UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA

SENTIMENTALNA ANALIZA SLOVENSKIH TVITOV S POMOČJO STROJNEGA UČENJA

VID JEROVŠEK

2020
ZAKLJUČNA NALOGA
JEROVŠEK

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

Sentimentalna analiza slovenskih tvitov s pomočjo strojnega učenja

(Sentimental analysis slovenian tweets using machine learning)

Ime in priimek: Vid Jerovšek
Študijski program: Računalništvo in informatika
Mentor: izr. prof. Jernej Vičič
Somentor: doc. dr. Branko Kavšek

Koper, september 2020

Ključna dokumentacijska informacija

Ime in PRIIMEK: Vid JEROVŠEK

Naslov zaključne naloge: Sentimentalna analiza slovenskih tvitov s pomočjo strojnega učenja

Kraj: Koper

Leto: 2020

Število listov: 39 Število slik: 10 Število tabel: 7

Število referenc: 27

Mentor: izr. prof. Jernej Vičič

Somentor: doc. dr. Branko Kavšek

Ključne besede: sentimentalna analiza, strojno učenje, linearna regresija, tviti, klasifikacija

Izvleček:
Zaključna projektna naloga predstavlja problem ugotavljanja sentimentalnega mnenja tvitov z uporabo strojnega učenja. Sentimentalnost je lahko napovedana na tri razrede: pozitivni, nevtralni in negativni, ali pa na dva razreda: pozivni in negativni. V delu so predstavljene nekatere možne metode strojnega učenja in nekateri dosedanji dosežki za slovenski jezik. V člankih pokažejo različne metoda pridobivanja podatkov, pred obdelavo teh in različne metode analiziranja. Predstavljen je tudi naš poskus analiziranja slovenskih tvitov z uporabo strojnega učenja in pridobljeni rezultati.

Key document information

Name and SURNAME: Vid JEROVŠEK

Title of final project paper: Sentimental analysis slovenian tweets using machine learning

Place: Koper

Year: 2020

Number of pages: 39 Number of figures: 10 Number of tables: 7

Number of references: 27

Mentor: Assoc. Prof. Jernej Vičič, PhD

Co-Mentor: Assist. Prof. Branko Kavšek

Keywords: sentimental analysis, machine learning, linear regression, tweets, classification

Abstract: This final project task presents the problem of determining the sentimental opinion of tweets using machine learning. Sentimentality can be predicted either into three classes: positive, neutral, and negative, or into two classes: inviting and negative. The paper presents some possible methods of machine learning and some achievements so far for the Slovene language. The articles show different methods of obtaining data, processing these and different possible methods for analysis. Our attempt to analyze Slovenian tweets using machine learning and the obtained results are also presented.

Zahvala

Zahvalil bi se mentorju izr. prof. dr. Jerneju Vičiču in somentorju doc. dr. Branku Kavšku za vso podporo, strokovno pomoč in usmeritve tako pri zaključnem delu, kot v času izobraževanja. Prav tako bi se zahvalil družini in prijateljem za vso podporo, ki so mi jo izkazali na izobraževalni poti.
Hvala!

Kazalo vsebine

Kazalo tabel

Kazalo slik in grafikonov

Seznam kratic

NL	natural language (naravni jezik)
NB	Naivni Bayes
JVM	Java virtual machine (Javina virtualna platforma
CPE	centralna procesna enota
SMS	short message service (kratka sporočilna storitev)
DM	direct message (neposredno sporočilo)
TXT	standardni format za besedilne datoteke
URL	uniform resource locator (enotni naslov vira)
CSV	comma-seperated value (Datoteka z vrednostmi, locenimi z vejico)
ARFF	Attribute-Relation File Format (format zapisa podatkov, ki ga uporablja WEKA)

1 Uvod

»Pero je močnejše od meča.« je pregovor, ki nam pove kolikšen vpliv ima prosta komunikacija (zlasti pisna). V življenju podamo veliko subjektivnih mnenj o različnih tematikah. V času informacijske dobe veliko osebnih mnenj napišemo na različna socialna omrežja. V začetku leta 2020 je več kot 4,5 milijarde ljudi uporabljalo internet. Obstaja več kot 3,8 milijarde aktivnih uporabnikov socialnih omrežij. To je za 9 odstotkov več kot prejšnje leto. Poleg najpopularnejšega Facebooka je popularen tudi Twitter, ki je na 13. mestu najbolj uporabljenih socialnih omrežij na svetu [1]. Največ uporabnikov je starih med 25 in 34 let. Za tem sta skupini 18-24 in 35-49 let [2].
Velika količina ljudi prinese veliko količino objav – podatkov. Zato dandanes izvajamo različne analize besedil, da bi iz podatkov pridobili čim večjo količino informacij. Takšne informacije veliko podjetij uporablja za načrtovanje produktov, da bi povečali dobiček. Drugi primer uporabnosti je med volilnim obdobjem. Z analizo lahko ugotovimo kaj trenutno meni javnost o političnih kandidatih [3].
V diplomskem delu bo predstavljena ena iz med metod analize besedila in sicer sentimentalna analiza besedila. Obstaja več metod analiziranja sentimenta. Najpogostejše so z uporabo sentimentalnih leksikonov, uporaba strojnega učenja in hibridi med njima. V tej nalogi se bomo osredotočili predvsem na metodi strojnega učenja, natančneje na uporabo Naivnega Bayesa(NB) in logične regresije.

1.1 Sentimentalna analiza

Ljudje smo družabna in sentimentalna bitja in zato vsi podajamo veliko svojih sentimentalnih mnenj. Sentimentalna analiza je serija metod, tehnik in orodij za detekcijo in analizo subjektivnih informacij kot so mnenje in odnos.
Pri analizi je potrebno biti pozoren tudi na perspektivo. Na isto mnenje lahko gledamo iz več različnih zornih kotov. Avtorjev (ang. opinion holder) pogled je lahko drugačno od bralčevega, ki prebere mnenje. Za primer lahko vzamemo stavek: »Cena delnice je zopet padla, kar je slabo za podjetje.« Avtor zapisa govori o negativni novici. Vendar bralci lahko gledajo na novico z več različnih zornih kotov. Za lastnike delnic je ta novica vsekakor slaba, vendar za potencialne kupce je lahko to dobra novica, saj bodo delnice lahko kupili po nižji ceni [4] [5].

Analiza sentimenta se lahko v splošnem izvaja na treh različnih nivojih: [5]

- Nivo dokumenta: Na tem nivoju je cilj analize določiti ali je dokument kot celota izraža pozivno ali negativno mnenje. Primer besedila je lahko mnenje kupca o produktu, ki ga je oddal na spletni strani. Interes podjetja je ugotoviti, ali je kupec zadovoljen z njihovem produktom oziroma ali je mnenje kot celota negativno ali pozitivno. Ta nivo analize predvideva, da je v dokumentu izraženo le mnenje o eni sami entiteti (produktu).
- Nivo stavka: Cilj na tem nivoju je določitev negativnega, nevtralnega ali pozitivnega sentimenta za vsak stavek posebej. Če stavek ne izraža subjektivnosti, je nevtralen.
- Nivo vidika: Je najkompleksnejši nivo od vseh treh. Na prejšnjih nivojih ne vemo, kaj točno je bilo dobro ali slabo. Na tem nivoju namesto jezikovnih gradnikov gledamo z vidika razpoloženja (pozitivnega ali negativnega) in tarče mnenja. S tem lahko za vsako mnenje natančno povemo na katero entiteto se nanaša in s tem povečamo natančnost. V stavku »Čeprav plača ni dobra, mi je to delovno mesto zelo všeč.« lahko opazimo, da je izraženo mešano mnenje. Če se osredotočimo na vsak vidik posebej, vidimo, da je izraženo negativno mnenje glede strežbe, pozitivno mnenje pa je izraženo za delovno mesto. S tem se zviša natančnost pri sentimentalni analizi, vendar je tak postopek tudi posledično ustrezno bolj kompleksnejši za analizo od prejšnjih.

