge: Sentimentalna analiza tweetov

Kraj: Koper

Leto: 2019

Število listov: 58 Število slik: 14 Število tabel: 20

Število referenc: 49

Mentor: doc. dr. Jernej Vičič

Ključne besede: tweeti, sentiment, podatkovno rudarjenje, strojno učenje, klasifikatorji

Izvleček: V tej diplomski nalogi je predstavljen problem ugotavljanja sentimentalnosti tweetov. Sentimentalnost je lahko napovedana glede na dva razreda: pozitivni in negativni, ali pa na tri razrede: pozitivni, nevtralni in negativni. Predstavljeni so trije raziskovalni članki, ki so problem poskusili rešiti z raznimi metodami podatkovnega rudarjenja in strojnega učenja. Te metode se med seboj razlikujejo v načinu pridobivanja podatkovne zbirke, predobdelavo teksta, izbira klasifikatorjev in ekstrakciji atributov. Predstavljena je tudi nova metoda, ki pri ugotavljanja sentimentalnosti uporabi več različnih slovarjev za določitev sentimentalnosti pa preprosto formulo.

Key words documentation

Name and SURNAME: Matic ADAMIČ

Title of the final project paper: Sentiment analysis of tweets

Place: Koper

Year: 2019

Number of pages: 58 Number of figures: 14 Number of tables: 20

Number of references: 49

Mentor: Assist. Prof. Jernej Vičič, PhD

Keywords: tweets, sentiment, data mining, machine learning, classifiers

Abstract: This diploma tackles the problem of finding sentiment value of tweets. Sentiment can be predicted based on two classes: positive and negative, or based on three classes: positive, neutral and negative. There are three research papers presented in this diploma, which use different data mining and machine learning methods to achieve this goal. These methods differ in acquiring datasets of tweets, text pre-processing, feature extraction and usage of different classifiers. A new method of classification is also presented, which uses a number of dictionaries and uses a simple formula in an attempt to define sentiment.

Zahvala

Rad bi se zahvalil mentorju doc. dr. Jerneju Vičiču za pomoč pri izdelavi diplomske naloge. Zahvaljujem se tudi Apoorvu Agarwalu za deljenje podatkovne zbirke in slovarja krajšav, ki sta bila uporabljena v članku Agarwal et al. [2] in v tej diplomi.

Kazalo vsebine

Kazalo tabel

Kazalo slik

Seznam kratic

URL	enotni naslov vira (angl. uniform resource locator)
API	programski vmesnik (angl. application programming interface)
CSV	format za besedilno datoteko, ki vsebuje z vejico ločene vrednosti (angl. comma separated values)
TXT	standardni format za besedilne datoteke
DAL	slovar vpliva besed na emocije (angl. Dictionary of Affect in Language)
POS	besedne vrste (angl. part of speech)
CPE	centralno procesna enota (angl. central processing unit - CPU)
JVM	angl. Java Virtual Machine

1 Uvod

Dandanes si življenje brez socialnih omrežij težko predstavljamo. Čas, ki jim ga ljudje posvečajo se v zadnjih letih stalno povečuje, še posebej med najstniki [29] [28], število uporabnikov pa vsako leto naraste za 9% [7]. Twitter se uvršča na 6. mesto najbolj popularnih socialnih omrežij, kjer registrirani uporabniki objavijo okoli 511.000 tweetov na minuto [3] [27].

Zaradi velike količine teh objav, lahko uporabimo različne tehnike analiz, s pomočjo katerih lahko razberemo javno mnenje o različnih temah, produktih, ali pa preprosto ugotovimo o kateri stvari ljudje trenutno radi debatirajo. S pametno analizo podatkov lahko ugotovimo o katerih političnih kandidatih se trenutno največ govori oziroma kaj o njih meni javnost. Analizo tweetov pa lahko uporabimo tudi za vpogled v zgodovino objav neke osebe in spreminjanja javnega mnenja o njej skozi čas. Podjetja pa jo lahko uporabijo za poizvedbo mnenja o novo izdanih produktih, s pomočjo katere lahko izboljšajo prodajo.

V prvem poglavju je predstavljeno strojno učenje in podatkovno rudarjenje, ki zajema vse metode, ki so uporabljene v predstavljenih člankih. V drugem poglavju je opisana vsa tehnologija in metode, ki so uporabljene pri naši rešitvi, ki sentimentalno opredeli tweete. V drugem poglavju pa so predstavljeni trije članki, opisana je njihova metodologija in način reševanja problema ugotavljanja sentimentalnosti. Zadnje poglavje pa predstavi našo rešitev. Tam je opisano delovanje programa, rezultati in nadaljnje možne izboljšave.

2 Podatkovno rudarjenje in strojno učenje

Pri pridobivanju, ugotavljanju sentimentalnosti in analizi strukture tweetov so uporabljene razne metode podatkovnega rudarjenja in strojnega učenja. Te nam pomagajo od koraka zbiranja podatkovne zbirke do ugotavljanja sentimentalnosti.

2.1 Strojno učenje

Strojno učenje je zelo širok pojem. Definiramo ga lahko na sledeč način: za računalnik lahko rečemo, da se je naučil nekaj novega, vsakič ko spremeni svojo strukturo, program ali podatke, glede na podane vhodne podatke, na tak način, da se njegovi rezultati v prihodnosti izboljšajo [24].

Strojno učenje je v zadnjih desetletjih postalo zelo popularno. Lahko bi rekli, da so to omogočili hitri napredki v razvoju računalništva, saj te prinašajo več zmogljivosti vsako leto. Po drugi strani, se približno vsako drugo leto količina podatkov podvoji. Kombinacija hitrejših računalniških sistemov in rast količine podatkov so omogočile področjem, kot so strojno učenje, da dokončno dosežejo svoj potencial uporabnosti v vsakdanjem življenju. Metode se uporabljajo marketingu, prevodih besedil in govora, ekonomiji, bančništvu itd [49].

2.1.1 Atributi

Atributi (angl. features) so osnovni deli podatkovnih zbirk, predstavljajo pa izmerljive lastnosti poljubnega objekta, ki ga želimo opisati in analizirati [8]. So ključen rezultat pri procesiranju in pripravi podatkov do mere, kjer se lahko nadaljuje v fazo učenja.

V podatkovnih zbirkah se pojavijo kot stolpci, v nasprotju z instancami, ki jih predstavljajo vrstice [8]. Atributi se lahko pojavljajo v različnih oblikah, kot število, tekst ali kombinacija obeh, pomembna pa je konsistentnost pri vsaki instanci. So pomemben del pri strojnem učenju, saj so to podatki, na katerih se modeli učijo.

2.1.2 Ekstrakcija atributov

Ekstrakcija atributov (angl. feature extraction) je proces, ki iz velike količine podatkov izlušči in povzame pomembne informacije. To tehniko uporabimo, ko je naša začetna podatkovna zbirka prevelika ali vsebuje redundantne podatke. Ta proces je odvisen od problema, ki ga rešujemo in od same podatkovne zbirke, saj kot rezultat vrne manjšo množico podatkov, ki opisujejo originalno podatkovno zbirko.

Cilj procesa je zmanjšanje podatkovne zbirke in procesiranje podatkov do take mere, da so lahko rezultati tega procesa uporabljeni kot vhod za poljuben model strojnega učenja [9].

2.1.3 Klasifikacija

Klasifikacija je proces ugotavljanja oziroma napovedovanja razreda poljubnega nabora povezanih podatkov. Te podatki ponavadi predstavljajo eno instanco podatkovne zbirke. Rezultat procesa pa je določen razred, v katerega klasifikator določi, da instanca spada. Poznamo več vrst klasifikacije, ena izmed njih je nadzorovano učenje (angl. supervised machine learning). Ta potrebuje podatkovno zbirko, ki je vnaprej pravilo klasificirana. Na taki podatkovni zbirki se algoritem lahko uči. Nadzorovano učenje se deli v dve kategoriji:

Klasifikacija (angl. classification) zajema vrsto problemov, pri katerem je velikost množice razredov končna oziroma omejena. Kot primer si lahko predstavljamo algoritem, ki glede na sliko, pove ali je na sliki vidno drevo ali ne. V tem primeru bi bila množica razredov omejena na da ali ne, brez vmesnih vrednosti.

Regresija (angl. regression) zajema vrsto problemov, pri katerih je velikost množice razredov neskončna ali z drugimi besedami, vrednost je zvezno število (angl. continuous value) [50]. Primer takega problema bi bila napoved cene poljubnega izdelka, saj je ta realno število.

2.1.4 Naive Bayes

Naive Bayes je algoritem, ki instancam pripisuje razrede, glede na njihove atribute. Večinoma se uporablja v sentimentalnih analizah, filtrih nezaželene pošte itd. Algoritem predpostavlja, da so vsi atributi med seboj neodvisni (kar pa ni vedno res). Kljub temu, je hiter in zlahka implementiran, osnovan pa je na Bayes teoremu:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{2.1}$$

Kjer je P(A|B) verjetnost dogodka A, kjer se je dogodek B že zgodil. V našem primeru je A vrednost razreda, B pa vrednost atributa.
Obstaja več vrst Naive Bayes algoritmov, med temi so:

- **Multinomial Naive Bayes:** večinoma uporabljen pri klasificiranju dokumentov, kjer so dokumenti označeni glede na njihovo vsebino, na primer: šport, politika, itd. Model napove razred glede na pogostost pojavitev besed v besedilu.
- **Bernoulli Naive Bayes:** podoben Multinomial Bayes, le da so vsi atributi logične vrednosti - 0 ali 1.
- **Gaussian Naive Bayes:** ko se v atributih pojavijo neprekinjene vrednosti (angl. continuous value) oziroma nediskretne vrednosti, algoritem predpostavi, da so vrednosti porazdeljeno po Gaussiovi krivulji [13].

2.1.5 Maximum Entropy

Maximum Entropy je klasifikator, osnovan na atributih. Tako kot Naive Bayes, je večinoma uporabljen pri sentimentalni klasifikaciji besedil, deluje pa po sledeči formuli:

$$P_{ME}(c|d,\lambda) = \frac{exp[\sum_i \lambda_i f_i(c,d)]}{\sum_{c'} exp[\sum_i \lambda_i f_i(c,d)]} \tag{2.2}$$

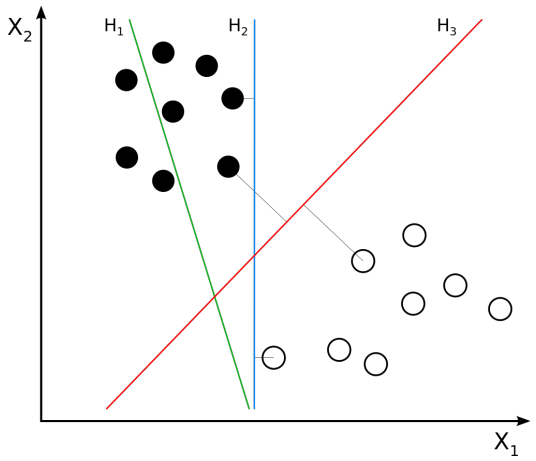
Kjer je c razred, d je instanca v podatkovni zbirki, λ pa je težni vektor.

Težni vektor λ je številka, ki določa pomembnost atributa pri nalogi klasificiranja, kjer visoka vrednost določa, da je atribut zelo pomemben [14]. Tako ta metoda za vsak atribut izračuna njegovo pomembnost in jih uporabi pri nalogi klasificiranja. Za razliko od Naive Bayes klasifikatorja, Maximum Entropy ne predpostavlja tega, da so vsi atributi med seboj neodvisni [39].

2.1.6 Support Vector Machines

Metoda podpornih vektorjev (angl. support vector machines) je algoritem, ki ustvari eno ali več ravnin v poljubno dimenzionalnem prostoru. Te so ustvarjene glede na atribute instanc, z njimi pa jim določi razred. Originalno, je SVM binarni klasifikator, obstajajo pa metode, ki so zmožne tudi večrazredne klasifikacije. Ravnine so postavljene na tak način, da so kar se da oddaljene od katerekoli instance in da obenem še

vedno kar se da dobro razmejujejo instance v dva ali več razredov. Tem se nato določi razred, glede na njihovo pozicijo in ravnino.



a

Slika 1: Primeri treh ravnin, ki ločujejo instance podatkov v dva razreda. polne kroglice predstavljajo prvi razred, prazne pa drugi razred

[a]Avtor: ZackWeinberg, narejeno po PNG verziji uporabnika: Cyc - Datoteka je bila pridobljena iz: Svm separating hyperplanes.png, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=22877598`

V primeru na sliki 1 so ravnine linearne funkcije v 2D prostoru. Premica H1 ne loči med razredoma, premica H2 loči med razredoma, vendar ne optimalno. H3 loči med razredoma in ima najbolj oddaljena od instanc obeh razredov.

