UNIVERZITET U NIŠU
ELEKTRONSKI FAKULTET

Danica Đorđević

Detekcija tema i prepoznavanje imenovanih entiteta

Web majning

Mentor: doc. dr Miloš Bogdanović
Student: Danica Đorđević, 1121

Niš, 2021. god.

Sadržaj

1. UVOD

Pojedinci i organizacije svakodnevno generišu veliku količinu podataka. Statistika kaže da je gotovo 80% postojećih tekstualnih podataka nestruktuisano, što znači da nisu organizovani na unapred definisan način, da se ne mogu pretraživati i gotovo je nemoguće njima upravljati [1]. Nestruktuisan tekst se javlja prilikom interakcija sa klijentima, razmene elektronske pošte, objavljivanja teksta na društvenim mrežama, itd. Ručna analiza ovako velike količine podataka, u današnje vreme, predstavlja veoma zahtevan, dugotrajan i obiman zadatak. Takođe, ručno sortiranje velikih količina podataka može dovesti do čestih grešaka i nedoslednosti.

Kompanije koje vrše analizu velikih količina podataka, danas koriste sisteme veštačke inteligencije za obavljanje tog posla. Ovakvi sistemi koriste metode tekst majninga (*eng. text mining*), s obzirom da je on presudan za organizaciju, kategorizaciju i detekciju relevantnih informacija iz sirovih podataka. Ti sistemi imaju sposobnost detekcije teme teksta, kao i prepoznavanja imenovanih entiteta, kako bi iz teksta ekstrahovale relevantne informacije. Modeli analize tema omogućavaju pregled velike količine podataka i identifikacije najčešćih i najvažnijih tema na lak, brz i potpuno skalabilan način [2]. Pomoću prepoznavanja imenovanih entiteta mogu se detektovati ključne informacije kako bi se bolje razumelo o čemu se radi u tekstu ili se mogu koristiti za prikupljanje važnih informacija za čuvanje u bazi podataka [3].

Ovaj rad će se baviti metodom detekcije teme teksta, kao i metodom detekcije imenovanih entiteta, u cilju boljeg razumevanja teksta i ekstrakcije korinih informacija iz istog.

2. TEKST MAJNING

Tekst majning, takođe poznat i kao analiza teksta, je proces pretvaranja nestruktuisanog teksta u struktuisane podatke za analizu. Tekst majning koristi tehnike za obradu prirodnog jezika (*eng. Natural Language Processing - NLP*), čime omogućava mašinama da razumeju ljudski jezik i automatski ga obrade.

Za preduzeća, velika količina podataka koja se generiše svakodnevno predstavlja i veliku prednost i izazov. Podaci, sa jedne strane, pomažu kompanijama da steknu pametan uvid u mišljenja ljudi o proizvodu ili usluzi. Sa druge strane, postoji dilema kako obraditi sve ove podatke. U ovoj situaciji tekst majning igra glavnu ulogu. Kompanije koriste tekst majning za automatizaciju mnogih svojih procesa. Transformišući podatke u informacije koje mašine mogu da razumeju, tekst majning automatizuje proces klasifikacije tekstova prema osećanjima, temama i namerama. Zahvaljujući tekst majningu, preduzeća mogu da analiziraju složene i velike skupove podataka na jednostavan, brz i efikasan način. U isto vreme, kompanije koriste ovaj moćan alat kako bi smanjile neke od svojih ručnih i repetitivnih zadataka, štedeći svojim timovima dragoceno vreme. Tekst majning algoritam mogao bi pomoći da se prepoznaju najpopularnije teme koje se pojavljuju u komentarima kupaca i način na koji ljudi misle o njima: da li su komentari pozitivni, negativni ili neutralni? Takođe, mogle bi da se otkriju glavne ključne reči koje su kupci naveli u vezi sa datom temom. Ukratko, tekst majning pomaže kompanijama da maksimalno iskoriste svoje podatke, što dovodi do boljih poslovnih odluka zasnovanih na podacima.

Tekst majning postiže ove rezultate uz pomoć mašinskog učenja. Mašinsko učenje je disciplina povezana sa veštačkom inteligencijom, koja se fokusira na kreiranju algoritama koji omogućavaju da mašine nauče zadatke na osnovu primera. Modele mašinskog učenja treba trenirati nad određenim skupom podataka, koji se naziva trening skup. Nakon treninga model će vršiti predikcije sa određenim nivoom preciznosti [1].

Važno je napraviti poređenje između sledećih pojmova: tekst majning, analiza teksta i analitika teksta. Tekst majning i analiza teksta se često koriste kao sinonimi. Međutim, analitika teksta je malo drugačiji koncept. I tekst majning i analitika teksta imaju za cilj rešavanje istih problema, ali korišćenjem različitih metoda. Oboje nameravaju da reše problem automatske analize sirovih tekstualnih podataka. Tekst majning identifikuje relevantne informacije u tekstu i stoga daje kvalitativne rezultate. Analitika teksta, međutim, fokusira se na pronalaženje obrazaca i trendova u velikim skupovima podataka, što rezultira više kvantitativnim rezultatima. Analitika teksta se obično koristi za kreiranje grafikona, tabela i drugih vrsta vizuelnih izveštaja. Tekst majning kombinuje pojmove statistike, lingvistike i mašinskog učenja da bi se stvorili modeli koji uče iz trening podataka i mogu predvideti rezultate na osnovu novih informacija, na osnovu njihovog prethodnog iskustva. Analitika teksta, s druge strane, koristi rezultate analiza izvedenih pomoću modela tekst majninga, za stvaranje grafikona i svih vrsta vizualizacija podataka. Izbor pravog pristupa zavisi od toga koja vrsta informacija je dostupna. U većini slučajeva, oba pristupa se kombinuju za svaku analizu, što dovodi do ubedljivijih rezultata [1].

U nastavku rada će biti obrađenje različite metode i tehnike, koje tekst majning koristi.

2.1. Metode i tehnike u tekst majningu

Postoje različite metode i tehnike koje se koriste u tekst majningu. Metode se mogu podeliti u osnovne i napredne metode. U osnovne metode tekst majninga se ubrajaju:

- **Učestalost reči** - Učestalost reči se može koristiti za identifikovanje termina ili koncepata koji se najčešće ponavljaju u skupu podataka. Pronalaženje najviše pomenutih reči u nestruktuisanom tekstu može biti posebno korisno kada se analiziraju recenzije kupaca, razgovori na društvenim mrežama ili povratne informacije kupaca. Na primer, ako se reči "skupo" i "precenjeno" često pojavljuju u recenzijama kupaca, to može ukazivati na to da se cene treba prilagoditi ciljnom tržištu.
- **Kolokacija** (*eng. collocation*) - Kolokacija se odnosi na niz reči koji se često pojavljuju jedna blizu druge. Najčešći tipovi kolokacija su bigrami (par reči koji će se verovatno spojiti) i trigrami (kombinacija tri reči). Identifikovanje kolokacija kao jedne reči poboljšava granularnost teksta, omogućava bolje razumevanje njegove semantičke strukture i, na kraju, dovodi do tačnijih rezultata tekst majninga.
- **Konkordancija** (*eng. concordance*) - Konkordancija se koristi za prepoznavanje određenog konteksta ili instance u kojoj se pojavljuje reč ili skup reči. Ljudski jezik može biti dvosmislen, što znači da se ista reč može koristiti u mnogo različitih konteksta. Konkordancija može pomoći u razumevanju tačnog značenja reči na osnovu konteksta.

Napredne metode tekst majninga uključuju klasifikaciju teksta. Klasifikacija teksta je postupak dodeljivanja kategorija ili oznaka (*eng. tags*) nestruktuisanim tekstualnim podacima. Ovaj zadatak obrade prirodnog jezika olakšava organizovanje i struktuiranje složenog teksta, pretvarajući ga u značajne podatke. Zahvaljujući klasifikaciji teksta, preduzeća mogu da analiziraju sve vrste informacija i da na brz i isplativ način dobiju dragocene uvide. Najvažniji zadaci klasifikacije teksta predstavljaju:

- **Analizu tema** – pomaže pri razumevanju glavnih tema teksta i organizovanju tekstualnih podataka (na primer, podaci sa označenim temama se mogu klasifikovati po njima),
- **Analizu sentimenta** – bavi se analizom emocija koje leže u osnovi bilo kog teksta. Analiza sentimenata pomaže pri razumevanju mišljenja i emocija u tekstu i njegovoj klasifikaciji kao pozitivog, negativnog ili neutralnog. Ova analiza se može koristiti pri analizi objava na društvenim mrežama, pregleda recenzija, itd.,
- **Detekciju korišćenog jezika** – omogućava klasifikaciju teksta na osnovu jezika koji je u tekstu korišćen. Jedna od korisnih primena jeste aplikacija koja automatski prosleđuje tikete pravom geografskom pogrućju,
- **Detekcija namera** – omogućava klasifikovanje teksta prema nameri koja stoji iza njega ili svrhe iza teksta. Ovo omogućuje kompanijama da lakše prepoznaju potencijalne klijente koji su zainteresovani za njihove usluge i da njima posvete svoji pažnju.

Tehnika koja se često primenjuje kod tekst majninga jeste ekstrakcija teksta. Ekstrakcija teksta je tehnika analize teksta koja iz teksta izdvaja određene podatake, kao što su ključne reči, imena entiteta, adrese, e-mail adrese itd. Korišćenjem ekstrakcije teksta kompanije mogu izbeći ručno sortiranje podataka da bi izvukle ključne informacije. Često,

može biti najefikasnije kombinovati ekstrakciju teksta i klasifikaciju teksta u istoj analizi. Osnovni zadaci tehnike ekstrakcije teksta su:

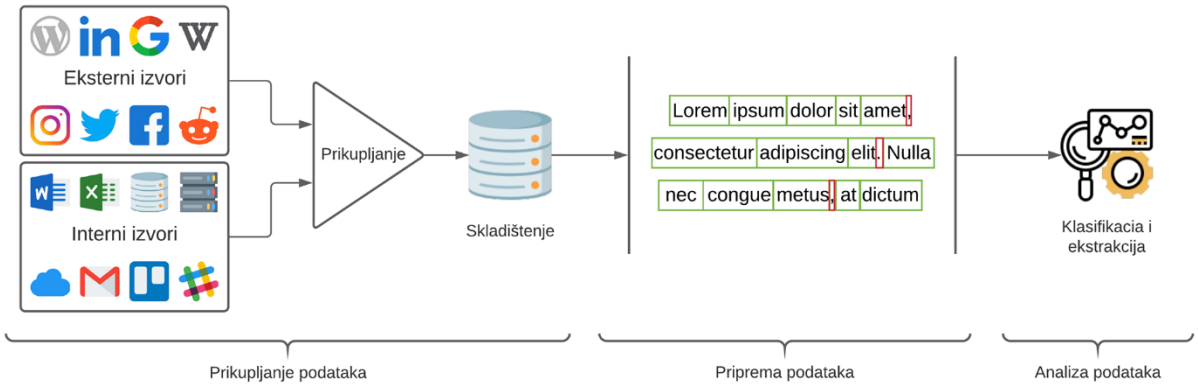
- **Izdvajanje ključnih reči** – omogućava indeksiranje podataka koji se pretražuju, rezimiranje sadržaja teksta ili označivanje teksta (tagovanje), s obzirom da ključne reči predstavljaju najrelevantnije pojmove u tekstu,
- **Prepoznavanje imenovanih entiteta** – omogućava identifikaciju i izvlačenje imena kompanije, organizacija, osoba i drugih imenovanih entiteta iz teksta,
- **Izdvajanje karakteristika** - pomaže u identifikovanju specifičnih karakteristika proizvoda ili usluga u skupu podataka. Na primer, ako se analiziraju opisi proizvoda, mogle bi da se detektuju karakteristike poput boje, marke, modela itd [1].

U nastavku će biti više reči o procesu tekst majninga.

2.2. Proces tekst majninga

Prilikom procesa tekst majninga podaci prolaze kroz tri glavna koraka:

- **Prikupljanje podataka** – ovaj korak uključuje prikupljanje internih (razmenjene poruke, razmenjena elektronska pošta, ankete, itd.) i/ili eksternih (objave sa društvenih mreža, novinski članci ili tekst sa bilo koje Web stranice) podataka i njihovo skladištenje,
- **Priprema podataka** – prilikom rukovanja tekstom, primenjuju se tehnike poput tokenizacije, lematizacije, uklanjanja stop reči, stemovanja, kako bi se reči pripremile i propustile kroz model mašinskog učenja,
- **Analiza podataka** – primene klasifikacije i ekstrakcije teksta [1].



Slika 1. Proces tekst majninga

Na slici 1 je prikazan proces tekst majninga. U nastavku će biti obrađeno prepoznavanje imenovanih entiteta i detekcija tema.

2.3. Detekcija tema

Detekcija tema je tehnika mašinskog učenja koja organizuje i razume velike kolekcije tekstualnih podataka, dodeljivanjem oznaka ili kategorija prema svakoj pojedinačnoj temi. Detekcija tema koristi obradu prirodnog jezika za razgradnju ljudskog jezika tako da se mogu pronaći uzorci i razaznati semantičke strukture u tekstovima, kako bi se izvuklo značanje i pomoglo u donošenju odluka na osnovu podataka. Dva najčešća pristupa za analizu tema uz mašinsko učenje su NLP modeliranje tema i NLP klasifikacija tema.

Modeliranje tema je tehnika mašinskog učenja bez nadzora. To znači da može da kreira obrasce i grupiše slične izraze bez prethodnog definisanja oznaka tema ili treninga podatka. Ova vrsta algoritma se može primeniti brzo i lako, ali postoji i mana - prilično su netačni.

Klasifikacija teksta, s druge strane, mora da zna teme teksta pre započinjanja analize, jer podaci moraju biti označeni da bi se trenirao klasifikator teme. Iako je u pitanju dodatni korak, klasifikatori tema se dugoročno isplaćuju i mnogo su precizniji od tehnika klasterizacije.

S obzirom da je u praktičnoj implementaciji projekta korišćeno modeliranje tema, ono će biti dalje obrađeno u ovom radu [2].

2.3.1. Modeliranje tema

Modeliranje tema rešava vrstu problema karakterisanu sledećim permisama: ako postoji skup tekstualnih dokumenata i cilj je saznati različite teme koje oni pokrivaju i grupisati ih prema tim temama. Način na koji ovi algoritmi rade pretpostavlja da je svaki dokument sastavljen od mešavine tema, a zatim pokušava da otkrije koliko je intenzivna prisutnost svake teme u datom dokumentu. To se postiže grupisanjem dokumenata na osnovu reči koje sadrže i uočavanjem korelacije između njih. Ovaj rad će obraditi najčešće korišćen algoritam: Latentna semantička analiza (LSA).