L. Bing omeni različne izraze, ki se pojavljajo dandanes kot sopomenke izraza sentimentalna analiza: rudarjenje mnenj, ekstrakcija mnenj, rudarjenje sentimenta, analiza subjektivnosti, analiza vplivnosti, analiza emocij in tako dalje. Najpogosteje delimo mnenja na dva načina: pozitivno – negativno in pozitivni – nevtralno - negativno. Objekt takšne analize je največkrat storitev ali produkt, čigar ocena je javno objavljena na internetu. Skozi leta je nastalo več načinov pristopa. Te se lahko razvrsti v tri glavne skupine: uporaba sentimentalnih leksikonov, uporaba strojnega učenja in hibridi med njima [5].

1.1.1 Sentimentalni leksikoni

Najpomembnejši indikatorji sentimenta so sentimentalne besede. To so besede, ki so pogosto uporabljene za izražanje pozitivnega ali negativnega mnenja. Na primer: super, čudovito, odlično so pozitivne sentimentalne besede, medtem ko pa zanič, slabo, grdo so negativne sentimentalne besede. Seveda samo besede niso dovolj, saj je problem veliko bolj kompleksen. L. Bing [5] je izpostavil naslednje probleme:

1. Pozitivna ali negativna sentimentalna beseda ima lahko različen pomen v različnih kontekstih. Beseda »brutalen« ima po navadi negativen sentiment, npr. »Pretep je bil brutalen."lahko pa pomeni pozitiven sentiment, npr. » Žur je bil brutalen.«

2. Poved z vsebovano sentimentalno besedo morda nima sentimentalnega pomena. Pojav je pogost v vprašalnih in pogojnih povedi, na primer: »Kateri telefon je dober?« in »Če najdem dobri telefon, ga bom kupil.« Obe povedi vsebujeta besedo dober, vendar ne izrazita pozitivnega ali negativnega sentimenta. To pa seveda ne pomeni, da vse povedi take vrste ne vsebujeta sentimenta.

3. Povedi s sarkazmom ne glede na vsebovanost sentimentalnih povedi je težko obravnavati. Sarkazem ni zelo pogost v kritikah izdelkov, so pa zelo pogoste v političnih mnenjih.

4. Veliko povedi brez sentimentalnih besed lahko nakazujejo na sentiment. Mnogo takih povedi je objektivnih, ki vsebujejo dejanska dejstva. Za primer vzemimo »Ta žarnica porabi veliko elektrike.« nakazuje na negativen sentiment o žarnici, saj porabi veliko energije.

Leksikone delimo na: [6]

- Splošni: V slovarju so označene besede na brezkontekstni pomen, kar pomeni, da ima takšen sentiment večini primerov.

- Domenski: Slovar je prilagojen posamezni domeni, kar lahko drastično poveča natančnost analize, saj ima lahko beseda različno sentimentalno vrednost, odvisno od domene v kateri je uporabljena [7].

2 Strojno učenje

Strojno učenje je izraz zelo širokega pomena. Ena iz med definicij se glasi: »Študija računalniških algoritmov, ki se avtomatsko izboljšujejo z izkušnjami.« Prvi jo je leta 1959 Arthur Samuel definiral kot iskanje rešitev na problem brez eksplicitno določenih poti ali načinov [8].
Pogosto je videna kot podvrsta umetne inteligence. Algoritem zgradi matematični model na podane vzorčne podatke, ki se imenujejo vadbeni nabor podatkov (ang. training set), da bi naredil odločitev ali napoved brez, da bi bil za to eksplicitno programiran. Takšni algoritmi se pogosto uporabljeni na različnih področjih, npr. filtriranje elektronske pošte, priporočanje izdelkov na spletu, itd [9].
V zadnjih desetletjih se je strojno učenje zelo populariziralo. Najbolj je k temu pripomogel hitri razvoj računalništva – vse zmogljivejši računalniki. Seveda pa je pomemben podatek, da se vsako leto število podatkov podvoji.

2.1 Atributi

Osnovne dele podatkovnih zbirk za strojno učenje imenujemo atributi (ang. features). Izbira informativnih, razlikovalnih in neodvisnih atributov je ključnega pomena za učinkovite algoritme pri prepoznavanju, klasifikaciji in regresiji. Atributi se v podatkovnih zbirkah pojavijo kot stolpci, največkrat pa zavzamejo numerične vrednosti. Druge večkrat uporabljene oblike pa so besedilo in grafi. Pomembno pa je, da se vsaki instanci (vrstici) pojavijo konsistentno, da se lahko model na njih uči [10].
V primeru prevelike podatkovne zbirke uporabimo proces imenovan ekstrakcija atributov (ang. feature extraction). Proces je potrebno prilagoditi vsakemu primeru, cilj pa je dobiti manjšo skupino podatkov, ki nimajo redundantnih podatkov in lahko predstavijo celostno zbirko podatkov [11]. Obstaja pa več splošnih tehnik, ki so pogosto uporabljene, nekatere pa so:

- ICA (ang. independent component analysis): je računalniška metoda za ločevanje multivariatnega signala na aditivne pod komponente. Pogost primer uporabe je pri poslušanju govora ene osebe v hrupni sobi [12].
- PCA (ang. principal component analysis): glede na zbirko točk v dveh, treh ali višjih dimenzijah je mogoče najprimernejšo črto definirati kot črto, ki zmanjša

povprečno razdaljo kvadrata od točke do črte. Naslednjo najprimernejšo linijo lahko podobno izberemo iz smeri, ki so pravokotne na prvo.

- MDR (ang. multifactor dimensionality reduction): je statistični pristop, ki se uporablja tudi pri avtomatskih pristopih strojnega učenja, za zaznavanje in karakterizacijo kombinacij atributov ali neodvisnih spremenljivk, ki medsebojno vplivajo na vplivno odvisnost ali spremenljivko razreda. Zelo pogosto uporabljena v genetiki [13].

- Autoencoder: je vrsta umetna nevronska mreža, ki se uporablja za nenadzorovano učenje učinkovitega kodiranja podatkov. Cilj mreže se je naučiti najti zastopajoče atribute in podatke, pri čemer ignorira »hrup« (ang. noise) v podatkih.

2.2 Klasifikacija

Proces ugotavljanja oziroma razporejanje podatkov v različne skupine imenujemo klasifikacija. To se ponovi za vsako instanco v podatkovni zbirki. Obstaja več vrst klasifikacij. Pri strojnem učenju uporabljamo nadzorovano učenje (ang. supervised learning). Koraki so naslednji:

- Določite vrsto vadbenega nabora podatkov. Najprej mora uporabnik določiti kakšno vrsto podatkov bo uporabil za učenje. Za primer teksta je lahko en znak, ena beseda ali pa celotna poved.

- Zberite vadbeni nabor podatkov. Nabor mora biti reprezentativen za dejanski problem. Zato morajo vhodne in izhodni objekte izbrati strokovnjaki na tem področju.

- Določite reprezentativne atribute za vhodne objekte naučene funkcije. Natančnost naučene funkcije je močno odvisna od tega koraka. Najpogosteje je objekt spremenjen v atributni vektor, ki predstavlja podatke vhodnega objekta.

- Zaženite naučen algoritem nad vadbenem naborom podatkov.

- Ocenite natančnost algoritma. Po končani optimizaciji parametrov izmerite natančnost algoritma na testnem naboru podatkov, ki je ločen od vadbenega nabora podatkov.

Poznamo več različnih klasifikatorjev. V nadaljevanju je opisanih nekaj najenostavnejših ali najpogosteje uporabljenih.

