Real or Not? NLP with Disaster Tweets

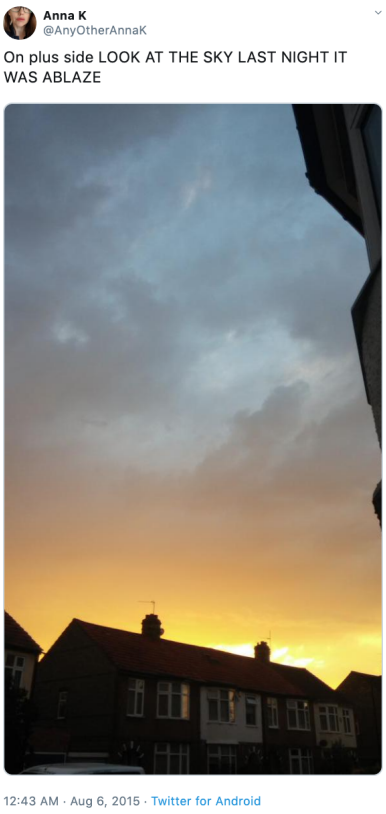
Dora Parmać, Bruno Fabulić, Nikola Vučković

18. lipnja 2020.

Sadržaj

1 Uvod

Twtitter, kao i mnoge društvene mreže, postao je jedan od glavnih komunikacijskih kanala. Temelji se na pisanju tweetova koji mogu biti dugački maksimalno 140 znakova na koje korisnici mogu odgovarati, proslijediti ih na svoj profil ili ih samo označiti oznakom sviđanja. Sadržaj tweetova kreće se od zabavnog, političkog do informativnog pa je jedan od najbržih načina saznavanja novih informacija. Sveprisutnost mobilnih telefona omogućuje ljudima da obavijeste druge u slučaju nesreće u danom trenutku, te ukažu na potencijalne opasnosti. No, nije uvijek sasvim jasno je li osoba izvještava o stvarnoj nesreći, ili je riječ o nečem drugom.



Autor ove objave eksplicitno koristi riječ "ablaze" što bi u doslovnom hrvatskom prijevodu značilo "u plamenu". U ovom slučaju, ključna riječ korištena je metaforički i svakoj osobi je jasno značenje ove objave. No, prilikom analize podataka i ključnih riječi, računalu ne mora biti jasno što je točno autor htio reći, pa nam ipak trebaju druge metode predikcije i analize.

Naš tim odabrao je ovaj projekt zato što nam se činio kao dobar uvod u NLP i zbog sveprisutnije tematike *fake news-a*. Najviše nas je zanimalo s kolikom točnošću možemo predviditi je li objavljeni tweet istinit ili lažan, tj. za dani tweet s kolikim postotkom možemo tvrditi da on izvještava o istinitoj nesreći i jesmo li na dobrom putu da stanemo na kraj *fake news-u*.

2 Opis problema

Podaci korišteni u ovom projektu preuzeti su s trenutnog *ongoing* natjecanja na *Kaggle-u*, poznatoj stranici za natjecanja u području *Machine learning-a*. Podaci se sastoje od 5 stupaca:

- **id**
- **keyword**
- **location**
- **text**
- **target**

Gdje nam *keyword* označava ključnu riječ u tweetu, *location* lokaciju, a *text* i *target* su nam tweet, tj. tekst u tweetu i zastavica je li dani tweet izvješatava o istinitoj nesreći.

Također uz podatke za treniranje imamo i podatke za testiranje na kojima će kasnije testirati naš dobiveni model.

Trenutno u podacima za treniranje imamo 7613 tweeta, dok u podacima za testiranje 3263 tweeta.

3 Algoritam

Random forest ili nasumične šume sastoje se od mnoštva stabala odluka. Broj stabala odluka u algoritmu postavljamo sa *n_estimators*. Svako pojedino stablo iz šume donosi svoju odluku. Zatim, kada su sva stabla donijela odluku, klasa koja ima najviše glasova postaje naša predikcija. Glavni koncept ove metode je znanje gomile. Također, predikcije ne ovise jedna o drugoj tako da i ukoliko pojedina stabla krivo predvide, većina će i dalje dati dobar odgovor.