2.3.1.1. Latentna semantička analiza

Latentna semantička analiza je tradicionalna metoda za modeliranje tema. Zasnovana je na principu koji se naziva hipoteza o distribucij. Hipoteza o distibuciji nalaže da reči i izrazi koji se javljaju u sličnim delovima teksta imaju slična značenja. Algoritam je zasnovan na frekvencijama reči u skupu podataka. Opšta ideja je da za svaku reč u svakom dokumentu izbroji učestalost te reči i da se grupišu dokumenti koji imaju visoke frekvencije istih reči. Učestalost reči ili pojma u dokumentu je broj koji pokazuje koliko se često reč pojavljuje u dokumentu. Učestalost se može izračunati jednostavnim brojanjem - ako se reč mačka pojavi 10 puta u dokumentu, onda je njena učestalost 10. Ovaj pristup se pokazao pomalo ograničenim, pa se *tf-idf (eng. Term Frequency–Inverse Document Frequency)* obično koristi. *Tf-idf* mera je proizvod logaritma frekfvence termina i inverzne frekvence dokumenata za posmatrani termin. Frekvenca dokumenata za dati termin *t* (df_t) je broj dokumenata u korpusu

koji sadrže posmatrani termin. Inverzna frekvenca dokumenata za posmatrani termin *t* se definiše kao:

$$idf_t = \log_{10} (N / df_t)$$

Gde je *N* ukupan broj dokumenata u korpusu. Inverzna frekvenca dokumenata je veća za termine koji se retko pojavljuju u korpusu. Formula po kojoj se *tf-idf* mera izračunava je:

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \times \log_{10}(N / df_t)$$

Tf-idf uzima u obzir koliko je reč učestana (u svim dokumentima) u odnosu na to koliko je učestana u određenom dokumentu, pa su češće reči rangirane više, jer se smatraju boljim „predstavljanjem“ dokumenta, čak i ako nisu najbrojnije.

Nakon proračuna učestalosti reči, kreira se matrica koja ima red za svaku reč i kolonu za svaki dokument. Svaka ćelija je izračunata učestalost za određenu reč u određenom dokumentu. Ova matrica naziva se matrica dokumenata. Iz nje se mogu kreirati matrica teme dokumenta i matrica teme pojma, koje povezuju dokumente sa temama i termine sa temama. Ove matrice prikazuju informacije o temama tekstova.

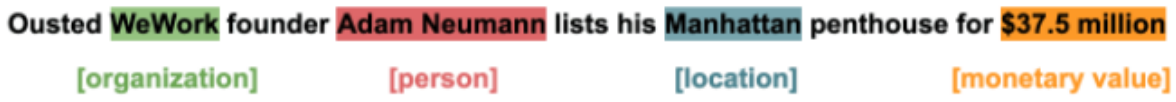
Način na koji se generišu ove matrice je dekompozicija matrice dokumenata na tri matrice upotrebom tehnike dekompozicije pojedinačne vrednosti, koja se naziva skraćeno SVD. Dekompozicija pojedinačne vrednosti (*SVD*) je linearni algebarski algoritam za faktorizaciju matrice u proizvod tri matrice *U * S * V*. Važno je da je srednja matrica S dijagonalna matrica singularnih vrednosti originala matrica. Za LSA, svaka pojedinačna vrednost predstavlja potencijalnu temu.

Skraćeni SVD bira najveće *t* singularne vrednosti i zadržava prvih *t* kolona U i prvih *t* redova V, smanjujući dimenzionalnost prvobitnog razlaganja. *t* će biti broj tema koje algoritam pronađe, pa je to hiperparametar kojem je potrebno podešavanje. Ideja je da se odaberu najvažnije teme, gde je U matrica dokument-tema, a V matrica-tema.

Vektori koji čine ove matrice predstavljaju dokumente izražene temama i pojmove izražene temama; mogu se meriti tehnikama poput kosinusne sličnosti za procenu [2].

2.4. Prepoznavanje imenovanih entiteta

Prepoznavanje imenovanih entiteta (*Named Entity Recognition - NER*), takođe se naziva identifikacija entiteta ili izdvajanje entiteta, je tehnika obrade prirodnog jezika (*Natural Language Processing - NLP*) koja automatski identifikuje imenovane entitete u tekstu i klasifikuje ih u unapred definisane kategorije. Entiteti mogu biti imena ljudi, organizacija, lokacija, vremena, količine, novčane vrednosti, procenti i još mnogo toga. Na slici 2 su prikazani kategorizovani imenovani entiteti u rečenici.



Slika 2. Imenovani entiteti u rečenici

Pomoću prepoznavanja imenovanih entiteta se mogu izvući ključne informacije za razumevanje o čemu se radi u tekstu.

Kada se čita tekst, prirodno se prepoznaju i kategorišu imenovani entiteti kao što su ljudi, organizacije, države, gradovi, lokacije itd. Na primer, u rečenici „Mark Zuckerberg je jedan od osnivača Facebook-a, kompanije iz Sjedinjenih Država“ mogu se identifikovati tri vrste entiteta:

1. Ličnost: Mark Zuckerberg
2. Kompanija: Facebook
3. Lokacija: Sjedinjene Države

Međutim, računari moraju da prvo prepoznaju entitete kako bi ih mogli kategorisati. To se postiže mašinskim učenjem i obradom prirodnog jezika (NLP). NLP proučava strukturu i pravila jezika i stvara inteligentne sisteme sposobne da ekstrahuju značenje iz teksta i govora, dok mašinsko učenje pomaže mašinama da se vremenom uče i poboljšavaju. Da bi naučio šta je entitet, NER (*eng. Named Entity Recognition - NER*) model mora biti u stanju da detektuje reč ili niz reči koje čine entitet (npr. Njujork, Novi Zeland) i da zna kojoj kategoriji entiteta pripada.

Dakle, prvo se moraju kreirati kategorije entiteta, kao što su: ime, lokacija, događaj, organizacija itd. Onda se u NER model unose relevantni trening podaci. Zatim, označavanjem uzoraka reči i fraza njihovim odgovarajućim entitetima, NER model uči kako da sam otkriva entitete [3].

3. IMPLEMENTACIJA

U ovom radu je urađena detekcija tema i prepoznavanje imenovanih entiteta, u cilju razumevanja tekstualnih podataka Donalda Trampa i Hilari Klinton, koji su preuzeti sa Tvitera [4]. Tviter je društvena mreža na kojoj ljudi međusobno komuniciraju i objavljuju kratke objave, koje se nazivaju tvitovi. Podaci su prikupljeni neposredno pre predsedničkih izbora u Americi 2016. godine. Ova analiza teksta pomaže u zapažanju najčešćih tema i imenovanih entiteta o kojima su Donald Tramp i Hilari Klinton govorili pred izbore. Za potrebe implementacije su korišćene biblioteke:

- spaCy,
- NLTK,
- matplotlib,
- pandas.

SpaCy je softverska biblioteka otvorenog koda za naprednu obradu prirodnih jezika, napisana na programskim jezicima Python i Cython. Za razliku od NLTK biblioteke, koja se široko koristi za nastavu i istraživanje, spaCy se fokusira na pružanje softvera, koji će se koristiti u produkcionim okruženjima [5].

NLTK biblioteka predstavlja skup biblioteka i programa za simboličku i statističku obradu prirodnog jezika (engleskog jezika) napisana na programskom jeziku Python. Razvili su ga Steven Bird i Edvard Loper na Odeljenju za računarstvo i informatiku na Univerzitetu u Pensilvaniji. NLTK biblioteka je namenjena podršci istraživanjima i nastavi u NLP-u ili blisko povezanim oblastima, uključujući empirijsku lingvistiku, kognitivnu nauku, veštačku inteligenciju, pronalaženje informacija i mašinsko učenje. NLTK se uspešno koristi kao nastavno sredstvo, kao individualno sredstvo za proučavanje i kao platforma za izradu

prototipova i izgradnju istraživačkih sistema. Postoje 32 univerziteta u SAD-u i 25 zemalja koje koriste NLTK na svojim kursevima. NLTK podržava funkcionalnosti klasifikacije, tokenizacije, stemovanja, označavanja, raščlanjivanja i semantičkog zaključivanja [6].

SpaCy biblioteka je korišćenja za prepoznavanje imenovanih entiteta jer ima bolje performanse u odnosu na NLTK biblioteku. SpaCy biblioteka ima više oznaka za imenovane entitete u odnosu na NLTK biblioteku, te tako preciznije označava imenovane entitete u tekstualnim podacima. Tabela podržanih oznaka za imenovane entitete biblioteka spaCy i NLTK su date na slici 3. Takođe, NLTK biblioteka je u nekim primerima pogrešno klasifikovala organizacije kao što su Google, Facebook, itd. SpaCy biblioteka je navedene organizacije označila oznakom ORG, dok je NLTK biblioteka ove organizacije označila oznakom PERSON, što prestavlja pogrešnu oznaku (iako obe biblioteke imaju oznake PERSON i ORG).

spaCy	NLTK
PERSON	PERSON
ORG	ORG
FAC	FACILITY
GPE	GPE
LOC	LOCATION
DATE	DATE
TIME	TIME
PERCENT	PERCENT
MONEY	MONEY
NORP	/
PRODUCT	/
EVENT	/
LAW	/
WORK_OF_ART	/
LANGUAGE	/

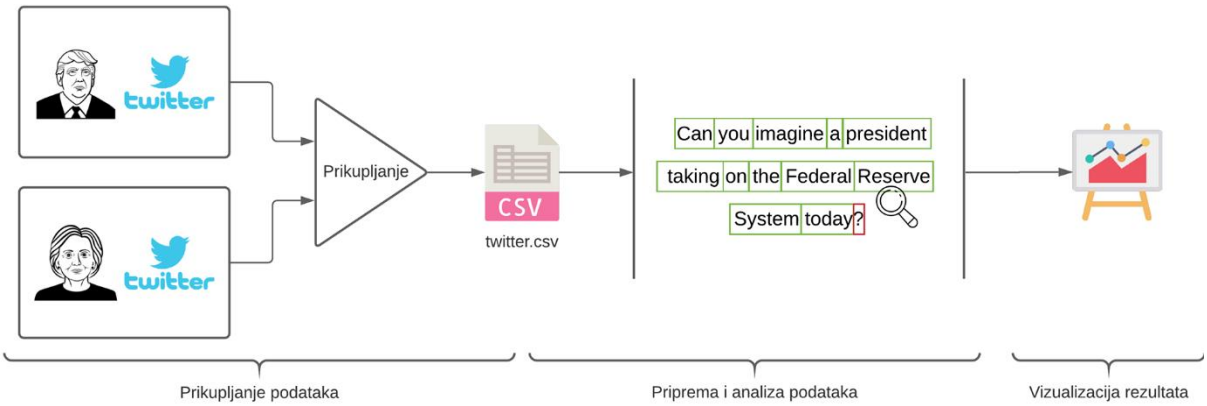
Slika 3. Podržane oznake za kategorizaciju imenovanih entiteta u bibliotekama spaCy i NLTK

NLTK biblioteka je korišćena za detekciju tema, s obzirom da su tokenizacija i stemovanje bolje implementirane u ovoj biblioteci, u poređenju sa spaCy bibliotekom. Prilikom tokenizacije reči, spaCy biblioteka ima bolje performanse u odnosu na NLTK biblioteku. S druge strane, prilikom tokenizacije rečenica, NLTK biblioteka nadmašuje performanse spaCy biblioteke. Loš učinak biblioteke spaCy u tokenizaciji rečenica rezultat je različitih pristupa. NLTK pokušava da tekst podeli na rečenice. Suprotno tome, spaCy pravi sintaksičko stablo za svaku rečenicu, robusniju metodu koja daje mnogo više informacija o tekstu, ali ima lošije performanse [7]. S obzirom da su tvitovi u obliku jedne ili više rečenica, odlučeno je da se za potrebe tokenizacije, stemovanja i detekcije tema koristi NLTK biblioteka.

Za vizuelizaciju rezultata, dobijenih tekstualnom analizom, korišćena je matplotlib biblioteka. Matplotlib je biblioteka za kreiranje statičkih, animiranih i interaktivnih grafova u Python-u. Ova biblioteka je izabrana zbog lakoće njenog korišćenja, raznovrsnosti i lakoće postavljanja.

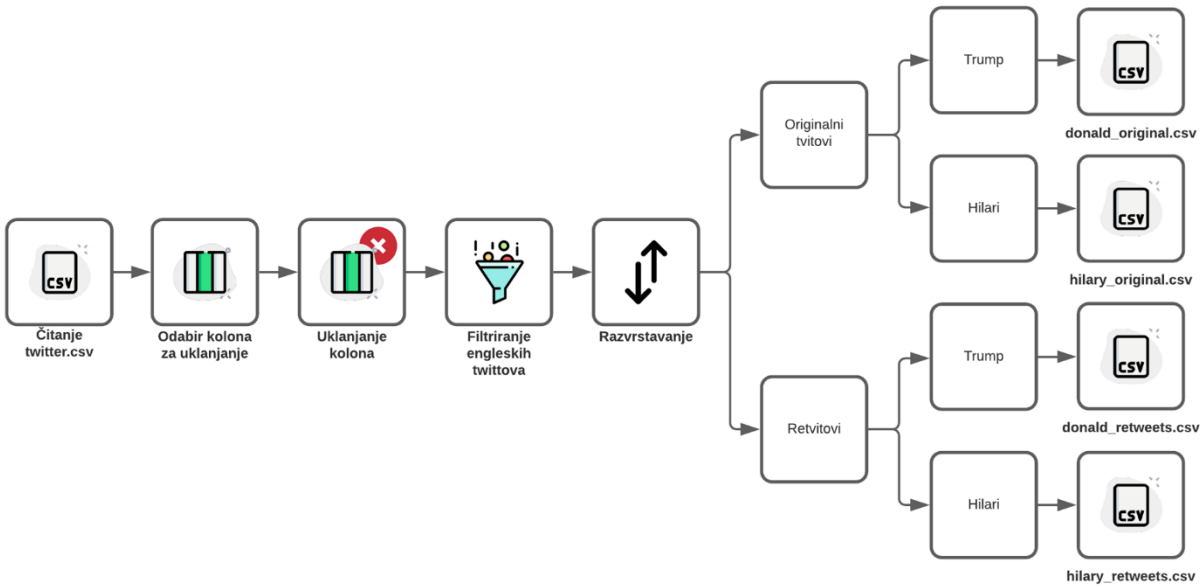
Za učitavanje i manipulaciju podacima je korišćena pandas biblioteka. Pandas biblioteka predstavlja moćan alat za manipulaciju podacima. Pandas ima za cilj da bude osnovni gradivni element na visokom nivou za obavljanje praktične analize podataka u stvarnom svetu u Python-u. Pored toga, ima širi cilj da postane najmoćniji i najfleksibilniji alat za analizu / manipulaciju podacima otvorenog koda, dostupan na bilo kom jeziku. Sadrži brz i efikasan objekat (*DataFrame*) za manipulaciju podacima sa integrisanim indeksiranjem. Sadrži i alate za čitanje i upis podataka u različitim formatima: CSV i tekstualne datoteke, Microsoft Excel, SQL baze podataka i brzi HDF5 format. Ova biblioteka ima visoko optimizovane performance i koristi se u širokim akademskim i komercijalnim krugovima, uključujući i finansije, ekonomiju, statistiku, Web analitiku, marketing, itd [8].

Skup podataka koji je korišćen u ovom projektu ima 8448 redova i 28 atributa. Implementirani projekat podatke nakon prikupljanja, najpre priprema, zatim analizira i na kraju vizuelizuje. Tok obrade podataka je prikazan na slici 4.