Metoda SVM se pogosto uporablja pri kategorizaciji raznih besedil, slik, ročno napisanih znakov itd. [45].

2.1.7 Penn Treebank

Penn Treebank je množica oblikoslovnih oznak, ki jih lahko pripisujemo besedam. Te oznake zajemajo besedne vrste, kot so: glagol, pridevnik, samostalnik, itd. Označujejo tudi stvari kot so: spol, čas, itd. [1]. Primeri oznak so podani v tabeli 1.

2.1.8 Partial Tree Kernel

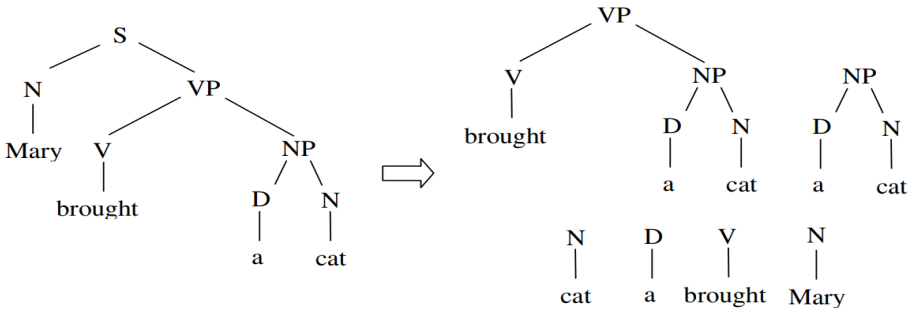
Tree Kernel so podatkovne strukture, ki se pogosto uporabljajo pri procesiranju besedil. Z drevesno strukturo opisujejo zgradbo povedi, kjer se ta tokenizira po presledkih. Vsak žeton (angl. token) je tako beseda, te se nahajajo v listih, medtem, ko vozlišča

Tabela 1: Primeri TreeBank oznak za označevanje angleških besed

Število oznake	Oznaka	Opis
2.	CD	Cardinal number (število)
8.	JJR	Adjective, comparative (pridevnik, primernik)
17.	POS	Possessive ending (izražanje svojilnosti)
20.	RB	Adverb (prislov)
24.	SYM	Symbol (simbol)

opisujejo sintaktično strukturo povedi, besede pa so oblikoslovno označene. Označene so z PennTreebank naborom.

Med drevesnimi strukturami je možna primerjava, ki določa podobnost med dvema stavkoma. To je doseženo s primerjanjem vseh možnih poddreves med njima. Na sliki 2 je podan primer stavka, na desni strani pa so vidna poddrevesa [21].



Slika 2: Primer procesirane povedi: "Mary brought a cat" in njena poddrevesa.

2.1.9 N-grami

N-grami so vsa zaporedja dolžine n objektov, ki se pojavijo v poljubnem besedilu. Te objekte lahko predstavljajo besede, črke, oznake, številke, itd. N-gramom dolžine 1 pravimo unigrami, dolžine 2 bigrami in dolžine 3 trigrami [43]. V besedilu, kjer je 5 besed, ki so med ločene s presledki, bi tako z metodo trigramov dobili množico treh elementov, kateri bi vsebovali po tri besede.
Primer takega besedila: "Danes je zelo lepo vreme".
Rezultat metode so trigrami: {"Danes je zelo", "je zelo lepo", "zelo lepo vreme"}.

2.1.10 Morfo-sintaktično označevanje

Morfo-sintaktično označevanje (angl. part-of-speech tagging) je proces slovničnega označevanja besed v besedilu. Proces besede kategorizira glede na kontekst uporabe

in jih opredeli v skupine, kot so pridevniki, glagoli, simboli itd. Pripisuje jim lahko tudi stvari kot so čas, število, spol itd. Naborom oznak pravimo morfo-sintaktični deskriptorji (angl.moprho-syntactic descriptors). Te množice oznak so različne glede na jezik besedila [12].

2.2 Podatkovno rudarjenje

Podatkovno rudarjenje je proces, pri katerem se odkriva znanja iz poljubne podatkovne zbirke. Proces je sestavljen iz več korakov, vsak mora biti pazljivo načrtovan in izveden. Koraki zajemajo (niso pa edini): zbiranje podatkov, predobdelavo podatkov, normalizacija podatkov, analiza podatkov, odkrivanje vzorcev in pridobivanje novega znanja. V vsakem koraku se uporabi različne tehnike obdelave podatkov [33].

2.2.1 Primer podatkovnega rudarjenja

Znan primer uporabe podatkovnega rudarjenja so trgovine. Kot primer lahko vzamemo kartice, ki jih različne trgovine ponujajo svojim kupcem. Kartice svojim uporabnikom doprinesejo nižje cene različnim izdelkom, z njimi pa trgovine avtomatizirajo proces beleženja kupcev in njihovih nakupov. Tako lahko s pomočjo podatkovnega rudarjenja in strojnega učenja iz velikih količin nabranih podatkov pridobijo nova znanja o navadah njihovih strank. S temi informacijami si lahko pomagajo pri odločanju časovnih intervalov, kjer znižajo cene določenim produktom [35].

2.2.2 Distant supervision

Da bi naš algoritem strojnega učenja oziroma model naučili opaziti razliko med različnimi razredi, potrebujemo podatkovno zbirko, na kateri se bo algoritem učil. Distant supervision je tehnika zbiranja take podatkovne zbirke. Ker ponavadi želimo imeti čim večjo in čim bolj uravnovešeno podatkovno zbirko, je dobra ideja, da za zbiranje teh podatkov uporabimo tehniko, ki proces avtomatizira. Glavna prednost teh tehnik je seveda hitrost zbiranja, saj ne potrebujejo človeškega posredovanja. Metoda nam instance podatkov avtomatsko klasificira v določene razrede. Ta način doprinese napake (angl. noise) v našo podatkovno zbirko, saj lahko hitro pride do nekega odstotka napačno klasificiranih instanc, ta je lahko večji ali manjši, kar je treba pri učenju modelov upoštevati [10].

3 Uporabljena tehnologija in programska oprema

Za pomoč pri ugotavljanju sentimentalnosti so uporabljene različne programske opreme in tehnologije.

3.1 Java

Java je zelo razširjen in priljubljen splošno namenski programski jezik, ki se je prvič pojavil leta 1995. Največja prednost in slogan jezika je "Write once, run anywhere.". Ideja jezika je, da se program napiše enkrat in uporabi na katerem koli sistemu, ki je sposoben poganjati Java SE Development Kit oziroma JDK [34].

Programski jezik je:

- objektno usmerjen (angl. object oriented), ki snovi na razredih
- večniten (angl. multi-threaded), podpira istočasno izvajanje računalniške kode
- neodvisen od platforme (angl. platform independent), ki izvršuje program

3.1.1 Predmetno usmerjen jezik

Vsa programska koda nekega programa se nahaja znotraj razredov. Razredi vsebujejo procedure oziroma funkcije, ki izvršijo del kode in pa podatkovna polja oziroma atribute, kjer so shranjene vse informacije, ki jih predmet (angl. object) potrebuje za njegovo delovanje. Namen razredov je, da nudijo način pisanja programske kode, ki je pregledna in razčlenjena. To pomeni, da vsak razred skrbi in rešuje en problem, kar se navezuje na drugo idejo predmetno usmerjenih jezikov, ki je zmožnost ponovne uporabe in modularnost programske kode. [44].

3.1.2 Večnitno

Večnitnost je zmožnost izvrševanja različne programske kode istočasno. V Java programskem jeziku je to doseženo s promočjo niti (angl. threads). Te so napisane kot

razredi in so zadolžene za izvajanje vnaprej napisane kode. Glavni vzrok in prednost za pisanje večnitne programske kode je hitrost, saj istočasno izvajanje kode v nekaterih primerih doprinese hitrejše izvajanje programa [42].

3.1.3 Neodvistnost platforme

Platformo sestavljata operacijski sistem in strojna oprema. Java doseže neodvisnost z uporabo izvajalnega okolja JVM. Vsak Java program se prevede v "byte code", kjer so zapisani ukazi za virtualni CPE. JVM si lahko predstavljamo kot virtualni računalnik. Zadolžen je za izvajanje prevedenih ukazov. To doseže tako, da ponovno prevede ukaze, ki so zapisani v "byte code" v strojni jezik, ki je odvisen od fizične CPE enote. Ker obstaja več različnih računalniških sistemov z različnimi CPE enotami, obstaja tako tudi več različnih JVM okolj. Ta so narejena za različne platforme in so zadolžena za prevod v strojno kodo in izvajanje prevedenih programov [38] [6].

3.2 Twitter

Twitter je priljubljeno socialno omrežje, kjer registrirani uporabniki objavljajo kratka sporočila, imenovani "tweeti". Twitter se je prvič pojavil leta 2006, začetna ideja storitve pa je bila komunikacija preko SMS sporočil z omejeno skupino ljudi. Socialno omrežje se je že leto dni zatem začelo širiti, okoli leta 2009 pa je prišlo do eksponentne rasti uporabnikov, ta se je umirila leta 2015 [47]. Twitter se uvršča med najbolj popularne socialna omrežja, s približno 330 milijonov aktivnih uporabnikov v Juliju 2019 [32].

Uporabniki lahko med seboj komunicirajo s tweeti, ki so po prevzetem javni. Tweete lahko uporabniki objavljajo tudi samo svojim sledilcem (angl. followers), to so drugi uporabniki, ki so naročeni na njihove objave. Registrirani uporabniki lahko druge objavljene tweete "všečkajo", re-tweetajo – citirajo tweet v svojem tweetu in po želji dodajo še svoje besedilo, ali komentirajo oziroma odgovorijo na specifičen tweet. Komunikacija je mogoča tudi s privatnimi sporočili, imenovanimi direktna sporočila (angl. direct message) ali na kratko DM. Ta sporočila so privatna in vidna le pošiljatelju [31] [20].

3.2.1 Tweets

Tweeti so osnovni način komunikacije na Twitter socialnem omrežju. Sporočila so bila do 7. novembra 2017 omejena na 140 znakov, od takrat naprej pa so lahko dolga do 280 znakov. Ta kratka sporočila pa niso omejena le na tekst, saj lahko vsebujejo tudi

videe ali slike. Te se od leta 2017 ne štejejo k omejitvi znakov [47].

Zaradi sproščene narave omrežja in omejitve dolžine sporočil, se v tweetih pojavljajo neformalni izrazi, okrajšave, slang, emotikoni, emojiji in Twitter specifične besede, kot so "hashtagi", "@" oznake, itd. Sporočila niso imuna napačno črkovanim besedam ali besedam, ki vsebujejo večkrat ponovljene črke, kar je dobro upoštevati pri kakršnih koli analizah.

Slika 3: Primer tweeta, ki ga je objavil ameriški dnevnik New York Times

Emotikon

Emotikoni so zaporedje znakov, ki skupaj ponavadi predstavljajo izraz obraza. Prvič so se pojavili leta 1982, kot oznake za šale, uporabljen :-), in ne-šale, uporabljen :-(. Popularnost so dosegli v devetdesetih, kjer so bili pogosto uporabljeni v takrat mladem internetu in telefonskih sporočilih SMS [16].

Pogosto se nahajajo na koncu kratkih sporočil, kjer ustvarjatelj sporočila opiše trenutna čustva ob pisanju sporočila. Emotikoni so v uporabi še danes, čeprav jih počasi zamenjujejo novejši emoji znaki.

Emoji

Emojiji, tako kot emotikoni, ponavadi predstavljajo obrazne izraze. Za razliko od emotikonov, emojiji niso sekvence znakov temveč majhne slike, ki so v besedilu pred-

stavljeni kot en znak.

Emojiji so se prvič pojavili leta 1997 na Japonskem, kot 12x12 pikslov velike slike, po celotnem svetu pa so se razširili šele po letu 2003 [16]. Originalno je bilo izdanih 90 emojijov, zanimivo pa je, da med njimi ni bilo nobenih takih, ki bi izražali čustva ali obraznih izrazov – te so prišli šele leta 2003 [4]. Emojiji se še vedno posodabljajo in razvijajo, nazadnje je bilo Junija 2018 dodanih 157 novih, skupno pa jih je že skoraj 3000 [11].

Hashtag

Hashtagi (prevod: ključnik) so vrste metapodatkovnih oznak (angl. tag). Oznake so dodeljene poljubnim informacijam, na primer besedam v tekstu, slikam, itd. Te oznake poljubne stvari opisujejo, z njimi pa je omogočeno hitrejše in lažje iskanje v večjih zbirkah podatkov [46].

Hashtagi so pogosto uporabljeni na socialnih omrežjih, med njimi je tudi Twitter, kjer so se prvič pojavili leta 2007. Uporabniki hashtage ponavadi uporabljajo na koncu svojih kratkih objav, te pa se navajajo na sporočilo in označujejo njegovo temo. Vedno se začnejo z znakom "#", kateremu nato sledi beseda ali več besed, ki niso ločene s presledkom, temveč je vsaka beseda zapisana z začetno veliko črko (na primer #NewYear), ni pa nujno [23].

