**Predlog projekta**

**Osnovi računarske inteligencija**

Profesori: Aleksandar Kovačević, Aleksandra Kaplar

Studenti: Boris Bodvanski RA120-2022

Sergej Mika SV24-2022

# Naziv teme

**Inteligentni sistem za klasifikaciju i analizu mejlova (Bag-of-Words + Naivni Bajes vs BERT modeli: DistilBERT, BERT-base, RoBERTa)**

**Definicija problema**

Problem koji se rešava jeste **automatska klasifikacija mejlova** na osnovu njihovog sadržaja i meta-podataka. Ogromna količina mejlova koja se svakodnevno šalje uključuje i:

**Spam mejlove** – neželjene poruke koje se masovno šalju sa ciljem reklame, prevare, phishing napada ili širenja malvera. Tipično sadrže ponude tipa „Win money now“, Limited offer“ ili linkove ka sumnjivim sajtovima.

**Ham mejlove** – legitimne, korisne poruke koje korisnik zaista želi da primi (poslovna i privatna komunikacija, obaveštenja, fakture, lične poruke).

Zadatak sistema je da, na osnovu teksta i dodatnih obeležja, dodeli svakoj poruci odgovarajuću klasu i da pritom bude dovoljno precizan da se izbegnu dve kritične greške:

**False negative (propust spama)** – opasno jer korisnik i dalje vidi maliciozne poruke.

**False positive (legitiman mejl označen kao spam)** – problematično jer korisnik može da propusti važnu poruku.

# Motivacija

**Automatizacija filtracije:** velik broj organizacija i privatnih korisnika prima stotine mejlova dnevno - sistem koji pouzdano filtrira spam i phishing štedi vreme i smanjuje rizik od kompromitacije.

**Sigurnost:** prepoznavanje phishing poruka smanjuje rizik od krađe lozinki i socijalnog inženjeringa.

# Skup podataka

1. [Venky Spam Mails Dataset](https://www.kaggle.com/datasets/venky73/spam-mails-dataset)

* Broj mejlova: 5.170
* Format: CSV sa 3 kolone (ID, Klasa: spam/ham, Tekst)
* Upotreba: Prvi dataset za standardno testiranje svih modela i cross-dataset evaluaciju

1. [Enron Spam Dataset](https://www.kaggle.com/datasets/marcelwiechmann/enron-spam-data)

* Broj mejlova: 33.716
* Format: CSV sa 4 kolone (ID, Subject, Message, Klasa, Datum)
* Upotreba: Drugi dataset za standardno testiranje svih modela i cross-dataset evaluaciju

**Način pretprocesiranja podataka**

* **Učitavanje podataka** – dataset sadrži tri kolone: redni broj mejla, labelu (ham/spam) i tekst mejla (subject + telo).
* **Čišćenje teksta** – uklanjanje nepotrebnih karaktera, višestrukih praznina i novih redova, pretvaranje u mala slova.
* **Tokenizacija i Bag-of-Words** – razbijanje teksta na reči i kreiranje **Bag-of-Words reprezentacije** (tj. brojanje učestalosti svake reči u dataset-u).
* **Kodiranje labela** – ham je 0, spam je 1, radi kompatibilnosti sa Naive Bayes algoritmom.
* **Podela dataset-a:**
* **Standardna evaluacija:** 80% treniranje, 20% testiranje za svaki dataset
* **Cross-dataset evaluacija:** treniranje na jednom celom dataset-u, testiranje na drugom celom dataset-u
* **BERT modeli + tokenizacija:** BERT Tokenizacija
  + Dodavanje [CLS] i [SEP] tokena
  + Kreiranje attention mask
  + Kodiranje labela ham = 0, spam = 1

## **Metodologija – način rešavanja problema**

1. Naive Bayes + Bag-of-Words

* **Cilj:** Klasifikacija mejlova na spam i ham koristeći Bag-of-Words + Naive Bayes.
* **Ulaz:** Tekst mejla (subject + telo)  
  **Izlaz:** Klasa mejla (spam ili ham)
* **Koraci procesa:**
  + **Prikupljanje i učitavanje dataset-a** – učitavanje CSV-a u DataFrame.
  + **Pretprocesiranje teksta** – čišćenje, standardizacija, pretvaranje u Bag-of-Words.
  + **Treniranje Naive Bayes modela** – koristi se jer je pogodan za rad sa frekvencijama reči.
  + **Evaluacija modela** – testiranje na neviđenim mejlovima iz test skupa.