Slika 4. Tok obrade podataka u implementiranoj aplikaciji

Podaci su prikupljeni i skladišteni u *twitter.csv* fajlu. Ovaj fajl sadrži tvitove i Hilari Klinton i Donalda Trampa. Ove podatke treba dalje pročistiti tj. pripremiti za analizu. Tok prečišćavanja podataka je dat na slici 5.



Slika 5. Tok pripreme i prečišćavanja podataka

S obzirom da postoji veliki broj nerelevantnih atributa, selektovane su kolone koje su nerelevantne za analizu i izbačene su. Izbačeni atributi su prikazani na slici 6.

```
labels_to_drop = ['in_reply_to_screen_name',
                  'in_reply_to_status_id',
                  'in_reply_to_user_id',
                  'is_quote_status',
                  'longitude',
                  'latitude',
                  'place_id',
                  'place_full_name',
                  'place_name',
                  'place_type',
                  'place_country_code',
                  'place_country',
                  'place_contained_within',
                  'place_attributes',
                  'place_bounding_box',
                  'source_url',
                  'truncated',
                  'entities',
                  'time',
                  'retweet_count',
                  'favorite_count',
                  'extended_entities']
```

Slika 6. Atributi koji su selektovani za izbacivanje

Svaki red sadrži sledeće preostale atribute:

- **id** – atribut koji jedinstveno identifikuje svaki tvit,
- **handle** - atribut koji označava da li je tvit postavio Donald Tramp ili Hilari Klinton,
- **text** - atribut koji predstavlja tekstualni sadržaj tvita,
- **is_retweet** - atribut koji označava da li je tvit postavljen od strane neke druge osobe (osobe koja nije Hilari Klinton ili Donald Tramp),
- **original_author** - atribut koji predstavlja ime osobe koja je postavila tvit, ako je *is_retweet* atribut postavljen na *True*. Ako je *is_retweet* atribut postavljen na *False*, vrednost ovog atributa je prazna,
- **lang** - atribut koji predstavlja jezik u kojem je tvit napisan.

Pre daljeg prečišćavanja podataka, treba uzeti u obzir broj jedinstvenih vrednosti koje imaju atributi. U ovom skupu podataka, atributi *text*, *lang*, *handle* i *original_author* imaju sledeći broj jedinstvenih vrednosti (slika 7):

```
Atribut 'handle' ima 2 jedinstvenih vrednosti
Atribut 'text' ima 6434 jedinstvenih vrednosti
Atribut 'original_author' ima 279 jedinstvenih vrednosti
Atribut 'lang' ima 8 jedinstvenih vrednosti
```

Slika 7. Jedinstvene vrednosti atributa

Može se videti da atribut hadle ima 2 jedinstvene vrednosti, a to su HillaryClinton (3226 redova ima ovu vrednost za atribut *handle*) i realDonaldTrump (3218 redova ima ovu vrednost za atribut *handle*). Ovo se može videti i na slici 8:

```
HillaryClinton     3226
realDonaldTrump    3218
```

Slika 8. Distribucija vrednosti atributa *handle*

Treba uzeti u obzir činjenicu da Donald Tramp i Hilari Klinton postavljaju tvitove na engleskom jeziku, ali i na drugim jezicima, kao što se može videti na slici iznad. Na slici 9 se može videti da ima tvitova napisanih na španskom, francuskom, danskom, finskom i ostalim jezicima.

```
en     6248
es      105
und      82
da        3
fr        2
tl        2
fi        1
et        1
```

Slika 9. Distribucija vrednosti atributa *lang*

Ovaj projekat je implementiran za potrebe analize teksta napisanog na engleskom jeziku, te tako iz skupa podataka treba izbaciti sve tvitove koji su napisani na nekom drugom jeziku.

Zatim treba izvršiti podelu tvitova prema tome da li su tvitovi napisani od strane Donalda Trampa ili Hilari Klinton lično, ili su napisani od strane neke druge osobe. Ako je tvit napisan od strane Donalda Trampa ili Hilari Klinton lično, onda će vrednost atributa *original_author* biti prazno. S druge strane, ako je tvit napisan od strane neke druge osobe, a Donald Tramp ili Hilari Klinton su ih samo retvitovali, onda će vrednost atributa *original_author* biti ime osobe koje originalno napisala ovaj tvit. Podela tvitova prema tome da li su tvitovi originalni ili retvitovani je prikazana na slici 10.

```
""" Split dataset into original tweets and reteets """
df_original = df[df.original_author.isnull() == True]
df_retweets = df[df.original_author.isnull() == False]
```

Slika 10. Podela tvitova prema vrednosti atributa *original_author*

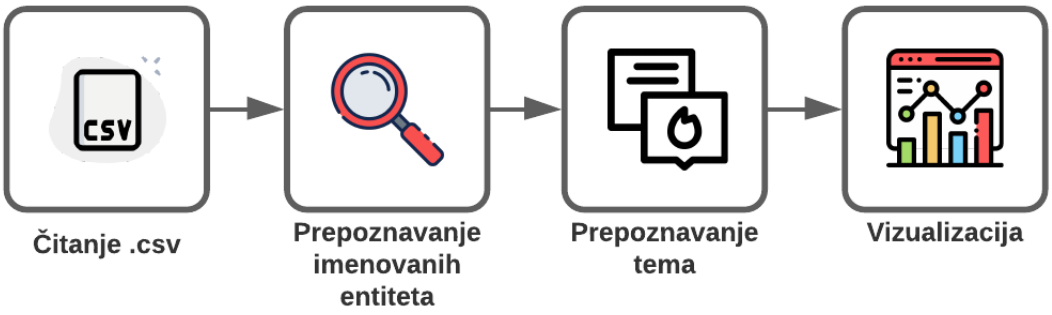
Nakon podele na originalne i neoriginalne tvitove, treba dalje podeliti tvitove prema tome ko ih je postavio. Ako je tvit postavila Hilari Klinton, vrednost atributa *handle* će biti HilaryClinton. Ako je tvit postavio Donald Tramp, vrednost atributa *handle* će biti realDonaldTrump. Podelu tvitova prema tome ko ih je postavio, treba odraditi i nad originalnim tvitovima i nad neoriginalni (retvitovanim) tvitovima. Ova podela je prikazana na slici 11.

```
""" Split original tweets into hilarys and donalds """
donald_tweets = df_original[df_original.handle == 'realDonaldTrump']
hilary_tweets = df_original[df_original.handle == 'HillaryClinton']

""" Split retweeted tweets into hilarys and donalds """
donald_retweets = df_retweets[df_retweets.handle == 'realDonaldTrump']
hilary_retweets = df_retweets[df_retweets.handle == 'HillaryClinton']
```

Slika 11. Podela tvitova prema vrednosti atributa *handle*

Ovim je priprema podataka završena i može se preći na analizu teksta. Za svaki tvit u skupu podataka je urađeno prepoznavanje imenovanih entiteta, a zatim su izdvojene i teme tvita. Tok analize tvitova je prikazan na slici 12.



Slika 12. Tok procesa analize tvitova

Prepoznavanje imenovanih entiteta je realizovano uz pomoć spaCy biblioteke. SpaCy biblioteka poseduje funkciju *nlp* koja kao ulaz uzima tvit, a vraća listu prepoznatih entiteta i oznaka. Oznake govore kojoj grupi prepoznati entitet pripada. Grupe mogu biti: osobe, organizacije, geografska područja, novac, vreme, datum, itd. Funkcija koja realizuje prepoznavanje imenovanih entiteta je prikazana na slici 13.

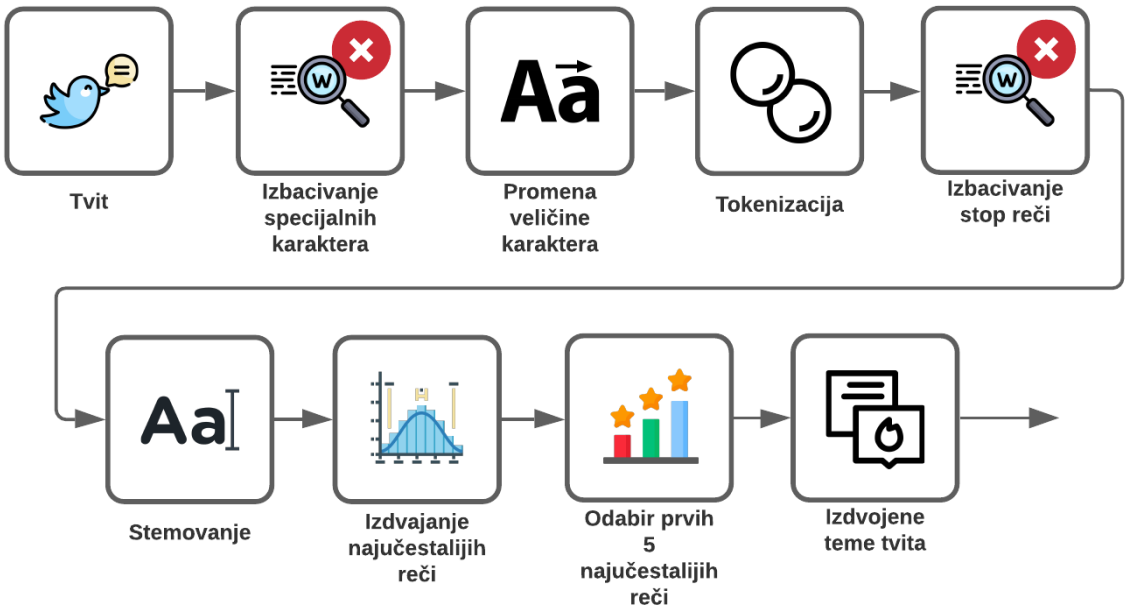
```
def _get_entities_list(tweets):
    entities_list = []
    url_list = []
    for index, row in tweets.iterrows():
        tweet = row['text']
        # tweet = _remove_links(tweet)
        recognized_entities = nlp(tweet)
        # url_list_tmp=[]
        # for i, token in enumerate(recognized_entities):
        #     if token.like_url:
        #         token.tag_ = 'URL'
        #         url_list_tmp.append((str(token), "URL"))

        # url_list.append(url_list_tmp)

        entities_list.append([(X.text, X.label_)
                              for X in recognized_entities.ents])
    return entities_list, url_list
```

Slika 13. Implementacija prepoznavanja imenovanih entiteta

Nakon prepoznavanja imenovanih entiteta treba detektovati temu tvita. Za detekciju teme je korišćena NLTK biblioteka. Tok detekcije tema tvita je prikazan na slici 14.



Slika 14. Tok procesa detekcije tema tvita

Iz svakog tvita se najpre izbacuju svi specijalni karakteri i brojevi, zatim se vrši promena veličine slova, tako da su sve reči napisane malim slovima. Nakon toga, treba izvršiti tokenizaciju teksta, zatim izbaciti stop reči i nakon toga izvršiti stemovanje. Za stemovanje se može koristiti Poterov stemer, a može se vršiti i lematizacija. U projektu je korišćen Poterov stemer. Stemovane reči se zatim prosleđuju funkciji *FreqDist*, koja pronalazi raspodelu učestanosti za svaku reč i sortira ih u opadajući redosled. Prvih pet najučestanijih reči, koje se pojavljuju u tvitu, određuju temu tvita. Detektovanje tema je implementirano u funkciji prikazanoj na slici 15.

```
def _get_text_tags_list(tweets):
    topic_tags_list = []
    for index, row in tweets.iterrows():
        stop_words = set(stopwords.words('english'))
        tweet = row['text']
        # tweet = _remove_links(tweet)
        tweet = re.sub("[^a-zA-Z]", " ", tweet).lower()

        word_tokens = word_tokenize(tweet)
        filtered_sentence = [w for w in word_tokens if not w in stop_words]
        stemmed_words = []
        stemmer = PorterStemmer()
        #stemmer = WordNetLemmatizer()
        for word in filtered_sentence:
            word = stemmer.stem(word)
            #word = stemmer.lemmatize(word)
            stemmed_words.append(word)

        fdist = FreqDist(stemmed_words)
        most_frequent_words = fdist.most_common(5)
        topic_tags_list.append(most_frequent_words)
    return topic_tags_list
```

Slika 15. Implementacija detekcije tema

Ovim je analiza teksta završena i rezultati se mogu dalje vizuelizovati. Za vizuelizaciju je iskorišćen grafikon tipa *horizontal bar*.

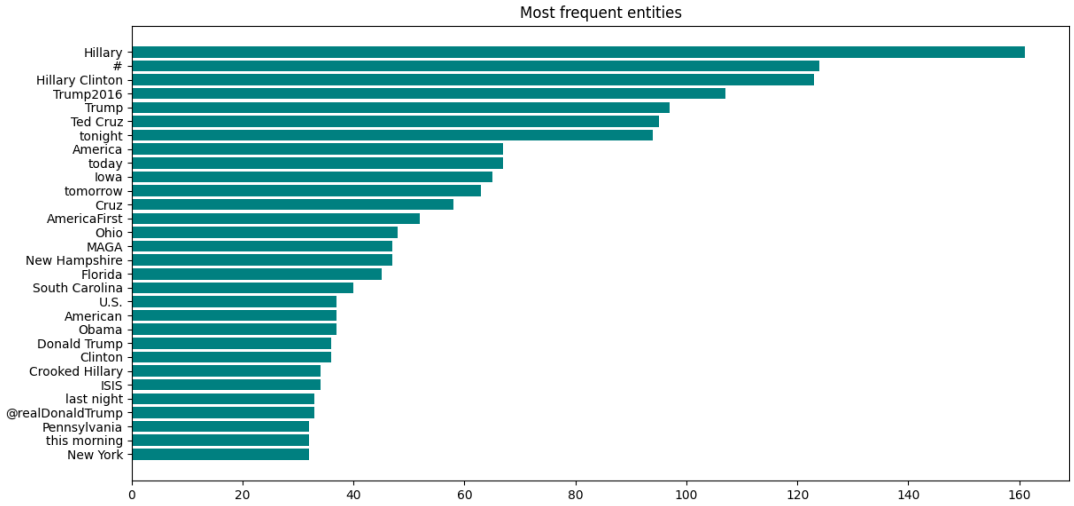
3.1. Analiza rezultata

Najpre je izvršena vizuelizacija originalnih tvitova Donalda Trampa. Na slici 16 su prikazani najčešće korišćeni imenovani entiteti, koji pripadaju bilo kojoj grupi entiteta. Sa slike 16 se može videti da je Donald Tramp najčešće pominjao svoju protivnicu Hilari Klinton, zatim Teda Kruza, senatora Teksasa, koji je u vreme izbora 2016. godine bio Trampov politički protivnik. Teda Kruza je Donald Tramp najčešće pominjao u tvitovima kao "lažnjivi Ted Kruz", a Hilari Klinton takođe pominje najčešće u negativnom kontekstu, govoreći "nepoštena Hilari" (*eng. crooked Hillary*). Tramp je takođe često pominjao i naziv svoje kampanje *Trump2016* i državu Ajovu. U državi Ajova je te godine Hilari Klinton održala "najgori" nastup demokratske partije. Tako loš nastup u državi Ajova nije imao nijedan član demokratske partije još od 1980. godine. Tramp je osvojio izbore u Ajovi sa 51,2% glasova, dok je Hilari Klinton dobila 41,7% [9]. Takođe, tvitovi Donalda Trampa često pominju Ajovu i Teda Kruza zajedno. Donald Tramp je često tvitovao o navodnoj prevari Teda Kruza prilikom biranja lokalnih političkih predstavnika u državi Ajova. Zatim se može zapaziti da je Donald Tramp dosta pominjao Ameriku i slogan *AmericaFirst*. *AmericaFirst* se odnosi na program politike predsedničke kampanje Donalda Trampa. Takođe, može se videti da pored Ajove, Donald Tramp dosta pominje i države Pensilvanija,

Južna Karolina, Florida, Ohajo, Njujork i Nju Hempšir. Klintonova je pobedila na izborima u Nju Hempširu i Njujorku. Dok je Donald Tramp osvojio više glasova u Pensilvaniji, Južnoj Karolini, Ohaju i Floridi [10] [11] [12] [13] [14]. Tramp, takođe, pominje i organizaciju ISIS. Najčešće povezuje ovu organizaciju sa Hilari Klinton i Barakom Obamom. To se može videti u više tvitova, kao što su:

„ISIS gained tremendous strength during Hillary Clinton's term as Secretary of State. When will the dishonest media report the facts!“

„Obama's disastrous judgment gave us ISIS, rise of Iran, and the worst economic numbers since the Great Depression!“



Slika 16. Najčešće korišćeni imenovani entiteti bilo koje kategorije u originalnim tvitovima Donalda Trampa

Na slici 17 su prikazani najčešće korišćeni imenovani entiteti koji pripadaju grupi PERSON. Pored gore navedenih osoba, sa slike se može videti da je Donald Tramp pominjao i Marka Rubia, senatora Floride. Marka Rubia je pominjao u negativnom kontekstu i nazivao ga prevarantom u svojim tvitovima. To se može videti u sledećem tvitu:

“I will be using Facebook and Twitter to expose dishonest lightweight Senator Marco Rubio. A record no-show in Senate, he is scamming Florida.”