Target

Cilj (angl. target) je znak za označevanje drugih uporabnikov Twitter omrežja. To je omogočeno s pomočjo "@" znaka, po katerem sledi uporabniško ime naslovnika, kar omogoča uporabnikom specifično naslavljanje sporočil drugim uporabnikom. Njihov namen je, da naslovljene uporabnike o objavi obvesti.

Krajšave in sleng

Zaradi omejitve dolžine besedila, ki ga je mogoče objaviti, se na Twitter socialnem omrežju uporablja veliko okrajšav in kratic. Te niso specifično uporabljene samo na Twitter omrežju, vendar na internetu na splošno. Zaradi sproščene narave samih objav, se seveda pojavlja tudi veliko slenga.

URL povezava

Enotni naslov vira (angl. uniform resource location) je referenčni naslovi do spletnih virov, ki specificirajo njihov naslov v računalniškem omrežju. Kličemo jih tudi spletni

naslov ali URL. Sledijo specifični sintaksi, ki jim določa strukturo [48].

3.2.2 Twitter API

Programski vmesniki (angl. application programming interface) preko računalniških omrežij omogočajo komunikacijo med različnimi aplikacijami [22]. Twitter ponuja svoj API, katerega lahko uporabljajo registrirani uporabniki. Ima velik nabor funkcij, kot so objavljanje, iskanje, sortiranje, pretakanje (angl. stream) tweetov v realnem času itd. [36].

3.3 Twitter4J

Twitter4J je neodvisna knjižnica napisana v Javi, ki omogoča komunikacijo s Twitter API. Je popolnoma kompatibilna s Twitter API verzijo 1.1. Nastavi se jo z uporabo internetnega protokola OAuth, ki omogoča varno komunikacijo med uporabnikom in strežnikom. Potreben je tudi Twitter uporabniški profil, z nastavljeno in odprto aplikacijo, ki ponuja dostop do Twitter API [37].

Sama knjižnica ponuja veliko funkcionalnosti, med njimi so: objavljanje, brisanje, iskanje in prenašanje naključnih tweetov, itd. [5].
Glavni razlog za uporabo knjižnice v diplomski nalogi je olajšana komunikacija s Twitter API in zbiranjem podatkov.

3.4 Datoteke CSV

Format datoteke, kjer so vrednosti med seboj ločene z vejico (angl. comma-separated values) je vrsta datoteke v kateri so shranjeni podatkovni zapisi. Te so urejeni kot navadno besedilo, v katerem vsaka vrstica predstavlja en podatkovni zapis. Vsak je lahko sestavljen iz več vrednosti, te pa so lahko pojavijo v obliki teksta, števil, znakov, itd. Vrednosti so med seboj ločene z vejico, v vsaki vrstici pa so vrednosti urejene v enakem zaporedju. To omogoča enostaven način branja vrstic in njihovih vrednosti v katerem-koli programskem jeziku. Format je tako enostaven za branje s strani računalnika kot tudi s strani uporabnika, saj je podprt s strani praktično vseh programov za urejanje besedil.

3.5 Slovar vplivnosti besed

Slovar vplivnosti besed (angl. dictionary of affect) je zbirka besed, ki meri prisotnost emocij, ki jih čutijo ljudje ko slišijo ali vidijo določeno besedo. Slovar vsebuje 8.742 besed, vsaki so pripisane tri vrednosti:

- "pleasantness" (prijaznost), ki meri prijetnost besede
- "activation" (občutek), ki meri kako močno beseda vzbuja občutke
- "imagery" (predstavljivost), ki meri ali si je besedo lahko predstavljati

Besede so bile izbrane iz korpusa 1.000.000 besed, ki je bil ustvarjen iz časopisov in revij iz leta 1960. Vsaka beseda, ki se je pojavila vsaj desetkrat in se je nahajala vsaj v dveh različnih virih, je bila dodana v slovar. Ta je bil nato primerjan z štirimi tipi besedili, ki so jih ustvarili ljudje in z veliko zbirko mladinske literature. Dodane so bile tudi unikatne besede iz teh besedil. Tipi besedil so bili: povzetki zgodb s strani študentov, intervjuji o nasilju, opisi čustev s strani mladostnikov, in študentski eseji.

Vsi podatki o "pleasantness", "imagery" in "activation" za vse besede so bili zbrani v drugi polovici devetdesetih. Pri pridobivanju teh podatkov je sodelovalo 200 ljudi, izmed katerih je bila večina študentov obeh spolov. Vsaki kategoriji je pripisana ena izmed treh vrednosti: 1, 2 ali 3.

- "pleasantness": 1 pomeni neprijetno, 3 pomeni prijetno
- "activation": 1 pomeni pasivno, 3 pomeni aktivno
- "imagery": 1 pomeni, da beseda ne prikliče slike na misel, 3 pomeni, da jo zlahka prikliče

Vsaka beseda je bila ocenjena približno osemkrat za kategoriji "pleasantness" in "activation" ter približno petkrat za "imagery". Statistika je podana v tabeli 2, primeri besed pa so vidni v tabeli 3.

Tabela 2: Statistična analiza besed

	Povprečje	Standardna deviacija
Pleasantness	1,84	0,44
Activation	1,85	0,39
Imagery	1,94	0,63

Slovar je bil testiran na 16 novih, naključno izbranih vzorcih besedil, prav tako kot na korpusu s 350.000 angleškimi besedami. Pri teh slovar pokrije 90% vseh najdenih besed [41] [40].

Tabela 3: Primer besed v slovarju DAL

Beseda	Pleasantness	Activation	Imagery
absurd	1,0000	1,5000	2,2000
accept	2,4444	1,5000	1,8000
multiply	1,4000	2,0000	1,6000
rash	1,7143	1,8750	2,8000

4 Dosedanje metode klasifikacije tweetov

Izvedenih je bilo že kar nekaj raziskav in objavljenih veliko člankov. Primeri teh člankov so Agarwal et al. [2], Pak and Paroubek [25] in Go et al. [15]. Opisane so njihove metode reševanja problema in njihovi rezultati.

4.1 Agarwal et al., 2011

4.1.1 Uvod

Ustvarjena so bila različna orodja za ekstrakcijo in procesiranje atributov. Z različnimi modeli so tweeti klasificirani v tri razrede: pozitivni, nevtralni, negativni in v dva razreda: pozitivni ter negativni. Eksperimentirajo s tremi modeli:

- unigram model (model na osnovi besed)
- feature based model (model osnovan na atributih)
- tree kernel based model (model osnoval na drevesni predstavitvi teksta)

Pri modelu osnovanem na atributih, so uporabili atribute, ki so bili uporabljeni že v prejšnjih raziskavah, dodajo pa še nekaj novih. Skupno ima model približno 100 atributov. S tree kernel modelom predstavijo novo drevesno strukturo, ki opisuje strukturo tweeta. Unigram model pa je uporabljen kot referenčni model, ker se je v prejšnjih raziskavah izkazal kot dober model za sentimentalno analizo tweetov. Eksperimentirali so tudi s kombinacijami različnih modelov, in sicer:

- Unigram z njihovimi novimi atributi
- Tree Kernel z njihovimi novimi atributi

4.1.2 Podatkovna zbirka

Pridobijo 11.875 tweetov iz komercialnega vira. Zbirka vsebuje klasificirane tweete v treh razredih: pozitivni, nevtralni in negativni. Tweeti so bili zbrani s pomočjo Twitter API, pri zbiranju pa niso bili uporabljeni nobeni kriteriji. Vsi tweeti, ki niso bili

originalno v angleškem jeziku, so bili prevedeni s pomočjo Google Translate. Jezikovno poenoteno zbirko so nato ročno klasificirali v omenjene tri razrede, z dodatnim razredom "junk". V ta razred so bili opredeljeni tweeti, ki jih niso mogli klasificirati – izkazalo se je, da je bila večina takih tweetov tistih, ki so bili prevedeni.

Iz novonastale zbirke odstranijo vse tweete, ki so bili označeni kot "junk", teh je bilo okoli 3.000. Na koncu zberejo uravnovešeno zbirko 5.127 tweetov, kjer je 1.709 tweetov iz vsakega izmed treh razredov.

4.1.3 Uporabljeni slovarji

Uporabijo več različnih slovarjev, ki besedam v besedilu pripisuje numerične vrednosti ali pa z njihovo pomočjo odkrivajo specifične besede, za katere so menili, da so pomembne.

Slovar vplivnosti besed

Veliko atributov je osnovanih na sentimentalnosti besed. Besedam pripišejo sentimentalno vrednost, ki je realno število. Ta jim določa prijetnost (angl. pleasantness). Pri tem si pomagajo z DAL slovarjem, ki ga razširijo z WordNet. Če beseda, ki ji pripisujejo prijaznost, ni najdena v DAL slovarju, poiščejo njene sopomenke v WordNet. Če je sopomenka najdena v DAL, se originalni besedi pripiše njena prijaznost. V primeru, da beseda niti njene sopomenke niso najdene v DAL slovarju, beseda ostane neoznačena.

Besedam v DAL so pripisane vrednosti med 1 in 3, te vrednosti normalizirajo med 0 in 1. Postavijo meje, ki razmejujejo besede v pozitivne, negativne in nevtralne razrede. Če ima beseda vrednost nad 0,8 je pozitivna, če ima vrednost pod 0,5 pa je negativna. Ostale vrednosti spadajo v nevtralni razred.

S to tehniko, so uspešno pripisali sentimentalno vrednost 88.9% besedam. Izmed teh je 81,1% besed bilo takoj najdenih v DAL slovarju, ostalih 7.8% pa je sopomenk, ki so bile naknadno pridobljene iz WordNet.

Slovar emotikonov

Emotikon slovar nabrane emotikone opredeli v pet razredov, ki določajo stopnjo sentimentalnosti. Sestavljen je iz 170 različnih emotikonov, seznam pa je pobran iz Wikipedije.

Tabela 4: Primeri iz slovarja emotikonov

Emotikon	Sentimentalnost
:D C:	Zelo pozitivna
:-) :) :o) :] :3 :c)	Pozitivna
:|	Nevtralna
:-(:(:c :[	Negativna
D8 D; D= DX v.v	Zelo negativna

Slovar krajšav

Slovar okrajšav je sestavljen iz različnih krajšav in slenga, ki ga pogosto srečujemo na internetu. Vsebuje 5.184 vnosov, kjer je vsaki okrajšavi pripisana knjižna beseda ali izraz. Primeri so podani v tabeli 5.

Tabela 5: Primeri iz slovarja okrajšav

Krajšave	English expansion
gr8, gr8t	great
lol	laughing out loud
rotf	rolling on the floor
bff	best friend forever

Slovar funkcijskih besed

Funkcijske besede (angl. stop-words) so besede, ki se pogosto pojavljajo v danem jeziku in ne doprinesejo veliko informacij pri procesiranju in analizi dokumentov. Take besede se ponavadi iz besedila odstrani ali pa označi za nepomembne zaradi pohitritve procesiranja. Slovar pridobijo iz spletnega vira. Primeri besed so podani v tabeli 6.

4.1.4 Predobdelava tweetov

Vsi tweeti so podobdelani na sledeč način:

- Zamenjajo vse emotikone z njihovo sentimentalno vrednostjo iz slovarja emotikonov
- Zamenjajo vse URL naslove z oznako $||U||$
- Zamenjajo vse "target"(na primer: @blabla) z oznako $||T||$

- Zamenjajo vse zanikalne besede z (na primer: "no", "not", "n't") z oznako NOT
- Okrajšajo vse besede, ki vsebujejo ponavljajoče zaporedje črk, na primer: "coo-ool" se okrajša v "coool". Pri tem obdržijo en odvečen ponavljajoči znak, da se kasneje loči med "raztegnjenimi" besedami in normalnimi besedami.

Tabela 6: Primeri iz slovarja funkcijskih besed

Beseda
abroad
almost
come
greetings
him
little
specify
to
under
etc

Tokenizacija

Za tokenizacijo tweetov uporabijo Stanford Tokenizer Klein in Manning [17]. Vse besede, ki niso bile predobdelane in so najdene v WordNet so štete kot angleške besede. Za označevanje ločil so uporabljene standardne oznake iz Penn Treebank nabora. Vse ostale neoznačene besede so štete kot tuje besede ali ostali simboli (na primer: "coool" in "zzz"). Tabela 7 prikazuje statistično analizo vseh žetonov oziroma besed.