2.3 Naive Bayes

V strojnem učenju so Naivni Bayes klasifikatorji družina enostavnih »verjetnostnih klasifikatorjev«. Temeljijo na Bayesovem teoremu z predpostavko o močni (naivni) neodvisnosti med atributi. Bili so temeljito preučeni že v 60tih letih prejšnjega stoletja. Dandanes so še vedno popularni pri klasifikaciji besedila, na primer določitev ali je pošte legitimna ali nezaželena (ang. spam). Drugi primer je ali besedilo spada pod šport ali pod politiko in tako dalje.
Njegovi popularnost pripisujemo zaradi hitrosti in enostavni implementaciji. Pomembna je tudi lastnost, da se lahko razširi na večjo količino podatkov. Odloča se po preprosti formuli:

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

P(A—B) je verjetnost dogodka A, kjer se je dogodek B že zgodil. V našem primeru je A vrednost razreda, B pa vrednost atributa. Obstaja več vrst algoritmov Naivnega Bayesa. Najgosteje uporabljeni so:

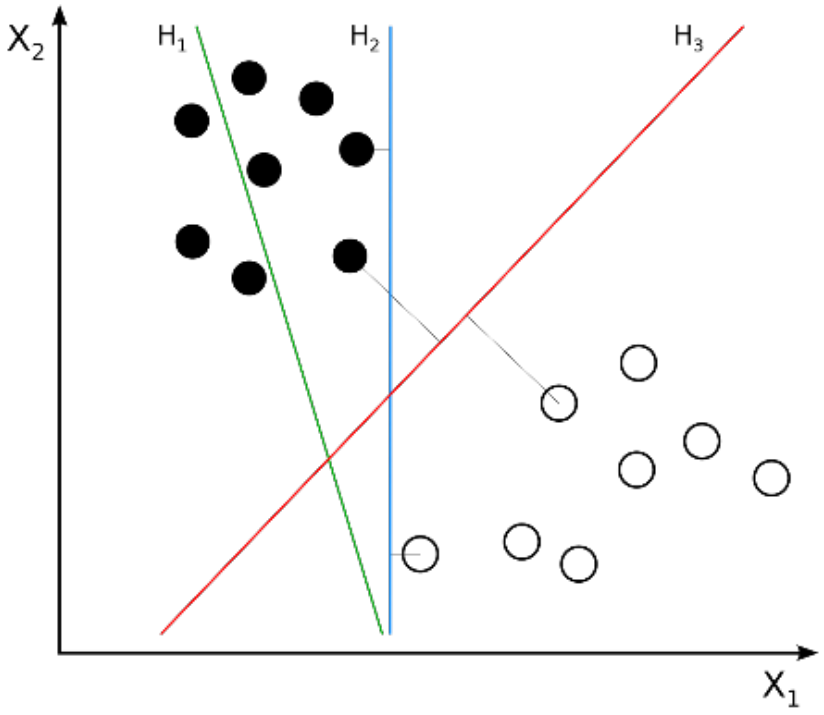
- Multinomial Naive Bayes: večinoma uporabljen pri klasificiranju besedila, kjer je besedilo označeno glede na njihovo vsebino, na primer: šport, politika, itd. V našem primeru se bo odločal o sentimentalni vrednosti besedila. Model napove razred glede na pogostost pojavitev besed v besedilu.

- Bernoulli Naive Bayes: podoben prejšnjemu, vendar so vsi vhodni atributi zavzamejo eno od dveh logičnih vrednosti – 0 ali 1. Pogosto uporabljen v kratkih besedilih, kjer se določi vrsto besedila s pojavitvijo določene besede.

- Gaussian Naive Bayes: uporabljen kadar se ukvarjamo z neprekinjenimi vrednosti, nad katerimi predpostavimo, da so razporejeni po normalni (Gausovi) krivulji.

2.4 Suport Vector Machines

Metoda podpornih vektorjev je algoritem, ki v n-dimenzionalnem prostoru nariše vektor, ki razdeli podatke v dve skupini. Na začetku je obstajala le binarna verzija, sedaj pa se uporablja način, kjer se večji problem razdeli na več manjših binarnih problemov. Rešitev je boljša, če je razdalja med vektorjem in podatki večja.



Slika 1: Možne vektorske rešive[2]; Rešitev H1 (zelena) ni primerna, H2 in H3 sta primerni. Vendar je pri H3 vsota pravokotnih razdalj večja in je posledično boljša rešitev.

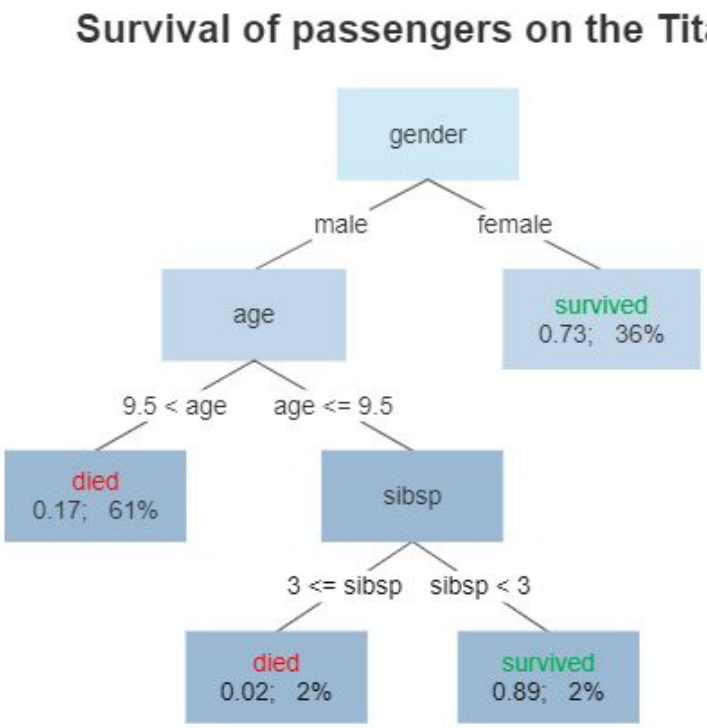
Pogosto se uporablja v različnih primerih:

- Kategorizacija besedila: močno zmanjša potrebo po označevanju vadbenega nabora besedila. Največji tekmec Naivnega Bayesa.
- Klasifikacija slik: raziskave so pokazale višjo natančnost od tradicionalnih metod.
- Prepoznava ročno pisanih črk.
- Uporaba v biologiji in drugih znanostih: uporabljeni za klasifikacijo proteinov z do 90% natančnostjo [14].

2.5 Decision tree learning

Učenje z uporabo odločitvenih dreves je statistični pristop do klasifikacije podatkov. Uporablja odločitvena drevesa, da opiše vhodne objekte. Vejitve so na podlagi vhodnih atributov. Na zadnjem nivoju odločitvenega drevesa so zapisane rezultati klasifikacije.

[2]Vir:`https://en.wikipedia.org/wiki/Support_vector_machine#/media/File:Svm_separating_hyperplanes_(SVG).svg`



Slika 2: Primer odločitvenega drevesa[4];Prikazano je odločitveno drevo, ki prikazuje klasifikacijo preživelih in umrlih ob potopu ladje Titanik. Za vsako kategorijo so še zapisani verjetnost preživetja in delež potnikov spadajočih v to kategorijo.

Na primeru zgornje slike lahko opazimo klasifikacijo na preživele in preminule ob potopitvi ladje Titanik. Veliko verjetnost, da si preživel, če si bil ženskega spola ali pa moški mlajši od 9.5 let in imel manj kot 3 sorojence.
Metoda ima veliko prednosti in nekaj slabosti.
Prednosti so:

- enostavne za razumevanje in interpretacijo,
- lahko uporabimo numerične in podatke razpredeljene na kategorije,
- potrebno je malo predpriprave podatkov,
- lahko je razumeti razloge za dobljene rezultate,
- dobro deluje z velikimi količinami podatkov.