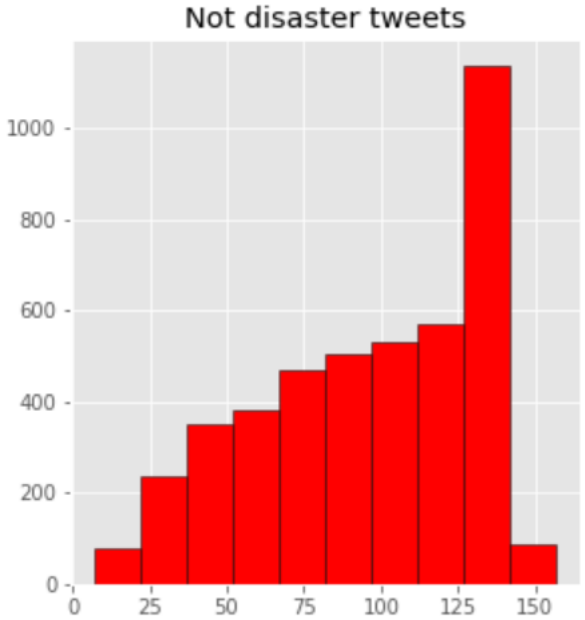
Metoda potpornih vektora (engl. Support Vector Machine, SVM) spada pod nadzirane metode učenja koji za dani set podataka predviđa kojoj klasi pojedini podatak pripada. Za razliku od drugih klasifikacijskih algoritama, metoda potpornih vektora traži najbolju razdvajajucu hiperravninu. Intuicija traže nja najudaljenije hiperravnine od samih klasa jest pojam generalizacije. Generalizacija je sposobnost modela da točno klasificira (ili predviđa ukoliko se radi o regresivnom problemu) svaku novu instancu nevidenu u skupu na kojem je model učen.

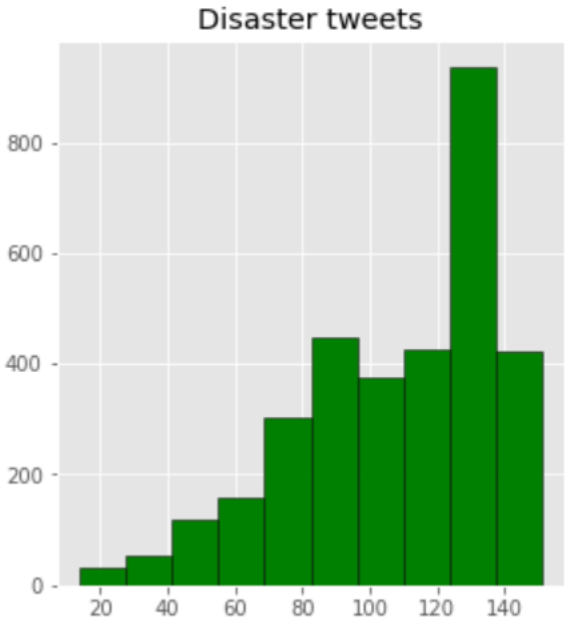
Također implementirali smo baseline model u kojem smo pretpostavili da su svi tweetovi lažni te smo proučavali koliko su naši modeli bolji od osnovne pretpostavke.

4 Analiza podataka

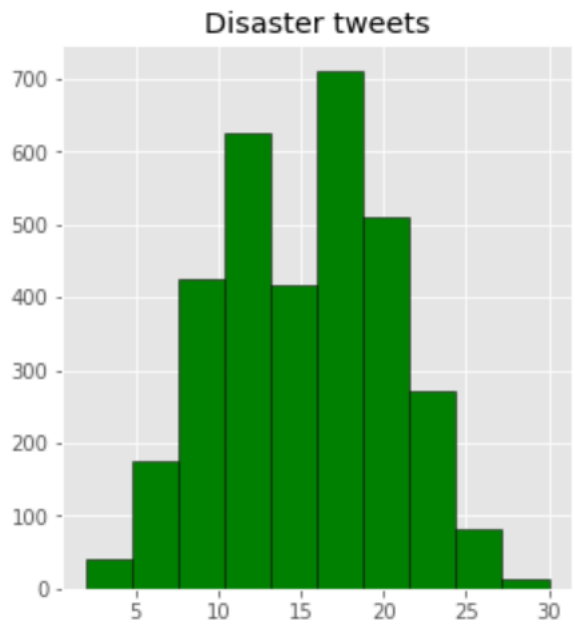
Podatke za treniranje i testiranje smo podjelili u dvije kategorije: *Not disaster tweets* i *Disaster tweets*, ovisno o tome je li dani tweet istinit ili lažan.

Prvu kategoriju koju smo proučavali je broj slova u tweetu. Očekivali smo da će tweetovi u *Disaster tweets* biti tweetovi s više slova, tj. da su tweetovi koji izvještavaju o nesreći napisani detaljnije i opširnije, npr. tweetovi od medijskih kuća.