4.1.5 Atributi

Atribute opredelijo v več različnih skupin, ki jih uporabijo pri napovedovanju razredov. Skupno določijo 50 različnih tipov atributov, ki jih določijo s pomočjo celotnega tweeta in iz zadnje tretjine tweeta. Skupaj tako ustvarijo 100 atributov za vsak tweet. Tabela 8 predstavlja vse vrste atributov, te poimenujejo "senti-atributi". Razdelijo jih v tri večje skupine:

- Atributi, ki štejejo prisotnosti različnih oznak, njihova vrednost je naravno število
- Atributi, ki zajamejo sentimentalnost besed iz DAL slovarja, njihova vrednost je realno število

- Atributi, ki beležijo prisotnost ločil in z veliko napisan tekst, njihova vrednost je 0 ali 1

Vsako izmed teh skupin razdelijo v dve kategoriji:

- Polarni atributi
- Nepolarni atributi

Atribut je polarni, če lahko njegovo polarnost izračunamo s pomočjo DAL ali emotikon slovarja. V drugem primeru je nepolarni, to so tisti, ki nimajo pripisane sentimentalne vrednosti. Te dve kategoriji razdelijo še na dve pod-kategoriji:

- POS
- Ostali

POS kategorija se nanaša na atribute, ki zajemajo statistiko o POS oznakah.

Tabela 7: Statistična analiza vsebine tweetov po tokenizaciji

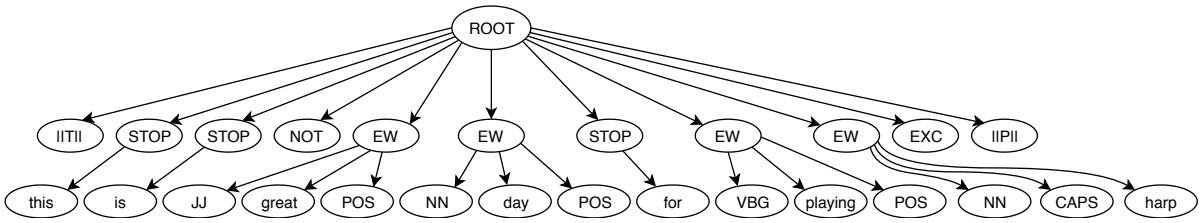
Število žetonov	79.152
Število funkcijskih besed	30.371
Število angleških besed	23.837
Število ločil	9.356
Število besed z veliko začetnico	4.851
Število hashtagov	3.371
Število klicajev	2.228
Število zanikalnih besed	942
Število ostalih žetonov	9047

4.1.6 Tree kernel

Ustvarjena je drevesna predstavitev tweeta za združevanje različnih kategorij atributov. Za določanje podobnosti dreves je uporabljen Partial Tree Kernel. Ta izračuna podobnost dveh dreves s primerjavo vseh njunih poddreves. Primer drevesa je viden na sliki 4.

Tabela 8: Tabela senti-atributov. Razdeljeni so v več nivojev in skupin, ki jim določajo vrsto števila. Atribut f_8 zajema vsoto sentimentalnih vrednosti besed, ki so se pojavile pred besedo, ki je bila označena kot samostalnik (NN), prislov (RB), pridevnik (JJ) ali glagol v osnovni obliki (VB).

$\mathbb{N}$	Polarni	POS	Število (+/-) POS oznak (JJ, RB, VB, NN)	f_1
		Ostalo	Število zanikalnih besed, pozitivnih besed, negativnih besed	f_2
			Število zelo-negativnih, zelo-pozitivnih, pozitivnih in negativnih emotikonov	f_3
			Število (+/-) hashtagov, besed z veliko začetnico, besede s klicaji	f_4
	Nepolarni	POS	Število JJ, RB, VB, NN	f_5
		Ostalo	Število slengovskih izrazov, latinskih znakov, besed v iz slovarjev, besed	f_6
			Število ciljev (@), hashtagov, URL naslovov, znakov za novo vrstico	f_7
$\mathbb{R}$	Polarni	POS	Seštevek polaritet prejšnjih besed za POS oznake JJ, RB, VB, NN	f_8
		Ostalo	$\sum$ polaritete vseh besed	f_9
	Nepolarni	Ostalo	Procent besedila napisanega z velikimi črkami	f_{10}
$\mathbb{B}$	Nepolarni	Ostalo	Klicaj ali z veliko napisan tekst	f_{11}



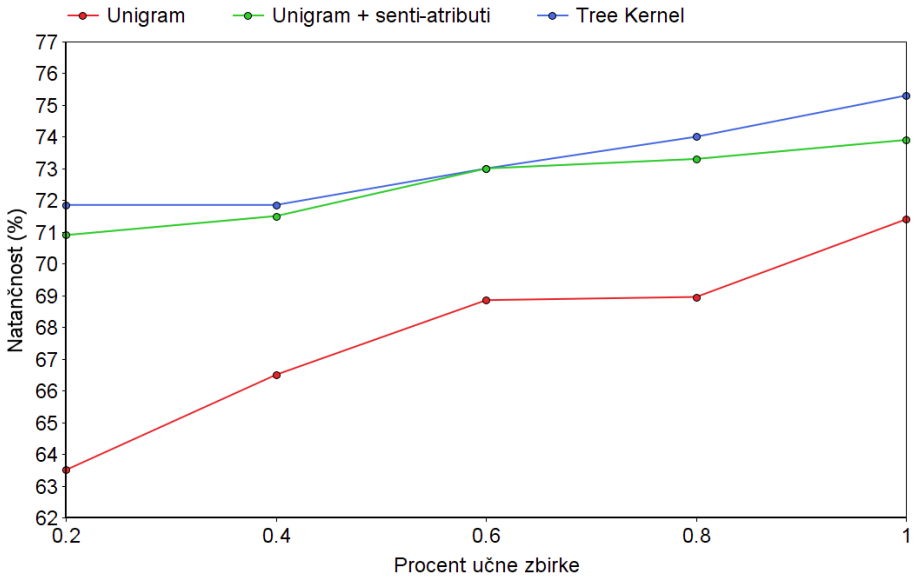
Slika 4: Primer drevesne strukturo po procesiranju tweeta: "@Fernando this isn't a great day for playing the HARP! :)"

4.1.7 Eksperimentiranje in rezultati

Predstavljeni so eksperimenti, kjer uporabijo določen model in s časom dodajajo različne skupine atributov, da bi videli kako različni atributi vplivajo na rezultate modelov. Rezultate merijo tudi z metode F-measure Manning in Schutze [19]. Pri vseh eksperimentih je uporabljen klasifikator SVM, uporabijo pa metodo 5-kratnega prečnega preverjanja (angl. k-fold cross validation). Ta metoda razdeli podatkovno zbirko na k enako velikih množic. Klasifikator se nato uči na eni izmed k množic, ostale množice (teh je $k-1$) pa so uporabljene kot testne zbirke. Korak se ponovi za vsako množico, na koncu pa se izbere instanco klasifikatorja, ki je imel najboljši rezultat.

Klasifikacija v dva razreda

Pri dvorazredni klasifikaciji pozitivnih in negativnih tweetov uporabijo uravnoteženo zbirko podatkov s 3.418 primeri tweetov, od tega je pol pozitivnih in pol negativnih. Graf na sliki 5 prikazuje razmerje med večanjem natančnosti in večanje učne zbirke, rezultati modelov pa so vidni v tabelah 9 in 10.



Slika 5: Prikaz krivulje učenja pri večanju učne podatkovne zbirke za razlčne modele

Tabela 9: Rezultati postopno dodanih atributov pri klasificiranju v dva razreda. Unigram modelu so postopoma dodani senti-atributi. Prikazana je tudi natančnost pri klasificiranju pozitivnih in negativnih tweetov.

Atributi	Natančnost (%)	F-measure	
		Pozitivni	Negativni
Unigram (referenčni model)	71,35	71,13	71,50
$+f_5, f_6, f_7, f_{10}, f_{11}$	70,10	69,66	70,46
$+f_1, f_8$	74,84	74,40	75,20
$+f_2, f_3, f_4, f_9$	75,39	74,81	75,86

Tabela 10: Rezultati klasificiranja v tri razrede, podana sta natančnost modelov in standardna deviacija

Model	Povprečna natančnost (%)	Standardna deviacija (%)
Unigram	71,35	1,95
Senti-atributi	71,27	0,65
Tree Kernel	73,93	1,50
Unigram + Senti-atributi	75,39	1,29
Tree Kernel + Senti-atributi	74,61	1,43

Klasifikacija v tri razrede

Klasifikacija v tri razrede: pozitivni, nevtralni in negativni. Uporabljena je uravnotežena zbirka podatkov velikosti 5.127 tweetov, kjer je v vsakem razredu tretjina tweetov. Rezultati so vidni v tabeli 11 in 12.

Tabela 11: Rezultati postopno dodanih atributov pri klasificiranju v tri razrede. Podane so točnosti napovedovanja za vse tri razrede.

Atributi	Natančnost (%)	F-measure		
		Pozitivni	Nevtralni	Negativni
Unigram (referenčni model)	56,80	56,86	56,58	56,20
$+f_5, f_6, f_7, f_{10}, f_{11}$	56,91	55,12	59,84	55,00
$+f_1, f_8$	59,86	58,42	59,82	59,82
$+f_2, f_3, f_4, f_9$	60,50	59,41	60,15	61,86

Tabela 12: Rezultati modelov pri klasificiranju v tri razrede, podana sta natančnost in standardna deviacija

Model	Povprečna natančnost (%)	Standardna deviacija (%)
Unigram	56,58	1,52
Senti-atributi	56,31	0,69
Kernel	60,60	1,00
Unigram + Senti-atributi	60,50	2,27
Kernel + Senti-atributi	60,83	1,09

4.1.8 Ugotovitve

Njihovi eksperimenti so pokazali, da Twitter specifični atributi, kot so hashtagi, emotikoni, itd. pripomorejo pri doseganju višje natančnosti sentimentalne analize, vendar ne veliko. Atributi, ki vsebujejo sentimentalnost besed, skupaj s POS oznakam, pa pripomorejo največ. Iz rezultatov klasifikacije v dva razreda zaključijo, da so najpomembnejši atributi tisti, ki zajemajo sentimentalnost besed. Vsi ostali atributi pa igrajo manjšo vlogo pri izboljšavi klasifikacije. Še več, eksperiment z unigram modelom ki je vseboval le pomembne atribute, je minimalno slabši od unigram modela, ki vsebuje vse atribute. To implicira, da je bolj pomembna dobra ekstrakcija atributov, ne pa njihova količina.

Uporaba Tree kernel modela z dobro izbranimi atributi, se je izkazala za najboljšo rešitev pri klasifikaciji tweetov. Obenem so pokazali tudi, da klasificiranje tweetov ni tako drugačno od klasificiranja drugih tipov teksta, saj so uporabili podobne metode.

4.2 Pak in Paroubek, 2010

4.2.1 Uvod

V raziskavi pokažejo avtomatiziran proces zbiranja podatkovne zbirke sestavljene iz tweetov, ki jo nato statistično analizirajo glede na besedne vrste. S korpusom ustvarijo model, ki klasificira tweete v tri razrede: pozitivne, nevtralne in negativne. Pokažejo, da se njihov model odreže bolje kot drugi modeli v tistem času.

4.2.2 Podatkovna zbirka

Ustvarjen je korpus 300.000 tweetov v angleškem jeziku. Te so bili zbrani zbrani brez človeškega posredovanja s pomočjo Twitter API. Zbirka je uravnoteženo razdeljena glede na vsebino tweetov v tri razrede: pozitivni, negativni in objektivni.

Pozitivni tweeti vsebujejo pozitivne emotikone, ki izražajo srečo, veselje, itd.
Negativni tweeti vsebujejo negativne emotikone, ki izražajo žalost, slabo počutje, itd.
Objektivni tweeti pa so zbrani iz objav popularnih časopisov in revij, kot so New York Times, Washington Post, itd. Takih virov je 44.

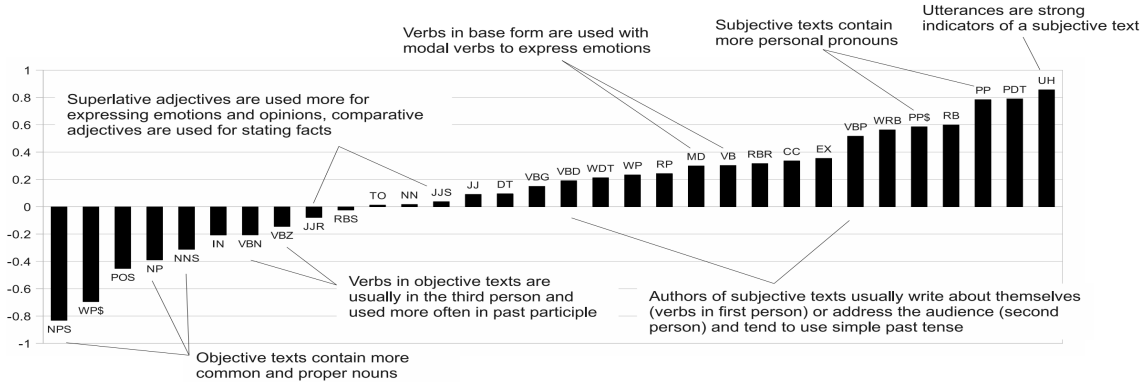
Ker je bil vsak tweet v tistem času dolg največ 140 znakov, predpostavljajo da so tweeti eno-povedni in zato tudi, da se vsi prisotni emotikoni nanašajo na celoten tweet – in obratno.

