Obrada prirodnog jezika – tweet-ovi katastrofa

Juraj Brigljević i Ana Marija Zebić

1. Uvodni opis problema

**1.1 Koji problem pokušavamo riješiti?**

Problem kojim se mi bavimo jest obrada prirodnog jezika (*eng. natural language processing*) u tweetovima o katastrofama. Zadatak je zadan na platformi [*Kaggle*](https://www.kaggle.com/competitions/nlp-getting-started/overview). Ponuđeno je desetak tisuća tweetova s oznakama ovisno o tome jesu li govorili o katastrofalnim događajima (npr. požar, potres, teroristički napad) ili ne.

**1.2 Kojim skupom podataka se koristimo?**

Kako je *Kaggle* platforma koja služi za dijeljenje raznih problema strojnog učenja, svaki od njih sadrži i pripremljen skup podataka. U našem slučaju, taj skup je već djelomično obrađen pa je svakom tweetu pridruženo još par parametara. Njih smo detaljnije proučili u bilježnici s eksploratornom analizom.

2. Cilj i hipoteze istraživanja problema

Na *Kaggleu* je moguća objava rješenja i pritom se ona uspoređuju ovisno o uspješnosti klasifikacije. Nažalost, ovaj problem je vječno dostupan i u potpunosti javan pa ne postoji vjerodostojna ljestvica s najboljim rezultatima. Srećom, ljudi često objavljuju svoja rješenja i uspješnost na stranici, pa smo uspjeli pronaći nekolicinu od njih. Promatrajući najviše rangirane, s obzirom na korisničko glasanje, rezultati se kreću oko 80% uspješnosti. Nama će cilj biti pokušati osvojiti što višu ocjenu, nadamo se iznad 70%. Također, pokušat ćemo proučiti i kompleksnije modele iz područja dubokog učenja koji se specifično koriste u svrhe proučavanja NLP problema.

3. Pregled dosadašnjih istraživanja

**3.1 Koje su postojeće metode kojim se problem rješavao i koji su nedostaci postojećih metoda?**

4. Materijali, metodologija i plan istraživanja

**4.1 Na koji način (kojim pristupom) ćete probati riješiti problem?**

Za početak, problemu ćemo pristupiti na jednostavan način, osnovnim algoritmima za učenje pod nadzorom. Ovisno o njihovoj uspješnosti namjeravamo pokušati poboljšati njihove rezultate korištenjem kompliciranijih metoda, npr. ansambla. Osim toga pokušat ćemo proučiti metode iz područja dubokog učenja koje se prvenstveno koriste za NLP probleme, transformere.

**4.2 Kako ćete prikupiti podatke?**

Kao što smo već napomenuli, podaci su priloženi uz sam problem.

**4.3 Kojim metodama/algoritmima/tehnikama/alatima se mislite koristiti?**

Na temelju osnovne analize podataka koju smo proveli, pokušat ćemo napraviti neki model u području algoritama najbližih susjeda. Njega ćemo pokušati poboljšati korištenjem nekih metoda iz ansambla. Osim toga proučit ćemo i XGBoost biblioteku s obzirom na to da ima vrlo visoku uspješnost na problemima s *Kagglea*. Što se tiče transformera, vjerojatno ćemo se pozabaviti jednim od mogućih modela, npr. BERT, XLNET ili LSTM.

**4.4 Kako mislite ocijeniti uspješnost rezultata svoga projekta (interpretacija)?**

Uspješnost našeg projekta ovisi o tome u kolikoj mjeri će naša rješenja dobro označavati testne primjerke u ponuđenom skupu podataka. Cilj nam je bar s jednim od njih se približiti ostalim visoko rangiranim rješenjima. Osim toga, želimo usporediti neke od jednostavnijih i općenitih modela s ovim posebnijima i specifično razvijenih za NLP probleme. Zanima nas daju li osnovni modeli zadovoljavajuća rješenja i do koje mjere kompleksniji i posebno razvijeni modeli uspijevaju poboljšati, ako uopće, te rezultate.

5. Očekivani rezultati predloženog projekta

**5.1 Što očekujete predati kao konačni rezultat projekta?**

6. Popis literature

[1] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second edition

[2] XGBoost: https://xgboost.readthedocs.io/en/stable/index.html