Analiza sentimenta objava na Twitteru vezanih za koronavirus

Ivaninić, Alesia

Sveučilište u Rijeci – Odjel za informatiku

Informacijsko komunikacijski sustavi

Alesia Ivaninić

Analiza sentimenta objava na Twitteru vezanih za koronavirus

Diplomski rad

Mentor: Izv. prof. dr. sc. Ana Meštrović

Rijeka, srpanj 2021.

ODJEL ZA INFORMATIKU
SVEUČILIŠTE U RIJECI

Rijeka, 15.6.2021.

Zadatak za diplomski rad

Pristupnik: Alesia Ivaninić

Naziv diplomskog rada: Analiza senimenta objava na Twitteru vezanih uz koronavirus

Naziv diplomskog rada na eng. jeziku: Sentiment analysis of Twitter posts related to coronavirus

Sadržaj zadatka:

Pandemija uzrokovana koronavirusom donosi brojne izazove ne samo za područje medicine već i za druga područja znanosti. Jedan od izazova predstavlja krizna komunikacija u društvenim medijima koja je popraćena infodemijom. Analiza tekstulanih poruka u drštvenim medijima može dati bolji uvid u kriznu komunikaciju. U okviru toga, od posebnog su interesa algoritmi koji omogućavaju automatsko označavanje sentimenta poruka koje se objavljuju na društvenim mrežama. Zadatak diplomskog rada je istražiti i analizirati neke od postojećih pristupa za analizu sentimenta. Nadalje, potrebno je primijeniti postupke na skupove poruka s Twittera vezanih uz koronavirus te prikazati rezultate analize.

Mentor:

Izv. prof. dr. sc. Ana Meštrović

Voditelj za diplomske radove:

Izv. prof. dr. sc. Ana Meštrović

Zadatak preuzet: 25.6.2021.

(potpis pristupnika)

Sadržaj

Sažetak

U prvom dijelu rada ukratko su opisane korištene tehnologije te sama analiza sentimenta.

Drugi dio bavi se analizom sentimenta tweetova vezanih za sveprisutnu pandemiju koronavirusa. Najprije je prikazan način prikupljanja podataka za daljnju upotrebu. Dio podataka preuzet je s Kaggle baze, dok je drugi dio prikupljen koristeći Twitter API.
Zatim je prikazan i način pripreme podataka i njihove obrade za daljnju analizu te vizualizaciju rezultata. Za obradu podataka, korišten je programski jezik Python. Kao najvažnije biblioteke, ističu se „pandas“ za pripremu podataka i baratanje njima. Zatim slijedi „textblob“ za analiziranje sentimenta i dobivanje vrijednosti polariteta. A na samom kraju, korištena je biblioteka „matplotlib“ pomoću koje su prikazani rezultati kroz grafikone.

Nakon prikupljanja podataka i njihove obrade, uslijedila je vizualizacija rezultata. Ovaj dio rada podijeljen je u tri manja dijela, te su rezultati prikazani za svaki od skupova podataka zasebno.

Ključne riječi

Analiza sentimenta, obrada prirodnog jezika, Twitter, tweet, koronavirus, COVID-19, pandemija

Abstract

The first part of the paper briefly describes used technologies and sentiment analysis itself.

The second part deals with sentiment analysis of tweets related to the coronavirus pandemic. First, the method of data collection for further processing is shown. Part of the data was downloaded from the Kaggle database, while the other part was collected using the Twitter API.
Then, the method of data preparation and processing for further analysis and visualization of results is presented. Data was processed using the Python programming language. The most important libraries used are "pandas" for data preparation and manipulation. Then „textblob“ for analyzing sentiment and obtaining polarity values. And at the very end, the "matplotlib" library was used to present the results through graphs.

After data collection and processing, the result visualization followed. This part of the paper is divided into three smaller parts, and the results are presented for each of the datasets separately.

Keywords

Sentiment analysis, natural language processing, Twitter, tweet, coronavirus, COVID-19, pandemic

1. Uvod

Kroz povijest se pojavljivalo više zaraznih bolesti poput kuge, kolere i španjolske gripe. Takve epidemije bolesti utjecale su na društvo, kroz brojne izgubljene živote te gubitke u ekonomiji. Krajem 2019. godine započela je epidemija koronavirusa svjetskih razmjera. Bolest prouzrokovana ovim virusom, skraćeno se naziva COVID-19. Prema podacima sa stranice Worldometer, do sada je zabilježeno 183 157 621 slučajeva zaraze te 3 965 562 smrtna slučaja [1]. Postoji više mišljenja o stvarnom porijeklu ovog virusa koja se mogu razvrstati u dvije glavne kategorije – virus je prirodan ili virus je umjetno stvoren.

Zbog rasprostranjenosti virusa diljem svijeta, kao i zbog nejasnoća vezanih uz njegovo podrijetlo, ova je pandemija veoma česta tema razgovora.

Obzirom da danas jako veliki broj ljudi koristi društvene mreže, one mogu biti riznice korisnih i zanimljivih informacija općenito, a još i više kad je u pitanju pojava koja utječe na cijeli svijet. Imajući uvid u stavove ljudi, te njihove osjećaje vezane uz neku temu, moguće je lakše planirati sljedeće korake ili predvidjeti kako će društvo reagirati na novosti koje se uvode.

Jedan od načina praćenja stavova na društvenim mrežama jest analiza sentimenta. Ovom analizom mogu se automatski detektirati određeni osjećaji i stavovi te se kasnije razvrstati u neku kategoriju. Ona također pomaže raznim organizacijama da imaju saznanja o stavovima u društvu te da primjereno reagiraju. Kako bi rezultati bili čim precizniji, za provedbu analize, potrebno je koristiti čim veći skup podataka.

U ovom radu analizirati će se sentiment objava na Twitteru vezanih uz epidemiju koronavirusa. Slične analize objavljene su i u radovima: [2,3,4,5].

Nakon uvodog poglavlja najprije je dan teorijski pregled područja analize sentimenta i korištenih tehnologija.
Zatim su opisani skupovi podataka korišteni za provedbu analize sentimenta. Jedan od skupova podataka samostalno je prikupljen korištenjem Twitter API-ja i Pythonove „tweepy“ biblioteke, dok su preostala tri skupa preuzeta s Kaggle stranice.
Nakon preuzimanja podataka, obzirom na njihovu nestrukturiranost, bilo ih je potrebno pročistiti i pripremiti za daljnje korištenje.
Kad su podaci bili spremni, izvršen je proces analize sentimenta korištenjem „textblob“ biblioteke.
Na kraju rada, slijedi vizualizacija dobivenih rezultata te njihov opis, te zaključna razmatranja.

2. Analiza sentimenta

Prema Bing Liu, analiza sentimenta (eng. *Sentiment analysis*), koja se naziva i rudarenjem mišljenja (eng. *Opinion mining*), područje je znanosti koje se bavi analizom mišljenja i osjećaja te procjenom stavova ljudi o nekoj temi kao što su proizvodi, usluge, organizacije, pojedinci ili problemi koji se dotiču veće mase ljudi [6].

Analiza sentimenta je kompleksni proces iz domene obrade prirodnog jezika (eng. *Natural language processing, NLP*) koji uključuje pet faza.

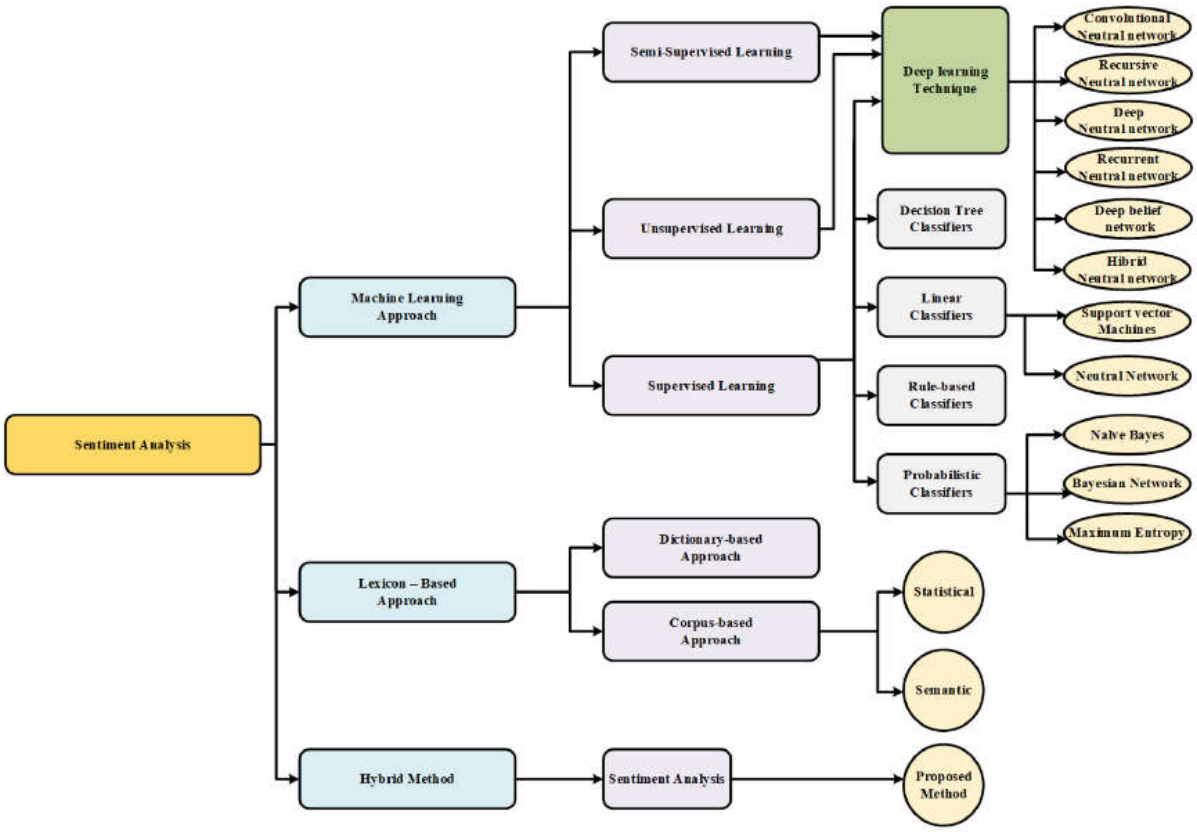
- Prikupljanje podataka napisanih od strane ljudi na raznim blogovima, forumima i društvenim mrežama. Prikupljeni podaci su nestrukturirani, napisani na razne načine koristeći različite vokabulare i žargone. Iz tog razloga, podatke je najprije potrebno pripremiti.
- Priprema teksta sastoji se od čišćenja podataka prije analize. Dijelovi koji nisu tekstualnog oblika i nisu relevantni za analizu, se uklanjaju iz prikupljenih podataka.
- Kod otkrivanja osjećaja, zadržavaju se subjektivne rečenice (rečenice koje sadrže osobna uvjerenja i stavove), a objektivne rečenice (sadrže činjenice) se odbacuju.
- Prilikom klasifikacije osjećaja, subjektivne rečenice se klasificiraju kao pozitivne odnosno negativne.
- Prezentacija rezultata, odnosno pretvaranje nestrukturiranih podataka u značajne informacije jest glavni cilj analize sentimenta. Krajnji rezultati se prikazuju pomoću grafova.

Tekst se klasificira obzirom na polaritet izraženog sentimenta (pozitivan, neutralan, negativan), polaritet posljedica nekih događaja, slaganje i neslaganje s određenom temom i slično.