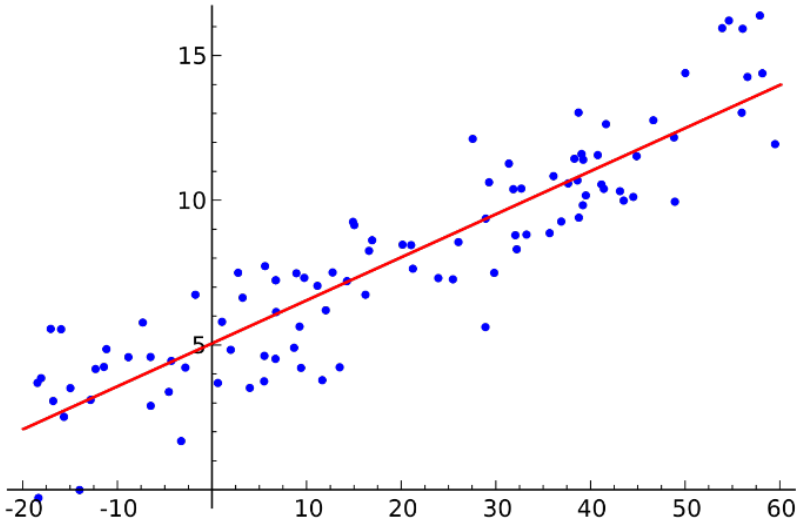
Slabosti pa so:

[4]Vir:`https://en.wikipedia.org/wiki/Decision_tree_learning#/media/File:Decision_Tree.jpg`

- lahko so zelo nerobustna – z majno spremembo podatkov v fazi učenja se lahko rezultati zelo spremenijo
- problem iskanja najoptimalnejše rešitve je NP-complete
- drevesa lahko postanejo preveč kompleksna in se prilagodijo vadbenim podatkom

2.6 Regression analysis

Regresivna analiza predstavlja veliko različnih statističnih metod za ocenjevanje odnosa med različnimi atributi. Najpogostejša oblika je linearna regresija, kjer se izriše premica, ki najbolj ustreza različnimi matematičnimi pogoji. V primeru sentimentalne analize se nariše premica za vsak razred. Na to se vsi podatki klasificirajo kot je najbližja premica [15].



Slika 3: Primer linearne regresije [6]; Na sliki je narisana premica, ki se najbolje prilega vsem podatkom. Najboljše prileganje pomeni najmanjša vsota pravokotnih oddaljenosti točk od premice.

[6] Vir:https://en.wikipedia.org/wiki/Regression_analysis#/media/File:Linear_regression.svg

3 Uporabljena orodja in tehnologije

Za sentimentalno analizo slovenskih tvitov z uporabo strojnega učenja smo uporabili različne orodja in tehnologije. Za pridobitev podatkov in analizo smo uporabili Python, nato pa Javo. Opis je v tem poglavju.

3.1 Python

Python je splošno namenski visokonivojski programski jezik. Jezik podpira proceduralne, objektno orientirane in funkcijske pristope. Referenčno implementacijo Phytona je CPhyton, ki jo razvija in vzdržuje globalna skupnost programerjev. Pythonovi tolmači (ang. Interpreter) so na voljo za mnogo različnih operacijskih sistemov [19]. Python je zasnovan, da v svojem jedru nima veliko funkcij, je pa zato zelo razširljiv z uporabo različnih modulov. Zaradi velikega števila funkcij za obdelavo besedila, je zelo popularen tudi pri obdelavi naravnega jezika. V naši nalogi smo uporabili knjižnice sys, argparse, re, in tweepy, ki nam omogoča dostop do Twitter API.

3.2 Java

Java je visokonivojski programerski jezik, ki se prvič pojavi leta 1995. Je splošno namenski in eden od najpogosteje uporabljenih jezikov. Zasnovan je na principu »write once, run anywhere«, kar pomeni, da se lahko isti program lahko uporablja na katerikoli napravi ali sistemu, ki podpira Javo, brez ponovnega prevajanja. Druge glavne lastnosti Jave so [20] [21]:

- objektno usmerjenost (ang. object oriented),
- večnitnost (ang. multi-threaded),
- neodvisnost računalniškega okolja (ang. platform independet).

3.2.1 Objektno usmerjenost

Java uporablja koncept razredov. Znotraj se nahaja vsa koda programa. Razredi vsebujejo opise podatkovnih polj, ki vsebujejo vse informacije za delovanje razreda in metode, ki izvedejo zaželeni del kode. Prednost takšne oblike je preglednost in razčlenjenost, saj vsak razred rešuje le en podproblem, kar nam omogoča možnost večkratne uporabe in modularnost programske kode [22].

3.2.2 Večnitnost

Večnitenje (ang. multi-threading) je zmožnost izvajanja več zaporednih delov kode hkrati. V Javi jo dosežemo z uporabo niti (ang. threads). Program se razdeli več med seboj neodvisnih delov zaporedne kode in se izvede istočasno. To je zelo koristno pri zahtevnejših ponavljajočih nalogah, saj nam lahko skrajša čas izvajanja za n-krat. V Javi je potrebno vsako nit določiti kot svoj razred.

3.2.3 Neodvisnost računalniškega okolja

Računalniško okolje (ang. platform) je skupek tehnologij (strojna oprema in operacijski sistem), na kateri se gradijo vse aplikacije, procesi, ... Java ima svojo platformno neodvisno JVM (Java virtual machine), ki omogoča prenosnost programov na kakršnokoli kombinacijo strojne opreme in operacijskega sistema [23].
Vsi Java programi se prevedejo v bitno kodo (byte code), kjer so zapisani ukazi za virtualni CPE. JVM na uporabnikovi napravi prevede bitno kodo v strojne ukaze, ki jih podpira uporabljeni procesor. To je pomembno, saj različni procesorjih lahko imajo različne strojne ukaze ali pa so le ti različno zakodirani. V ta namen obstaja več različnih JVM, vsaka namenjena svoji platformi.

3.3 Twitter

Twitter je ameriška mikroblogarska in socialno omrežna storitev, ustanovljena leta 2006. Na njej uporabniki objavljajo sporočila znana pod imenom »tviti (ang. tweets)«. Ideja Twitterja je bila komunikacija skozi SMS (ang. Short Message Service) za omejeno skupino ljudi. Sedaj so privzeto sporočila javna vsem, lahko pa se objavitelj odloči in omeji samo na svoje sledilce (ang. followers) ali pošlje sporočilo posamezniku preko DM (ang. direct messege). Ko uporabnik nekaj objavi, se besedilo pojavi na objaviteljevemu »zidu (ang. feed)« ter pri vseh njegovih sledilcih.
Dolžina sporočila je bila omejena na 140 znakov, leta 2017 pa se le ta povečala na 280 znakov. Tviti pa niso omejeni le na besedilo, vsebujejo lahko tudi multimedijske vse-

bine kot so slike in videi. Ta od leta 2017 ne odštevajo omejitev znakov na sporočilu. Lahko imajo opis do 480 znakov.
Lastnosti kot so neformalnost in omejenost količine napisanega besedila so vodile k pojavitvi mnogih okrajšav, slenga, emotikov, različnih oznak (, @,. . .) in tako dalje. Neformalnost je hitro opazna, saj veliko ljudi ne preveri črkovanja besed in so zato besedila polna pravopisnih napak, kar je potrebno upoštevati pri analizi besedila [24].

3.3.1 Okrajšave in sleng

Ena iz med glavnih lastnosti tvitov je pojavitev okrajšav in slenga v besedilu. Te so se pojavile zlasti zaradi neformalnosti in omejitve dolžine. Oboje seveda ni omejeno le na twitter, ampak je pogost pojav tudi pri vseh drugih socialnih omrežjih. Obstajajo neformalni dogovori kaj okrajšava pomeni. Seveda pa ima lahko več besed enak krajši zapis, zato je potrebno poznati kontekst besedila, da lahko izvemo, kaj je avtor želel z okrajšavo besede.