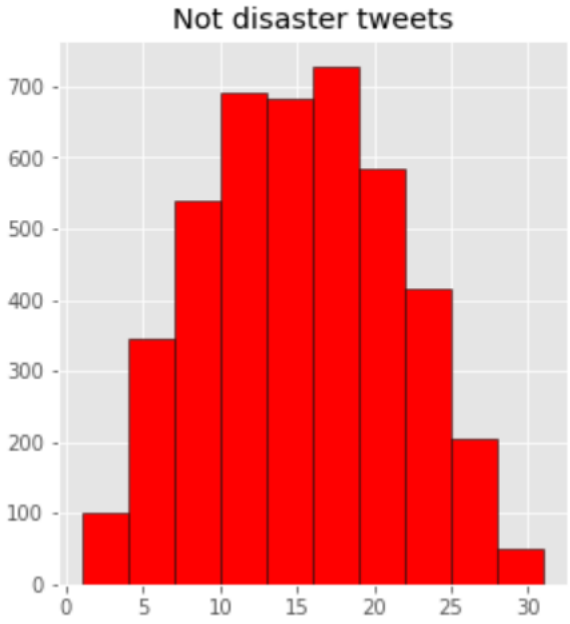


Slika 1: Broj slova u *tweet-u*



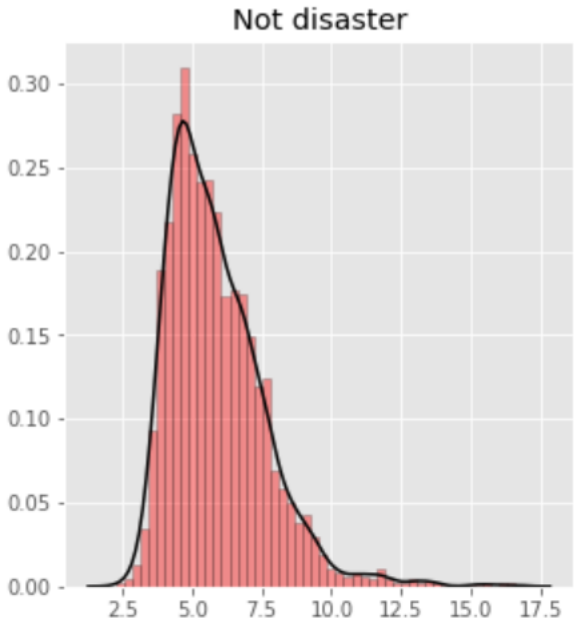
Slika 2: Broj riječi u *tweet-u*

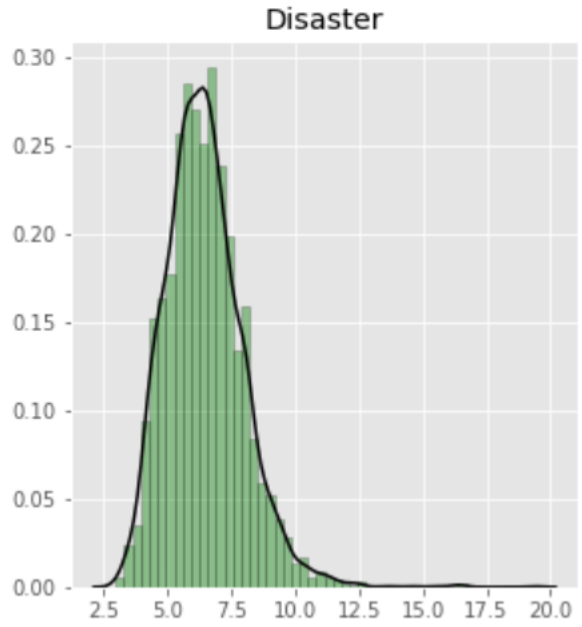
Iz grafova smo zaključili da se distribucije razlikuju kod *Not disaster tweets* i *Disaster tweets* te da će nam ovaj *feature* biti koristan.



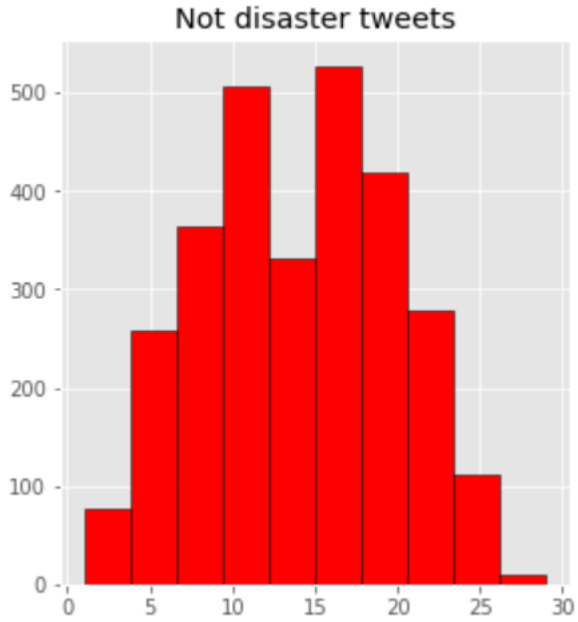
Kod pučavanja broja riječi u *tweetovima* analognim razmišljanjem pretpostavili smo da će broj riječi u *Disaster tweets* biti veći nego u *Not disaster tweets* te smo pomoću grafičkog prikaza odlučili uzeti ovaj *feature*.