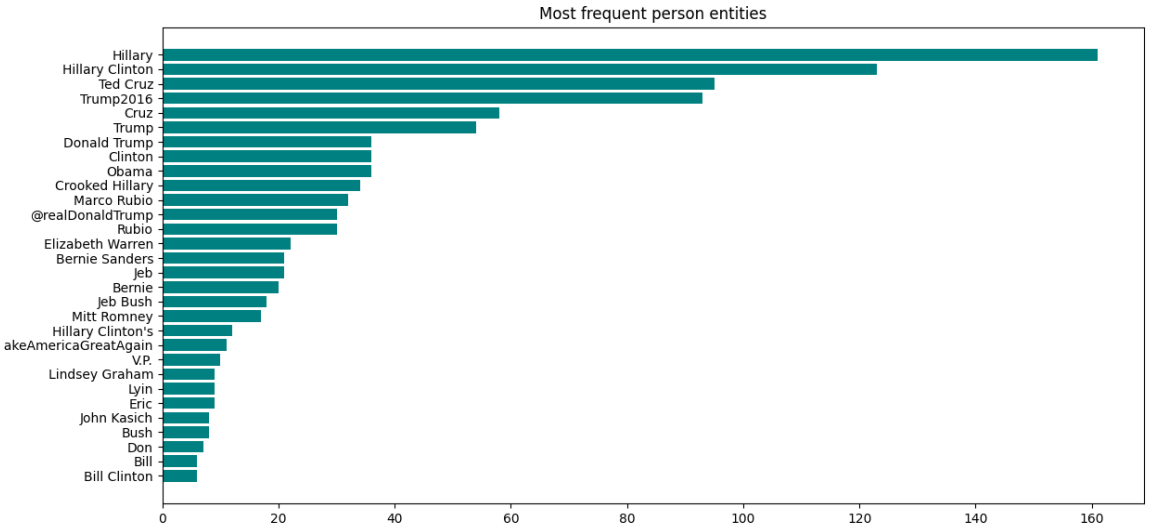
Može se videti da je Donald Tramp pominjao često u tvitovima Elizabet Varen, koju je najčešće nazivao “blesava Elizabet Varen”. Tvitovi o Elizabet Varen su deo “tviter rata” koji su njih dvoje vodili. Sa slike 17 se može videti da ju je Donald Tramp proznao u više od 20 navrata. Prepiska je počela nakon što je Elizabet nazvala Trampa propalitetom i šarlatanom. Na šta je Tramp odgovorio brojnim tvitovima i nazivao je rasistom, nasilnikom, pa ju je čak i optužio da laže o svom poreklu. To se može videti u sledećim tvitovima:

“Crooked Hillary is wheeling out one of the least productive senators in the U.S. Senate, goofy Elizabeth Warren, who lied on heritage.”

“Goofy Elizabeth Warren, sometimes referred to as Pocahontas, pretended to be a Native American in order to advance her career. Very racist!”

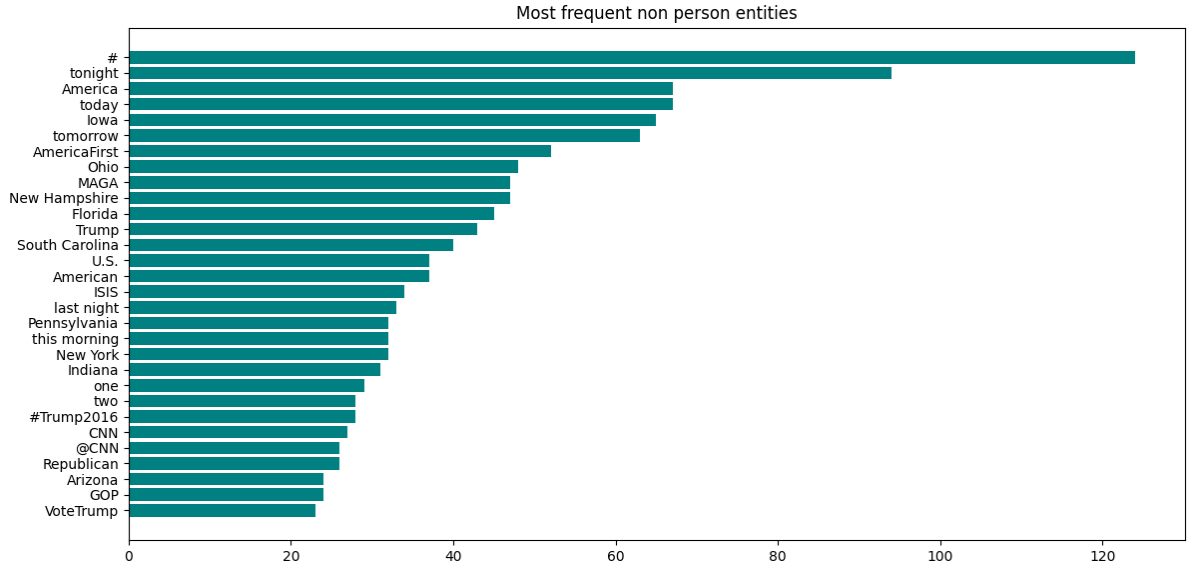
Dalje se može videti da je Tramp govorio i o Džebu Bušu i o Berniju Sandersu, takođe senatoru u SAD-u. Tramp je pominjao i Mita Romnea, nazivajući ga u većini tvitova

“propalim kandidatom”. Takođe je pominjao u negativnom kontekstu i svoje protivnike Lindsei Grejam, Džona Kasika i Bil Klintona.



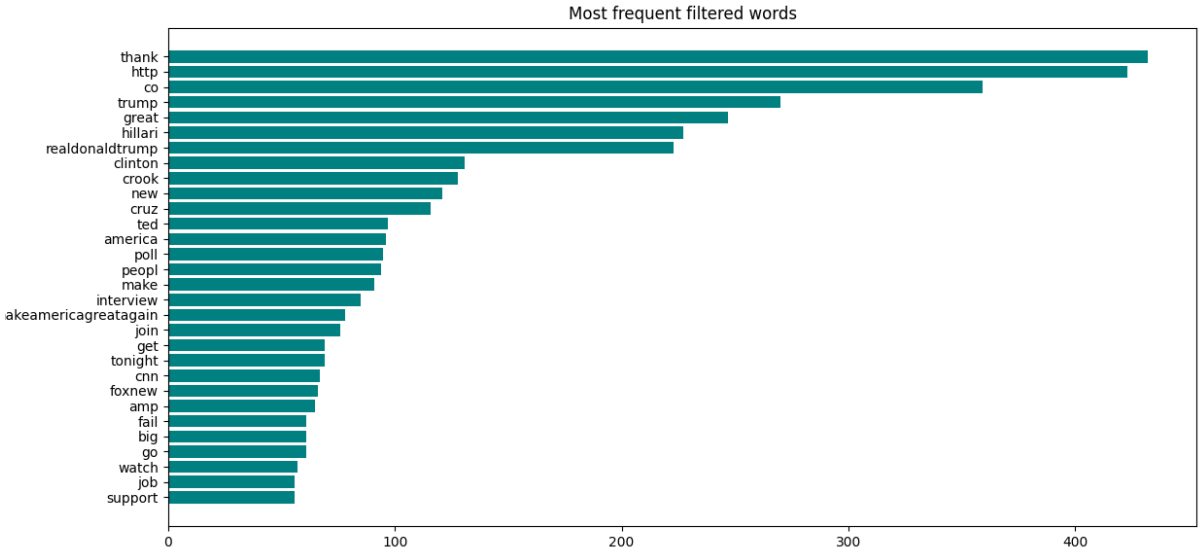
Slika 17. Najčešće korišćeni imenovani entiteti kategorije PERSON u originalnim tvitovima Donalda Trampa

Na slici 18 se mogu videti najčešći imenovani entiteti, korišćeni od strane Donalda Trampa, koji ne pripadaju grupi PERSON. Na ovoj slici se vidi da je jako često koristio skraćenicu MAGA, što predstavlja skraćenicu od *Make America Great Again*. Takođe je koristio i skraćenicu GOP (*eng. Grand Old Party*), koja predstavlja drugo ime za Republikansku partiju. Može se videti da je Tramp prozivao i CNN. CNN-u je zamerao što ga pominju samo u negativnom kontekstu i optuživao ga da propagira laži.



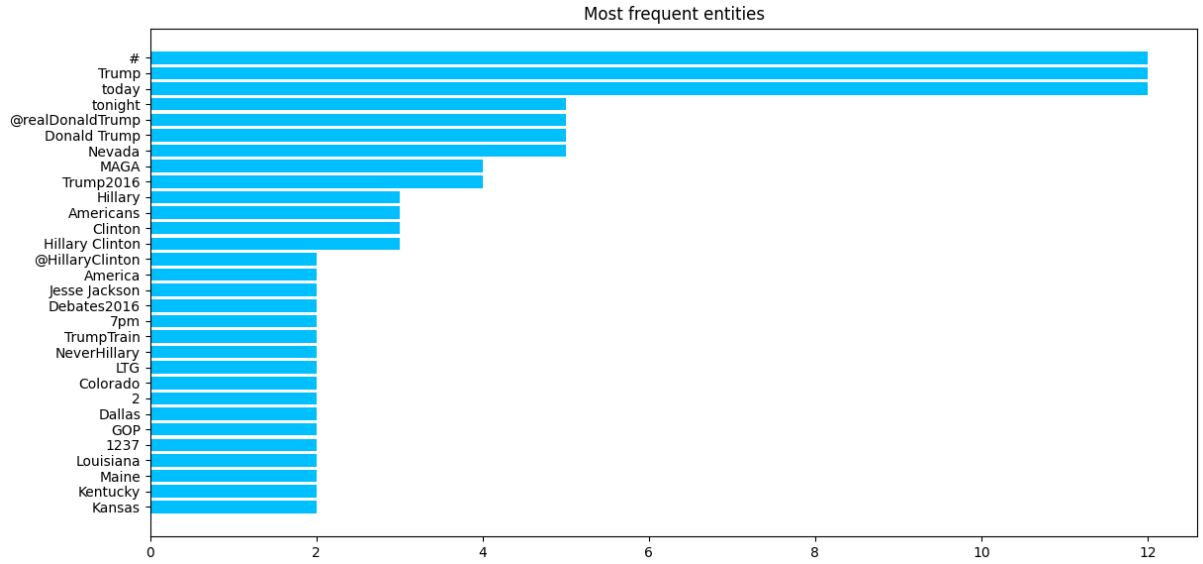
Slika 18. Najčešće korišćeni imenovani entiteti koji nisu kategorije PERSON u originalnim tvitovima Donalda Trampa

Na slici 19 se mogu videti najčešće teme Donaldovih tvitova. Sa slike 19 se može videti da je Donald Tramp često postavljao linkove stranica, videa i novinskih članaka. Glavne teme tvitova su mu bili “nepoštena Hilari”, Ted Kruz, sam Donald Tramp, kao i televizijski kanali CNN i FoxNews.



Slika 19. Najčešće teme originalnih tvitova Donalda Trampa

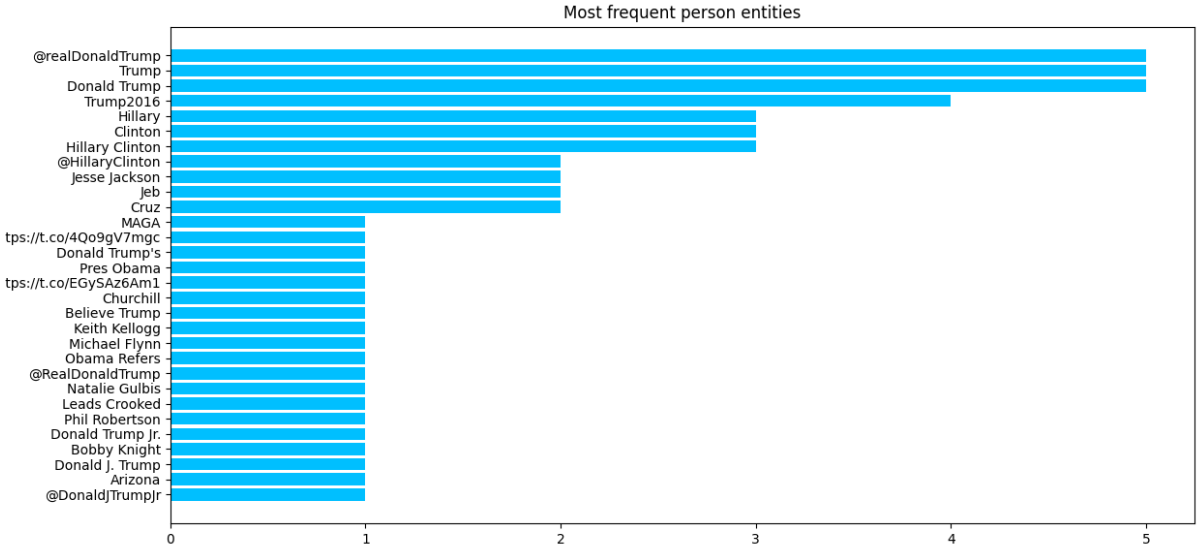
Kada su u pitanju retvitovani tvitovi Donalda Trampa, na slici 20 su prikazani najčešće korišćeni entiteti iz bilo koje kategorije. Sa slike 20 se može videti da je on najviše retvitovao tvitove koji govore o njemu, njegovoj kampanji (*MAGA*, *Trump2016*, *TrampTrain*), o debatama, Hilari Klinton i kampanji protiv Hilari (*NeverHillary*)