4.2.3 Analiza podatkovne zbirke

Z uporabo TreeTagger Schmid, 1994 [30] označijo vse besede v korpusu, zanima pa jih kako pogosto se pojavljajo različne oznake glede na razrede. Za primerjavo oznak med dvema poljubnima zbirkama izračunajo naslednjo vrednost za vsako oznako v obeh zbirkah, na primer med pozitivnimi in negativnimi tweeti.

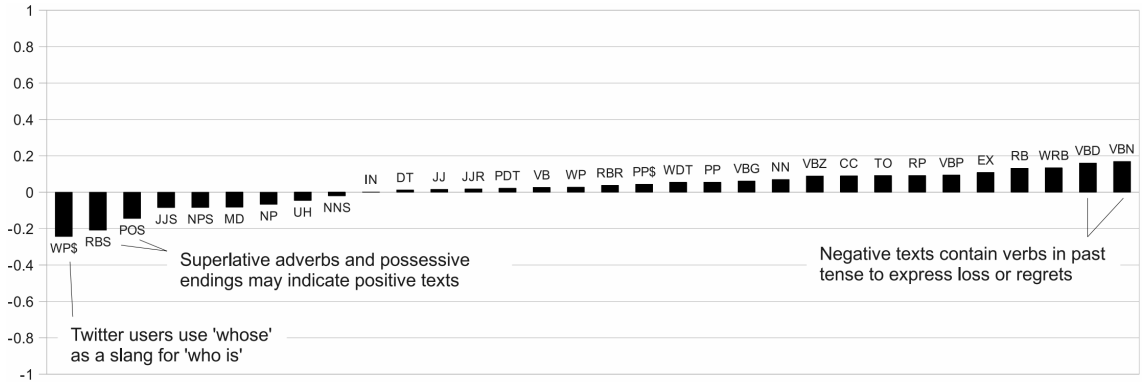
$$P_{1,2}^{T} = \frac{N_1^T - N_2^T}{N_1^T + N_2^T} \tag{4.1}$$

Kjer sta N_1^T in N_2^T števili oznak T, ki predstavljata kolikokrat se je oznaka T pojavila v razredu 1 in razredu 2 (na primer med razredoma pozitivnih in objektivnih tweetov).



Slika 6: Graf prikazuje pogostost uporabe različnih besed med objektivno in subjektivno zbirko. Vidi se, da objektivni tweeti vsebujejo več samostalnikov občnih in lastnih imen (angl. common and proper nouns) (NPS, NP, NNS), medtem ko se pri subjektivnih tweetih bolj pogosto pojavljajo osebni zaimki (angl. personal pronouns) (PP, PP$). Avtorji subjektivnih tweetov ponavadi sebe opisujejo v prvi osebi, občinstvo pa v drugi osebi (VBP), uporabljen pa je preteklik (angl. simple past tens((VBD) namesto preteklega deležnika (angl. past participle) (VBN), ki je uporabljen v objektivni zbirki. Subjektivni tweeti ponavadi vsebujejo tudi glagole v osnovnih oblikah (angl. base form of verbs) (VB), kar je smiselno glede na to, da se pogosto uporabljajo tudi modalni glagoli (angl. modal verbs) (MD). Opazi se tudi to, da se pridevniki v presežniški obliki (angl. superlative adjectives) (JJS) uporabljajo bolj pogosto pri ekspresiji čustev in mnenj, primerniki (angl. comparative adjectives) (JJR) pa se uporabljajo pri izražanju dejstev in navajanju informacij. Prislovi (angl. adverbs) (RB) se večinoma pojavljajo v subjektivnih tweetih v kombinaciji z glagoli (angl. verb).

Graf na sliki 6 prikazuje vrednosti P^T za vse POS oznake, kjer je ena zbirka subjektivna, ta je mešanica pozitivnih in negativnih tweetov. Druga zbirka je objektivna, ta je sestavljena iz objektivnih tweetov. Iz grafa se vidi, da POS oznake niso razporejeno enakomerno med dvema zbirkama, kar pomeni, da so razlike lahko uporabljene kot atributi pri klasifikaciji. Izkaže se tudi to, da izjave (angl. utterances) (UH), močno implicirajo, da je tweet subjektivni. Izjave so najmanjši del besedila, ki nosi svoj pomen. Graf na sliki 7 kaže razlike med POS oznakami med pozitivnimi in negativnimi tweeti.



Slika 7: Graf prikazuje vrednosti P^T za negativne in pozitivne zbirke. Izkaže se, da ima pozitivna zbirka veliko svojilnih zaimkov, ki se začnejo na "wh" (angl. wh-pronoun), kot je "whose" (WH$), kar je nepričakovano. Po podrobnejšem pregledu korpusa se izkaže, da se "whose" uporablja kot slang oziroma kratica za "who is", na primer: "dinner & jack o'lantern spectacular tonight! :) whose ready for some pumpkins??". Še en indikator pozitivnih tweetov je uporaba prislovov v presežniški obliki (angl. superlative adverbs) (RBS), kot sta na primer: "most" in "best". Prepoznani pa so lahko tudi po uporabi izražanju svojilnosti (angl. posessive ending) (POS). V nasprotju s pozitivno zbirko, negativna zbirka vsebuje več glagolov v pretekliku (VBN, VBO), ker veliko avtorjev izrazi negativna čustva. Primeri najbolj pogostih glagolov so: "missed", "bored", "gone", "lost", "stuck", "taken".

4.2.4 Predobdelava tweetov

Vsem tweetom odstranijo:

- URL povezave
- Uporabniška imena, ki so prepoznane z uporabo "@" simbola
- Twitter specifične oznake,na primer "RT", ki označuje re-tweet
- Emotikone

Tekst razčlenijo in tokenizirajo po presledkih in ločilih, pri tem pazijo, da ohranijo besede kot so "don't" kot ena beseda. S tem postopkom dobijo nabor besed, izmed katerih odstranijo funkcijske besede, specifično podajo primere besed "a", "an" in "the".

Nato ustvarijo n-grame. Zanikalne besede (na primer: "no" in "not") so "zlepljene" skupaj z naslednjo in predhodno besedo.
Primer tweeta z vsebino "I do not like fish" vsebuje tri bigrame: "I do+not", "do+not like" in "not+like fish".

4.2.5 Uporaba klasifikatorjev in atributi

Uporabijo več klasifikatorjev, najbolje pa se je odrezal Multinomial Naive Bayes za katerega podajo tudi rezultate. Ta je osnovan na atributih, ki zajemajo n-grame in POS oznake.

4.2.6 Uporabljeni metodi za izboljšanje modelov

Da bi izboljšali natančnost klasifikacije, predpostavljajo, da bi morali odstraniti pogoste n-grame, to so tisti, ki ne nakazujejo močne sentimentalnosti ali objektivnosti povedi. Taki n-grami se pojavijo v vseh treh pod-skupinah podatkovne zbirke. Da bi take n-grame našli in jih odstranili, uporabijo dve metodi.

Prva metoda

Metoda sloni na izračunu entropije verjetnostne porazdelitve n-grama v različnih podatkovnih zbirkah. Entropija verjetnostne porazdelitve pojavitev n-gramov v različnih podatkovnih zbirkah z različnimi sentimentalnostmi. (angl. computing the entropy of a probability distribution of the appearance of an n-gram in different datasets (different sentiments)). Po formuli Shannonove entropije:

$$entropija(g) = -\sum_{i=1}^{N} p(S_i|g) \log p(S_i|g) \tag{4.2}$$

Kjer je N število razredov, v tem primeru so trije. Visoka vrednost entropije implicira, da je distribucija n-grama v različnih pod zbirkah (glede na sentimentalnost) skoraj enakomerna. Tak n-gram ne pripomore veliko pri klasifikaciji tweeta. Majhna vrednost entropije pa implicira, da se n-gram pojavlja večkrat v eni pod zbirki, kot pa v drugih, zato tak n-gram lahko uporabijo pri razlikovanju med razredi. Da bi izboljšali natančnost modela bi radi uporabili le n-grame z majhno vrednostjo entropije. Natančnost lahko nadzirajo z omejitvijo praga, ki ga določa spremenljivka θ. To bi zmanjšalo priklic (angl. recall) modela, ker zmanjšajo število atributov.

Druga metoda

V drugi metodi je predstavljen nov izraz "salience", to je vrednost, ki je izračunana za vsak n-gram, po formuli:

$$salience(g) = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1 - \frac{min(P(g|s_i)), P(g|s_j)}{max(P(g|s_i), P(g|s_j)} \tag{4.3}$$

Formula vrne vrednost med 0 in 1. Majhna vrednost predstavlja majhno "salience" in take n-grame penalizirajo. Tako kot pri entropiji, lahko nadzirajo natančnost modela s spreminjanjem spremenljivke θ, ki določa prag.
Z uporabo formul entropije in "salience", dobijo končno enačbo, ki izračuna logaritmirano oceno verjetnosti:

$$L(s|M) = \sum_{g \in G} \log(P(g|s) \cdot if(f(g) > \theta, 1, 0)) + \sum_{t \in G} \log(P(t|s)) \tag{4.4}$$

Kjer je $f(g)$ entropija ali šalience"n-grama, θ pa določa prag.

4.2.7 Rezultati

Testirali so model na zbirki tweetov, ki so bili ročno označeni. Uporabili so enako testno podatkovno zbirko kot v članku Go et al. [15]. Podrobnosti o zbirki so predstavljene v tabeli 13.

Tabela 13: Opis testne podatkovne zbirke

Sentiment	Število primerkov
Pozitivni	108
Nevtralni	75
Negativni	33
Skupaj	216

Najprej so testirali, kako velikost n-gramov ($n = 1, 2, 3$) vpliva na natančnost modela. Iz grafa na sliki 8 se vidi, da je imel najboljšo natančnost model ki je uporabljal bigrame.

Pojasnjujejo, da bigrami podajo dobro razmerje med pokritjem besedila (ki ga nudijo unigrami) in zajemanjem vzorcev, ki izražajo sentimentalnost (ki ga nudijo trigrami). Eksperimentirali so tudi z "lepljenjem" zanikalnih besed pri ustvarjanju n-gramov, rezultati so vidni na grafu na sliki 10. Podajajo tudi vpliv večanja podatkovne zbirke na natančnost modela. Da bi to zmerili so uporabili metodo F-measure Manning in Schutze [19]:

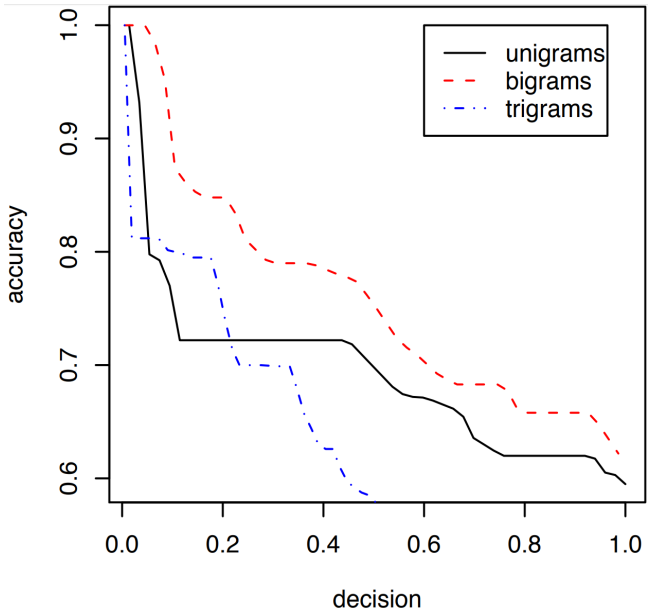
$$F = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 \cdot precision \cdot recall} \tag{4.5}$$

Kjer je "precision" natančnost, "recall" priklic, "accuracy" točnost in "decision" vrednost, prikazuje kako velik del podatkovne zbirke je bil kalsificiran.