Postoje tri razine klasifikacije:

- Na razini dokumenta, određuje se izražava li cijelokupno korisnikovo mišljenje negativan ili pozitivan sentiment.
- Na razini rečenice, ukoliko je subjektivna, klasificira se kao pozitivna ili negativna.
- Na razini aspekta, korisnik može imati mišljenje različitog sentimenta ako izražava mišljenje iz više gledišta

Postoje tri pristupa klasifikaciji analize sentimenta opisana u nastavku.



Slika 1 - Različiti pristupi analize sentimenta [7]

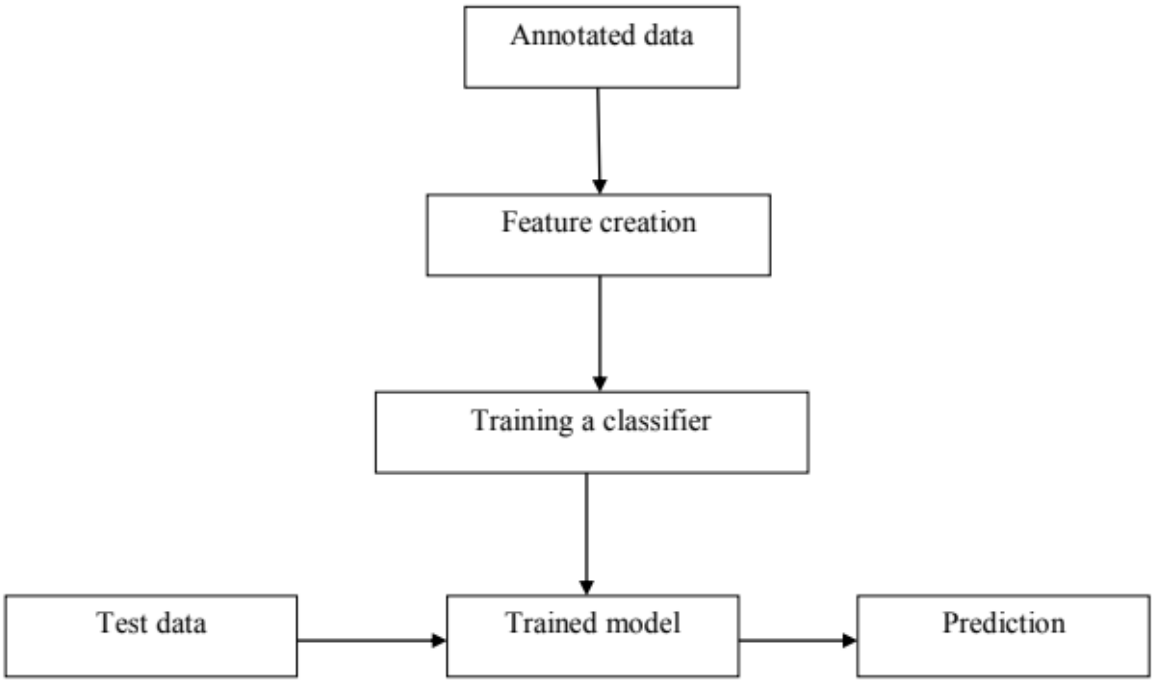
Pristup temeljen na strojnom učenju (eng. *Machine learning, ML*) sastoji se od skupa za treniranje (služi za učenje različitih karakteristika nekog teksta) i skupa za testiranje (koristi se provjeru točnosti nekog klasifikatora). Najčešće se koristi nadzirano strojno učenje. Postoji više algoritama koji se koriste za provjeru točnosti klasifikatora, a neki od njih su linearna regresija (eng. *Linear regression*), logistička regresija (eng. *Logistic regression*), stabla odluke (eng. *Decision tree*), Naivni Bayes (eng. *Naive Bayes*) i potporni vektori (eng. *Support Vector Machine, SVM*).

Također, u zadnjih nekoliko godina razvijaju se i pristupi temeljeni na dubokom učenju (eng. *Deep learning*), koji daju još bolje i preciznije rezultate. Neki od modela dubokog učenja jesu konvolucijske neuralne mreže (eng. *Convolution Neural Network, CNN*), duboke neuronske mreže (eng. *Deep Neural Network, DNN*), ponavljajuće neuronske mreže (eng. *Recurrent Neural Network*) i BERT model.

Najčešće korišteni klasifikatori kod tekstualnih podataka jesu naivni Bayes i metoda potpornih vektora [7]. Naive Bayes klasifikator predviđa vjerojatnost klase iduće riječi na temelju raspodjele riječi u dokumentu. Klasifikator radi tako što zanemaruje položaj riječi u dokumentu. Metoda potpornih vektora, čiji je cilj odrediti linearne separatore koji najbolje razdvajaju različite klase, koristi se kod klasifikacije kraćih tekstova [8].

Za evaluaciju točnosti modela, koristi se nekoliko mjera. *Mjera preciznosti* (eng. *precision*) mjeri točnost klasifikatora kao omjer točno izvađenih mišljenja i ukupnog broja mišljenja. *Mjera opoziva* (eng. *recall*) mjeri potpunost i osjetljivost klasifikatora kao omjer točno izvađenih mišljenja i ukupnog

broja anotiranih mišljenja. *F-mjera* kao jedinstvena ocjena modela dobivena kombinacijom mjera preciznosti i opoziva. I posljednja, *mjera točnosti* (eng. *accuracy*), kao omjer točno predviđenih pojavljivanja i ukupnog broja pojavljivanja.



Slika 2 - Pristup temeljen na strojnom učenju [8]

Osim pristupa temeljenog na strojnom učenju, postoji i *pristup temeljen na rječniku*.
Ovaj pristup koristi rječnik, odnosno leksikon, s ocjenom sentimenta pojedine riječi i povezuje taj rječnik s ulaznim skupom podataka te određuje polaritet. Sentiment se najčešće mjeri na skali u rasponu od -1 do 1, gdje -1 označava negativan, a 1 pozitivan sentiment.
Leksikon sentimenta, rječnik je koji sadrži popis riječi i njima pridružen polaritet. Neki od njih su AFINN, Bing Liu, MPQA, SentiWordNet, VADER i TextBlob [9].

U ovom pristupu, polaritet riječi u tekstu određuje se usklađivanjem riječi iz ulaznog teksta s leksikonima sentimenta, te se na kraju zbrajanjem polariteta dobiva sveukupni sentiment koji se potom klasificira kao pozitivan, neutralan ili negativan sentiment.
Nedostatak ovog pristupa je to što u obzir ne uzima poredak riječi u rečenici, nego samo njihovo pojavljivanje.

Osim dva prethodno navedena pristupa, postoji i *hibridni pristup* koji kombinira pristupe temeljene na strojnom učenju i rječniku.

Analiza sentimenta ima širok spektar primjene, a neki od njih su:

- U poslovanju, za analizu mišljena kupaca o nekom proizvodu, analizu reputacije neke tvrtke ili pojedine marke proizvoda i slično.
- U politici se koristi za predviđanje rezultata glasovanja, određivanje pozicije kandidata za vrijeme izbora
- Kod društvenih događanja, za praćenje događaja u stvarnom svijetu, analiziranje mišljenja prilikom uvođenja novih zakona.
- U financijama se koristi za predviđanje kretanja cijena proizvoda i vrijednosti dionica te za predviđanje rizika prilikom ulaganja.

U ovom radu, sentiment se klasificira na razini dokumenta (pojedinog tweeta) te se koristi pristup temeljen na leksikonu. Analizirat će se, odnosno pratiti javno mišljenje korisnika Twittera o aktualnoj pandemiji koronavirusa u svijetu.

3. Opis podataka

U ovom poglavlju opisana su tri različita skupa podataka vezanih uz pandemiju COVID-19 prikupljena s Twittera. Većina podataka je preuzeta su sa stranice Kaggle – najveće svjetske zajednice za podatkovnu znanost (eng. *data science*), dok je za prvi podatkovni bilo potrebno prikupiti podatke sa Twittera kako bi se dobio skup koji će omogućiti usporedbu podataka.

3.1 Prvi skup podataka

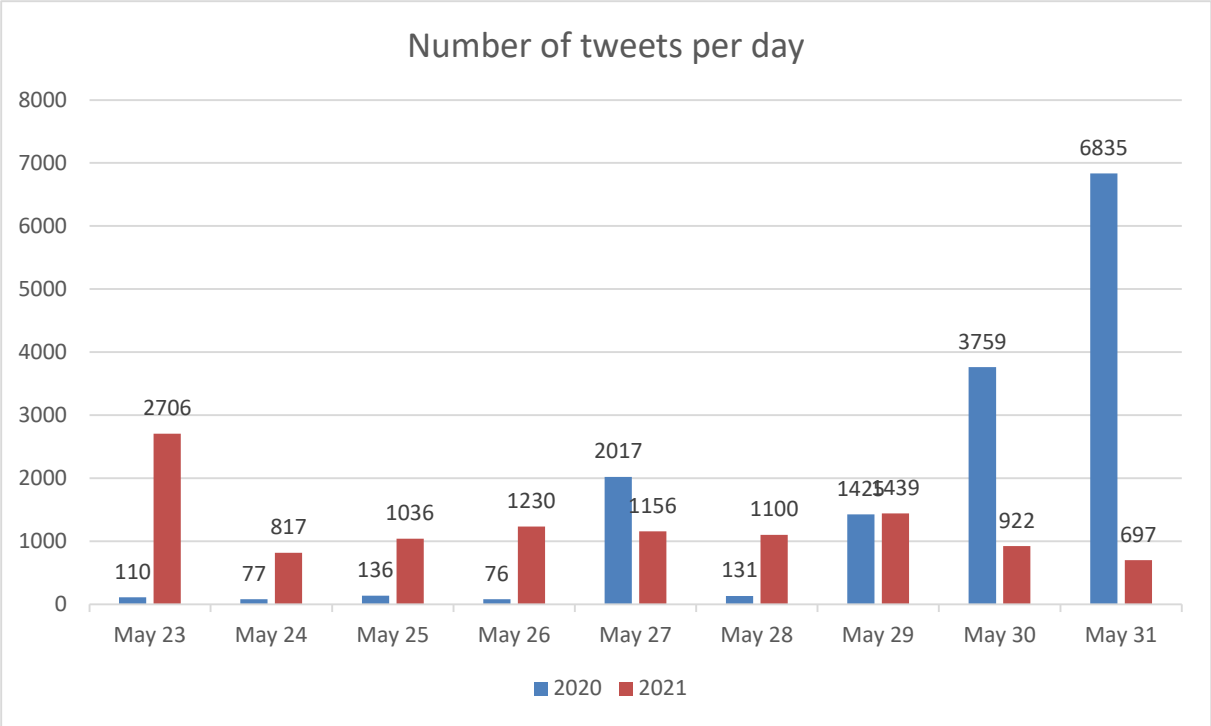
Prvi skup podataka se sastoji od dva podskupa koja će se uspoređivati.

Prvi podskup sadrži tweetove vezane za koronavirus skupljane u razdoblju od 23. do 31. svibnja 2020. godine. Ovaj skup je preuzet gotov sa stranice Kaggle [10]. Filtrirani su podaci samo za razdoblje od 9 dana u svibnju 2020. godine te se sastoji od ukupno 14 566 tweetova. Ukupan broj riječi je 138 414, a prosječan broj riječi pojedinog tweeta je 9.50. Ukupan broj znakova je 1 132 265, dok je prosječan broj znakova po tweetu 77.73. U Tablici 1 je prikazan sažeti prikaz informacija prvom skupu podataka.