3.3.2 Emotikon in emoji

Ljudje smo socialna bitja in radi izrazimo svoje mnenje in razpoloženje. Vendar v neformalnih besedilih ne izrazimo tega vedno le z besedami. Pogost način izražanja so emotikoni in emojiji [25].
Emotikoni so zaporedje znakov iz katerega je mogoče razbrati človeški obraz in izraz čustev na njem. Pojavili so se v 1982, po tem, ko nekdo ni dojel, da je sporočilo šala. Zato je Dr. Scott E. Fahlman predlagal, da se šale in ne šale označijo z dvema zaporedjema - za šale : -) in za nešale : -(. Kmalu je sistem postal zelo popularen, zlasti v SMS sporočilih. Sedaj se delijo na tri glavne stile. Na »zahodnjaškem« načinu pomemben cel obraz, je na »vzhodnjaškem« poudarek predvsem na očeh. Tretji način je »2channel«, ki je predvsem popularen na Japonskem. Zanj je značilno, da so emotikoni lahko bistveno večji in celo narisani v več vrsticah.
Emoji so slikovni prikaz čustev na obrazu. Pojavili so se prvič na Japonskem leta 1999 v obliki 12*12 pikslov sliki. S pojavitvijo pametnih telefonov se je uporaba le teh drastično povečala in postala vsakdanjost v neformalnih sporočilih.

3.3.3 Oznake

Tviti so znani po vključevanju različnih oznak v besedilo. Vsaka oznaka ima svojo funkcijo in pomen. Najpogostejši in najpomembnejši na twiterju sta ključnik (ang. hashtag (#)) in cilj (ang. target (@)).
Ključniki so tip metapodatkovnega označevanja na socialnih omrežjih. Uporabnikom

omogoča označevanje besedil, kar omogoča lažje iskanje drugim uporabnikom pri iskanju podobnih vsebin na enako temo. Sestavljen je iz začetnega znaka »#«, ki mu sledi neprekinjen niz znakov (npr. #DanesJeLepDan).
Cilj je znak za označevanje oseb oziroma uporabnika. Uporablja se, kadar se želi uporabnika obvestiti, da si ga omenil v svoji objavi ali odgovoril na njegovo objavo. Uporablja se znak »@« in nato sledi uporabniško ime drugega uporabnika (npr. @Vid-Jerovsek).

3.3.4 Ostala terminologija

V spodnji tabeli je napisana še preostala uporabna terminologija [24].

Tabela 1: V tabeli je predstavljena ostala pogosta terminologija s slovenskimi prevodi.

Izraz	Slovenski prevod	Pomen
Follow	Sledi	Sledenje nekomu pomeni, da se tebi prikažejo vsi zapisi tiste osebe na tvojem »zidu«.
Who to follow list	Priporočeni za sledenje	Avtomatsko priporočilo twiterja, komu slediti glede na tvoj dosedanji profil.
Unfollow	Prenehanje s sledenjem	Prenajte slediti nekomu, s čimer se njegove objave ne prikažejo na tvojem »zidu«.
Block	Blokiraj	Če ne želiš, da te neka oseba sledi, ga lahko odstraniš in preprečiš ponovno sledenje.
Retweet, RT	Posreduj	Tvit, ki si ga videl deli z svojimi sledilci.
Reply	Odgovori	Odzovi se na tvit drugega.
mentions	Omenitve	Tukaj se pokažejo vsi tviti v katerih si označen.
Direct message	Direktno sporočilo	Zasebno sporočilo, katerega lahko vidi samo prejemnik.
Shortened URLs	Skrajšan URL	Zaradi predolgega spletnega naslova, je dan krajši naslov, ki preusmeri na želeno stran.

3.3.5 Twitter API

Twitter ponuja tviter aplikacijski programski vmesnik (ang. application programming interface), katerega lahko registrirani uporabniki uporabljajo za različne namene. Nabor funkcij je velik, najpogosteje uporabljene so iskanje, pretakanje (ang. streaming), sledenje, objavljanje, sortiranje, ... V naši nalogi smo ga predvsem uporabili za pridobivanje izbranih tvitov.

3.4 CSV datoteka

Datoteka z vrednostmi, ločenimi z vejico (ang. comma-seperated values), je navadna besedilna datoteka, ki vsebuje seznam podatkov. Pogosto se takšne datoteke uporabljajo za izmenjavo podatkov med različnimi aplikacijami. Na primer baze podatkov pogosto podpirajo CSV datoteke [26].
Struktura datoteke je preprosta. V prvi vrstici zapišemo imena atributov ločene z vejico in nato v naslednje vrstice zapišemo vrednosti atributov instanc v enakem zaporedju.

3.5 ARFF datoteka

Datoteka ARFF (ang. Attribute-Relation File Format) je ASCII besedilna datoteka, ki opisuje seznam instanc podatkov, ki si delijo nabor atributov. Podatki so zapisani kot v CSV datoteki, le da je pred tem še definirano, katere atribute imamo in kakšna je njihova oblika. Datoteke ARFF je razvil projekt Machine Learning na Oddelku za računalništvo Univerze v Waikatoju (Nova Zelandija) za uporabo s programsko opremo za strojno učenje Weka [27].

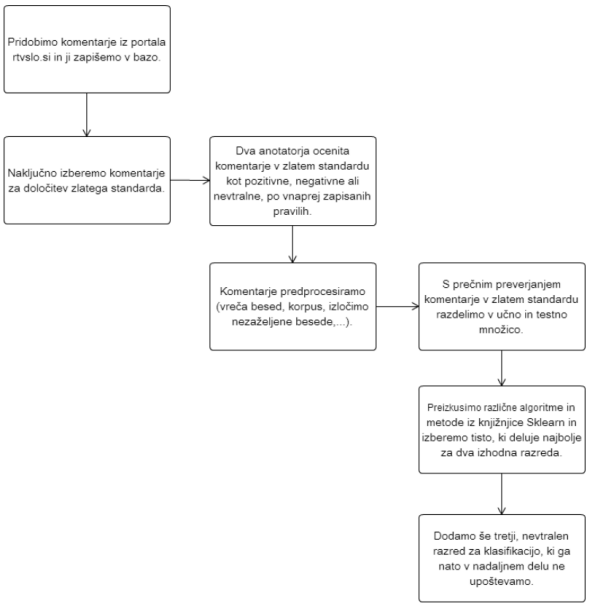
4 Dosedanja sorodna dela za slovenski jezik

Izvedenih je bilo že kar veliko raziskav na temo sentimentalne analize. Vendar je večino teh raziskav narejeno za angleški in kitajski jezik. V tem poglavju pa bodo opisane nekatere raziskave narejene za slovenski jezik.

4.1 Brina Škoda 2013

V diplomski nalogi se Brina Škoda osredotoči na klasifikacijo komentarjev portala rtvslo.si.
Najprej so pridobili komentarje in izbrali 500 med njimi. Nato sta jih dva označevalca (ang. annotators) označila in porazdelila v tri kategorije (pozitivni, nevtralni, negativni). Za tem je izločila vse nevtralne komentarje in tiste, kjer se označevalca nista strinjala. Ostalo so 301 komentarji. Zaradi pretežno negativnih komentarjev, so uporabili le komentarje iz športa, ki je bila edina kategorija z enakomerno porazdelitvijo. Proces so ponovili in pridobili so 511 označenih komentarjev iz kategorije športa. Pri pred procesiranjem podatkov so izločili še besede, ki imajo nevtralen pomen, ločila in števila.
Izbrali so 5 algoritmov strojnega učenja in jih primerjali med sabo. Komentarje so klasificirali v dva razreda in dobili rezultate prikazana na sliki 5.

Izkazalo se je, da ima najboljšo natančnost metoda podpornih vektorjev (SVM). Kasneje so še dodali nevtralno kategorijo in določili mejo verjetnosti 70% in točnost metode se je zvišala na 82% [16].