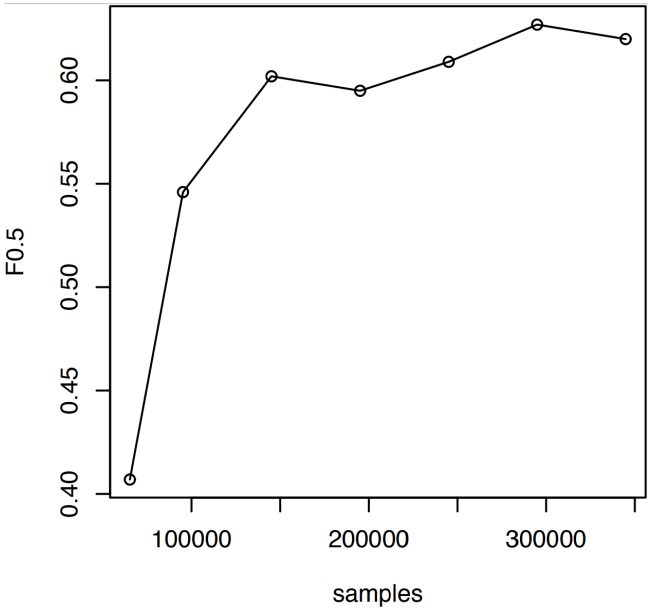
Pri njihovih evaluacijah, zamenjajo natančnost (angl. precision) s točnostjo (angl. accuracy) in priklic (angl. recall) z "decision", ker imajo več razredov ne pa binarno klasifikacijo. Formula je potem:

$$F = (1 + \beta^2)\frac{accuracy \cdot decision}{\beta^2 \cdot accuracy \cdot decision} \tag{4.6}$$

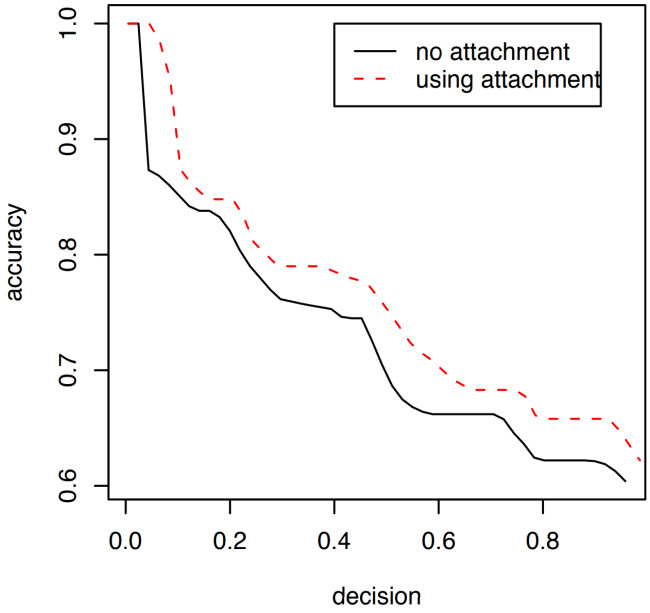
Kjer je $\beta = 0.5$. V tem eksperimentu ne filtrirajo n-gramov. Rezultat je viden na grafu na sliki 9, na katerem se vidi, da z se večanjem velikosti učne zbirke poveča natančnost modela, vendar le do neke mere. Graf na sliki 11 kaže, da so uporabo z"sailence" dosegli boljše rezultate kot pa z entropijo.



Slika 8: Primerjava klasifikacijske natančnosti pri uporabi unigramov, bigramov in trigramov



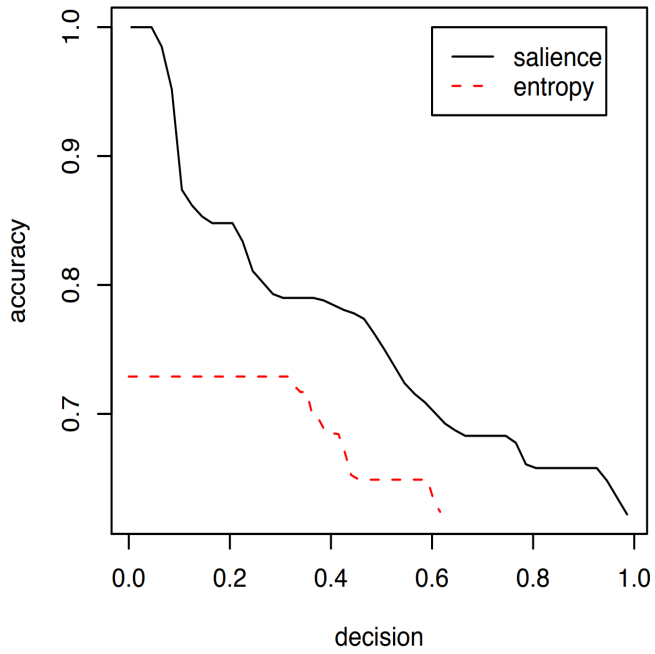
Slika 9: Učinek večanja učne podatkovne zbirke na oceno $F_{0.5}-$



Slika 10: Vpliv "lepljenja" zanikalnih besed, polna črta predstavlja rezultate kjer metoda "lepljenja" besed ni bila uporabljena, črtkana črta pa ko je bila

4.2.8 Ugotovitve

V tej raziskavi je predstavljena metoda za avtomatsko kreacijo korpusa, ki je lahko uporabljen za treniranje modela. Uporabili so TreeTagger za POS označevanje besed in opazovali različne distribucije besednih vrst med pozitivnimi, objektivnimi in negativnimi tweeti. S pomočjo teh oznak in analize so pokazali, da določene POS oznake implicirajo sentimentalnost. Njihov najboljši model je zmožen prepoznavanja senti-



Slika 11: "salience" in entropija pri odstranjenih skupnih/pogostih n-gramov

mentalnosti tweetov, ta je Multinomial Naive Bayes klasifikator, ki uporabi n-grame in POS oznake kot atribute.

4.3 Go et al., 2009

4.3.1 Uvod

Za treniranje modelov, supervised learning metoda navadno potrebuje ročno označene instance podatkovne zbirke. Tako podatkovno zbirko je težko dobiti ali ustvariti, zato podobno kot v članku Pak in Paroubek [25], uporabijo metodo distant supervision, kjer tweete klasificirajo glede na prisotnost emotikonov. Uporabijo več različnih klasifikatorjev in metod ekstrakcije atributov. Vse te modele preizkusijo nad podatkovno zbirko, v kateri ni nujno, da se v vseh instancah pojavi emotikon.

Njihova definicija sentimentalnosti

Sentimentalnost definirajo kot "a personal positive or negative feeling". V tabeli 14 podajo nekaj primerov. Velikokrat je nejasno, ali tweet vsebuje sentimentalnost. V takih primerih uporabijo "litmus" test: če bi se tweet lahko pojavil na prvi strani časopisa kot naslov članka, ali kot poved na Wikipediji, potem tweet obravnavajo kot nevtralni. V tej raziskavi, nevtralne tweete ne obravnavajo pri fazi učenja in testiranja. Uporabijo samo pozitivne in negativne tweete.

Tabela 14: Primeri tweetov, ki so bili pridobljeni z različnimi poizvedbami. Za vsak razred je podan en tweet.

Sentiment	Poizvedba	Tweet
Pozitivni	jquery	dcostalis: Jquery is my new best friend.
Nevtralni	San Francisco	schuyler: just landed at San Francisco
Negativni	exam	jvici0us: History exam studying ugh.

4.3.2 Podatkovna zbirka

Tweeti so zbrani s pomočjo Twitter API in sicer z metodo distant supervision. Klasificirani so glede na prisotnost emotikonov. Če se v tweetu pojavi pozitivni emotikon, ga označijo kot pozitivnega, in kot negativnega če se pojavi negativni emotikon. Tabela 15 vsebuje vse emotikone, ki so jih uporabili pri zbiranju. S tem postopkom zberejo 800.000 negativnih in 800.000 pozitivnih tweetov, skupno ustvarijo uravnoteženo podatkovno zbirko veliko 1.600.000 tweetov, ki jo uporabijo za treniranje modelov.

Za testiranje modela uporabijo novo podatkovno zbirko, ki je ustvarjena s pomočjo Twitter API. To storijo s poizvedovanjem po besedah. Tweeti morajo vsebovati specifične besede ali oznake, ki so iz različnih domen in različnih tem. Primeri teh besed so najdeni v Tabeli 5. Zbrani tweeti so bili nato ročno klasificirani glede na vsebino – klasificirajo neodvisno od prisotnosti emotikonov.

Tabela 15: Seznam emotikonov uporabljen pri zbiranju podatkovne zbirke

Emotikoni povezani z ":)"(pozitivni)	Emotikoni povezani z ":("(negativni)
:)	:(
:-)	:-(
:)	: (
:D	
=)	

4.3.3 Predobdelava tweetov

Učna podatkovna zbirka je predobdelana na naslednji način:

- Emotikoni, predstavljeni v tabeli 15 so odstranjeni
- Uporabniška imena so odstranjena, te so prepoznana s simbolom "@"

- Tweeti, ki vsebujejo pozitivne in negativne emotikone, so odstranjeni, ker so se hoteli izogniti temu, da bi bili pozitivni atributi označeni kot del negativih tweetov in obratno
- Re-tweeti so odstranjeni, te so prepoznani so po oznaki "RT"
- Tweeti, ki vsebujejo emotikon ":P" so odstranjeni

Tweeti z emotikonom ":Pšo odstranjeni, ker je v času pisanja članka obstajala napaka v Twitter API. Ta je namreč vračal tweete, v katerih sta bila prisotna ali ":P" ali ":)" emotikon, čeprav je bila zahtevana le prisotnost emotikona ":(".

Posledice

Predobdelava zbirke ima posledico, da se število atributov po njihovi ekstrakciji zmanjša za približno 50%. Rezultati so predstavljeni v spodnji tabeli.

Tabela 16: Posledica predobdelave podatkovne zbirke in njen vpliv na število atributov

Način zmanjšanja atributov	Število atributov	Procent originalne zbirke
Noben	794.876	100.00%
Uporabniška imena	449.714	56.58%
URL povezave	730.152	91.86%
Ponavljajoče črke	773.691	97.33%
Vse	364.464	45.85%

4.3.4 Atributi in modeli

Uporabijo več načinov ekstrakcije atributov. Med njimi so: unigrami, bigrami, unigrami in bigrami, ter POS oznake. Vse testirajo z Naive Bayes, Maximum Entropy in SVM algoritmi.

4.3.5 Rezultati in ugotovitve

Vsi rezultati so podani v tabeli 17. Podane so tudi ugotovitve uporabe različnih atributov.

Unigrami

Unigrami so najpreprostejši način, kako pridobiti atribute iz teksta. Vsaka beseda predstavlja en atribut. Njihovi rezultati so podobni kot v prejšnjih raziskavah in sicer

81% pri Naive Bayes, 80% pri Maximum Entropy in 82.9% natančnost pri Support Vector Machine modelu.

Bigrami

Bigrami so uporabljeni za zaznavanje zanikanih fraz, kot so "not good" in "not bad". Te v njihovih eksperimentih niso izboljšali natančnosti kot dodatni atributi pri unigramih. Ker so bigrami raztreseni se natančnost zmanjša pri obeh Maximum Entropy in SVM klasifikatorjih. Problem raztresenosti se lahko vidi na primeru: "@stellargirl I looooovvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right". Maximum Entropy je poda enaki verjetnosti za pozitivni in negativni razred, ker ni bigrama ki bi nagnil tehtnico sentimentalnosti v pozitivno ali negativno smer. Splošno, bigrami sami, kot atributi niso zelo uporabni. Bolje jih je združiti z unigrami.

Unigrami in bigrami

Unigrami in bigrami skupaj prinašajo boljše rezultate, kot samo unigrami. Napovedi se izboljšajo iz 81.3% na 82.7% pri Naive Bayes, iz 80.5% na 82.7% za Maximum Entropy, pri SVM pa se poslabšajo iz 82.2% na 81.6%.

POS oznake

POS oznake uporabijo, ker ima lahko ista beseda več pomenov glede na kontekst. Po njihovih ugotovitvah, POS oznake niso zelo uporabne, kar se ujema z raziskavo Pang in Lee [26]. Naive Bayes in SVM sta imela slabše rezultate, Maximum Entropy pa se je za zanemarljivo vrednost izboljšal v primerjavi z rezultati unigramov.

Tabela 17: Rezultati modelov

Atributi	Naive Bayes	Maximum Entropy	Support Vector Machines
Unigram	81.3	80.5	82.2
Bigram	81.6	79.1	78.8
Unigram + Bigram	82.7	83.0	81.6
Unigram + POS	79.9	79.9	81.9

5 Lastna implementacija

5.1 Uvod

Rešitev je implementirana s pomočjo programskega jezika Java in z uporabo knjižnice Twitter4J. Ta omogoča komunikacijo s Twitter API za klasifikacijo tweetov, ki se prenašajo v realnem času. Tweeti so lahko opredeljeni v tri razrede (pozitivni, nevtralni in negativni) ali pa v dva razreda (pozitivni in negativni).