Tablica 1 - osnovne informacije o prvom skupu podataka

	Broj tweetova	**Ukupan broj riječi**	**Prosječan broj riječi**	**Ukupan broj znakova**	**Prosječan broj znakova**
Svibanj 2020. godine	14 566	138 414	9.50	1 132 265	77.73
Svibanj 2021. godine	11 103	105 511	9.50	832 675	75.00

U drugom podskupu se nalaze tweetovi vezani za koronavirus skupljeni za razdoblje od 23. do 31. svibnja 2021. godine. Ovaj skup je samostalno prikupljen s Twittera koristeći Twitter API. Skup se sastoji od ukupno 11 103 tweetova. Ukupan broj riječi je 105 511, a prosječan broj riječi pojedinog tweeta je 9.50. Ukupan broj znakova je 832 675, dok je prosječan broj znakova po tweetu 75.00.



Slika 3 - Broj tweetova po danima (prvi skup podataka)

Na prethodnom grafu (Slika 3) prikazana je količina tweetova u promatranom razdoblju. U svibnju 2020. godine, najviše objavljenih tweetova, njih 6835, bilo je posljednjeg dana u mjesecu, dok je u podskupu iz 2021. godine najviše objavljenih tweetova bilo 23. svibnja (2 706).

U prvom podskupu bilo je nekoliko „skokova“ u količini tweetova po danima, pet dana broj tweetova bio je ispod 150, dok u preostala četiri dana broj tweetova varira između 1 425 i 6 835. Dok je u drugom podskupu broj tweetova bio između otprilike 700 i 1 500, osim jednog dana kada se taj broj povećao na 2 705.

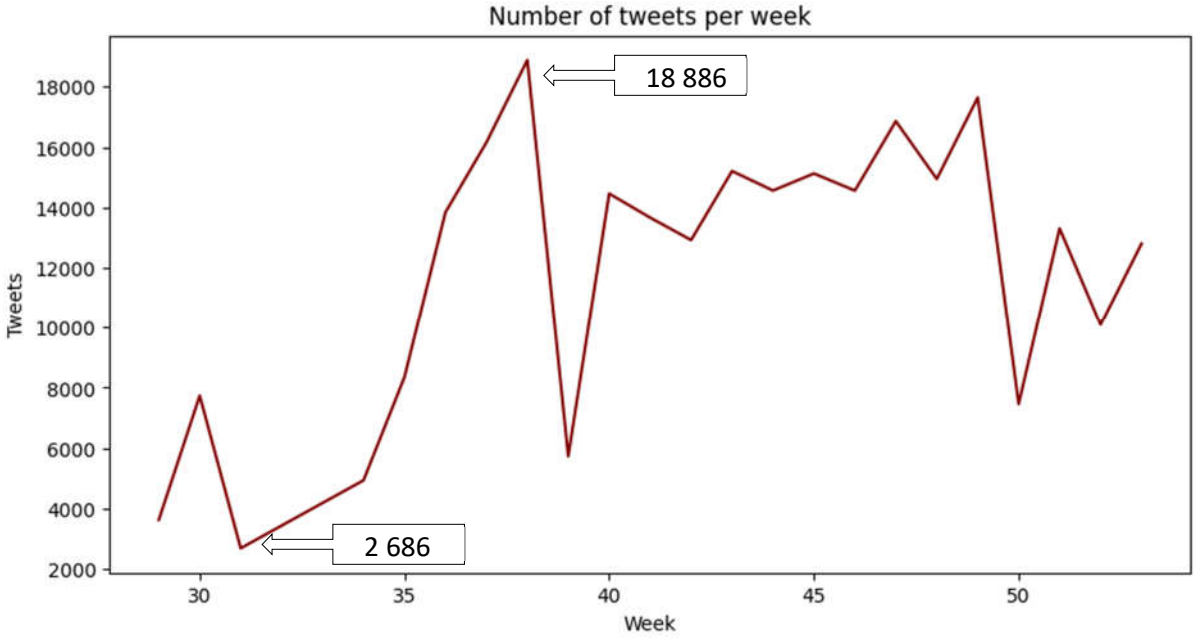
3.2 Drugi skup podataka

Drugi skup podataka je također preuzet s Kagglea [11]. Sastoji se od 275 536 tweetova koji u sebi sadrže oznaku *#Covid19* i *#coronavirus*. Ovaj skup je prikupljen u razdoblju od srpnja 2020. do siječnja 2021. godine. Sastoji se od 2 300 888 riječi i 17 864 973 znakova. Prosječan broj riječi po tweetu je 8.35, a prosječan broj znakova 64.84. Tablica 2 prikazuje sažeti pregled informacija o ovom skupu podataka.

Tablica 2 - osnovne informacije o drugom skupu podataka

Broj tweetova	Ukupan broj riječi	Prosječan broj riječi	Ukupan broj znakova	Prosječan broj znakova
275 536	2 300 888	8.35	17 864 973	64.84

Slika 4 prikazuje kretanje broja tweetova po tjednima. Najmanje ih je objavljeno u 31. tjednu, između 27. srpnja i 2. kolovoza, njih 2 686. Dok ih je najviše objavljeno u 38. tjednu, od 14. do 21. rujna, kad ih je bilo 18 886.



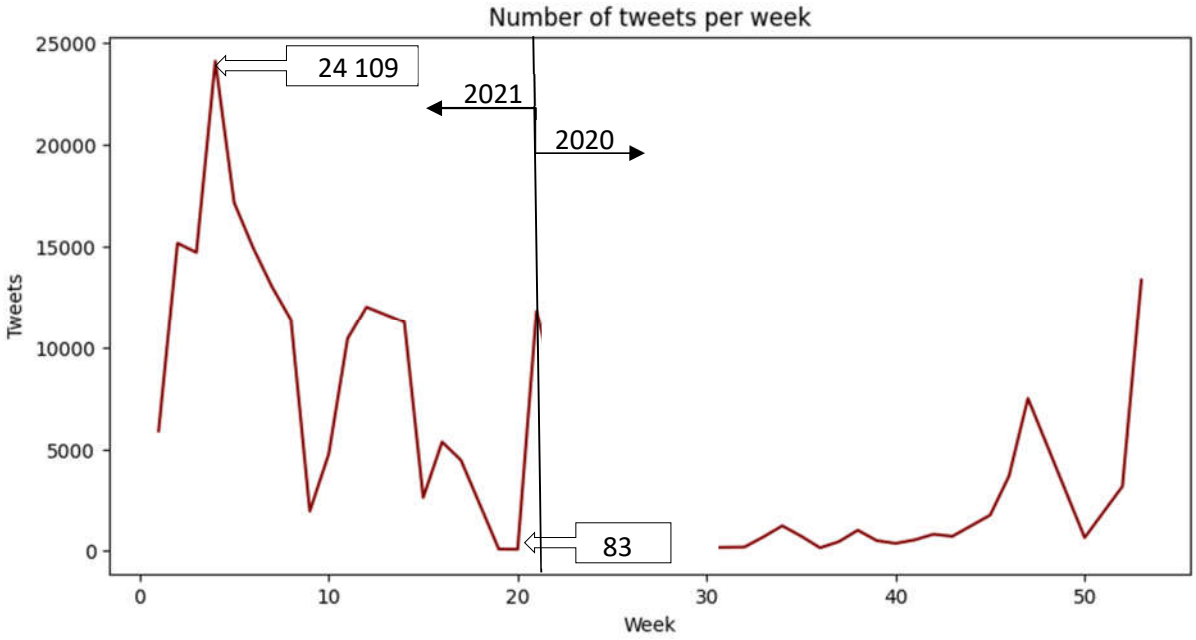
Slika 4 - Broj tweetova po tjednima (drugi skup podataka)

3.3 Treći skup podataka

Posljednji skup podataka je također preuzet s Kagglea [12], a sastoji se od 211 665 tweetova koji sadrže oznaku *#covidvaccine*. Prikupljanje tweetova za ovaj skup podataka je započelo u kolovozu 2020. godine, te se prikupljaju i dalje, a u ovom radu analizirali su se podaci prikupljeni do svibnja 2021. godine. U trenutku pisanja rada, ukupan broj riječi bio je 2 070 267, a prosječan broj riječi po tweetu bio je 9.46. Ukupan broj znakova bio je 16 434 502, s prosječnim brojem znakova 75.08. U Tablici 3 prikazane su osnovne informacije o trećem skupu podataka.

Tablica 3 - osnovne informacije o trećem skupu podataka

Broj tweetova	Ukupan broj riječi	Prosječan broj riječi	Ukupan broj znakova	Prosječan broj znakova
211 665	2 070 267	9.46	16 434 502	75.08



Slika 5 - Broj tweetova po tjednima (treći skup podataka)

Slika 5 prikazuje kretanje broja tweetova po tjednima. Najmanje ih je objavljeno u 20. tjednu 2021. godine, između 17. i 23. svibnja, njih 83. Dok je najviše objavljenih tweetova zabilježeno u 4. tjednu 2021. godine, od 25. do 31. siječnja, kad ih je bilo 24 109.

4. Analiza sentimenta tweetova vezanih za pandemiju koronavirusa

4.1 Prikupljanje podataka

Za prikupljanje podataka korišten je Twitter API (eng. application programming interface). Podaci su filtrirani tako da sadrže ključne riječi na engleskom jeziku koje su povezane s koronavirusom (npr. „covid19“, „lockdown“, „pandemic“ i „covidvaccine“).

Nakon učitavanja potrebnih biblioteka – „tweepy“ za prikupljanje podataka s Twittera i „pandas“ za manipulaciju tim podacima, postavljeni su pristupni podaci za Twitter API. Pomoću Twitter API modula, dobiveni su ključevi i tokeni za njihovu implementaciju u kodu:

```
auth = tw.OAuthHandler("___", "___")
auth.set_access_token("___", "___")
auth.secure = True
api = tw.API(auth, wait_on_rate_limit = True)
```

Zatim se postavlja ključna riječ (eng. *hashtag*) po kojoj će se objave pretraživati te početni datum od kojeg se pretražuju objave:

```
search_words = ['#covid19']#,'#covid19','#coronavirus',
date_since = '2021-01-01'
```

Potom se definira tablica (eng. dataframe) u koju će se željeni podaci spremiti. Za potrebe ovog rada, kodom u nastavku, spremili su se samo korisničko ime i lokacija, te tekst tweeta i vrijeme objave.

```
tweet_df = pd.DataFrame(data=tweet_details,columns=['geo','text','user','location','time'])
pd.set_option('max_colwidth',800)
print(tweet_df)
```

I na kraju, prikupljeni podaci su izvezeni u Excel tablicu:

```
tweet_df.to_excel (r'/home/alesia/diplomski/alesia_COVID19vaccine_new.xlsx', index=False, header=True)
```

4.2 Implementacija postupka analize sentimenta

Najprije su učitane potrebne biblioteke, od kojih se kao najvažnije ističu: „pandas“ za početnu obradu podataka, „textblob“ za analizu sentimenta i dobivanje vrijednosti polariteta po tweetovima i „matplotlib“ za krajnju vizualizaciju rezultata te prikaz grafikona.