Slika 4: Diagram poteka sentimentalne analize [2]; Na sliki je predstavljeno, kako je v njenem delu potekala pot od zbiranja podatkov do končanega projekta.

		točnost	natančnost	priklic	f1
TF-IDF	UNIGRAMI				
	SVM	73%	72%	70%	71%
	MaxEnt	72%	74%	66%	69%
	NKK	54%	54%	77%	59%
	MNB	71%	76%	58%	66%
	BNB	60%	55%	83%	66%
	BIGRAMI				
	SVM	74%	71%	72%	71%
	MaxEnt	74%	73%	67%	70%
	NKK	53%	52%	76%	57%
	MNB	73%	75%	61%	67%
	BNB	60%	53%	88%	66%
Vreča besed	UNIGRAMI				
	SVM	69%	65%	78%	71%
	MaxEnt	69%	65%	77%	71%
	NKK	58%	54%	88%	66%
	MNB	71%	77%	58%	66%
	BNB	57%	53%	93%	67%
	BIGRAMI				
	SVM	67%	61%	79%	69%
	MaxEnt	69%	63%	80%	70%
	NKK	55%	52%	89%	63%
	MNB	72%	77%	57%	65%
	BNB	51%	48%	97%	64%

Slika 5: Rezultati v zaključni nalogi Brine Škoda [4];Na sliki je narisana tabela rezultatov raziskave.

[2]Vir: ŠKODA Brina: Rudarjenje razpoloženja na komentarjih rtvslo.si, str :40
[4]Vir: ŠKODA Brina: Rudarjenje razpoloženja na komentarjih rtvslo.si, str :50

4.2 Rok Martinc 2013

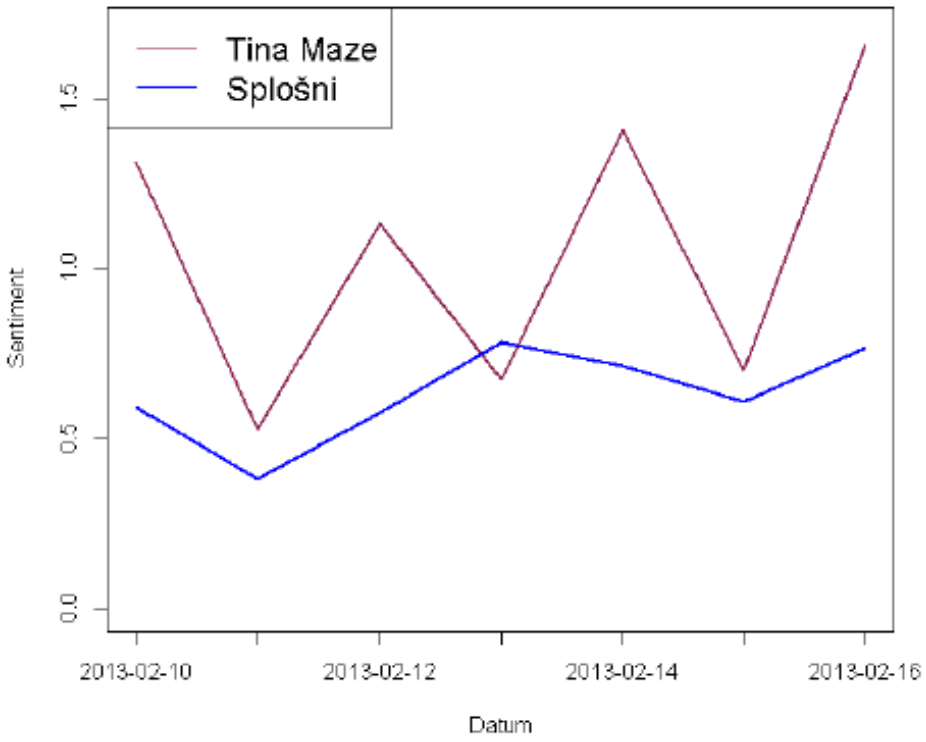
V magistrskem delu se je Martinc leta 2013 osredotočil na analizo mnenja družabnega omrežja Twitter. V tej raziskavi za razliko od prejšnje ni uporabil strojnega učenja temveč leksikonski pristop. Naredil je slovar sentimentalnih besed in jih označil od -5, za zelo negativne ter do 5, za zelo pozitivne. Spodnja slika prikazuje primere besed za lažjo predstavo.

ocena	besede
5	hura
4	hud, mojstrovina, najboljši
3	drzen, fascinanten, hvali
2	ekskluziven, eleganten, gojiti
1	bog, diamant, dogovoriti
-1	ambivalenten, anti, bankir
-2	aretacija, aroganten, bes
-3	brezbrižen, depresiven, dolgočasen
-4	faker, jezni, kreten
-5	baraba, nepridiprav, svinja

Slika 6: Primer sentimentalnega slovarja [6];Primeri besed iz slovenske tabele AFINN-111 s pripadajočo sentimentno oceno.

V besedilu so ovrednotili besede na vrednost sentimenta zapisanega v prej sestavljenem slovarju. V kolikor je bila vsota pozitivna, so označili besedilo kot pozitivno, negativno vsoto pa so označili kot negativni sentiment. Z uporabo te preproste metode so poskušali analizirati podatke o ameriških volitvah, podporo slovenski vladi in sentiment do Tine Maze med svetovnim prvenstvom. Spodaj je podana slika rezultatov pridobljenih o Tini Maze [17].

[6]Vir:KADUNC Klemen: Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega učenja, str: 30



Slika 7: Sentiment do smučarke Tine Maze med svetovnim prvenstvom.[7]

[7]Vir:MARTINC Rok :Merjenje sentimenta na družbenem omrežji Twitter: izdelava orodja ter evaluacija, str:44

4.3 Mateja Volčanšek 2015

V diplomskem delu so se za razliko od prejšnjih osredotočili na analizo formalnih besedil- člankov. Izbrali so leksikalno metodo analize. Primaren cilj naloge je bil izdelava kakovostnega sentimentalnega slovarja namenjen slovenskemu jeziku. Osnova je bil angleški leksikon General Inquirer[8]. Slovar so prevajali tako ročno kot z avtomatskim prevajalnikom. Končna verzija slovarja Beta sestavlja 1669 pozitivnih in 1912 negativnih besed. Za evaluacijo slovarja so ročno označili 5000 besedil.

	napovedani razred			
pravi razred	pozitiven	nevtralen	negativen	vsota
pozitiven	692	409	248	1349
nevtralen	618	558	427	1603
negativen	347	635	1066	2048
vsota	1657	1602	1741	5000

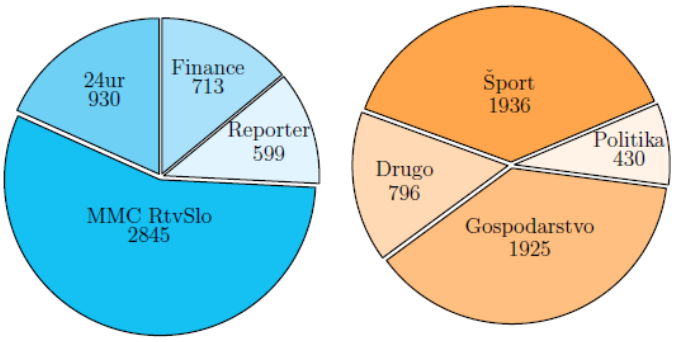
Slika 8: Matrika zmot Mateje Volčanšek[10]; Na sliki so prikazani razultati, ki jih je v svojem Diplomskem delu dobila Mateja Volčanšek. Od 5000 testiranih besedil je bilo pravilno klasificirano le 2316.

Z rezultati niso bili zadovoljni, saj ima slovar natančnost 46%, kar je le 5% boljše od večinskega [4].

4.4 Klemen Kadunc 2016

Kadunc je v svoji diplomski nalogi želel narediti klasifikator in potrebna orodja za analizo neformalnih uporabniških zapisov z uporabo strojnega učenja. Izdelali so lasten sentimentalni leksikon, ki temelji na angleškem slovarju SentiWordNet. Slovar so prevedli ročno s pomočjo uporabe različnih slovarjev kot je PONS[11]. Izdelali in ročno označili so svoj korpus uporabniških komentarjev. Korpus ima 5087 komentarjev. Spodnja slika prikazuje sestavo korpusa po mestu pridobljenega komentarja in kategorija komentarja.

