**ANALIZA I PREDIKCIJA ISTINITOSTI TWITTER OBJAVA**

1. **Opis problema**

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.

Pogledajmo primjer:

|  |  |
| --- | --- |
| tweet_screenshot.png | 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. |

1. **Cilj I hipoteze rješavanja problema**
   1. **Postojeće metode rješavanja**

* **Random forests -** jedna od metoda nadziranog učenja pogodna za rješavanje problema klasifikacije i aproksimacije funkcije. Prilikom učenja, gradi se mnogo stabala odluke na različitim podskupovima skupa za učenje. Predikcija ciljne varijable se računa kao model razreda kod klasifikacije, odnosno srednja vrijednost kod regresije, od dobivenih predikcija pojedinačnih stabala. Algoritam kasnije predviđa vrijednost odgovora s obzirom a novo uneseni prediktor.
* **Metoda potpornih vektora (SVM) –** ako pretpostavimo da je dan skup od n-dimenzionalnih podataka takvih da jedan dio podataka pripada jednoj klasi, a drugi drugoj, cilj je da od svih mogućih (n-1)-dimenzionalnih ravnina između dvije klase, odaberemo onu za koju je udaljenost do najbližeg primjera (podatka) sa svake strane maksimalna.
* **K-nn algoritam (k-najbližih susjeda)–** računa udaljenost između novog primjera I svih primjera iz skupa za učenje, odredi k-najbližih susjeda I pridjeli novom primjeru klasu koja je najčešća između k-najbližih susjeda.

**Nedostatatci**:

* RF: slabiji rezultati kada je veći broj kategorija za klasifikaciju što uzrokuje sporije vrijeme izvršavanja,
* SVM: neprikladan za velike skupove podataka. Ukoliko broj uzoraka prijeđe broj trening uzoraka, neće dati dobre rezultate. Nema probabilističkog objašnjenja za klasifikaciju.
* K-nn: ukoliko imamo veliki dataset, efikasnost i brzina se smanjuje. Zahtijeva homogenost podataka, optimalan broj susjeda i radi loše predikcije ukoliko dođe do disbalansa podataka.
  1. **Materijali, metodologija i plan istraživanja**

Naš zadatak je napraviti model koji analizira Twitter objave i može razlučiti koje su od njih stvarne nesreće, a koje ne.

Za pristup problemu, koristimo metodu obrade prirodnog jezika, potpodručje računarske znanosti, obrade informacija I umjetne inteligencije koje proučava interakciju između računala I ljudskih (prirodnih) jezika, posebno na načine kako procesirati I analizirati velike količine podataka iz prirodnog jezika pomoću računala.

Glavne datoteke koje će nam trebati su:

* Train.csv
* Test.csv
* Sample\_submission.csv

Svaki uzorak u train I test datoteci sadrži slijedeće informacije:

* text - sadržaj Twitter objave ili tweeta
* keyword - ključnu riječ (koja može biti prazna)
* location - lokaciju s koje je tweet poslan (također može biti prazna)
* id za svaku objavu
* target – 1 ako je stvarna nesreća, 0 ako nije

Ukoliko je analiza objave uspješna, tj. radi se o stvarnoj nesreći, program bi trebao vratiti 1, a u suprotnom 0.

Kao skup podataka koristimo dataset od 7,000 klasificiranih Twtitter objava.

Predviđeno je koristiti razne pythonove pakete kao npr.:

* numpy (Numerical Python) - linearnom algebrom,
* pandas (Python Data Analysis Library) - rad sa csv. datotetkama, analiza I organizacija podataka
* nltk (Natural Language Processing) - lingvistička analiza, kategorizacija teksta, morfološku analizu, čišćenje teksta od tzv. Stop riječi (primjer u engleskom su određeni I neodređeni članovi poput “the”, “a”, “an”, “in”…)
* re (Regular Expression) – niz znakova koji opisuje ruge nizove znakova u skladu s određenim sintaksnim pravilima
* scikit-learn (ekstenzija scipyja – Scientific Python) – machine learning biblioteka s algoritmima za klasifikaciju, regresiju (nadzirano učenje), grupiranje (ili clustering; nenadzirano učenje)

Ocjena uspješnosti rezultata projekta je ukupan broj pogođenih predikcija stvarnih nesreća u test dijelu za svaki od modela.

Cilj nam je usporediti model dobiven primjenom Random Forest te K-nn algoritma. Razlog za odabir ta dva modela je što je SVM neprikladan za velike količine podataka, dok velika količina podataka K-nn samo usporava. Očekujemo da metoda Random forest bude bolja jer imamo samo dvije kategorije za klasifikaciju te su nam podaci za train blago neuravnoteženi, što je mana za K-nn algoritam.

1. **Popis literature:**

* <https://www.kaggle.com/c/nlp-getting-started/overview>
* Strojno učenje, PMF Zagreb