Nakon toga, učitan je i skup podataka nad kojim će se provoditi analiza sentimenta:

```
df = pd.read_excel('df_2021.xlsx')
```

Sljedeći, važan korak jest čišćenje podataka. Najprije je trebalo izbrisati duple retke. Nakon toga su slova promijenjenjena u mala početna slova te su izbrisani dijelovi teksta poput hiperlinkova, spominjanja, oznaka, novih redaka. Osim ovih uobičajenih nepotrebnih dijelova, uklonjeni su i izrazi poput „b'“ i „amp“ koji su se pojavljivali u korištenim skupovima podataka. Također, uklonjene su zaustavne riječi [1](eng. *stop words*):

```
df = df.drop_duplicates()
```

```
def cleaning(text):

    text = text.lower()
    text = re.sub(r"b\'", '', text)
    text = re.sub(r"b\"", '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    #text = re.sub('https', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('rt', '', text)
    text = re.sub('amp', '', text)
    text = re.sub('[’“”…]', '', text)
    text = re.sub(r"@[A-Za-z0-9]+", "", text)
    text = re.sub(r"#[A-Za-z0-9]+", "", text)
    text = re.sub(r"#", "", text)
    text = re.sub(r"RT[\s]+", "", text)
    text = re.sub(r"https?:\/\/\S+", "", text)
    text = re.sub('\w*\d\w*', '', text)

    text = EMOJI_PATTERN.sub(r'', text)

    text_tokens = word_tokenize(text)
    tokens_without_sw = [
        word for word in text_tokens if not word in STOP_WORDS]
    filtered_sentence = (" ").join(tokens_without_sw)
    text = filtered_sentence

    return text
```

[1] Zaustavne riječi (eng. stop words) su najčešće korištene riječi u nekom jeziku koje ne daju previše informacija o tekstu. To su riječi poput veznika, zamjenica i prijedloga. Primjeri za engleski jezik: „a“, „an“, „the“.

U nastavku je prikazan dio koda pomoću kojeg je izvučeno 10 najčešće korištenih riječi u pojedinom skupu podataka. Frekventnost riječi kasnije je prikazana kroz oblak riječi kako bi se vidjele najčešće teme objava na Twitteru.

```
wd = df["text"]
word_count = Counter(" ".join(wd).split())
word_count = Counter(" ".join(wd).split()).most_common(10)
word_frequency = pd.DataFrame(word_count, columns = ['Word', 'Frequency'])
print(word_frequency)
```

Zatim je definirana funkcija za izračunavanje sentimenta u tweetovima. Dobivene vrijednosti polariteta spremljene su u početni skup podataka.

Također, definirana je i funkcija za dodijeljivanje kategorije sentimentu. Pa je tako za vrijednost manju od nula, sentiment označen kao „Negative“, za vrijednost jednaku nuli kao „Neutral“, te za vrijednost polariteta veću od nule dodijeljena je vrijednost „Positive“.

U nastavku slijede prethodno opisane„get_sentiment“ i „analysis“ funkcije:

```
def get_sentiment(df, text):
    return df[text].map(lambda txt: TextBlob(txt).sentiment.polarity)

df['textblob_sentiment'] = get_sentiment(df, 'text')
```

```
def analysis(score):
    if score < 0:
        return "Negative"
    elif score == 0:
        return "Neutral"
    else:
        return "Positive"

df["analysis"] = df["textblob_sentiment"].apply(analysis)
```

Zatim slijedi dio koda pomoću kojeg su izračunati brojevi riječi odnosno znakova u pojedinom objavljenom tweetu:

```
df['words'] = [len(x.split()) for x in df['text'].tolist()]

df['characters_nb'] = df.text.apply(len)
```

Nakon toga izračunati su i ukupni brojevi riječi i znakova koji se nalaze u skupu podataka. Također, izračunate su i prosječne vrijednosti broja riječi i znakova te sentimenta u pojedinom tweetu:

```
total_wd = df['words'].sum()
print("Total no of words", total_wd)

avg_wd = df['words'].mean()
print("Average no of words per tweet", avg_wd)

total_chr = df['characters_nb'].sum()
print("Total no of characters", total_chr)

avg_chr = df['characters_nb'].mean()
print("Average no of characters per tweet", avg_chr)

avg_sent = df['textblob_sentiment'].mean()
print("Average sentiment total", avg_sent)
```

Sve izračunate vrijednosti su na kraju prikazane kroz grafikone poput linijskog, stupčanog i kružnog na slikama 8, 9, 10, 11, 12, 13, 15, 16, 18 i 19. Za prikaz rezultata korištena je „matplotlib“ biblioteka.

5. Rezultati postupka analize sentimenta

U ovom poglavlju prikazani su rezultati postupka analize sentimenta. Rezultati su prikazani zasebno za svaki od skupova podataka.

5.1 Rezultati za prvi skup podataka

U nastavku (Tablica 4) je prikazano deset najčešće korištenih riječi i ukupan broj njihovih pojavljivanja u prvom skupu podataka.

Tablica 4 - 10 najčešće korištenih riječi (prvi skup podataka)

Svibanj 2020. godine		**Svibanj 2021. godine**	
Riječ	**Broj pojavljivanja**	**Riječ**	**Broj pojavljivanja**
cases	2999	covaxin	2584
india	1844	vaccine	2553
indiafightscorona	1476	moderna	1918
lockdown	1081	dose	1494
total	1002	lockdown	1083
active	883	got	805
people	828	vaccinated	758
surrounding	785	sputnikv	726
new	778	vaccines	718
may	701	first	717

Vidljivo je kako je 2020. godine veći fokus bio općenito na koronavirusu, nove slučajeve zaraze te zaključavanje (eng. *lockdown*), dok je 2021. godine fokus na terminima vezanim za cijepljenje, proizvođačima cjepiva te koronavirusu i zaključavanju.

Slijedi prikaz deset najčešće korištenih riječi prema sentimentu objava u kojima su te riječi korištene.

Iz Tablice 5 vidi se da se kod pozitivnog sentimenta često koriste riječi koje izražavaju sreću i nadu – „mind“, „beautiful“ i „avdheshanandg“ (Swami Avdheshanand Giri je indijski hinduistički duhovni vođa). Dok je kod negativnog sentimenta česta uporaba riječi vezanih za rast broja zaraženih, odnosno slučajeva, također se često spominje indijska profesorica Shamika Ravi.

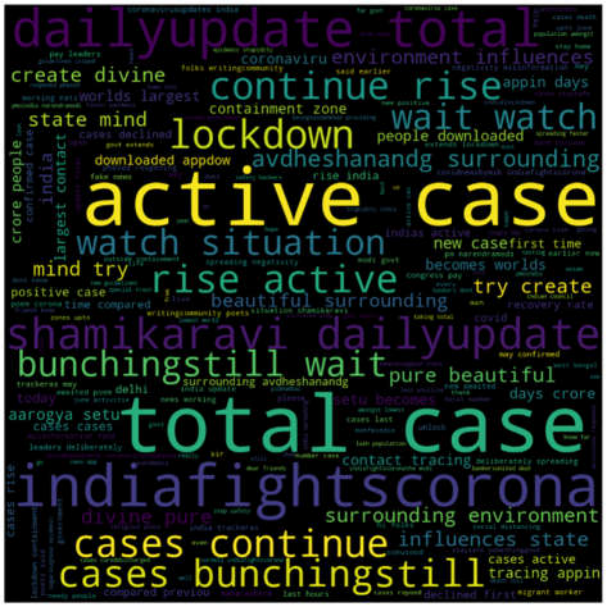
Tablica 5 - 10 najčešće korištenih riječi ovisno o sentimentu (svibanj 2020. godine)

Neutralni sentiment - riječi	Broj pojavljivanja	Pozitivni sentiment - riječi	Broj pojavljivanja	Negativni sentiment – riječi	Broj pojavljivanja
indiafightscorona	996	cases	1095	Cases	1069
india	874	surrounding	781	Active	573
cases	835	new	747	Total	463
lockdown	694	india	690	situation	421
people	508	state	479	watch	409
days	321	mind	429	Rise	409
coronavirus	316	beautiful	412	dailyupdate	408
may	284	environment	410	continue	408
corona	274	try	407	shamikaravi	406
setu	267	avdheshanandg	403	Wait	404

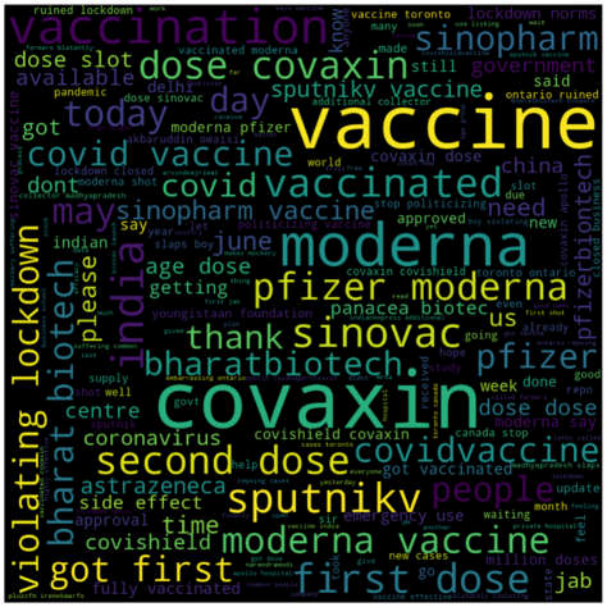
Tablica 6 - 10 najčešće korištenih riječi ovisno o sentimentu (svibanj 2021. godine)

Neutralni sentiment - riječi	Broj pojavljivanja	Pozitivni sentiment - riječi	Broj pojavljivanja	Negativni sentiment - riječi	Broj pojavljivanja
covaxin	1573	vaccine	1054	vaccine	288
vaccine	1214	moderna	828	covaxin	266
moderna	921	covaxin	745	lockdown	216
dose	799	first	695	moderna	169
lockdown	537	dose	580	dose	115
vaccinated	456	got	438	people	112
sputnikv	410	available	356	due	97
doses	359	lockdown	330	closed	93
vaccines	356	vaccines	299	stop	83
pfizer	355	effective	268	toronto	81

Iz Tablice 6 vidljivo je da se u pozitivnom kontekstu često koriste pozitivne riječi vezane uz cjepiva, na primjer „got“, „available“ i „effective“. S drugre strane, u negativnom kontekstu često su upotrebljavane riječi vezane za zaključavanje – „lockdown“, „closed“ i „stop“.



Slika 6 - Oblak riječi (svibanj 2020. godine)

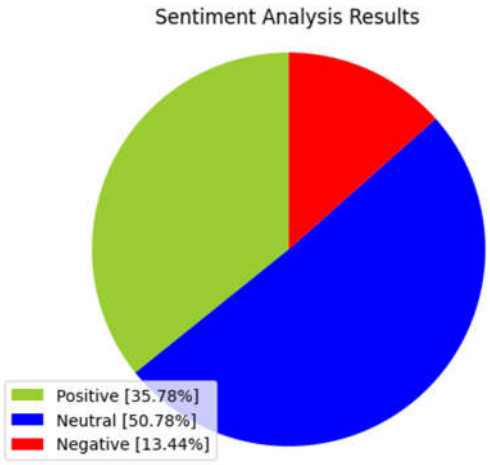


Slika 7 - Oblak riječi (svibanj 2021. godine)

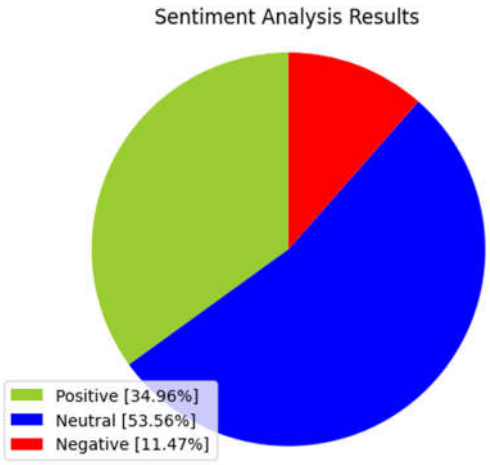
Prema oblaku riječi iz 2020. godine (Slika 6) korisnici Twittera su najviše koristili riječi vezane za slučajeve zaraze (aktivne i ukupne) i zaključavanju. Također ističu se i neke pozitivne riječi poput „divine“ i „pure beautiful“ te riječi vezane za aplikaciju za praćenje novih slučajeva zaraze.