[8]Dostopno na: `http://www.wjh.harvard.edu/~inquirer/homecat.htm`
[10]Vir: VOLČAŠEK Mateja: Leksikalna analiza razpoloženja za slovenska besedila, str: 44
[11]Dostopno na: `https://sl.pons.com/prevod`



(a) Po spletnih virih (b) Po kategorijah

Slika 9: Razporeditev analiziranih komentarjev [13]; Na sliki je prikazana razporeditev pridobljenih komentarjev uporabljenih pri analizi v delu Klemna Kadunc.

S poskusi so ugotavljali, kakšna konfiguracija predpriprave značilk in izbira leksikalnega vira je najboljša. Takšna konfiguracija je imela za več kot 10% izboljšanje v primerjavi s klasičnem modelom in kar več kot 30% višja od večinskega načina. Spodaj so prikazani rezultati njihovega dela [18].

Klasifikator	CA	mera F_1			
		pos	*neg*	*neu*	povp.
osnova	54,6	59,0	55,2	48,8	54,3
LR	63,6	68,1	61,3	61,6	63,7
SVM	63,2	69,0	62,1	58,6	63,2
MNB	**65,5**	68,6	66,8	60,6	**65,3**
BNB	60,1	65,0	56,7	58,4	60,0

Slika 10: Rezultati v delu Klemna Kadunc. [15]; Na sliki so podani rezultati svojega diplomskega dela dobljeni strani Klemna Kadunc.

[13]Vir: KADUNC Klemen: Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega učenja, str: 72

[15]Vir: KADUNC Klemen: Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega učenja, str: 105

5 Implementacija

V tem poglavju bo opisana naša implementacija sentimentalne analize slovesnih tvitov z uporabo strojnega učenja. Naša naloga je, da z uporabo strojnega učenja sentimentalno opredelimo pridobljene slovenske tvite.

5.1 Podatki

Podatki so bili pridobljeni s pomočjo Janes-Tweet korpusa[1]. Korpus vsebuje skoraj 10 milijonov tvitov. Zbrani so bili od približno 9 tisoč uporabnikov, ki tvitajo večinoma v slovenščini, v obdobju od junija 2013 do junija 2016. Korpus je strukturiran v posamezne tvite, skupaj z njihovimi metapodatki. Tviti v korpusu so tokenizirani, stavek segmentiran, beseda normalizirana, morfosintaktično označena, lematizirana in označena z imenovanimi entitetami. Korpus je zaradi Twitter-ovih pogojev storitve distribuiran v kodirani različici. Za dekodiranje je potrebo uporabiti pythonov program tweetpub (na voljo in dokumentiran na https://github.com/clarinsi/tweetpub). Pridobljenih je bilo 1650 tvitov,od tega 362 sentimentalno pozitivnih, 919 nevtralnih in 369 negativnih. Zaradi prevelike neenakosti v porazdeljenosti smo izbrisali nekaj nevtralnih tvitov.

5.2 Slovarji

Izdelali smo dva preprosta sentimentalna slovarja za slovenščino. Pri izdelavi smo se zgledovali na angleški slovar, zgrajen s strani Hu in Liu [5]. Njun slovar ima 4782 sentimentalno negativnih angleški besed in 2006 pozitivnih. Naš izdelani slovar ima po prevajanju in izločanju podvojenih besed 3502 negativnih in 1524 pozitivnih sentimentalnih besed. Zapisana sta v neg.txt in poz.txt tekstovni datoteki in sicer vsaka beseda v svoji vrstici.
Izdelali smo tudi slovarja pozitivnih in negativnih emotikonov. Vključenih je 37 negativnih in 61 pozitivnih emotikonov.

[1] Dostopno na:`https://www.clarin.si/repository/xmlui/handle/11356/1142?locale-attribute=sl`

5.3 Pred obdelava podatkov

Pridobljeni dekodirani podatki so bili zapisani v tekstovni datoteki. Z uporabo lastnega javanskega programa smo iz tekstovne datoteke izvlekli in sestavili celotno besedilo tvita in pripadajoči sentiment. Program se na to sprehodi skozi vsak tvit in prešteje pojavitev pozitivnih sentimentalnih besed, najdenih v poz.txt datoteki. V primeru, da se uporabi tudi slovar emotikonov, ki deluje enako kot prej omenjene slovar in se prišteje vsota k prejšnji številki.
Za tem naredi enako še za negativne besede (najdene v neg.txt). Ko je ta postopek končan, se urejeni podatki zapišejo v CSV datoteko. Na koncu se CSV datoteka pretvori še v ARFF datoteko.

5.4 Rezultati

V tem delu so diplomske naloge so zapisani rezultati sentimentalne analize nad istimi tviti z različnimi parametri in metodami.
Uporabili smo dve različni metodi strojnega učenja in sicer Naivni Bayes ter linearno regresijo. Primerjali smo tudi, kako vpliva slovar emotikonov na rezultate ter razliko uspešnosti pri razvrščanju v tri(pozitivno, nevralno, negativno) in v dve(pozitivno, negativno) kategoriji.
Za preverjanje natančnosti smo uporabili dve metodi preverjanja naučenega modela in sicer desetkratno prečno preverjanje in delitev seta podatkov. Pri delitvi seta podatkov smo uporabili 66% za učenje našega modela in ostalo za testiranje naučenega modela.

5.4.1 Trirazredna klasifikacija

V naslednji tabeli so prikazani rezultati testa razvrščanja v tri skupine in brez uporabe slovarja emotikonov. Vidimo, da je natančnejša linearna regresija s povprečno natančnostjo 49,27%. Naivni Bayes je dosegel le malo nižjo natančnost napovedovanja in sicer s povprečjem 48,46%. Nadaljevali smo enako kot v prejšnjem primeru, le da smo pri predpripravi podatkov uporabili tudi slovar emotikonov.

Tabela 2: V tabeli so predstavljeni dobljeni rezultati za trirazredno klasifikacijo brez uporabe slovarja emotikonov.

	Naivni Bayes	Linearna regresija
10-kratno prečno preverjanje	48,16	50,05
66 delitev	48,75	48,48
skupaj	48,46	49,27

Rezultati so pokazali, da je uspešnejši v tem primeru Naivni Bayes, vendar je natančnost manjša od prejšnjega modela in ima povprečno natančnost 45,69%. Medtem, ko je desetkratno prečno preverjanje pokazalo primerljive rezultate, je delilni način preizkuševanja pokazal bistveno slabše rezultate.

Tabela 3: V tabeli so predstavljeni dobljeni rezultati za trirazredno klasifikacijo z uporabo slovarja emotikonov.

	Naivni Bayes	Linearna regresija
10-kratno prečno preverjanje	48,16	47,13
66 delitev	43,21	43,21
skupaj	45,69	45,17

Pri obeh algoritmih lahko opazimo, da je največ napačnih klasifikacij prišlo pri klasifikaciji pozitivnih in negativnih tvitov, ki so bili klasificiral kot nevtralni.

Tabela 4: V matriki zmot so zapisani rezultati Naivnega Bayesa pri pogojih 10-kratnega prečnega preverjenja, trirazredna klasifikacija in brez uporabe slovarja emotikonov. a = neutral, b = positive, c = negative.

	a	b	c
a	232	62	36
b	156	170	36
c	181	79	109

Zgoraj je prikazana matrika zmot (ang. confuzion matrix) za primer Naivnega Bayesa brez uporabe emotikonskega slovarja.

5.4.2 Dvorazredna klasifikacija

Naše raziskovanje smo nadaljevali s klasifikacijo sentimenta v dva razreda (pozitivni, negativni). Najprej smo odstranili iz naših pridobljenih podatkov vse tvite označene nevtralno. Za tem pa smo ponovili analizo po enakem načinu, kot pri trirazredni klasifikaciji.