5.2 Podatkovne zbirke

Zbrani sta bili dve podatkovni zbirki:
Prva podatkovna zbirka je ustvarjena iz zbirke članka Agarwal et al. [2]. Z nami so delili podatkovno zbirko, v kateri je bilo 5.127 tweetov. Vsaka instanca zbirke je bila sestavljena iz sentimentalnosti in identifikacijske številke tweeta. S pomočjo programske opreme Twitter4J, je bila storjena poizvedba za vsako identifikacijsko številko. Na ta način je bila sestavljena nova podatkovna zbirka, ki je vsebovala besedilo tweeta in sentimentalno vrednost. Zaradi zastarelosti podatkovne zbirke (ustvarjena je bila leta 2011) veliko tweetov ni bilo več javno dostopnih ali pa so bili izbrisani. Zaradi starosti samih tweetov, emojiji niso prisotni. Končna podatkovna zbirka vsebuje 2.188 tweetov, izmed tega je 705 negativnih, 783 pozitivnih in 700 nevtralnih.

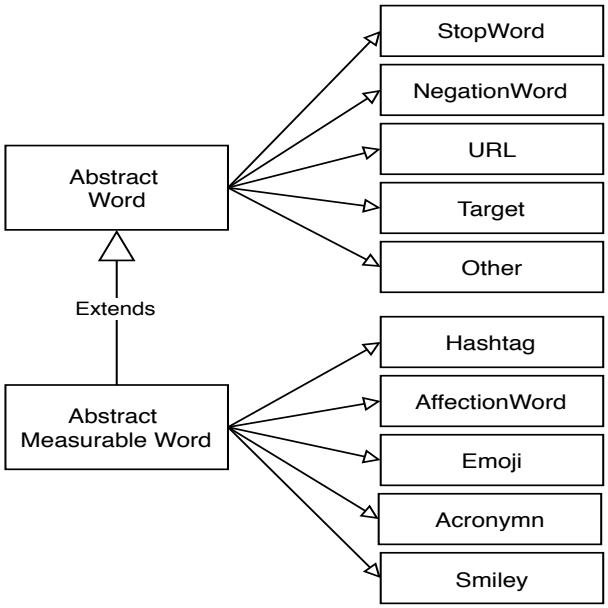
Druga podatkovna zbirka je bila pridobljena z metodo distant supervision. S pomočjo Twitter4J je bilo zbranih 120.120 tweetov, ki so bili avtomatsko klasificirani glede na prisotnost emotikonov. Če je tweet vseboval ":)" je bil označen kot pozitiven, če je vseboval ":(" pa kot negativen. Zbiranje je potekalo približno 16 ur, saj Twitter API omejuje prenašanje podatkov v realnem času (angl. real-time stream). Podatkovna zbirka vsebuje 48.679 negativnih in 71.441 pozitivnih tweetov.

5.3 Opis implementacije

Program za svoje delovanje uporablja več slovarjev, s katerimi prepozna besede v tweetu. Tweeti so tokenizirani po presledkih, pred tem pa je besedilo skrčeno v eno vrstico. Vsaka beseda je opredeljena v svoj razred, glede na slovarje. Definirana sta dva abstraktna razreda:

- AbsWord, ki opisuje poljubno besedo
- AbsMeasurableWord, ki opisuje vse besede, ki jim lahko pripišemo sentimentalno vrednost

Diagram implementacije razredov je viden na sliki 12.



Slika 12: Diagram implementacije razredov

Program lahko nastavimo, da v realnem času prenaša naključne tweete in jih ob prejemu avtomatsko klasificira. Možna je tudi uporaba poljubne poizvedbe. Pri procesu prenašanja tweetov je nastavljena omejitev: prenašajo se le tweeti napisani v angleškem jeziku. Razlog za to so angleški slovarji, s katerimi se sami tweeti procesirajo.
Opisi razredov, ki so implementirani iz abstraktnega razreda "AbsWord":

- **StopWord** razred določa besede, ki ne prinesejo veliko informacij. Te so v besedilu najdene s pomočjo slovarja funkcijskih besed.
- **NegationWord** razred določa besede, ki negirajo sentimentalno vrednost naslednje besede. Te so najdene s pomočjo slovarja zanikalnih besed.

- **URL** razred določa URL povezave. Te so prepoznane s pomočjo regularnega izraza (angl. Regular expression). Izraz je sledeč [1]

```
(https?:\\/\\/(?:www\\.|(?!www))[a-zA-Z0-9][a-zA-Z0-9-]
+[a-zA-Z0-9]\\.[^\\s]{2,}|www\\.[a-zA-Z0-9][a-zA-Z0-9-]
+[a-zA-Z0-9]\\.[^\\s]{2,}|https?:\\/\\/(?:www\\.|(?!www
))[a-zA-Z0-9]+\\.[^\\s]{2,}|www\\.[a-zA-Z0-9]+\\.[^\\s]
{2,})
```

- **Target** razred določa uporabniška imena, te so prepoznana po začetnem znaku "@".
- **Other** razred določa besede, ki niso prepoznane. Te so ponavadi neprepoznani emojiji, emotikoni, kratice in besede, ki vsebujejo slovnične napake.

Opis razredov, ki so implementirani iz abstraktnega razreda "AbsMeasurableWord":

- **Hashtag** razred določa hashtage, te so prepoznani po začetnem znaku "#".
- **AffectionWord** so besede, ki so najdene v DAL slovarju.
- **Emoji** razred določa emojije. To so vsi emojiji, ki so najdeni v besedilu tweeta in se nahajajo v emoji slovarju.
- **Acronymn** razred določa vse krajšave in slengovski izraze, ki so najdeni s pomočjo slovarja kratic.
- **Smiley** razred določa vse emotikone, te so najdeni s pomočjo emotikon slovarja.

5.4 Procesiranje in učenje

Po vsakem končanem procesiranju tweeta, program ustvari seznam besed, ki so definirane glede na pripadajoči razred s pomočjo slovarjev. V procesu določanja sentimentalnosti celotnega tweeta, se algoritem sprehodi čez seznam besed. Če se v besedilu pojavi zanikalna beseda, se naslednji besedi negira sentimentalno vrednost, če je ta definirana. Besede, ki imajo vsaj polovico črk napisanih z velikimi črkami se sentimentalnost poveča za faktor 1.3. Po sledeči formuli se določi vrednost *Sentiment*:

[1] Vir: foufus: `https://rubular.com/r/eGPe4bGlwMd98E`

$$Sentiment(t) = \sum_{i=0}^{N} W_i \tag{5.1}$$

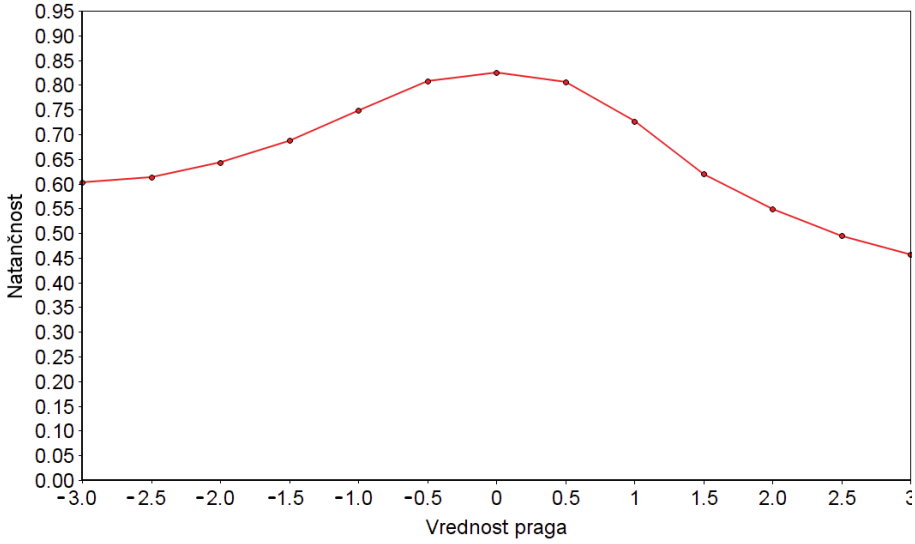
Kjer je N število besed v tweetu t, W_i pa poljubna beseda, ki je prisotna v seznamu besed. Glede na vrednost podane formule in vrednosti pragov, je določena sentimentalnost.

5.4.1 Določanje praga

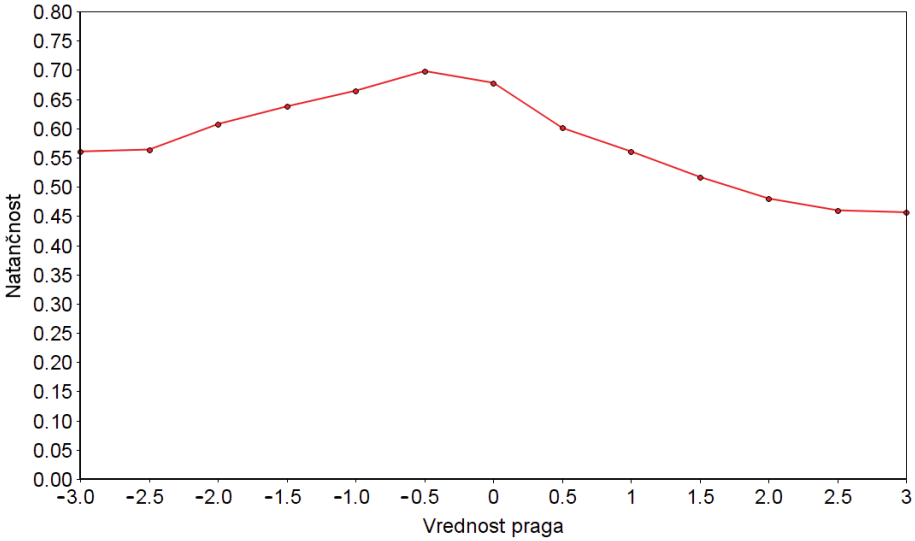
Da bi lahko tweete klasificirani v tri razrede (negativni, nevtralni in pozitivni) ali pa v dva razreda (negativni in pozitivni) so določene vrednosti, ki označujejo prag (angl. threshold). Te so izračunane iz obeh podatkovnih zbirk. Praga t^- in t^+ sta določeni s pomočjo **prve** podatkovne zbirke iz katere je bila ustvarjena je bila učna podatkovna zbirka, ki je zajemala 66% naključno izbranih tweetov. Praga sta določena iz omenjene zbirke, ker ta vsebuje tweete že opredeljene v tri razrede, medtem ko **druga** podatkovna zbirka vsebuje tweete klasificirane v dva razreda. Proces določanja vrednosti t^- in t^+ je zajemal spreminjanje njunih vrednosti in opazovanjem natančnosti klasifikacije. Izbrani sta bili vrednosti, ki sta najbolje napovedali razred prvi podatkovni zbirki. Vrednost t^- razmejuje vrednosti med negativnimi in nevtralnimi tweeti, vrednost t^+ pa med nevtralnimi in pozitivnimi.

Vrednost spremenljivke t^0 je bila določena s pomočjo **druge** podatkovne zbirke. Učna podatkovna zbirka in vrednost t^0 sta bila izbrana na enak način, kot v prvem primeru. Vpliv spreminjanja vrednosti praga na natančnost je viden na grafu na sliki 14. Zaradi presenetljivo dobrih rezultatov (glede na preprostost algoritma), je bila metoda testirana tudi nad podatki **prve** podatkovne zbirke. Proces učenja je ostal enak, spreminjanje praga na tej zbirki pa je viden na grafu na sliki 14. Pred fazo učenja in testiranja, so bili nevtralni tweeti odstranjeni. Optimalna vrednost praga je približno enaka kot pri učenju na **drugi** podatkovni zbirki. Končne vrednosti pragov so sledeče:

- $t^0 = 0,5$. Prag razmejuje med pozitivnimi in negativnimi tweeti. Če je vrednost formule 5.2 nad vrednostjo t^0, je instanca klasificirana v pozitivni razred, sicer v negativni.
- $t^- = 0,05$ in $t^+ = -0,5$. Vrednosti razmejujeta med pozitivnimi, nevtralnimi in negativnimi tweeti. Če je vrednost formule 5.2 manjša kot t^-, je instanca klasificirana v negativni razred. V primeru, da je večja od t^+ je klasificirana v pozitivni razred, sicer v nevtralni.



Slika 13: Posledica spreminjanja vrednosti praga t^0 pri učenju na natančnost klasifikacije v dva razreda pri drugi podatkovni zbirki



Slika 14: Posledica spreminjanja vrednosti praga t^0 pri učenju na natančnost klasifikacije v dva razreda pri prvi podatkovni zbirki, kjer so bili nevtralni tweeti odstranjeni

5.5 Slovarji

5.5.1 Slovar funkcijskih besed

Slovar zajema 103 primerov besed [2] [3]. Besede so shranjene v TXT formatu, kjer je vsaka beseda v svoji vrstici.

[2]Vir: Ranks: `https://www.ranks.nl/stopwords`
[3]Vir: sebleier: `https://gist.github.com/sebleier/554280`

5.5.2 Slovar zanikalnih besed

Slovar zajema 28 primerov besed [4]. Besede so shranjene v TXT formatu, kjer je vsaka beseda v svoji vrstici.