Slika 20. Najčešće korišćeni imenovani entiteti bilo koje kategorije u retvitovanim tvitovima Donalda Trampa

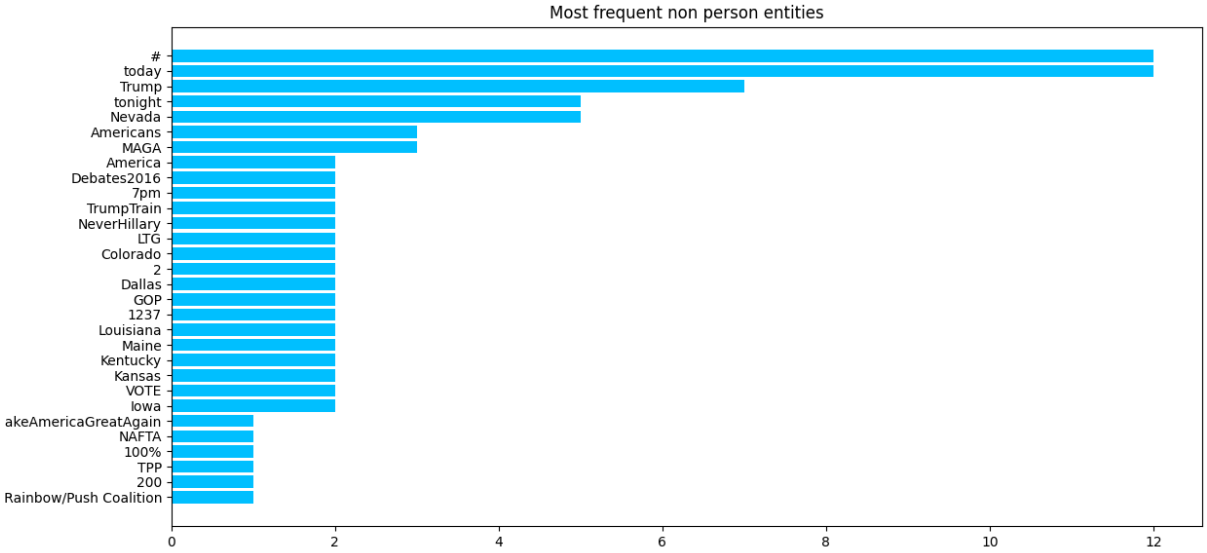
Na slici 21 se mogu videti najčešće korišćeni imenovani entiteti iz grupe PERSON, koji su se nalazili u tvitovima koje je Donald Tramp retvitovao. Može se videti da je najviše

retvitovao tvitove koji govore o njemu, o Hilari Klinton, Džebu Bušu, Tedu Kruzu, Džesiju Džaksonu, njegovoj kampanji (*MAGA, Believe Trump*). Takođe, može se videti da je pominjao i Vinstona Čerčila. Tramp je retvitovao tvitove koji su poredili njegov govor sa Čerčilovim i hvalili obojcu. U retvitovima je pominjan i Kit Kelog, Majk Flin, Natali Gulbis, Fil Robertson, Bobi Najt i sin Donalda Trampa (Donald Trump Jr.). Svi ovi ljudi podržavaju Donalda Trampa, tako da su retvitovani tvitovi o njima bili u pozitivnom kontekstu. Takođe, na slici se može videti i Leads Crooked. Ovo nije realna osoba, već je ovo još jedno ime, kojim Donald Tramp oslovljava Hilari Klinton, kada govori o tome da Hilari u nekoj od država ima prednost. Na grafikonu se može videti da su neki linkovi pogrešno prepoznati kao imenovani entiteti kategorije PERSON. Ti linkovi su zapravo linkovi do Trampovih govora i obraćanja.



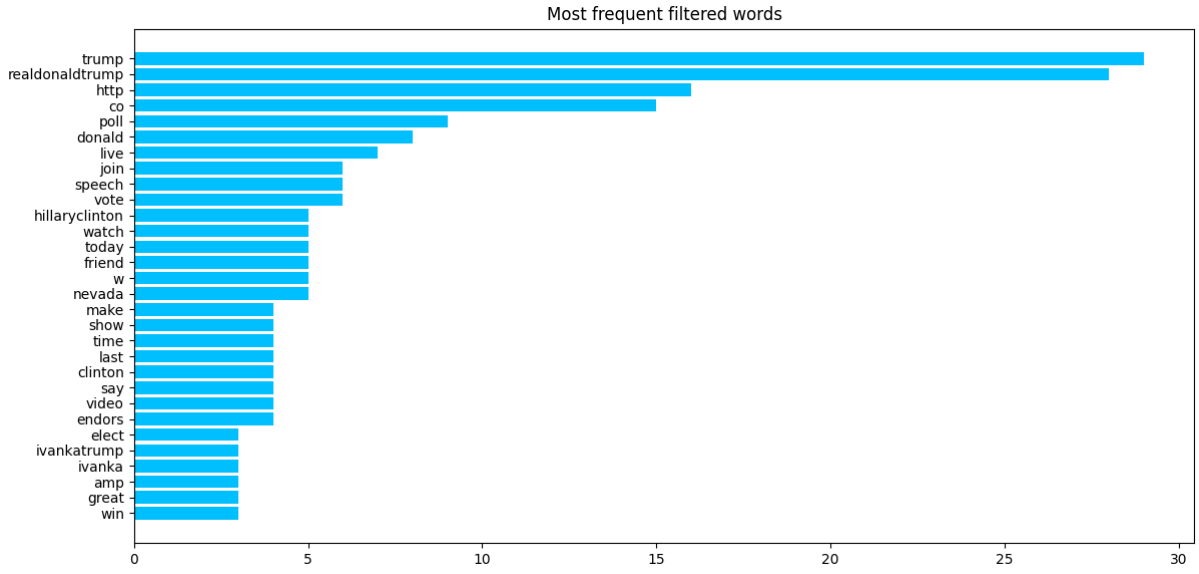
Slika 21. Najčešće korišćeni imenovani entiteti kategorije PERSON u retvitovanim tvitovima Donalda Trampa

Na slici 22 su prikazani najčešće korišćeni imenovani entiteti, koji ne pripadaju kategoriji PERSON, u retvitovima Donalda Trampa. Ovde su najčešće pominjani Donald Tramp, njegova kampanja, debate, Severnoamerički sporazum o trgovini, Republikanska partija, poziv na glasanje, negativne kampanje u vezi Hilari Klinton (*NeverHillary*) i različite države u SAD-u.



Slika 22. Najčešće korišćeni imenovani entiteti koji nisu kategorije PERSON u retvitovanim tvitovima Donalda Trampa

Na slici 23 su prikazane teme retvitova Donalda Trampa. I ovde su teme tvitova bile Donald Tramp, glasanje, birališta, Hilari Klinton, pobeda, izbori i Ivanka Tramp, čerka Donalda Trampa.



Slika 23. Najčešće teme retvitovanih tvitova Donalda Trampa

Nakon vizuelizacije tvitova Donalda Trampa, odrađena je vizuelizacija tvitova Hilari Klinton. Na slici 24 su prikazani najčešće korišćeni imenovani entiteti, iz bilo koje kategorije, u originalnim tvitovima Hilari Klinton. Zanimljivo je videti da Hilari Klinton zapravo najviše govori o sebi, dok Donald Tramp više govori o Hilari nego o sebi. Donald Tramp je sebe pomenuo nešto više od 100 puta, dok se Hilari sebe pomenula preko 500 puta. Hilari je najviše citirala sebe samu i govorila kako je ona spremna da bude nova predsednica SAD-a. Zatim, Hilari Klinton često pominje Ameriku, Donalda Trampa, prvu damu, predsednika SAD-a (u to vreme Barak Obama), Belu kuću, debatu, republikance i muslimane. Za

muslimane govori kako ih treba poštovati, a za republikance je govorila u negativnom svetlu. Neki od njenih tvitova o Republikanskoj partiji su:

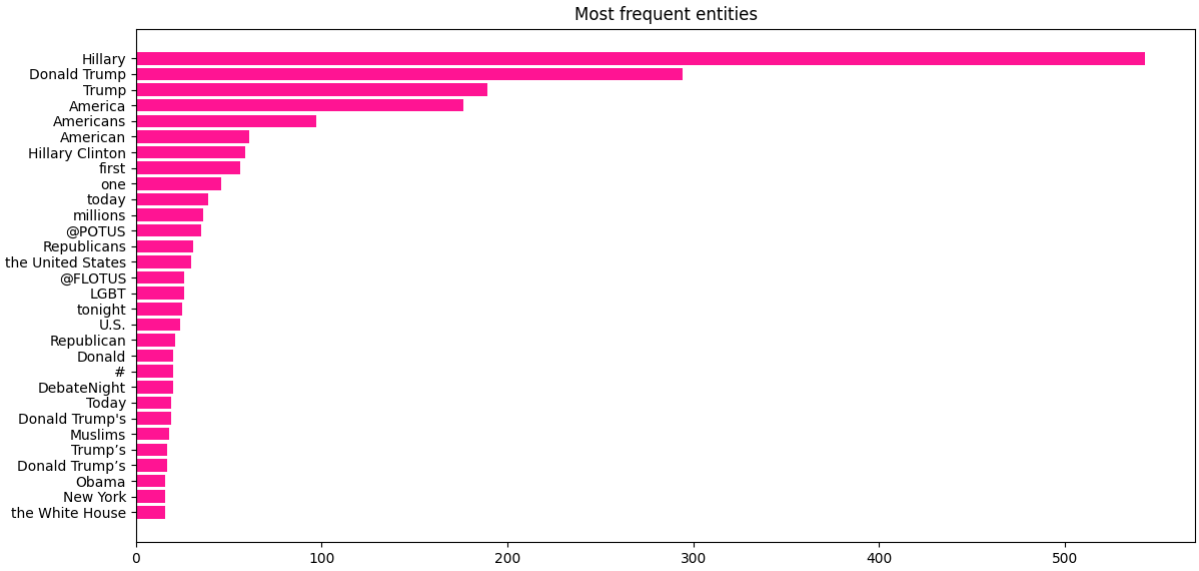
"We have a Republican nominee for president who incites hatred and violence like we've never seen before."

"It just seems that the economy does better under the Democrats than the Republicans."

Interesantno je zapaziti da je Hilari Klinton često pominjala LGBT zajednicu (čak 40 puta), dok je Donald Tramp u svojim originalnim tvitovima pomenuo ovu zajednicu tačno jedanput i to u tvitu u kome optužuje Hilari Klinton za oduzimanje slobode LGBT zajednici. To možemo videti u sledećem Donaldovom tvitu:

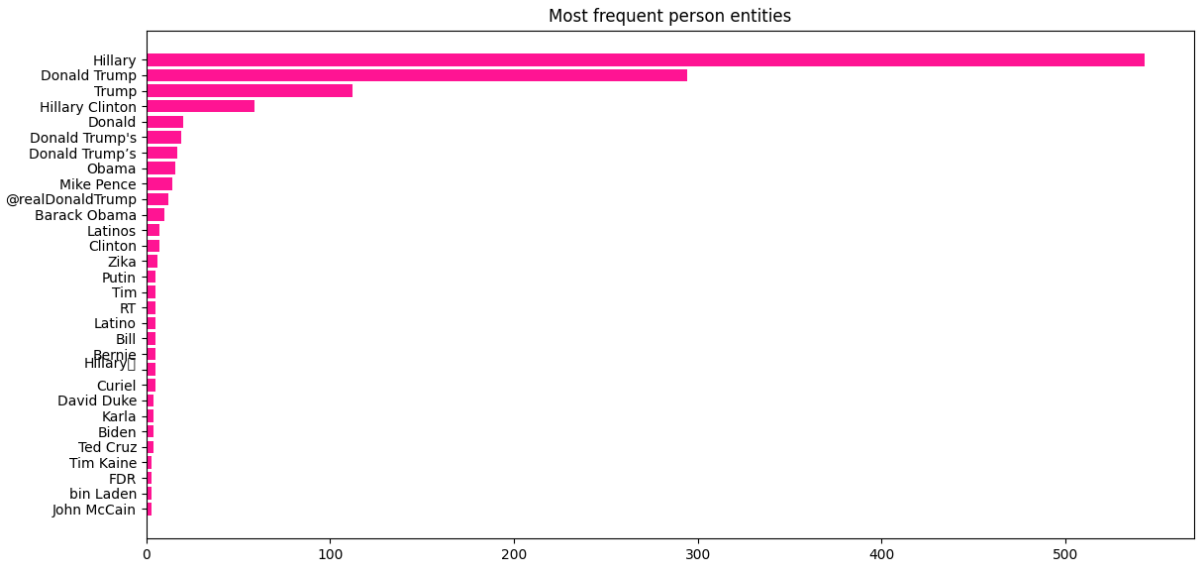
"Thank you to the LGBT community! I will fight for you while Hillary brings in more people that will threaten your freedoms and beliefs."

Takođe je interesantno da je većinu glasova LGBT zajednice odneo Donald Tramp te godine, iako je ovu zajednicu pomenuo svega jednom [15].



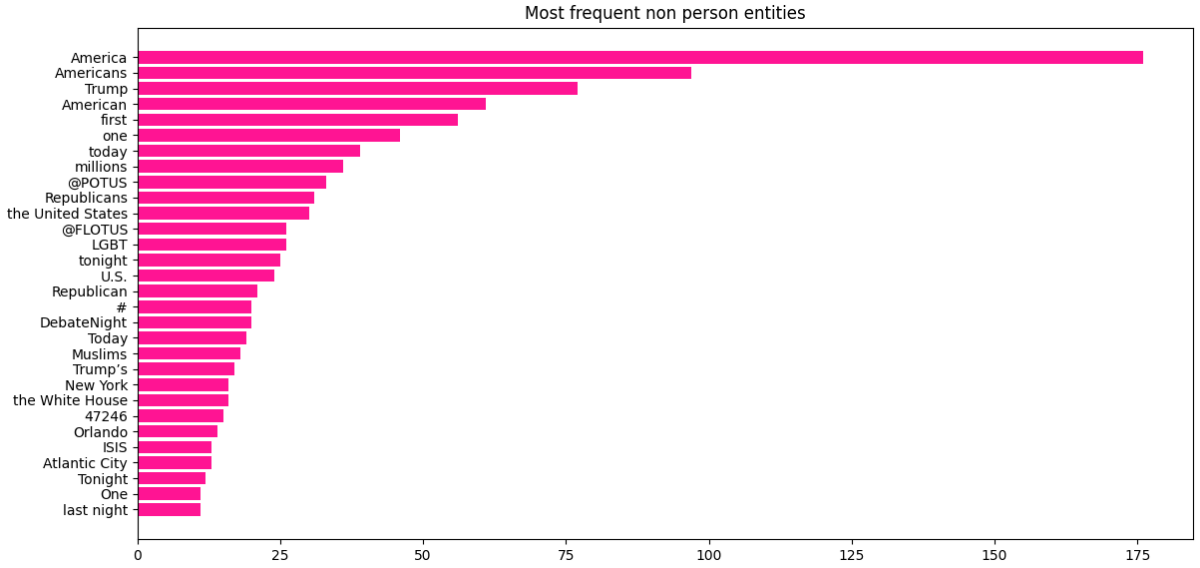
Slika 24. Najčešće korišćeni imenovani entiteti bilo koje kategorije u originalnim tvitovima Hilari Klinton

Na slici 25 se mogu videti najčešći imenovani entiteti kategorije PERSON, koji se pominju u originalnim tvitovima Hilari Klinton. Ovde se vidi da je ona najčešće sebe pominjala, zatim Donalda Trampa, Baraka Obamu, Majka Pensa, Latinoamerikance, Vladimira Putina, Teda Kruza, Bin Ladena, Džoa Bajdena, Frenklina Rozvelta (FDR) i Zika virus. Zanimljivo je to što je Hilari Klinton često govorila o Zika virusu (koji je u periodu 2015. i 2016. godine pogodio SAD) i o njemu napisala na desetinu tvitova, dok Donald Tramp ovaj virus nije nijednom spomenuo. Za Latinoamerikace je govorila da su važan deo društva i da ljudi trebaju da imaju više poštovanja prema njima. O Vladimiru Putinu je govorila u negativnom kontekstu. Takođe, zamerala je Donaldu Trampu zato što je hvalio predsednika Rusije. Nekoliko puta se i osvrtala na to kako su američke foke srušile Bin Ladena.