U svibnju 2021. godine (Slika 7) korisnici su najviše koristili riječi vezane za cijepljenje, pa se tako ističu riječi poput „vaccine“ i „covaxin“, zatim proizvođači cjepiva „moderna“, „pfizer“, „sinovac“ i „sputnikv“. Ističu se i riječi vezane za dozu cjepiva poput „dose“, „second dose“ te „got first“.

Kružni grafikoni u nastavku (Slika 8 i 9) prikazuju udio pozitivnih, negativnih i neutralnih tweetova. U oba skupa podataka je najveći udio neutralnih tweetova (50,78% i 53,6%), zatim pozitivnih oko 35% te otprilike 12% negativnih tweetova.



Slika 8 - Udio pojedinog sentimenta (svibanj 2020. godine)



Slika 9 - Udio pojedinog sentimenta (svibanj 2021. godine)

Tablice 7 i 8 prikazuju ukupni broj tweetova prema pojedinom sentimentu, te prosječne brojeve riječi i znakova. Najviše ima neutralnih tweetova u oba skupa podataka. Dok su tweetovi pozitivnog sentimenta oni s najmanjim prosječnim brojem riječi, odnosno znakova.

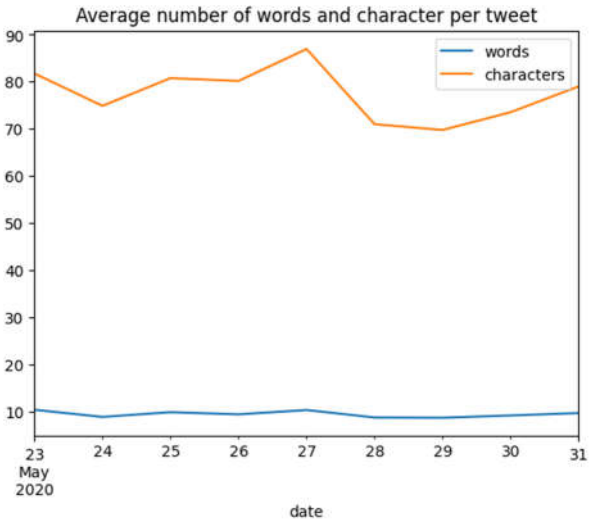
Tablica 7 - Količina tweetova i njihova dužina prema sentimentu (svibanj 2020. godine)

Sentiment	Ukupni broj tweetova	Prosječni broj riječi	Prosječni broj znakova
Negativni	1 958	10.31	80.09
Pozitivni	5 211	8.80	75.42
Neutralni	7 397	10.20	80.13

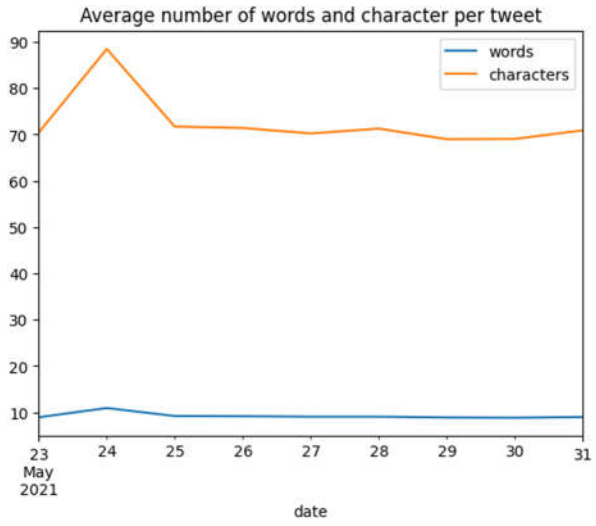
Tablica 8 - Količina tweetova i njihova dužina prema sentimentu (svibanj 2021. godine)

Sentiment	Ukupni broj tweetova	Prosječni broj riječi	Prosječni broj znakova
Negativni	1 274	10.89	82.46
Pozitivni	3 882	8.75	71.58
Neutralni	5 947	10.20	77.77

Na grafovima u nastavku (Slika 10 i 11) prikazano je kretanje prosječnog broja riječi odnosno znakova kroz vrijeme. Broj riječi je u oba skupa podataka oko 10, dok broj znakova varira između 70 i 90 po tweetu.

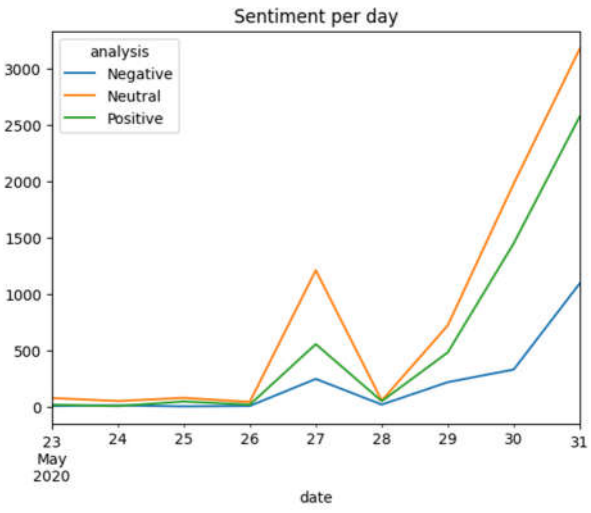


Slika 10 - Kretanje prosječnog broja riječi i znakova (svibanj 2020. godine)

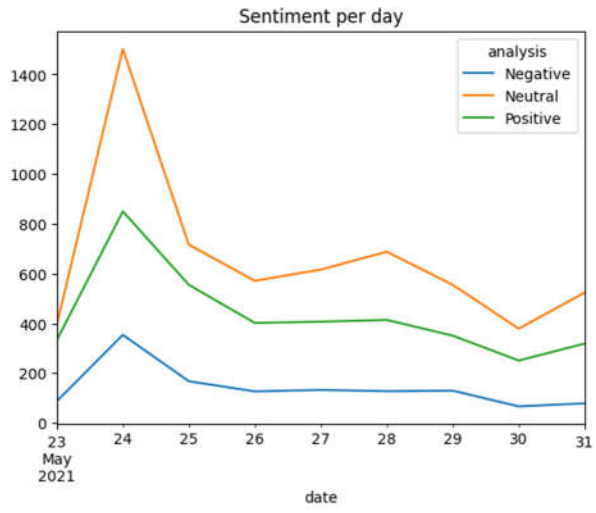


Slika 11 - Kretanje prosječnog broja riječi i znakova (svibanj 2021. godine)

Slike 12 i 13 prikazuju odnos sentimenta i količine tweetova kroz vrijeme. Omjer pojedinih sentimenta se niti u jednom skupu podataka nije previše mijenjao. Dok s dnevnom količinom tweetova to nije slučaj. Naime, u prvom podskupu se vidi rast broja tweetova s dva veća skoka u količini, dok se u drugom podskupu vidi pad broja tweetova.



Slika 12 - Sentiment po danima (svibanj 2020. godine)



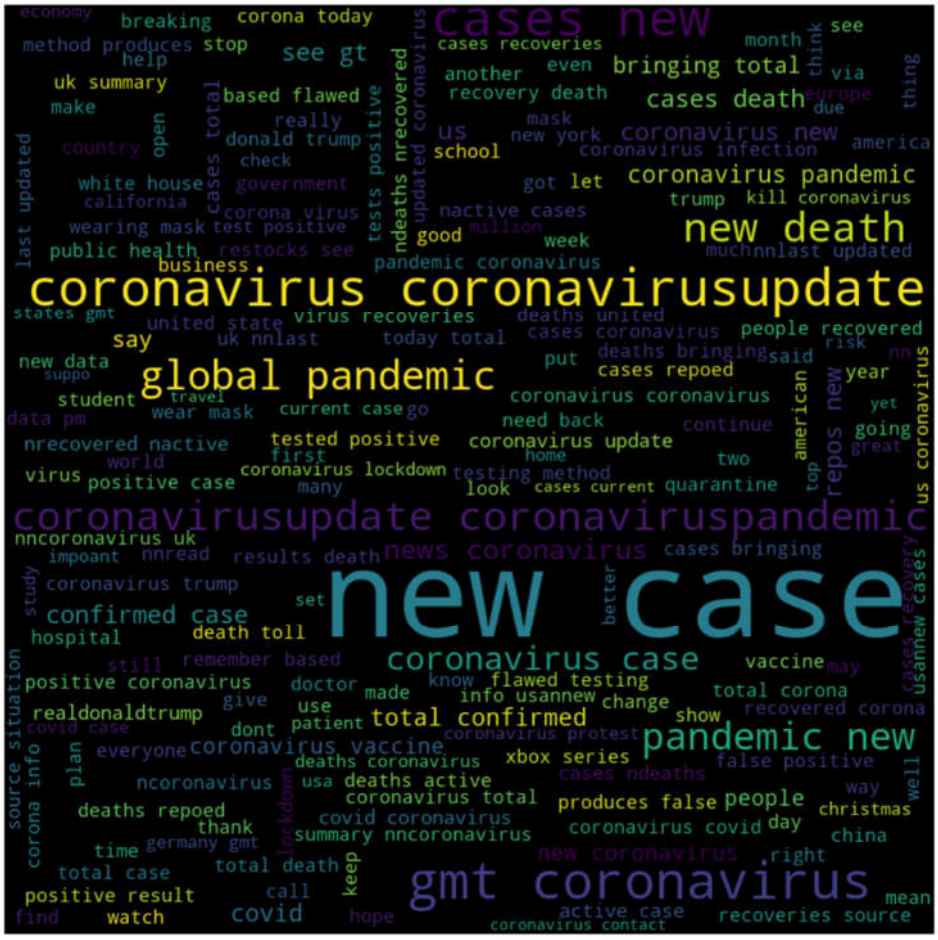
Slika 13 - Sentiment po danima (svibanj 2021. godine)

5.2 Drugi skup podataka

Tablica 9 prikazuje deset najčešće korištenih riječi i ukupan broj njihovih pojavljivanja za drugi skup podataka.
U razdoblju od srpnja 2020. do siječnja 2021. godine, najčešće korištena riječ je „coronavirus“, koja se spominje 115 132 puta, duplo češće od sljedeće dvije riječi, „cases“ i „new“.