5.5.3 Slovar vplivnosti besed

Slovar zajema 8.749 besed [5], kjer so vsaki besedi prepisane tri vrednosti: "pleasantness", "activation" in "imagery". Te vrednosti so med 1 in 3, ki so normalizirane med -1 in 1. Slovar je zapisan v CSV formatu, kjer so atributi podani v sledečem zaporedju: beseda, "pleasantness", "activation", "imagery".

5.5.4 Emoji slovar

Slovar je zgrajen iz obstoječega emoji slovarja [6] Novak et al. [18]. Njihov slovar vsebuje 969 emojijov, ki so shranjeni v CSV formatu. Atributi so podani v zaporedju: Unicode oznaka, število pojavitev, pozicija v besedilu, število pojavitev v negativnih besedilih, število pojavitev v nevtralnih besedilih, število pojavitev v pozitivnih besedilih, Unicode ime, Unicode blok.
Sentimentalna vrednost emojijov je izračunana po formuli:

$$\overline{s} = -1 \cdot p_- + 0 \cdot p_+ = p_+ - p_- \tag{5.2}$$

Kjer je s sentiment, p_- odstotek pojavitev v negativnih tweetih, p_+ pa odstotek pojavitev v pozitivnih tweetih.

Zgrajen je nov slovar, shranjen v CSV formatu, kjer so atributi podani v zaporedju: Unicode oznaka, opis, sentiment. Unicode oznaka v besedilu kodira emoji, opis je kratek opis emojija in vrednost je sentimentalna vrednost emojija, ta je podana z vrednostjo med -1 in 1.

5.5.5 Slovar emotikonov

Slovar zajema 140 različnih emotikonov iz različnih spletnih virov [7] [8]. Te so zapisani v CSV formatu, atributi so podani v sledečem zaporedju: emotikon, sentiment.

[4]Vir: Grammarly: `https://www.grammarly.com/blog/negatives/`

[5]Vir: God-helmet: `https://www.god-helmet.com/wp/whissel-dictionary-of-affect/index.htm`

[6]Vir: Novak et al. 2015: `http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html`

[7]Vir: sifei: `https://github.com/sifei/Dictionary-for-Sentiment-Analysis/blob/master/Emoticon/emoticonsWithPolarity.txt`

[8]Vir: Wikipedija: `https://en.wikipedia.org/wiki/List_of_emoticons`

Sentimentalnost je podana kot število med -1 in 1.

5.5.6 Slovar okrajšav in slenga

Slovar vsebuje 5263 primerov kratic in sleng besed in je enak kot iz članka Agarwal et al. [9]. Zaradi zastarelosti izrazov, je dodanih tudi nekaj novih izrazov in krajšav. Slovar je v CSV formatu, atributi so podani v zaporedju: krajšava, poln pomen. Krajšava predstavlja več vrst okrajšav, kratic ali sleng izrazov, poln pomen pa določa vse besede, ki so okrajšane. V primeru slenga pa določi knjižno besedo ali izraz.

5.6 Statistika n-gramov

Med pisanjem diplome je bilo s programom zbranih 120.000 tweetov, ki so bili zbrani in klasificirani s pomočjo metode distant supervision. Vsi tweeti, ki so vsebovali emotikon ":)" so bili označeni kot pozitivni, tweeti s prisotnim emotikonom ":(" pa kot negativni. V tabeli 19 in 18 so prikazani najbolj pogosti n-grami, ki so se pojavili v omenjenih razredih.

Tabela 18: Primeri najbolj pogostih unigramov, bigramov in trigramov iz pozitivne podatkovne zbirke

Unigrami	Bigrami	Trigrami
donkiss	donkiss onemorechance	donkiss onemorechance donkiss :(
onemorechance	onemorechance donkiss	onemorechance donkiss onemorechance
colours	now :)	hair 2 weeks
scientifically	ini adalah	mom now :)
coloured	coloured bleach	ini adalah tips
bleach	adalah tips	own hair 2
rambut	not scientifically	colours mom now
bercabang	ps retired	weeks once last
penjagaan	to colour	bercabang coloured bleach
adalah	bleach ini	last time ps

[9] Vir: Agarwal et al., 2011

Tabela 19: Primeri najbolj pogostih unigramov, bigramov in trigramov iz negativne podatkovne zbirke

Unigrami	Bigrami	Trigrami
namjoon	:(im	stay safe :(
charting	chinese and	all stay safe
babie	safe :(	im chinese and
pouted	left on	money left on
boating	im chinese	:(im chinese
ani	all stay	safe :(im
0xd0c0	china so	around china so
parentsfamily	guys all	guys all stay
photoprint	around china	chinese and ive
pc+	so precious	any money left

5.7 Rezultati, izboljšave in zaključek

Formula 5.2 s katero si pomagamo pri klasifikaciji sentimentalnosti je preprosta. Rezultati, ki jih dobimo so predstavljeni v tabeli 20. Te so v primerjavi s članki obravnavani v diplomi, so precej slabši pri trirazrednem klasificiranju, vendar prav tako dobri pri klasificiranju v dva razreda. Pri napovedovanju dveh razredov so bili rezultati podobni kot v članku Go et al. [15]. Upoštevati je treba, da se v naši implementaciji uporablja zelo preprost način klasifikacije. Po drugi strani, rešitve v predstavljenih člankih uporabljajo kompleksnejše metode, ki jih privedejo od 100 do 10.000 atributov na instanco. V našem primeru implementacija opredeli tweet glede na en sam atribut.

Za izboljšavo programa bi lahko dodali splošno namenski klasifikator, kot je Maximum entropy ali pa Naive Bayes. Metoda ekstrakcije atributov, ki trenutno vrača le eno število, bi lahko razširili na več atributov, kateri bi lahko zajemali tudi poljubno veliko n-grame in druge atribute, ki bi bolj natančno opisovali besedilo. Zavedati pa se moramo, da je jezik živa stvar, kot je razbrano iz tabele 19, saj opazimo, da so najbolj pogosti n-grami nanašajo na trenutno aktualne dogodke. Atributi osnovani na n-gramih tako ne bi pomagali pri klasifikaciji tweetov v prihodnosti, zato bi bilo treba nabor najbolj pogostih n-gramov skozi čas posodabljati. To pomeni, da bi moral imeti program avtomatiziran način zbiranja podatkovne zbirke in določanja seznama pogostih n-gramov. Problemi pri analiziranju besedil, kot so tweeti so tudi slovnične napake uporabnikov in zelo pogosta uporaba raznih kratic in slengovskih izrazov. Slovarje bi morali zato tako kot n-grame, periodično posodabljati.

Tabela 20: Rezultati klasifikacije v dva in tri razrede nad dvema podatkovnima zbirkama

	Natančnost (%)	
Podatkovna zbirka	Dvorazredna klasifikacija	Trirazredna klasifikacija
Prva (2.188 instanc)	68,21 (1.488 instanc)	51,76
Druga (120.120 instanc)	82,30	/

6 Literatura in viri

[1] Alphabetical list of part-of-speech tags used in the penn treebank project. Dostopano: 23.7.2019. *(Citirano na strani 5.)*

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011. *(Citirano na straneh IV, 15 in 34.)*

[3] I. Ahmad. The most popular social media platforms of 2019, 2019. Dostopano: 14.8.2019. *(Citirano na strani 1.)*

[4] D. H. Alonso. Emojipedia faq. Dostopano: 7.6.2019. *(Citirano na strani 11.)*

[5] Baeldung. Introduction to twitter4j, 2018. Dostopano: 5.4.2019. *(Citirano na strani 12.)*

[6] G. Bhardwaj. How java is platform independent, 2019. Dostopano: 18.7.2019. *(Citirano na strani 9.)*

[7] D. Chaffey. Global social media research summary 2019, 2019. Dostopano: 14.8.2019. *(Citirano na strani 1.)*

[8] DataRobot. Feature variables. Dostopano: 9.6.2019. *(Citirano na strani 2.)*

[9] DeepAI. Feature extraction. Dostopano: 9.6.2019. *(Citirano na strani 3.)*

[10] DeepDive. Distant supervision. Dostopano: 10.7.2019. *(Citirano na strani 7.)*

[11] Emojipedia. Emoji timeline, emojitimeline. Dostopano: 7.6.2019. *(Citirano na strani 11.)*

[12] S. Engine. Pos tags. Dostopano: 9.6.2019. *(Citirano na strani 7.)*

[13] R. Gandhi. Naive bayes classifier, 2018. Dostopano: 15.6.2019. *(Citirano na strani 4.)*

[14] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision, 2009. Dostopano: 15.6.2019. *(Citirano na strani 4.)*

[15] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision, 2009. *(Citirano na straneh 15, 27 in 41.)*

[16] C. Grannan. What's the difference between emoji and emoticons?, 2019. Dostopano: 7.6.2019. *(Citirano na straneh 10 in 11.)*

[17] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003. *(Citirano na strani 18.)*

[18] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič. Sentiment of emojis. *PLoS ONE*, 10(12):1–22, 2015. *(Citirano na strani 39.)*

[19] C. D. Manning, C. D. Manning, and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999. *(Citirano na straneh 20 in 27.)*

[20] S. C. Marketing. Twitter communication types. Dostopano: 6.6.2019. *(Citirano na strani 9.)*

[21] A. Moschitti. Making tree kernels practical for natural language learning, 2006. Dostopano: 20.8.2019. *(Citirano na strani 6.)*

[22] MuleSoft. What is an api? (application programming interface). Dostopano: 24.7.2019. *(Citirano na strani 12.)*

[23] D. Nations. What is a hashtag on twitter?, 2019. Dostopano: 22.7.2019. *(Citirano na strani 11.)*

[24] N. J. Nilsson. *Introduction to machine learning*. 1998. *(Citirano na strani 2.)*

[25] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010. *(Citirano na straneh 15 in 30.)*

[26] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002. *(Citirano na strani 33.)*

[27] A. Rasool. What happens within a minute over the online world? domo's annual infographic shows fascinating findings, 2019. Dostopano: 14.8.2019. *(Citirano na strani 1.)*

[28] Sahir. The use of social media drastically increasing in teens, 2018. Dostopano: 14.8.2019. *(Citirano na strani 1.)*

[29] S. Salim. How much time do you spend on social media? research says 142 minutes per day, 2019. Dostopano: 14.8.2019. *(Citirano na strani 1.)*

[30] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 1994. *(Citirano na strani 23.)*

[31] Shopify. Navigating twitter. Dostopano: 6.6.2019. *(Citirano na strani 9.)*

[32] Statista. Most popular social networks worldwide as of july 2019, ranked by number of active users (in millions). Dostopano: 11.7.2019. *(Citirano na strani 9.)*

[33] D. Team. Data mining tutorial – introduction to data mining (complete guide), 2018. Dostopano: 5.7.2019. *(Citirano na strani 7.)*

[34] TechoPedia. Java. Dostopano: 18.7.2019. *(Citirano na strani 8.)*

[35] A. Twin. Data mining, 2019. Dostopano: 5.7.2019. *(Citirano na strani 7.)*

[36] Twitter. Twitter developer documentation. Dostopano: 24.7.2019. *(Citirano na strani 12.)*

[37] Twitter4J. Twitter4j. Dostopano: 5.4.2019. *(Citirano na strani 12.)*

[38] M. Tyson. What is the jvm? introducing the java virtual machine, 2018. Dostopano: 18.7.2019. *(Citirano na strani 9.)*

[39] V. Vryniotis. Machine learning tutorial: The max entropy text classifier, 2013. Dostopano: 17.6.2019. *(Citirano na strani 4.)*

[40] C. Whissell. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *(Citirano na strani 14.)*

[41] C. Whissell. Whissell's dictionary of affect in language technical manual and user's guide. Dostopano: 8.6.2019. *(Citirano na strani 14.)*

[42] Wikipedia. Concurrency (computer science). Dostopano: 18.7.2019. *(Citirano na strani 9.)*

[43] Wikipedia. N-gram. Dostopano: 9.6.2019. *(Citirano na strani 6.)*

[44] Wikipedia. Object-oriented programming. Dostopano: 18.7.2019. *(Citirano na strani 8.)*

[45] Wikipedia. Support-vector machine. Dostopano: 10.7.2019. *(Citirano na strani 5.)*

[46] Wikipedia. Tag (metadata). Dostopano: 22.7.2019. *(Citirano na strani 11.)*

[47] Wikipedia. Twitter. Dostopano: 11.7.2019. *(Citirano na straneh 9 in 10.)*

[48] Wikipedia. Url. Dostopano: 9.6.201. *(Citirano na strani 12.)*

[49] Wikipedija. Machine learning applications. Dostopano: 20.8.2019. *(Citirano na strani 2.)*

[50] Wikipedija. Machine learning, types of learning algorithms. Dostopano: 20.8.2019. *(Citirano na strani 3.)*