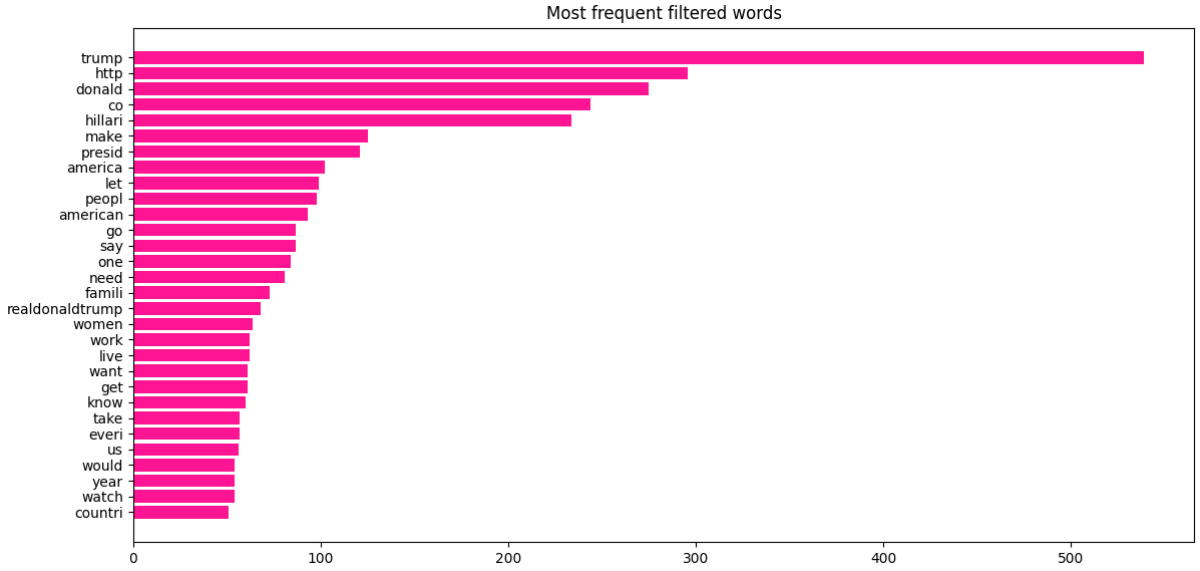
Slika 25. Najčešće korišćeni imenovani entiteti kategorije PERSON u originalnim tvitovima Hilari Klinton

Na slici 26 se mogu videti najčešći imenovani entiteti koji nisu iz kategorije PERSON, a pominju se u originalnim tvitovima Hilari Klinton. Ovde se vidi da je najčešće pominjala Ameriku, tadašnjeg predsednika (*POTUS*), republikance, tadašnju prvu damu (*FLOTUS*), LGBT zajednicu,debatu, Belu kuću i ISIS. Za ISIS je govorila kako će ih Amerika pobediti.



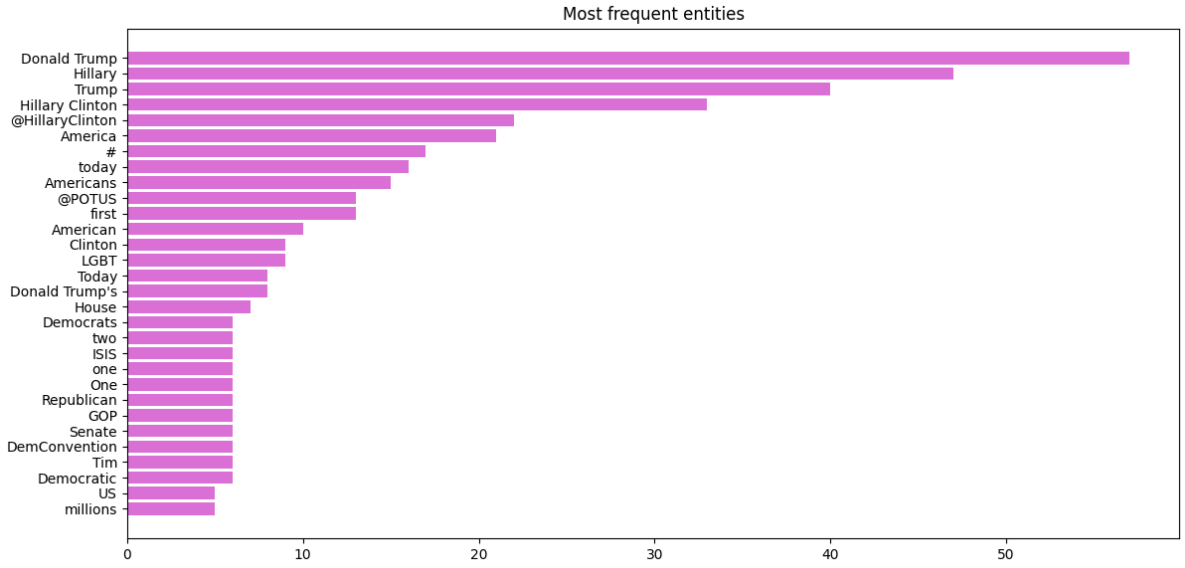
Slika 26. Najčešće korišćeni imenovani entiteti koji nisu kategorije PERSON u originalnim tvitovima Hilari Klinton

Na slici 27 se mogu videti teme originalnih tvitova Hilari Klinton. Najčešća tema joj je bio protivnik, Donald Tramp. Zatim je pominjala sebe, Ameriku, žene. Žene su bile česta tema Hilari Klinton. Ona je u mnogobrojnim tvitovima izražavala koliko su bitna prava žena i pričala koliko su žene moćne i jake. Ove teme su zastupljene i u tvitovima Donalda Trampa, ali ne u ovolikom broju, kao što je to slučaj sa tvitovima Hilari Klinton.



Slika 27. Najčešće teme originalnih tvitova Hilari Klinton

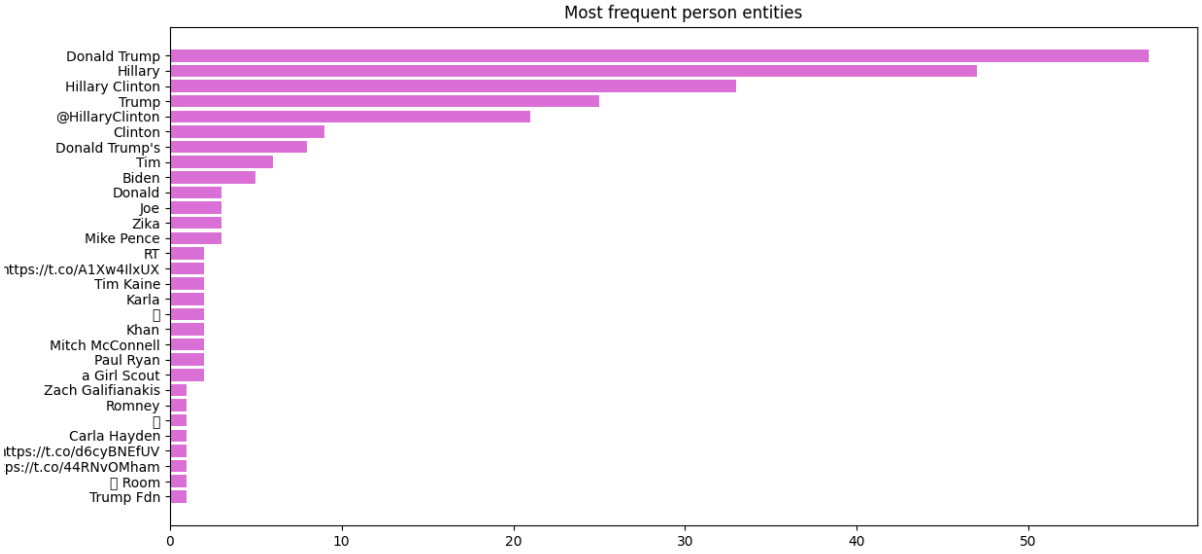
Na slici 28 se mogu videti najčešće pominjani imenovani entiteti, bilo koje kategorije, u retvitovima Hilari Klinton. Najčešće su pominjani Donald Tramp, Hilari Klinton, Barak Obama (POTUS), Amerika, LGBT zajednica, demokrate, senat i republikanci. Najčešće pominjani imenovani entiteti, bilo koje kategorije, retvitova Hilari Klinton su veoma slični kao i najčešće pominjani imenovani entiteti, bilo koje kategorije, njenih originalnih tvitova.



Slika 28. Najčešće korišćeni imenovani entiteti bilo koje kategorije u retvitovanim tvitovima Hilari Klinton

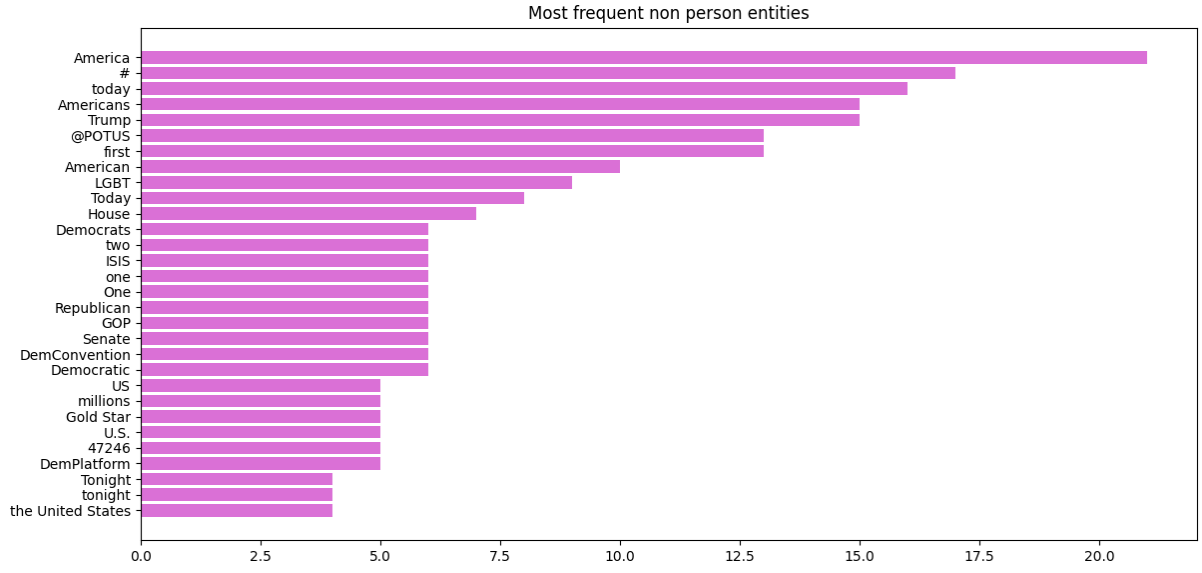
Na slici 29 se mogu videti najčešće pominjani imenovani entiteti kategorije PERSON u retvitovima Hilari Klinton. Najčešće su pominjani Donald Tramp, Hilari Klinton, Džo Bajden, Zika virus, Majk Pense, Tim Kejn i Karla Hajden. Na ovom grafikonu se može videti da su neki delovi tvitova pogrešno identifikovani. Može se videti da su neki linkovi

identifikovani kao imenovani entiteti kategorije PERSON. Ovi linkovi, zapravo, predstavljaju linkove do govora ili intervjua Hilari Klinton.



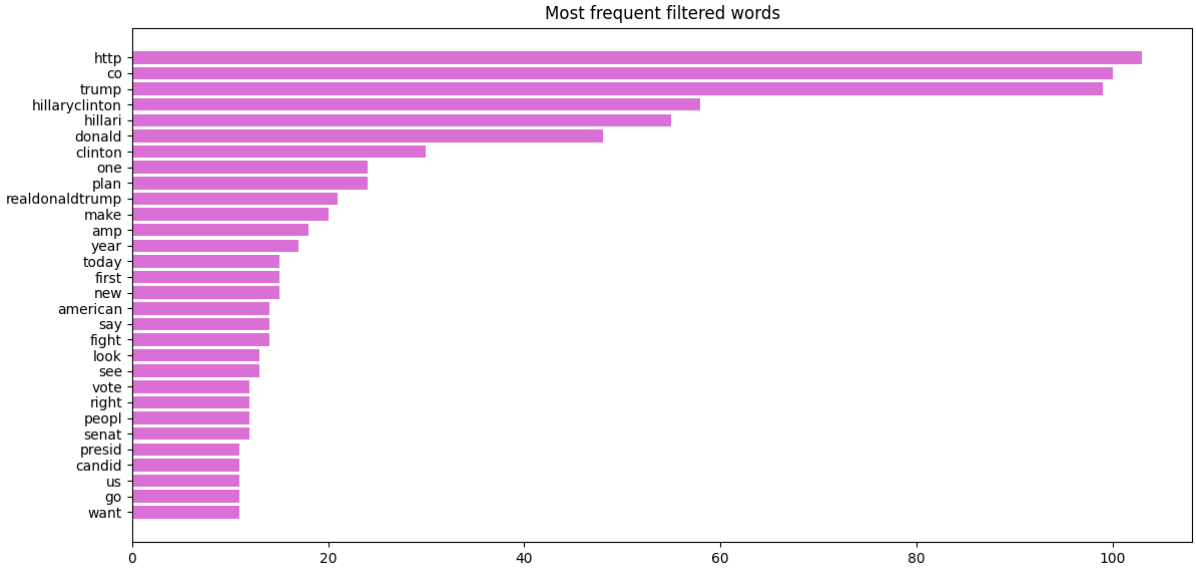
Slika 29. Najčešće korišćeni imenovani entiteti kategorije PERSON u retvitovanim tvitovima Hilari Klinton

Na slici 30 se mogu videti najčešće pominjani imenovani entiteti, koji ne pripadaju kategoriji PERSON, u retvitovima Hilari Klinton. Najčešće su pominjani Amerika, tadašnji predsednik (*POTUS*), LGBT zajednica, demokrate, kao i fondacija Gold Star. Gold Star je organizacija koju su osnovali članovi porodica, koje su svoje voljene izgubili u ratu u Iraku. Ovoj organizaciji je Hilari uputila na desetinu tvitova, zahvaljujući im za žrtve koje su pretrpele njihove porodice zarad slobode Amerike. Donald Tramp ovu organizaciju nije spomenuo nijednom.