Tablica 9 - 10 najčešće korištenih riječi (drugi skup podataka)

Srpanj 2020. – siječanj 2021. godine	
Riječ	**Broj pojavljivanja riječi**
coronavirus	115 132
cases	53 474
new	53 242
deaths	25 491
pandemic	23 124
news	17 383
global	14 094
covid	13 864
coronaviruspandemic	12 186
coronavirusupdate	11 835



Slika 14 - Oblak riječi (drugi skup podataka)

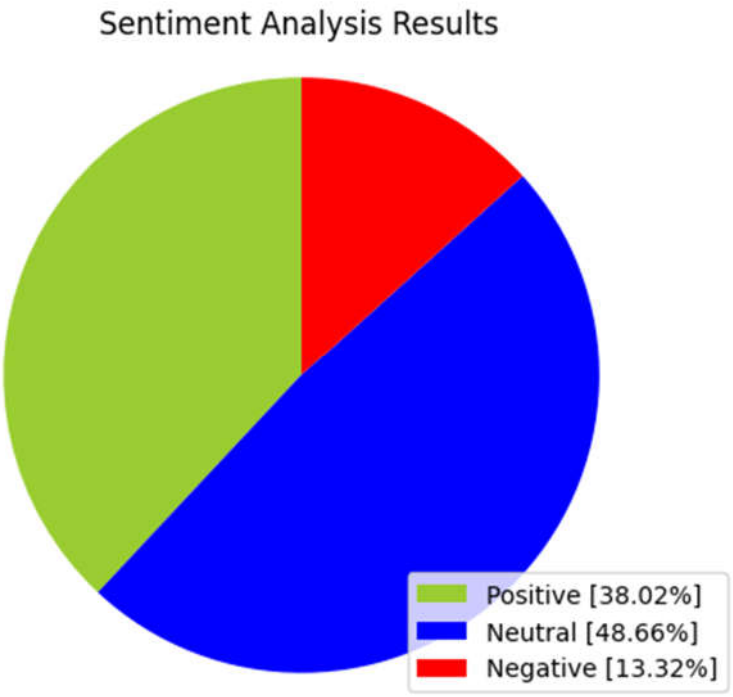
Iz oblaka riječi (Slika 14) vidljivo je da se najčešće koriste izrazi vezani za nove slučajeve zaraze, koronavirus općenito, globalnu pandemiju te nove slučajeve smrti.

Tablica 2 - Količina tweetova i njihova dužina prema sentimentu (drugi skup podataka)

Sentiment	Ukupni broj tweetova	Prosječni broj riječi	Prosječni broj znakova
Negativni	36 699	8.90	66.67
Pozitivni	104 768	7.83	62.56
Neutralni	134 069	8.83	67.11

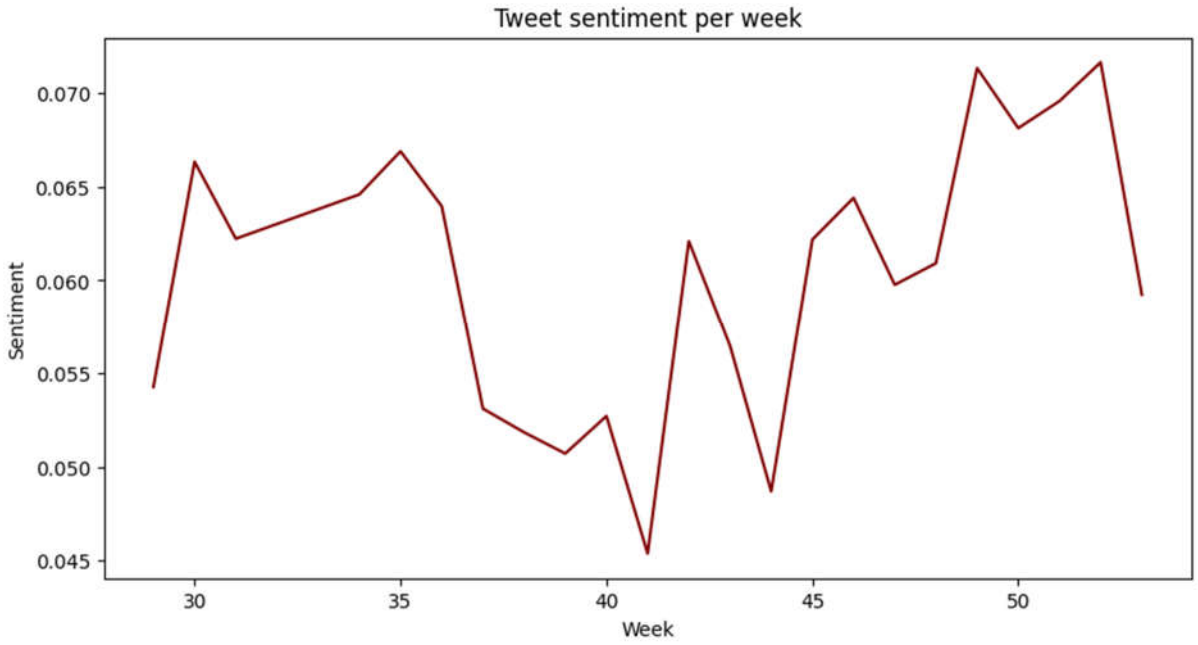
Prethodna Tablica 10 prikazuje ukupni broj tweetova prema pojedinom sentimentu, te prosječne brojeve riječi i znakova. Najviše ima neutralnih tweetova, a tweetovi pozitivnog sentimenta su najkraći.

Kružni grafikon (Slika 15) prikazuje raspodjelu pozitivnih, negativnih i neutralnih tweetova. Najveći je udio neutralnih tweetova 48,66%, zatim pozitivnih 38,02% te 13,32% negativnih tweetova.



Slika 15 - Udio pojedinog sentimenta (drugi skup podataka)

Graf u nastavku (Slika 16) prikazuje kretanje sentimenta po tjednima. Nakon 35. tjedna vrijednost sentimenta pada, zatim između 40. i 45. tjedna se događa nagli skok nakon kojeg vrijednost sentimenta opet počinje rasti do otprilike 50. tjedna kada opet počinje padati.



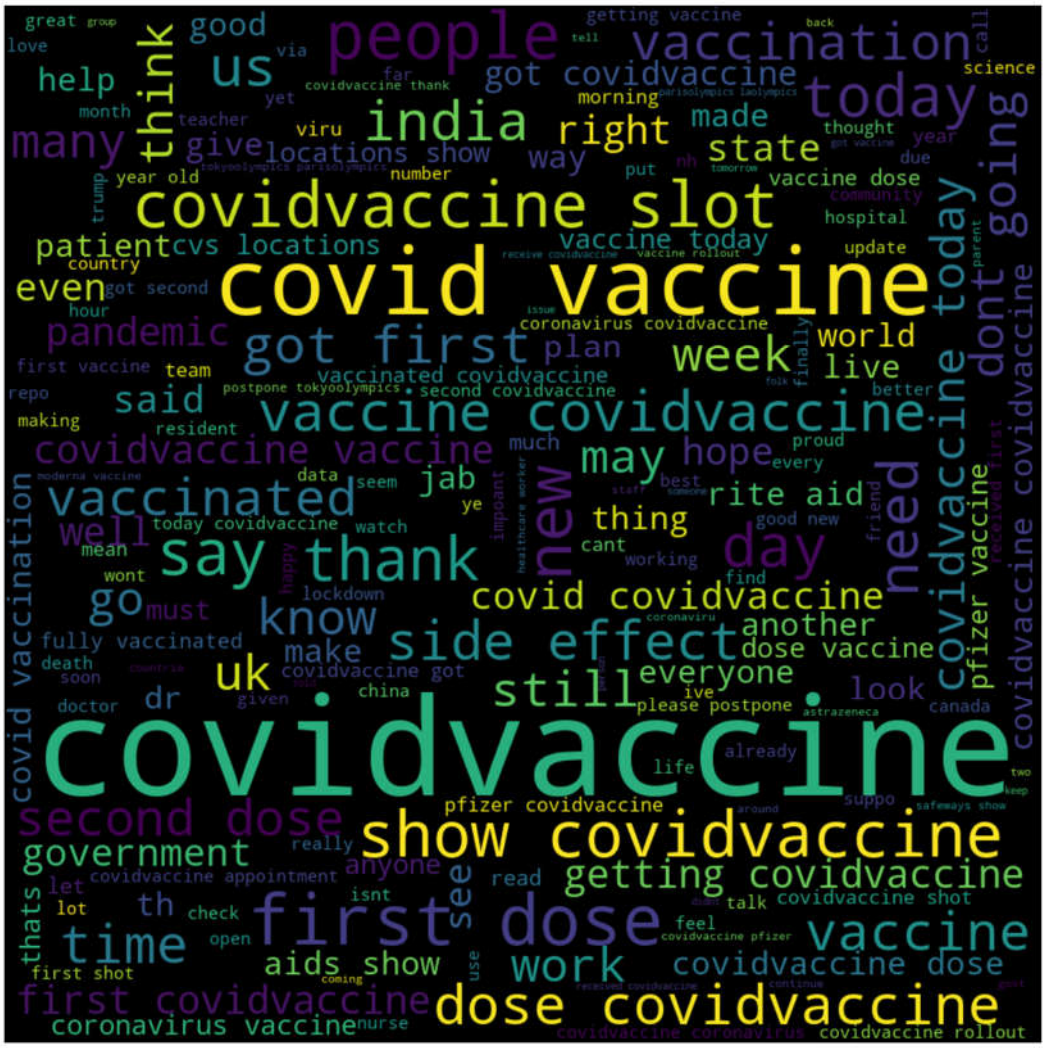
Slika 16 - Kretanje sentimenta po tjednima (drugi skup podataka – srpanj 2020. – siječanj 2021. godine)

5.3 Treći skup podataka

U Tablici 11 prikazano je deset najčešće korištenih riječi za razdoblje od kolovoza 2020. do svibnja 2021. godine. Najčešća riječ je „covidvaccine“, koja je spomenuta 116 774 puta. Ostale riječi su isto tako vezane za cijepljenje protiv koronavirusa.

Tablica 3 - 10 najčešće korištenih riječi (treći skup podataka)

Kolovoz 2020. godine – Svibanj 2021. godine	
Riječ	**Broj pojavljivanja riječi**
covidvaccine	116 774
vaccine	49 009
covid	19 945
get	19 598
first	14 249
people	12 975
fot	11 491
dose	11 227
today	11 144
vaccination	10 831



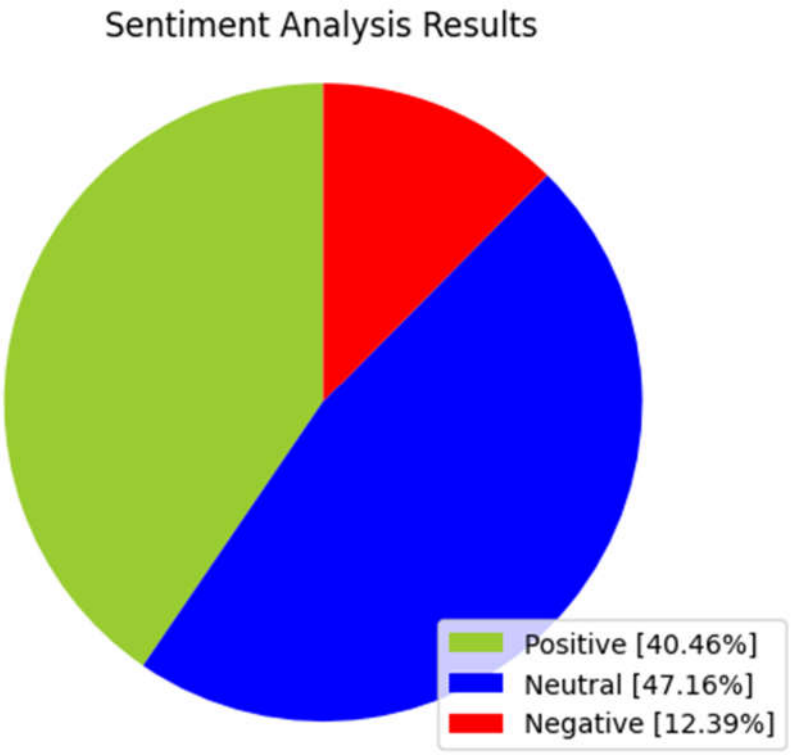
Slika 17 - Oblak riječi (treći skup podataka)

Oblak riječi (Slika 17) za razdoblje od kolovoza 2020. do svibnja 2021. godine prikazuje da su korisnici najčešće koristili riječi vezane za cijepljenje, najviše ističu riječi poput „covid vaccine“ i „vaccinated“. Ističu se i neke zemlje (Ujedinjeno Kraljevstvo, SAD i Indija). Isto tako, često su se koristile i riječi poput „side effect“, „think“, „aids“ te „please postpone“.