1. BERT modeli (DistilBERT, BERT-base, RoBERTa)

* **Cilj:** Klasifikacija mejlova koristeći pretrenirane BERT modele sa linearnim slojem za klasifikaciju.
* **Ulaz:** Tokenizovani mejlovi
* **Izlaz:** Verovatnoća da mejl pripada klasi spam (prag 0.5)
* **Koraci procesa:**
* **Učitavanje i čišćenje dataset-a** – Enron CSV fajl, uklanjanje nepotrebnih karaktera, konverzija u mala slova.
* **Tokenizacija BERT modela** – odgovarajući Tokenizer, dodavanje [CLS] i [SEP] tokena, kreiranje attention mask.
* **Kodiranje labela** – ham = 0, spam = 1.
* **Treniranje modela** – linearni sloj na vrhu pretreniranog BERT modela, treniranje na 80% podataka.
* **Evaluacija na test skupu** – testiranje na preostalih 20% podataka.
* **Cross-dataset evaluacija** – model treniran na Enron dataset-u može se testirati na Venky dataset-u za analizu sposobnosti generalizacije i obrnuto.

### 3. Obrada neuravnoteženih podataka

* Pre testiranja, proverava se distribucija spam/ham.
* Po potrebi se primenjuju
* **Class weighting** – model daje veću težinu manje zastupljenoj klasi
* **Oversampling** – dupliciranje primera iz manje zastupljene klase

## Način evaluacije

**Podela podataka:**

* **Standardno treniranje/testiranje:**
* **Venky dataset:** Svi modeli (NB, DistilBERT, BERT-base, RoBERTa) - 80% train, 20% test
* **Enron dataset:** Svi modeli (NB, DistilBERT, BERT-base, RoBERTa) - 80% train, 20% test
* **Cross-dataset evaluacija:**
* Treniranje na Venky (100%) → testiranje na Enron (100%)
* Treniranje na Enron (100%) → testiranje na Venky (100%)

**Metrike:**

* **Accuracy** – koliko mejlova je pravilno klasifikovano
* **Precision, Recall, F1-score** – posebno bitno kod spam klasifikacije
* **Confusion matrix** – za vizualizaciju performansi modela

**Upoređivanje metoda**

* Nakon evaluacije oba modela, prikazuje se **direktno poređenje rezultata (na istom datasetu).**

## Tehnologije

**Jezik:** Python

**Biblioteke:**

* pandas – rad sa CSV-om
* scikit-learn – CountVectorizer, MultinomialNB
* transformers – BERT modeli i tokenizer
* torch – PyTorch za treniranje DistilBERT modela
* nltk / spaCy – NLP pretprocesiranje
* matplotlib / seaborn – vizualizacija rezultata

## Relevantna literatura i resursi

* **Kaggle dataset:** [Spam Mails Dataset (venky73)](https://www.kaggle.com/datasets/venky73/spam-mails-dataset), [Enron Spam Dataset](https://www.kaggle.com/datasets/marcelwiechmann/enron-spam-data)
* **Bag-of-Words + Naive Bayes za spam klasifikaciju:**
  + [Kaggle Notebook: Spam Classification with Naive Bayes](https://www.kaggle.com/code/satarupadeb/na-ve-bayes-classification-spam-email-detection)
  + **Tutoriali NLP pretprocesiranja:** [“Introduction to Bag-of-Words and Text Classification”](https://www.datacamp.com/tutorial/naive-bayes-scikit-learn)
* [Building a Text Classification Model using DistilBERT](https://medium.com/@prakashram1327/building-a-text-classification-model-using-distilbert-703c1409696c)
* [HuggingFace RoBERTa model](https://huggingface.co/docs/transformers/en/model_doc/roberta)
* [HuggingFace BERT-base](https://huggingface.co/google-bert/bert-base-uncased)