Tabela 5: V tabeli so predstavljeni dobljeni rezultati za dvorazredno klasifikacijo brez uporabe slovarja emotikonov.

	Naivni Bayes	Linearna regresija
10-kratno prečno preverjanje	65,53	64,71
66 delitev	63,45	461,85
skupaj	64,49	63,28

Rezultati zapisani v zgornji tabeli nam pokažejo natančnost do 65,53 odstotkov pri Naivnem Bayesu. Malo slabši je bil pri danih parametrih model linearne regresije, sicer s povprečno 63,28% natančnostjo.

Tabela 6: V tabeli so predstavljeni dobljeni rezultati za dvorazredno klasifikacijo z uporabo slovarja emotikonov.

	Naivni Bayes	Linearna regresija
10-kratno prečno preverjanje	64,56	65,25
66 delitev	64,26	64,26
skupaj	64,41	64,76

Pri analizi z dodanim leksikonom emotikonov se pri dvorazredni klasifikaciji pokaže pričakovano izboljšanje rezultatov. Le ti so vseeno boljši v povprečju za manj kot 1-odstotek, s povprečjem 64,76 pri uporabi linearne regresije.

Tabela 7: V matriki zmot so zapisani rezultati Naivnega Bayesa pri pogojih 10-kratnega prečnega preverjenja, dvorazredna klasifikacija in brez uporabe slovarja emotikonov. a = positive, b = negative.

	a	b
a	299	63
b	189	180

Dvorazredna klasifikacija je pokazala, da ima največ napak pri klasifikaciji negativnih tvitov. V zgornji tabeli lahko razberemo, da je naš najboljši klasifikator razvrstil več kot polovico negativnih tvitov v razred pozitivnih. Razlogov za to je več, najpogostejši pa je uporaba sarkazma v besedilu.

6 Sklepna ugotovitev

V diplomskem delu smo obravnavali področje sentimentalne analize in uporaba strojnega učenja pri tem. Preučili smo razloge za popularnost tega področja med raziskovalici in čemu se le to uporablja v vsakdanjem življenju. Cilj praktičnega dela diplomske naloge je bila izdelava preprostega sentimentalnega slovarja ter izvedba analize nad zbranimi slovenskimi tviti z uporabo leksikona in s pomočjo različnih metod strojnega učenja. S serijo poskusov smo ovrednotili uporabo različnih kombinacij strojnega učenja in izdelanih sentimentalnih slovarjev.
Ugotovili smo, da pri strojnem učenju ni vedno dobro imeti več dodatnih slovarjev (npr. slovar emotikonov). Pokazali smo, da uporaba zelo preprostih sentimentalnih slovarjev ni učinkovita, saj so rezultati v primerjavi z bolj dodelanimi slovarji opazno slabši. Naša rešitev ima veliko potencialnih izboljšav. Ustvarjenje več domenskih sentimentalnih leksikonov bi izboljšala rezultate. Sistem, ki bi poskušal ugotoviti pravopisne napake in jih tudi poskusil odpraviti, bi zaradi obdelave neformalnih besedil lahko potencialno bistveno izboljšal rezultate. Potrebno bi bilo tudi dosledno posodabljati slovar, saj je jezik »živ« in se ves čas razvija.

7 Literatura in viri

[1] CHAFFEY, Dave, *Global social media research summary July 2020*, 2020. (Datum ogleda: 2. 6. 2020.) *(Citirano na strani 1.)*

[2] *Distribution of Twitter users worldwide as of July 2020, by age group*, 2020. (Datum ogleda: 2. 6. 2020.) *(Citirano na strani 1.)*

[3] MÄNTYLÄ, Mika V.; GRAZIOTIN, Daniel; KUUTILA, Miikka in The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers, Computer Science Review. *2018*, sprejeto v objavo. *(Citirano na strani 1.)*

[4] VOLČANŠEK, Mateja, *Leksikalna analiza razpoloženja za slovenska besedila*, 2015. *(Citirano na straneh 1 in 19.)*

[5] LIU, Bing, *Sentiment Analysis and Opinion Mining* , 2012. *(Citirano na straneh 1, 2 in 21.)*

[6] WANG, Leyi; XIA, Rui in Sentiment Lexicon Construction with Representation Learning Based on Hierarchical Sentiment Supervision, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. *2017*, sprejeto v objavo. *(Citirano na strani 3.)*

[7] HUANG, Sheng; NIU, Zhendong; SHI, Chongyang in Automatic construction of domain-specific sentiment lexicon based on constrained label propagation, Knowledge-Based Systems. *2014*, sprejeto v objavo. *(Citirano na strani 3.)*

[8] SAMUEL, Arthur L., Some studies in machine learning using the game of checkers. II—recent progress.. *IBM journal of research and development, 1959*, sprejeto v objavo. *(Citirano na strani 4.)*

[9] MITCHELL, Tom, *Machine Learning*, McGraw Hill.1997 *(Citirano na strani 4.)*

[10] BISHOP, Christopher M., *Pattern recognition and machine learning*, 2006. *(Citirano na strani 4.)*

[11] ALPAYDIN, Ethem, *Introduction to Machine Learning*, 2010. *(Citirano na strani 4.)*

[12] HYVARINEN, Aapo, Independent component analysis: recent advances. *Philos Trans A Math Phys Eng Sci., 2013*, sprejeto v objavo. *(Citirano na strani 4.)*

[13] MOTSINGER, Alison A.; RITCHIE, Marylyn D. in Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies, HUm Genomics. *2006*, sprejeto v objavo. *(Citirano na strani 5.)*

[14] GAONKAR, Bilwaj; DAVATZIKOS, Christos in Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification, Neuroimage. *2013*, sprejeto v objavo. *(Citirano na strani 7.)*

[15] FREEDMAN, David A., *Statistical Models: Theory and Practice*, 2009. *(Citirano na strani 9.)*

[16] ŠKODA, Brina, *Rudarjenje razpolo$_z$enjanakomentarjihrtvslo.si*
, 2013. *(Citirano na strani 15.)*

[17] MARTINC, Rok, *Merjenje sentimenta na družabnem omrežju Twitter: izdelava orodja ter evaluacija*, 2013. *(Citirano na strani 17.)*

[18] KADUNC, Klemen, *Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega učenja*, 2015. *(Citirano na strani 20.)*

[19] KUHLMAN, DAve, *A Python Book: Beginning Python, Advanced Python, and Python Exercises*, 2009. *(Citirano na strani 10.)*

[20] *What is Java technology and why do I need it?*, 2020. (Datum ogleda: 5. 6. 2020.) *(Citirano na strani 10.)*

[21] *Write once, run anywhere?*, 2002. (Datum ogleda: 5. 6. 2020.) *(Citirano na strani 10.)*

[22] LEWIS, John; LOFTUS, William, *Java Software Solutions Foundations of Programming Design 6th ed*, 2008. *(Citirano na strani 11.)*

[23] *Is the JVM (Java Virtual Machine) platform dependent or platform independent? What is the advantage of using the JVM, and having Java be a translated language?*, 2019. (Datum ogleda: 5. 6. 2020.) *(Citirano na strani 11.)*

[24] MACLEAN, Fiona; JONES, Derek; CARIN-LEVY, Gail; HUNTER Heather in Understanding Twitter, British Journal of Occupational Therapy. *2013*, sprejeto v objavo. *(Citirano na straneh 12 in 13.)*

[25] GRANNAN, Cydney, *What's the Difference Between Emoji and Emoticons?*, 2016. (Datum ogleda: 15. 6. 2020.) *(Citirano na strani 12.)*

[26] HOFFMAN, Cris, *What Is a CSV File, and How Do I Open It?*, 2018. (Datum ogleda: 17. 6. 2020.) *(Citirano na strani 14.)*

[27] *Attribute-Relation File Format (ARFF)*, 2002. (Datum ogleda: 20. 6. 2020.) *(Citirano na strani 14.)*