Klasifikacija lažnih vesti

Mateja Jovanović RA 160-2021

05.07.2024.

Definicija problema

U doba interneta, brzo širenje informacija donosi i problem lažnih vesti, koje mogu manipulisati javnošću i uticati na važne događaje poput izbora. Cilj ovog projekta je razviti alat koji će pomoću tehnika mašinskog učenja efikasno klasifikovati vesti kao "istinite" ili "lažne" i time identifikovati potencijalno štetan sadržaj pre nego što postane široko rasprostranjen.

Skup podataka

Za potrebe projekta se koristitio Fake News dataset sa Kaggle-a. Train.csv datoteka obuhvata oko 20.000 redova, sa približno jednakim brojem vesti označenih kao lažne i istinite. Dataset je na engleskom jeziku. Svaki red ima:

- ***id*** – jedinstveni identifikator svakog članka u datasetu
- ***title*** – naslov članka
- ***text*** – tekst samog članka, atribut koji će biti korišćen za treniranje modela
- ***author*** – autor teksta
- **label** – ciljno obeležje koje označava kategoriju članka
  - **1** (nepouzdan) - oznaka za članke koji su identifikovani kao potencijalno nepouzdani
  - **0** (pouzdan) - oznaka za članke koji su identifikovani kao pouzdani

Pristup rešavanja problema

Obučiti i evaluirati sledeće modele:

1. **Naive Bayes**
2. **TinyBERT**
3. **DistilBERT**

Evaluacija se vrši na test skupu istog dataseta, kao i na jednom potpuno drugom datasetu sa relativno sličnom tematikom.

Posmatramo metrike: *precision, recall, f1-score, accuracy*

Naive Bayes

Naivni Bajes je postigao dobre rezultate na testnom skupu originalnog dataseta na kojem je treniran.

Naive Bayes - test set			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.87	0.92	0.89
Fake	0.91	0.86	0.88
Accuracy	0.89		

Naive Bayes with preprocessing on test set			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.86	0.92	0.89
Fake	0.91	0.86	0.88
Accuracy	0.89		

Naive Bayes

Kada je testiran na drugom, sličnom datasetu, pokazao je solidnu sposobnost generalizacije.

Naive Bayes - different dataset			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.63	0.82	0.71
Fake	0.77	0.57	0.66
Accuracy	0.69		

Naive Bayes preprocessing - different dataset			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.62	0.83	0.71
Fake	0.78	0.54	0.66
Accuracy	0.68		

DistilBERT

Iako je DistilBERT postigao odlične rezultate na testnom skupu dataseta na kojem je treniran, testiranje na drugom, sličnom datasetu otkrilo je slabiju generalizaciju.

DistilBERT – test set			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.99	0.99	0.99
Fake	0.99	0.99	0.99
Accuracy	0.99		

DistilBERT – different dataset			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.62	0.04	0.07
Fake	0.53	0.98	0.69
Accuracy	0.53		

TinyBERT

Iako je TinyBERT postigao izuzetno dobre rezultate na testnom skupu dataseta na kojem je treniran, testiranje na drugom, relativno sličnom datasetu pokazalo je ne tako dobru generalizaciju.

TinyBERT – test set			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.99	0.99	0.99
Fake	0.99	0.99	0.99
Accuracy	0.99		

TinyBERT – different dataset			
Class	**Precision**	**Recall**	**F1 - Score**
True	0.42	0.05	0.08
Fake	0.52	0.94	0.67
Accuracy	0.51		

Analiza grešaka transformera

Performanse transformera pokazuju da su ovi modeli osetljivi na specifične karakteristike trening seta. Kada su trenirani na drugom datasetu i testirani na prvom, rezultati su bili bolji, što sugeriše potrebu za raznovrsnijim trening podacima.

- **False Negatives:** Transformeri su grešili u klasifikaciji istinitih vesti kao lažnih zbog strukturiranih narativa i uravnoteženog tona.
- **False Positives:** Lažne vesti su pogrešno klasifikovane kao istinte zbog emocionalnog tona, političke pristrasnosti i senzacionalističkog jezika.

Dalji rad

- **Raznovrsniji datasetovi:** Prikupljanje i korišćenje većeg broja datasetova kako bi se poboljšala sposobnost generalizacije modela. Raznovrsni podaci će omogućiti modelima da bolje prepoznaju različite stilove pisanja i kontekstualne razlike između pouzdanih i nepouzdanih vesti.
- **Promena fokusa sa “lažnih“ na “nepouzdane” :** Kategorizacija vesti na pouzdane i nepouzdane omogućava širi spektar procene. Vest može biti nepouzdana zbog više faktora kao što su neproverene informacije, loš izvor, senzacionalistički ton ili pristrasan prikaz. Ovo omogućava finije granulisanje u klasifikaciji i pomaže korisnicima da bolje razumeju nivo pouzdanosti vesti.