Slika 30. Najčešće korišćeni imenovani entiteti koji nisu kategorije PERSON u retvitovanim tvitovima Hilari Klinton

Na slici 31 se mogu videti najčešće teme retvitova Hilari Klinton. Najčešće teme su Donald Tramp, Hilari Klinton, planovi, predsednik, kandidat, glasanje i senat.



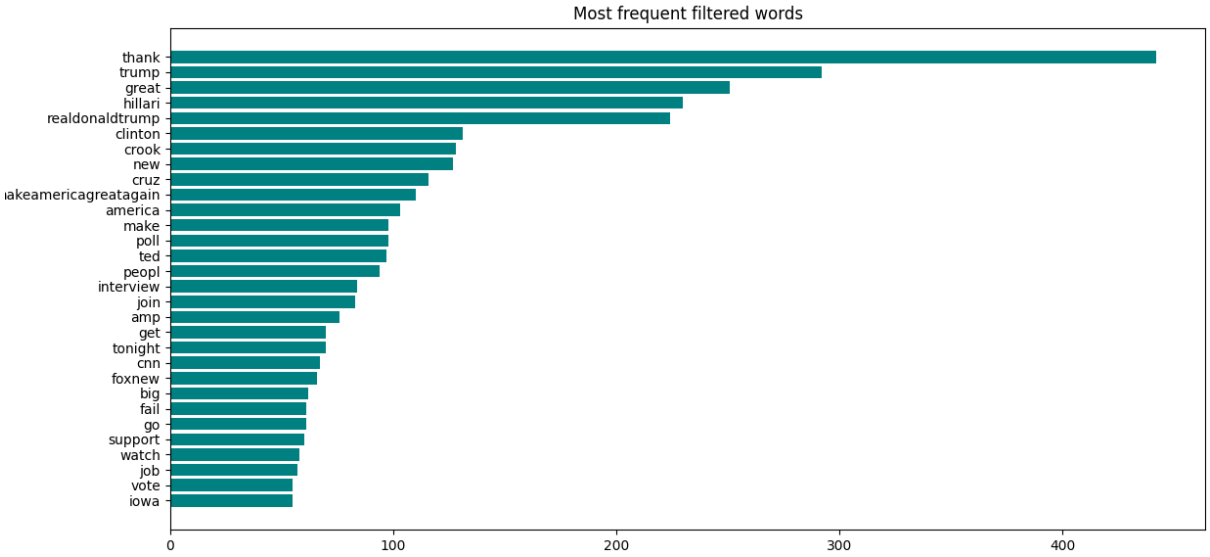
Slika 31. Najčešće teme retvitovanih tvitova Hilari Klinton

Na kraju analize se može videti da su imenovani entiteti, većinom, tačno prepoznati. U nekim slučajevima se vidi da su linkovi govora Hilari Klinton i Donalda Trampa greškom klasifikovani kao imenovani entiteti kategorije PERSON. Takođe, kao najčešće teme tvitova su detektovani “http” i “co”, koji predstavljaju delove linka.

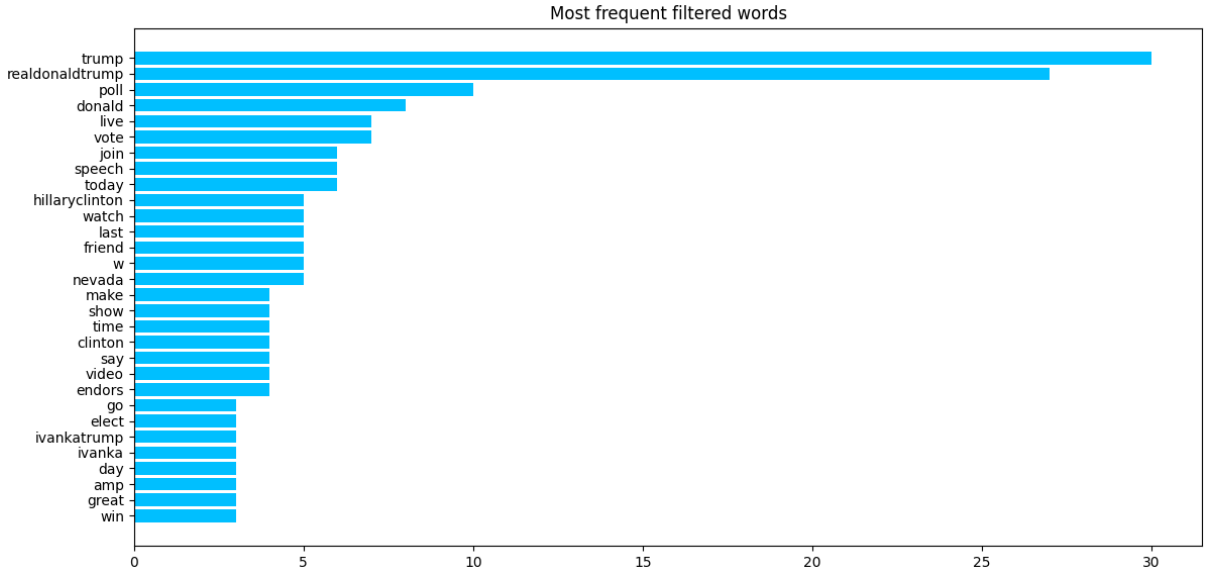
Ovakve greške, u detekciji tema, se mogu prevazići ako se u fazi pripreme i prečišćavanja podataka izbace svi linkovi iz teksta. Izbacivanje linkova se može realizovati korišćenjem regularnog izraza:

$$(https?:\/\/)?([\da\-z\.\-]+)\.([a\-z\.]\{2,6\})([\/\w\ \.\-]*) \quad [16]$$

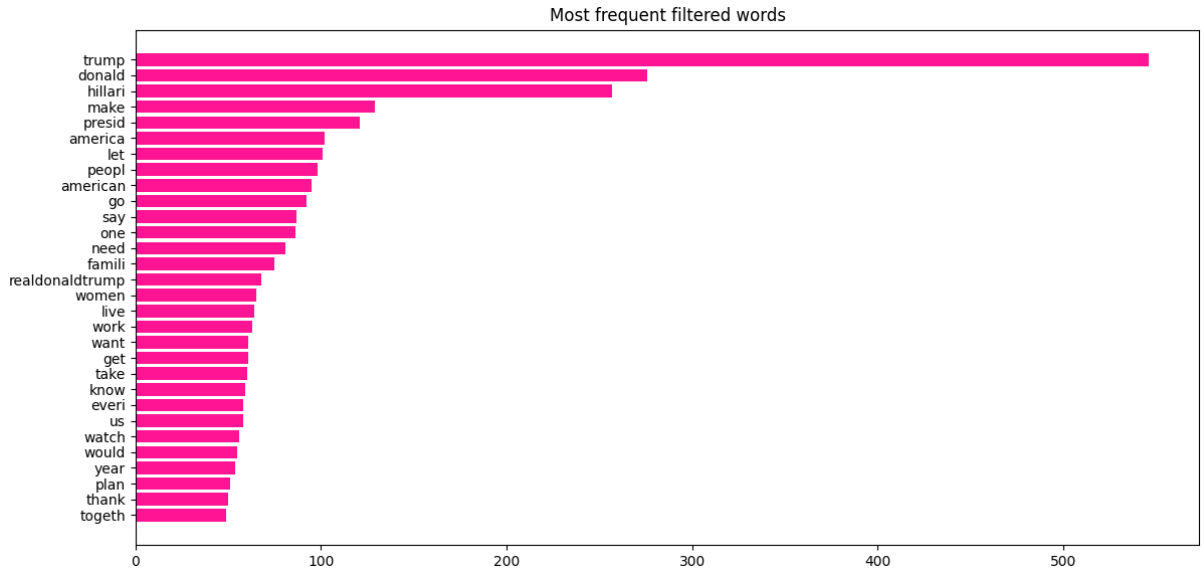
Nakon izbacivanja linkova, može se videti na slikama 32, 33, 34 i 35 da se “http” i “co” više ne pojavljuju kao teme u originalnim i retvitovanim tvitovima Hilari Klinton i Donalda Trampa.



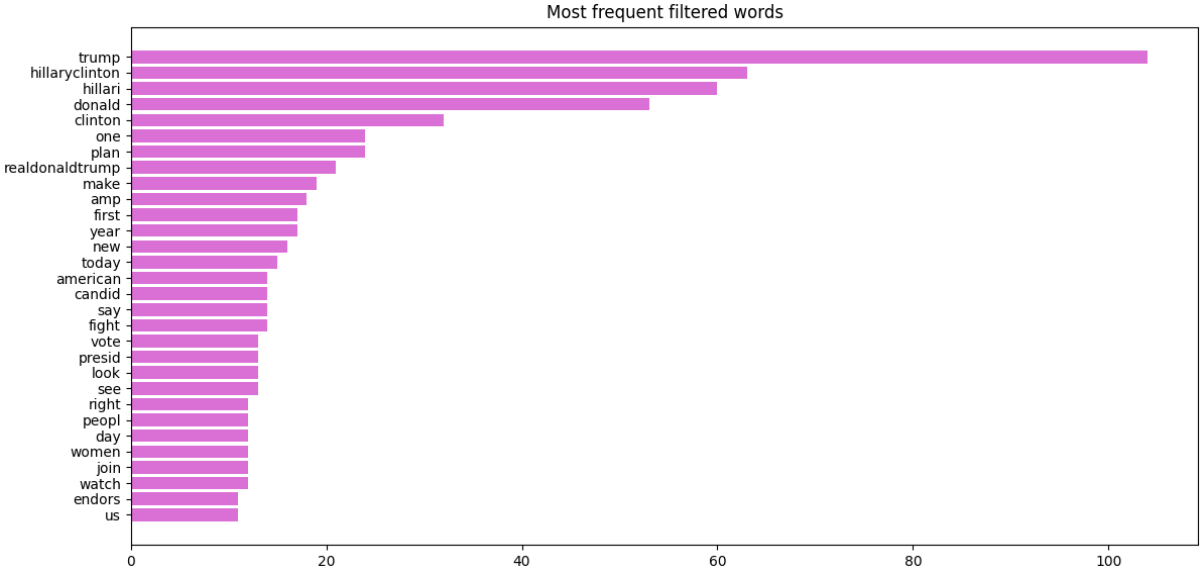
Slika 32. Najčešće teme originalnih tvitova Donalda Trampa, nakon izbacivanja linkova iz tvitova



Slika 33. Najčešće teme retvitovanih tvitova Donalda Trampa, nakon izbacivanja linkova iz tvitova

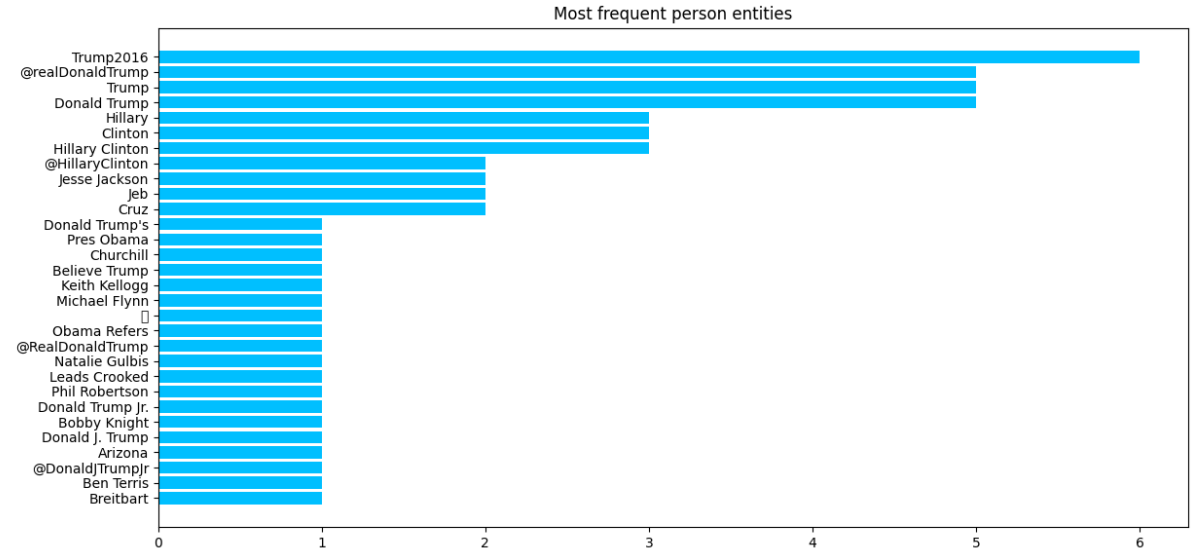


Slika 34. Najčešće teme originalnih tvitova Hilari Klinton, nakon izbacivanja linkova iz tvitova

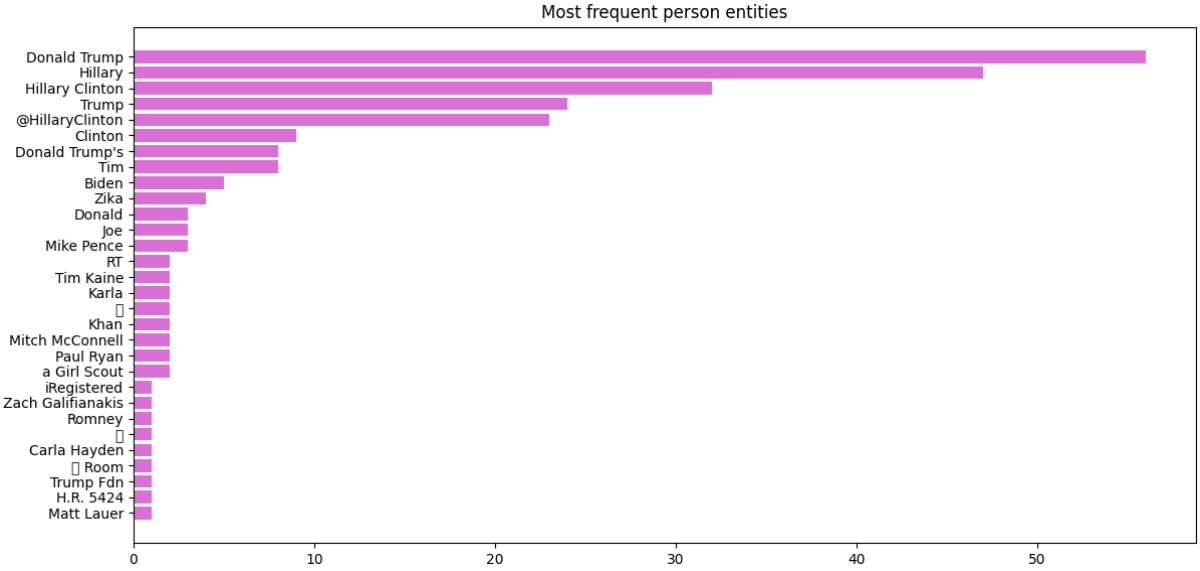


Slika 35. Najčešće teme retvitovanih tvitova Hilari Klinton, nakon izbacivanja linkova iz tvitova

Izbacivanje linkova treba uraditi i pre procesa prepoznavanja imenovanih entiteta, kako ne bi dolazilo do prepoznavanja linkova kao imenovanih entiteta tipa PERSON. Do ove greške je dolazilo prilikom prepoznavanja imenovanih entiteta u retvitovanim tvitovima Hilari Klinton i Donalda Trampa. Nakon izbacivanja linkova, na slikama 36 i 37, se može videti da više ne dolazi do pogrešnog prepoznavanja.