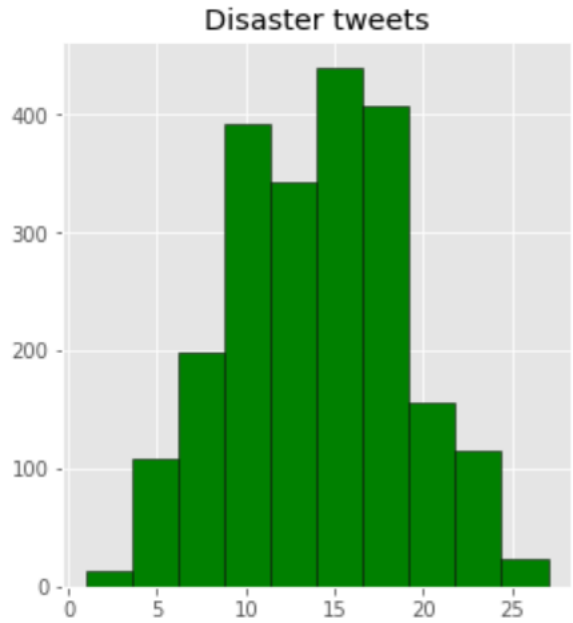
Iz grafova za prosječnu duljinu riječi *tweet-ova* vidimo da su im distribucije slične. Zato smo odlučili da ovaj *feature* nećemo koristiti u našem modelu.





Slika 3: Prosjecna duljina rijeci u *tweet-u*





Slika 4: Broj jedinstvenih riječi *tweet-u*

Detaljniju analizu moguće je vidjeti u predanoj bilježnici gdje smo još analizirali sljedeće *feature*: najčešće korištene riječi, broj ponavljanja interunkcijskih simbola te najčešće korišteni *bigram-i* (par susjednih riječi).

Na kraju smo za naš model odabrali sljedeće feature:

- Broj slova u tweet-u
- Broj riječi u tweet-u
- Broj *stop* riječi
- Broj jedinstvenih riječi
- Broj interpunkcijskih simbola
- Keywords

5 Čišćenje podataka

Također detaljnijom analizom podataka utvrdili smo da tweetovi sadrže previše neiskoristivnog teksta. U tu svrhu napisali smo funkcije za čišćenje tweetova. Koristili smo funkcije koje su iz tweetova izbacile: korištenja *emoji* emotikona i linkove za vanjske poveznice.

6 Rezultati učenja

Za rezultate učenja proučavali smo sljedeće dobivane podatke:

- *accuracy* - $\frac{\text{broj pogođenih}}{\text{ukupni broj podataka}}$
- *precision* - $\frac{\text{true positive}}{\text{true positive + false positive}}$
- *recall* - $\frac{\text{true positive}}{\text{true positive + false negative}}$

Za *baseline* model dobili smo *accuracy* = 0.570334, dok su *precision* i *recall* jednaki 0 jer smo pretpostavili da su svi tweetovi lažni.

6.1 Random forest

Korištenjem random forest algoritma dobili smo sljedeća mjerenja:

- *accuracy* - 0.7346
- *precision* - 0.783899
- *recall* - 0.527817

Vidimo da nam je točnost algoritma za čak 16.426% veća od baseline modela.

6.2 SVM

Korištenjem *SVM* algoritma dobili smo sljedeća mjerenja:

- *accuracy* - 0.693227
- *precision* - 0.662612
- *recall* - 0.582739

Vidimo da nam je točnost algoritma za 12.289% veća od baseline modela.

7 Zaključak i problemi

Iako smo dobili zadovoljavajući postotak korištenjem i *random forest* i *SVM* algoritama smatramo da postignute vrijednosti i dalje nisu dovoljno visoke da sa sigurnosću možemo tvrditi je li dani tweet istinit ili lažan. Nažalost, borba protiv *fake news-a* se nastavlja.

Smatramo da je najveći razlog još uvijek nedovoljno visokog postotka prognoziranja taj što su tweetovi u svim kategorijama bili relativno slični

te su tekstovi gramatički i semantički netočni. Kvalitetnijim čišćenjem teksta smatramo da bi se rezultati mogli unaprijediti, ali ipak smatramo da razlika ne bi bila prevelika. Isprobavanjem *KNN* ili *neuronskih mreža* bi se također postotak predviđanja mogao poboljšati.