Tablica 12 - Količina tweetova i njihova dužina prema sentimentu (treći skup podataka)

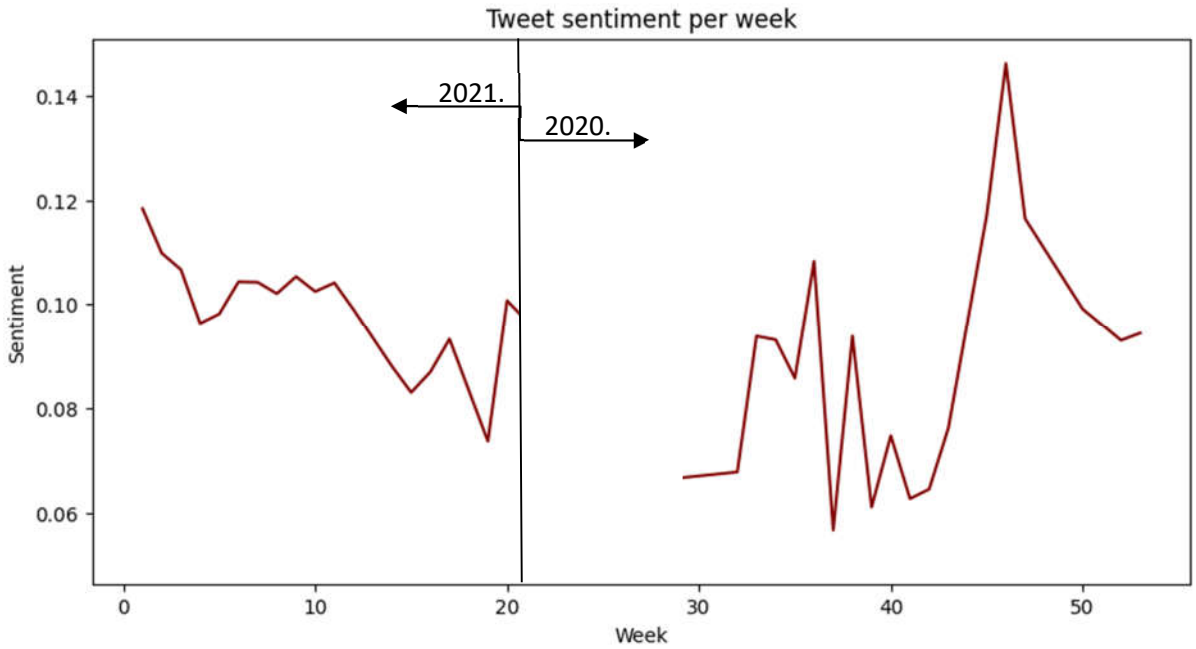
Sentiment	Ukupni broj tweetova	Prosječni broj riječi	Prosječni broj znakova
Negativni	27 109	10.40	79.33
Pozitivni	88 551	8.61	71.62
Neutralni	103 221	10.16	77.82

Prethodna Tablica 12 prikazuje ukupni broj tweetova prema pojedinom sentimentu, te prosječne brojeve riječi i znakova. Kao i u prva dva skupa podataka, ovaj također ima najviše neutralnih tweetova te su tweetovi pozitivnog sentimenta najkraći.



Slika 18 - Raspodjela pojedinog sentimenta (treći skup podataka)

Iz kružnog grafikona (Slika 18) vidljivo je da najveći udio imaju neutralni tweetovi (47,46%), zatim slijede pozitivni s 40,46% te negativni s 12,39%.



Slika 19 - Kretanje sentimenta po tjednima (treći skup podataka – kolovoz 2020. – svibanj 2021. godine)

Prethodni graf (Slika 19) prikazuje kretanje sentimenta po tjednima za treći skup podataka. Između 35. i 40. tjedna 2020. godine (otprilike između 24. kolovoza i 28. rujna) bilo je nekoliko skokova vrijednosti sentimenta, a zatim između 40. i 50. tjedna (otprilike između 28. rujna i 7. prosinca) vrijednost sentimenta naglo raste do otprilike 50. tjedna kada opet počinje padati. Pad vrijednosti sentimenta se nastavlja i u 2021. godini do 18. tjedna kad se ta vrijednost počinje ponovno uspinjati.

6. Zaključak

Nakon početka pandemije koronavirusa krajem 2019. godine, to je jako česta tema razgovora, što je i očekivano budući da bolest, kao i mjere za sprečavanje širenja bolesti, utječu na cjelokupnu populaciju.

Kako bi razne organizacije na vlasti imale uvid u osjećaje u društvu vezane za događaje u svijetu, te kako bi mogle ispravnije reagirati, provodi se analiza sentimenta. Jako dobar izvor podataka su društvene mreže poput Twittera. Twitter koristi mnogo ljudi diljem svijeta, tako da se s velikim skupom podataka može dobiti dosta dobar uvid u opće mišljenje u društvu.

Kod usporedbe posljednjih 9 dana u svibnju za 2020. i 2021. godinu, primjećeno je da se sentiment nije baš mijenjao, iako su teme u objavama nešto drugačije. U 2020. godini se najviše pričalo o slučajevima zaraze i zaključavanju – u to doba je zapravo završavao prvi val i počelo se popuštanjem mjera vezanih sa zaključavanjem. Dok je u 2021. godini fokus na cijepljenju i cjepivima – kada vlade pokušavaju ubrzati proces cijepljenja prije ljeta kako bi ljudi mogli slobodnije putovati.

U razdoblju od srpnja 2020. do siječnja 2021. godine, najviše se spominju općeniti izrazi vezani za pandemiju. Otprilike polovica objava ima neutralni sentiment, ali je čak četrdesetak posto objava pozitivnog sentimenta. Sentiment je u tom razdoblju većinom bio u laganom porastu, osim u periodu između 35. i 45. tjedna kada je sentiment padao uz nekoliko skokova. To je zapravo razdoblje kada je započeo drugi val, pa je bilo i očekivano da bi tada sentiment mogao biti nešto niži.

U razdoblju od kolovoza 2020. do svibnja 2021. godine, udio pojedinog sentimenta u objavama je otprilike isti kao i u prethodnom skupu podataka, iako je veći fokus objava bio na cijepljenju, cjepivima te njihovim proizvođačima. Sveukupni prosječni sentiment svejedno nije previsok – mjeri se u rasponu od 0,00 do 0,15 (na skali od -1,00 do 1,00). Svejedno, primjećuje se da je sentiment u blagom porastu sve do 50. tjedna 2020. godine, a nakon toga počinje naglo padati. Taj pad se, uz nekoliko skokova u vrijednostima, nastavlja do 20. tjedna 2021. godine. Razdoblje kada je sentiment počeo opadati je zapravo vrijeme otprilike dva tjedna prije Nove godine, a u to doba su vlade započele s pooštravanjem mjera prije božićnih i novogodišnjih praznika kako bi se smanjila putovanja i cirkulacija ljudi. Nakon toga započelo je i masovnije cijepljenje stanovnika, samim time se počelo uviđati više nuspojava, te su vlasti krenule s upozorenjima o uskraćivanju nekih mogućnosti za necijepljeni dio stanovništva. Svi ovi događaji su doprinijeli opadanju općeg sentimenta.

Analiza sentimenta važna je za planiranje budućih koraka, kao i za predviđanje reakcija u društvu. Kada bi popularnost Twittera kao društvene mreže bila rasprostranjenija u više dijelova svijeta (npr. prema Iqbalu, u ožujku 2021. godine, Facebook je imao 2,80 milijardi korisnika, dok je Twitter imao njih 0,33 [13]), krajnji rezultati analize sentimenta bili bi precizniji.

Popis literature

[1] Nepoznati autor. „COVID-19 CORONAVIRUS PANDEMIC“. Worldometer. Preuzeto: 1. srpnja 2021. https://www.worldometers.info/coronavirus/

[2] Boon-Itt, Sakun, and Skunkan, Yukolpat. „Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study“. JMIR Public Health Surveillanc. E 21987, no. 6 (2020)

[3] Chandrasekaran, Ranganathan et al. „Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study“. Journal of Medical Internet Research. E 22624, no. 22 (2020)

[4] Lwin, May Oo et al. „Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends“. JMIR Public Health Surveillanc. E 19447, no. 6 (2020)

[5] Xue, Jia et al. „Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach“. Journal of Medical Internet Research. E 20550, no. 22 (2020)

[6] Liu, Bing. „Sentiment Analysis and Opinon Mining“, San Francisco. Morgan & Claypool Publishers, 2012.

[7] Jain, Kamal. „Sentiment Analysis using Deep Learning“. Medium. Preuzeto: 5. srpnja 2021. medium.com/analytics-vidhya/sentiment-analysis-using-deep-learning-a416b230ca9a

[8] Thakkar, Harsh, and Patel, Dhiren. „Approaches for Sentiment Analysis on Twitter: A State-of-Art study“, Surat, India. Department of Computer Engineering, National Institute of Technology, 2015.