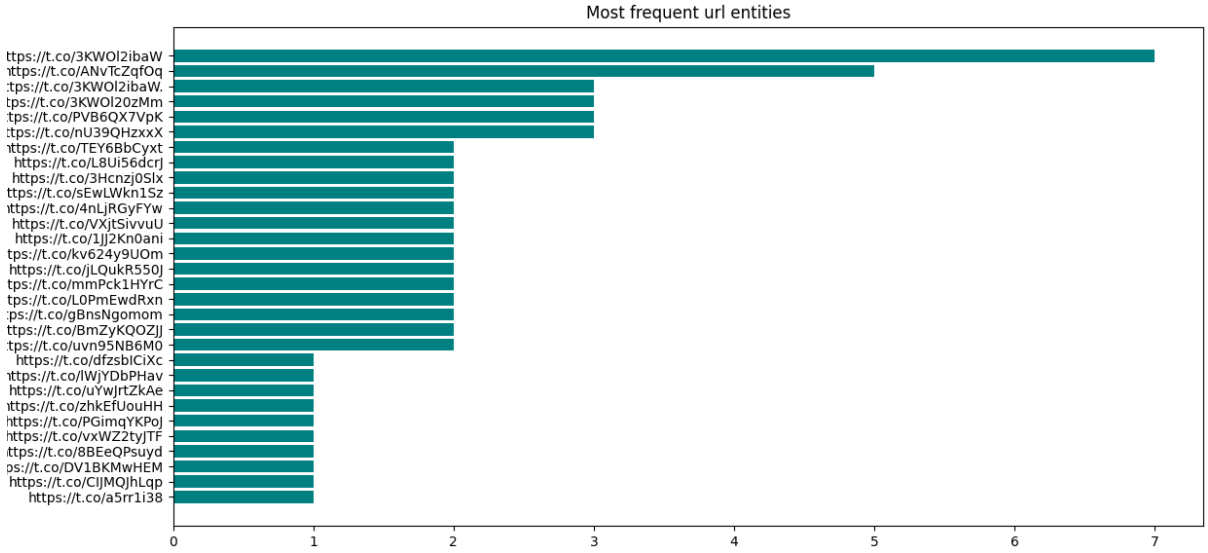


Slika 36. Najčešće korišćeni imenovani entiteti kategorije PERSON u retvitovanim tvitovima Donalda Trampa, nakon izbacivanja linkova

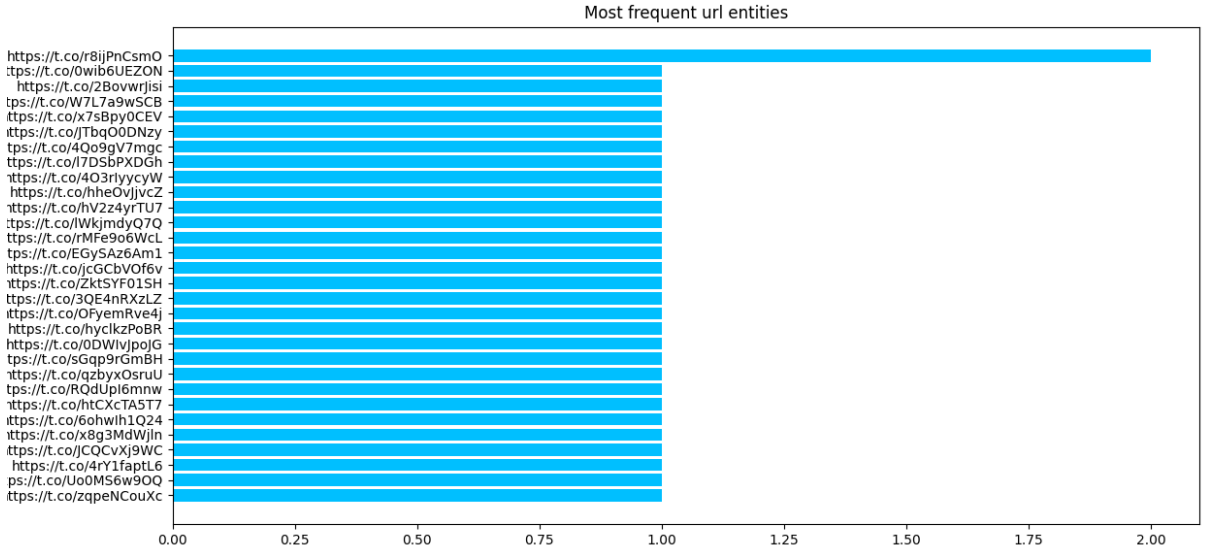


Slika 37. Najčešće korišćeni imenovani entiteti kategorije PERSON u retvitovanim tvitovima Hilari Klinton, nakon izbacivanja linkova

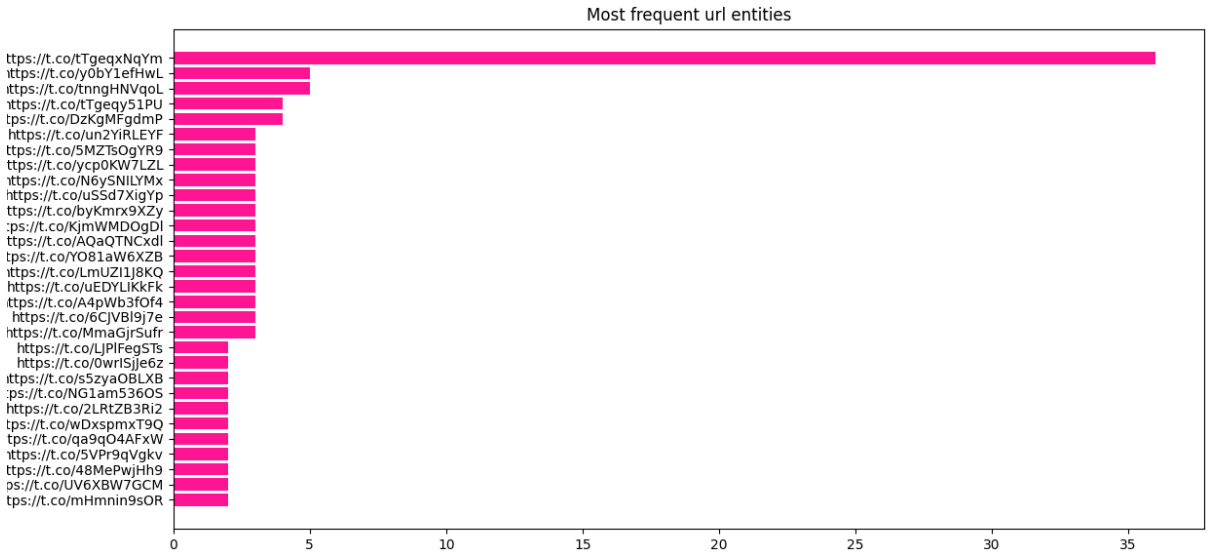
Drugi način, na koji se može rešiti problem pogrešnog prepoznavanja linkova kao imenovanih entiteta, jeste posebno klasifikovanje linkova. Nad svakim tokenom u listi prepoznatih imenovanih entiteta se treba izvršiti funkcija *like_url*. Ova funkcija vraća vrednost *True* ako token ima format URL-a, a *False* ako nema. Tokene koji imaju format URL-a treba izdvojiti u posebnu kategoriju. Zatim se mogu vizuelizovati i najčešće korišćeni linkovi, kao što je to prikazano na slikama 38, 39, 40, 41.



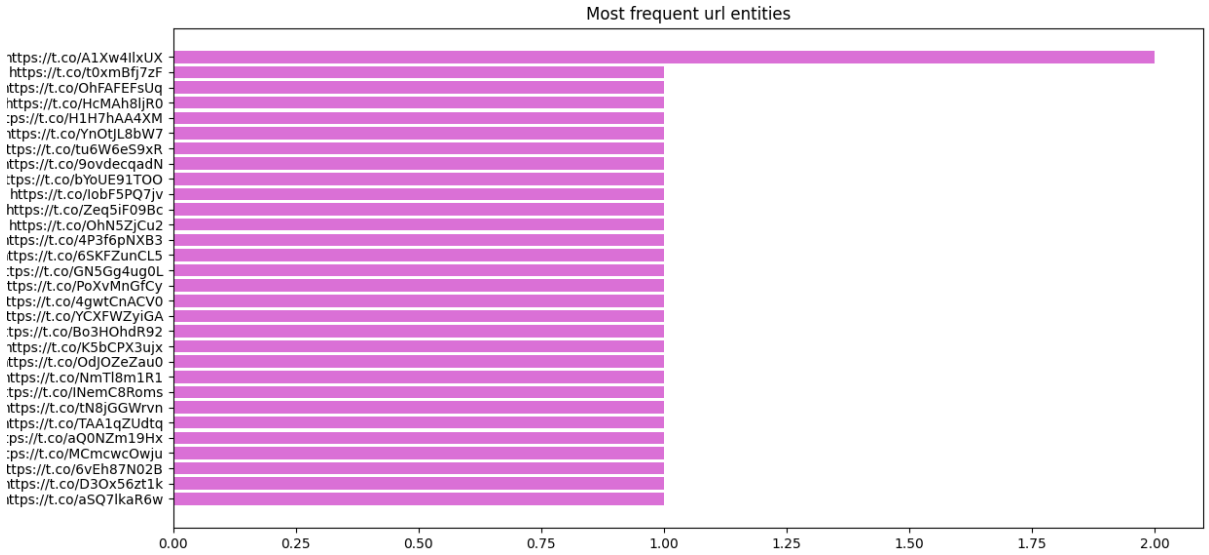
Slika 38. Najčešće korišćeni linkovi u originalnim tvitovima Donalda Trampa



Slika 39. Najčešće korišćeni linkovi u retvitovanim tvitovima Donalda Trampa



Slika 40. Najčešće korišćeni linkovi u originalnim tvitovima Hilari Klinton



Slika 41. Najčešće korišćeni linkovi u retvitovanim tvitovima Hilari Klinton

Sa vizualizacija najčešće korišćenih linkova se može videti da je Hilari Klinton u svojim originalnim tvitovima (slika 40) jedan link postavila čak više od 35 puta. Taj link vodi do sajta https://iwillvote.com/ , koji je namenjen glasanju.

Iz ove analize se može, takođe, zaključiti da ova dva predsednička kandidata dosta različito iskazuju svoja mišljenja. Donald Tramp je više govorio o svojim političkim protivnicima, dok je Hilari više govorila o svojim političkim istomišljenicima. Ova analiza pomaže u donošenju neki zaključaka. Na primer, iz ove analize bi se zaključilo da je Hilari Klinton osvojila većinu glasova LGBT zajednice, što je u realnosti drugačije. Donald Tramp je sa samo jednim pominjanjem ove zajednice, osvojio većinu glasova. Iz ove analize bi se, takođe, moglo zaključiti da je Hilari Klinton dobila većinu glasova žena jer se ona dosta više pričala o pravima žena i ohrabrivala ih. U realnosti, Hilari Klinton jeste osvojila većinu

ženskih glasova. Ove analize su zanimljive jer se iz njih mogu prepoznati određeni stavovi Hilari Klinton i Donalda Trampa, bez čitanja svih 8448 tvitova. Ovo ukazuje na to koliko je proces analize teksta važan i kolika je ušteda vremena prilikom njegovog korišćenja.

4. ZAKLJUČAK

Zbog sve većeg broja tekstualnih podataka, manuelno analiziranje je postalo veoma dug i zamoran proces, sklon greškama. Automatizacijom ovog procesa se brže i lakše obrađuju podaci, što omogućava kompanijama da izvlače iz tekstualnih podataka samo važne informacije i tako prate zahteve i ocene klijenata. Proces detekcije tema i imenovanih entiteta je važan proces prilikom obrade teksualnih podataka. Tema teksta, kao i imenovani entiteti koji se u tekstualnim podacima pojavljuju mogu mnogo toga reći kompanijama. Detekciju tema i prepoznavanje imenovanih entiteta ne moraju koristiti samo kompanije, već bilo koja osoba koja želi da iz teksta izvuče korisne ili interesantne informacije.

Za detekciju tema i prepoznavanje imenovanih entiteta se mogu koristiti već gotove biblioteke, od kojih su najpoznatije spaCy i NLTK. Ove dve biblioteke imaju svoje prednosti i manje. Ako je cilj što bolje i preciznije prepoznati imenovane entitete u tekstu, onda je spaCy pogodnija biblioteka za korišćenje, s obziom da ima više kategorija za imenovane entitete, bolje performanse i preciznije prepoznaje imenovane entitete. Sa druge strane, za potrebe detekcije tema teksta, NLTK biblioteka je uvek odličan izbor jer koristi proste algoritme odličnih performansi.

U ovom procesu može doći i do pogrešnog klasifikovanja tema i imenovanih entiteta, što može dovesti do pogrešnog tumačenja rezultata. Donošenje zaključaka na osnovu pogrešnih tumačenja može imati ozbiljne posledice. Sa druge strane, ove analize omogućavaju nikad lakšu kategorizaciju dokumenata, praćenje trendova, otkrivanja skrivenih informacija, efikasniju analizu i poboljšan kvalitet istraživanja.

5. LITERATURA

[1] https://monkeylearn.com/text-mining/

[2] https://monkeylearn.com/topic-analysis/

[3] https://monkeylearn.com/blog/named-entity-recognition/

[4] https://www.kaggle.com/benhamner/clinton-trump-tweets/home

[5] https://en.wikipedia.org/wiki/SpaCy

[6] https://en.wikipedia.org/wiki/Natural_Language_Toolkit

[7] https://medium.com/@akankshamalhotra24/introduction-to-libraries-of-nlp-in-python-nltk-vs-spacy-42d7b2f128f2#:~:text=NLTK%20is%20a%20string%20processing,and%20sentences%20are%20objects%20themselves.

[8] https://pandas.pydata.org/about/index.html

[9] https://en.wikipedia.org/wiki/2016_United_States_presidential_election_in_Iowa

[10] https://en.wikipedia.org/wiki/2016_United_States_presidential_election_in_New_York

[11] https://en.wikipedia.org/wiki/2016_United_States_presidential_election_in_New_Hampshire

[12] https://en.wikipedia.org/wiki/2016_United_States_presidential_election_in_Ohio

[13] https://en.wikipedia.org/wiki/2016_United_States_presidential_election_in_Florida

[14] https://en.wikipedia.org/wiki/2016_United_States_presidential_election_in_South_Carolina

[15] https://www.statista.com/statistics/632026/voter-turnout-of-the-exit-polls-of-the-2016-elections-by-sexual-orientation/

[16] https://code.tutsplus.com/tutorials/8-regular-expressions-you-should-know--net-6149