Predlog projekta iz predmeta Računarska inteligencija

Student: Tina Aničić

Broj indeksa: SV71/2022

1. Naziv teme

Klasifikacija spam i ham emailova

2. Definicija problema

Cilj ovog projekta je razviti sistem koji može automatski razlikovati neželjene (spam) i redovne (ham) poruke. Spam emailovi predstavljaju problem jer zauzimaju prostor, mogu sadržati maliciozne sadržaje i ometaju korisnike. Sistem treba da omogući automatsku identifikaciju tipa poruke na osnovu njenog sadržaja, pri čemu je fokus na postizanju visoke tačnosti i otpornosti na nove, sofisticirane spam poruke.

3. Motivacija

Neželjena pošta je široko rasprostranjena i utiče na produktivnost korisnika. Efikasan filter može smanjiti rizik od phishing napada i malvera, poboljšati korisničko iskustvo i olakšati upravljanje poštom. Detekcija spam poruka je i dalje aktuelan istraživački problem, naročito u eri kada spam poruke postaju sve sofisticiranije i teže za razlikovanje od legitimnih emailova.

4. Skup podataka

Za rešavanje ovog problema koristiće se obimniji i složeniji skupovi podataka. Konkretno, planirano je korišćenje **Enron Email Dataset-a** (veliki skup stvarnih emailova na engleskom jeziku), kao i **SpamAssassin Public Corpus** (raznovrsne spam i ham poruke različite strukture i kompleksnosti). Ovi skupovi sadrže hiljade email poruka različite strukture i kompleksnosti, što omogućava realističnije i zahtevnije testiranje modela.

- Enron Email Dataset: Sadrži oko 500.000 poruka, od kojih je nakon čišćenja oko 35.000 označeno kao spam, a ostatak kao ham. Poruke variraju od kraćih (par rečenica) do veoma dugačkih poslovnih mejlova. Prosečna dužina poruke je oko 150 reči. Skup nije balansiran (spam čini oko 17%). Uočene su i nedostajuće vrednosti u metapodacima (naslov, adresa pošiljaoca), dok sam tekst uglavnom postoji.

- SpamAssassin Public Corpus: Sadrži oko 6.000 poruka, od čega oko 1.800 spama i 4.200 ham poruka. Prosečna dužina mejla je oko 120 reči. Skup takođe nije balansiran, a prisutne su i duplikatske poruke koje je potrebno ukloniti.

5. Način pretprocesiranja podataka

- Čišćenje podataka od interpunkcije, brojeva i specijalnih znakova.
- Pretvaranje teksta u mala slova.
- Tokenizacija i lematizacija reči, uklanjanje stop reči.
- Obrada specifičnih tokena u emailovima (linkovi, adrese, brojevi).
- Transformacija teksta u numerički oblik korišćenjem:
  - Bag-of-Words
  - TF-IDF
  - Pretreniranih embedding modela (Word2Vec, BERT tokenizer)
- Za transformer modele koristiće se odgovarajući tokenizatori i fine-tuning tehnike

6. Metodologija

Rad počinje analizom i čišćenjem podataka, nakon čega se skup deli na trening i test deo.
Zatim se treniraju različiti klasifikacioni modeli: klasični (Naive Bayes, Logistic Regression, SVM) i duboki modeli (Transformer arhitekture poput BERT-a ili DistilBERT-a).
Na kraju se porede rezultati i bira model sa najboljim performansama.
Poseban deo metodologije uključuje i analizu otpornosti modela na nove vrste spam poruka (outlier detection).

7. Način evaluacije

Rezultati modela će se evaluirati pomoću metrika kao što su tačnost (accuracy), preciznost, odziv (recall) i F1-score.
Poseban fokus biće na pravilnom prepoznavanju spam poruka, jer je važno da model što ređe propusti neželjene poruke.
Dodatno će se koristiti i metričke evaluacije poput ROC-AUC i analiza confusion matrice, što omogućava detaljnije poređenje modela u scenarijima sa neuravnoteženim klasama.

8. Tehnologije

- Python

- scikit-learn
- pandas, numpy
- nltk ili spaCy za obradu prirodnog jezika
- PyTorch ili TensorFlow
- Hugging Face Transformers
- Jupyter Notebook za eksperimentisanje i vizualizaciju rezultata

9. Relevantna literatura

- Enron Email Dataset: https://www.cs.cmu.edu/~enron/
- SpamAssassin Public Corpus: https://spamassassin.apache.org/publiccorpus/
- scikit-learn dokumentacija: https://scikit-learn.org/stable/
- Hugging Face Transformers dokumentacija: https://huggingface.co/transformers/
- NLTK dokumentacija: https://www.nltk.org/
- I. Androutsopoulos i saradnici, *An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering,* ACM SIGIR 2000