[9] Sarkar, Dipanjan. „Emotion and Sentiment Analysis: A Practitioner’s Guide to NLP“. KDnuggets. Preuzeto: 5. srpnja 2021. kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html

[10] Band, Amey. „COVID-19 Tweets“. Kaggle. Preuzeto: 1. lipnja 2021. https://www.kaggle.com/ameyband/covid19-tweets

[11] Avasthi, Sandhya. „COVID19 tweets(July2020-December2020)„. Kaggle. Preuzeto: 1. lipnja 2021. kaggle.com/sandhyaavasthi/covid19-tweetsjuly2020december2020

[12] Kash. „Covid Vaccine Tweets“. Kaggle. Preuzeto: 1. lipnja 2021. https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets

[13] Iqbal, Mansoor. „Twitter Revenue and Usage Statistics (2021)“. Business of Apps. Preuzeto: 2. srpnja 2021. https://www.businessofapps.com/data/twitter-statistics/

Popis slika

Popis tablica

Prilog A: Kôd za prikupljanje podataka

```
# Uključivanje biblioteka

import os
import tweepy as tw
import pandas as pd
import json

# Twitter API autentifikacija

auth = tw.OAuthHandler("___", "___")
auth.set_access_token("___", "___")
auth.secure = True
api = tw.API(auth, wait_on_rate_limit = True)

# Postavljanje ključnih riječi i početnog datuma

search_words =
['#covid19']#,'#covid19','#coronavirus','#lockdown','#pandemic','#pl
andemic','#COVID19vaccine','#covidvaccine']
date_since = '2021-01-01'

# Prikupljanje objava s Twittera

for search_word in search_words:
    tweets = tw.Cursor(api.search, q= search_word, lang='en',
since=date_since, tweet_mode="extended",retry_count = 5, retry_delay
= 5).items(10000)
    tweet_details =
[[tweet.geo,tweet.full_text,tweet.user.screen_name,tweet.user.locati
on,tweet.created_at]for tweet in tweets]

# Spremanje podataka u pandas dataframe i Excel tablicu

tweet_df =
pd.DataFrame(data=tweet_details,columns=['geo','text','user','locati
on','time'])
pd.set_option('max_colwidth',800)
print(tweet_df)

tweet_df.to_excel
(r'/home/alesia/diplomski/alesia_COVID19vaccine_new.xlsx',
index=False, header=True)
```

Prilog B: Kôd za proces analize sentimenta

```
# Instalacija biblioteka

!pip install textblob
!pip install pycountry
!pip install langdetect

# Uključivanje biblioteka

from textblob import TextBlob
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os
import nltk
import pycountry
import re
import string
from wordcloud import WordCloud, STOPWORDS
from PIL import Image

import collections
from collections import Counter

from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

nltk.download('punkt')
nltk.download('stopwords')

# Učitavanje skupa podataka

df = pd.read_excel('df_srppro.xlsx')#, sheet_name="may_2020")

STOP_WORDS = stopwords.words()

# Postavljanje uzoraka za simbole

EMOJI_PATTERN = re.compile("["
                           u"\U0001F600-\U0001F64F"
                           u"\U0001F300-\U0001F5FF"
                           u"\U0001F680-\U0001F6FF"
                           u"\U0001F1E0-\U0001F1FF"
                           u"\U00002702-\U000027B0"
                           u"\U000024C2-\U0001F251"
                           "]+", flags=re.UNICODE)
```

```
# Brisanje duplih redaka

df = df.drop_duplicates()

# Definiranje funkcije za čišćenje podataka

def cleaning(text):
    text = text.lower()
    text = re.sub(r"b\'", '', text)
    text = re.sub(r"b\"", '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('rt', '', text)
    text = re.sub('amp', '', text)
    text = re.sub('[’“”…]', '', text)
    text = re.sub(r"@[A-Za-z0-9]+", "", text)
    text = re.sub(r"#[A-Za-z0-9]+", "", text)
    text = re.sub(r"#", "", text)
    text = re.sub(r"RT[\s]+", "", text)
    text = re.sub(r"https?:\/\/\S+", "", text)
    text = re.sub('\w*\d\w*', '', text)

    text = EMOJI_PATTERN.sub(r'', text)

    text_tokens = word_tokenize(text)
    tokens_without_sw = [
        word for word in text_tokens if not word in STOP_WORDS]
    filtered_sentence = (" ").join(tokens_without_sw)
    text = filtered_sentence

    return text

df['text'] = df['text'].str.encode('ascii',
'ignore').str.decode('ascii')

df['date'] = pd.to_datetime(df['date'])
df = df[df['date'].notna()]

# Primjena funkcije "cleaning"
# Ispis 10 najčešće korištenih riječi

if __name__ == "__main__":
    max_rows = 1000
    df["text"] = df["text"].astype(str).apply(cleaning)
    wd = df["text"]
```

```
    word_count = Counter(" ".join(wd).split())
    word_count = Counter(" ".join(wd).split()).most_common(10)
    word_frequency = pd.DataFrame(word_count, columns = ['Word',
'Frequency'])
    print(word_frequency)

# Definicija funkcije za izračunavanje vrijednosti sentimenta

def get_sentiment(df, text):
    return df[text].map(lambda txt:
TextBlob(txt).sentiment.polarity)

df['textblob_sentiment'] = get_sentiment(df, 'text')

# Definicija funkcije za dodijeljivanje tekstualne oznake sentimenta
(poz, neg, neu)

def analysis(score):
    if score < 0:
        return "Negative"
    elif score == 0:
        return "Neutral"
    else:
        return "Positive"

df["analysis"] = df["textblob_sentiment"].apply(analysis)

# Izračun ukupnog broja pojedinog sentimenta te njegovog udjela

series = pd.Series(df.analysis)
negs = series.str.count("Negative").sum()
poss = series.str.count("Positive").sum()
neus = series.str.count("Neutral").sum()
print("no of negs: ", negs)
print("no of poss: ", poss)
print("no of neus: ", neus)


tot_analysis = df['analysis'].count()
print("Total no of analysis", tot_analysis)

neg_percent = (negs/tot_analysis)*100
pos_percent = (poss/tot_analysis)*100
neu_percent = (neus/tot_analysis)*100

print("percent of negs: ", neg_percent, "%")
print("percent of poss: ", pos_percent, "%")
print("percent of neus: ",  neu_percent,"%")
```

```
# Izračun broja riječi i znakova u svakom tweetu

df['words'] = [len(x.split()) for x in df['text'].tolist()]
df['characters_nb'] = df.text.apply(len)

# Izračun ukupnog/prosječnog broja riječi/znakova
# Grupiranje vrijednosti prema datumu / mjesecu

total_wd = df['words'].sum()
print("Total no of words", total_wd)

avg_wd = df['words'].mean()
print("Average no of words per tweet", avg_wd)

total_chr = df['characters_nb'].sum()
print("Total no of characters", total_chr)

avg_chr = df['characters_nb'].mean()
print("Average no of characters per tweet", avg_chr)

avg_sent = df['textblob_sentiment'].mean()
print("Average sentiment total", avg_sent)

df['no1']=1

sent_day = df.groupby(df.date).mean()
sd = df.groupby(df.date).sum()
print(sd)

df['month'] = pd.DatetimeIndex(df['date']).month
import calendar
df['month'] = df['month'].apply(lambda x: calendar.month_abbr[x])
month_nums = df.groupby(df.month).sum()

print("\n")
print("Values per day", sent_day)

# Kružni grafikon, udio sentimenta

labels = ['Positive ['+str(round(pos_percent,2),)+'%]', 'Neutral
['+str(round(neu_percent,2))+'%]','Negative
['+str(round(neg_percent,2))+'%]']
sizes = [pos_percent, neu_percent, neg_percent]
colors = ['yellowgreen', 'blue','red']
patches, texts = plt.pie(sizes,colors=colors, startangle=90)
plt.style.use('default')
plt.legend(labels)
```

```
plt.title("Sentiment Analysis Results")
plt.axis('equal')
plt.show()

# Oblak riječi

allwords = " ".join([twts for twts in df["text"]])
wordCloud = WordCloud(width = 900, height = 900, random_state = 21,
max_font_size = 119).generate(allwords)
plt.figure(figsize=(20, 20), dpi=80)
plt.imshow(wordCloud, interpolation = "bilinear")
plt.axis("off")
plt.show()

# Prosječne vrijednosti grupirane prema sentimentu

df['text_len'] = df['text'].astype(str).apply(len)
df['text_word_count'] = df['text'].apply(lambda x:
len(str(x).split()))

text_len =
round(pd.DataFrame(df.groupby("analysis").text_len.mean()),2)

print(text_len)

tl = text_len.plot()
tl.set_title("Average number of characters grouped by sentiment")
tl

no_word =
round(pd.DataFrame(df.groupby("analysis").text_word_count.mean()),2)

nw = no_word.plot()
nw.set_title("Average number of words grouped by sentiment")
nw

# Prikaz kretanja sentimenta po danima

by_day_sentiment = df.groupby([pd.Grouper(key='date', freq='D'),
'analysis']).size().unstack('analysis')

ds=by_day_sentiment.plot()
ds.set_title("Sentiment per day")

wp = sent_day.words.plot()
sent_day.characters_nb.plot()

labels = ['words','characters']
```

```
wp.legend(labels)
wp.set_title("Average number of words and character per tweet")

wp = sent_day.words.plot()

wp.set_title("Average number of words per tweet")
wp

wp = sent_day.characters_nb.plot()

wp.set_title("Average number of characters per tweet")
wp

# Kreiranje tablice s jedinstvenim datumima

dates_unique = df.date.unique()
dates_unique

sd['dates'] = dates_unique

# Kreiranje tablice s jedinstvenim nazivima mjeseci (za 2. i 3.
dataset)

#srp-pro & kol datasets

months_unique = df.month.unique()
months_unique
month_nums['months'] = months_unique
month_nums

# Prikaz broja objavljenih tweetova po danima

fig, ax = plt.subplots(figsize=(16,9))
ax.barh(sd.dates, sd.no1)

ax.grid(b=True, color='grey', linestyle='-.', linewidth=0.5,
alpha=0.2)

for i in ax.patches:
  plt.text(i.get_width()+0.2, i.get_y()+0.5,
str(round((i.get_width()),2)),
           fontsize=10, fontweight='bold', alpha=0.2)
ax.set_title('Number of tweets per day')

plt.show()

# Prikaz broja objavljenih tweetova po mjesecima
```

```
fig, ax = plt.subplots(figsize=(16,9))
ax.barh(month_nums.months, month_nums.no1)

ax.grid(b=True, color='grey', linestyle='-.', linewidth=0.5,
alpha=0.2)

for i in ax.patches:
  plt.text(i.get_width()+0.2, i.get_y()+0.5,
str(round((i.get_width()),2)),
           fontsize=10, fontweight='bold', alpha=0.2)
ax.set_title('Number of tweets per month')

plt.show()

# Prikaz broja objavljenih tweetova po tjednima

df4 = df.loc[:, ["no1"]]
df4["week_no"] = pd.to_datetime(df["date"]).dt.week

df_weekly = (
    df4
    .groupby("week_no")
    .sum()
    .reset_index()
)

column = df_weekly["no1"]
max_value = column.max()
print(max_value)

column = df_weekly["no1"]
min_value = column.min()
print(min_value)

fig = plt.figure(figsize = (10, 5))

plt.plot(df_weekly.week_no, df_weekly.no1, color ='maroon')
plt.xlabel("Week")
plt.ylabel("Tweets")
plt.title("Number of tweets per week")

plt.show()

df3 = df.loc[:, ["textblob_sentiment"]]
df3["week_no"] = pd.to_datetime(df["date"]).dt.week

df_weekly = (
    df3
```

```
    .groupby("week_no")
    .mean()
    .reset_index()
)

fig = plt.figure(figsize = (10, 5))

plt.plot(df_weekly.week_no, df_weekly.textblob_sentiment, color
='maroon')
plt.xlabel("Week")
plt.ylabel("Sentiment")
plt.title("Tweet sentiment per week")
plt.show()

# Ispis 10 najčešće korištenih riječi prema sentimentu

wd1 = df.loc[df['analysis'] == 'Neutral']
wd1.head()

www = wd1["text"]
word_count1 = Counter(" ".join([str(i) for i in
www]).split()).most_common(10)
word_frequency1 = pd.DataFrame(word_count1, columns = ['Word',
'Frequency'])
print(word_frequency1)

wd2 = df.loc[df['analysis'] == 'Positive']
wd2.head()

www2 = wd2["text"]
word_count2 = Counter(" ".join([str(i) for i in
www2]).split()).most_common(10)
word_frequency2 = pd.DataFrame(word_count2, columns = ['Word',
'Frequency'])
print(word_frequency2)

wd3 = df.loc[df['analysis'] == 'Negative']
wd3.head()

www3 = wd3["text"]
word_count3 = Counter(" ".join([str(i) for i in
www3]).split()).most_common(10)
word_frequency3 = pd.DataFrame(word_count3, columns = ['Word',
'Frequency'])
print(word_frequency3)
```