KLASIFIKACIJA TEKSTUALNIH PODATAKA – FARM ADS DATASET

1. Uvod

Cilj ovog projekta je primena metoda klasifikacije nad tekstualnim skupom podataka Farm Ads dataset. U projektu je izvršeno preprocesiranje tekstualnih podataka, transformacija atributa, redukcija dimenzionalnosti, treniranje više klasifikacionih modela i analiza njihovih performansi.

Posebano je bitno poređenje performansi modela treniranih nad originalnim skupom atributa (TF-IDF reprezentacija) i nad redukovanim skupovima atributa (Chi2 selekcija i SVD redukcija dimenzionalnosti).

2. Opis podataka

Korišćen je Farm Ads dataset koji sadrži 4143 tekstualna oglasa. Ciljna promenljiva ima vrednosti {-1, 1} i označava da li oglas pripada kategoriji "Farm Ads".

Nakon učitavanja podataka izvršeno je mapiranje ciljne promenljive:

- -1 -> 0 (nije Farm Ads)
- 1 -> 1 (jeste Farm ads)

Tekstualni podaci su reprezentovani kao sparse matrica velike dimenzionalnosti. Nakon primene TF-IDF transformacije broj atributa iznosi 54877. Svaki red predstavlja jedan oglas, dok kolone predstavljaju reči iz vokabulara.

3. Obrada podataka

3.1 Podela na trening i test skup

Podaci su podeljeni na:

- 80% trening skup
- 20% test skup

Korišćena je stratifikovana podela kako bi odnos klasa ostao očuvan u oba skupa.

3.2 TF-IDF transformacija

Originalni podaci predstavljaju frekvenciju pojavljivanja reči u oglasima. Nad tim podacima primenjena je TF-IDF transformacija, kojom se rečima dodeljuju težine u zavisnosti od njihove važnosti u dokumentu ali i u svim oglasima zajedno. TF označava koliko se puta određena reč pojavljuje u jednom oglasu, dok IDF smanjuje značaj reči koje se pojavljuju u velikom broju oglasa.

TF-IDF smanjuje uticaj čestih ali manje informativnih reči, dok povećava značaj reči koje bolje razlikuju klase.

Rezultat TF-IDF transformacije je sparse matrica visoke dimenzionalnosti (54877 atributa), gde svaki red predstavlja jedan oglas, a svaka kolona predstavlja težinu određene reči.

3.3 Redukcija atributa

Radi poređenja performansi modela primenjene su dve metode redukcije:

3.3.1 Chi2 selekcija atributa

Korišćena je metoda SelectKBest sa Chi2 statistikom.
Zadržano je 20000 najinformativnijih atributa. Ova metoda predstavlja filter pristup selekciji atributa, jer koristi statističku povezanost između atributa i ciljne promenljive bez treniranja modela.

3.3.2 TruncatedSVD

Primenjena je SVD redukcija dimenzionalnosti sa 300 latentnih komponenti. SVD transformacija smanjuje dimenzionalnost podataka projektovanjem originalnog prostora u prostor manjih dimenzija koji zadržava najveći deo varijanse.

Multinomial Naive Bayes model nije primenjen nad SVD reprezentacijom jer TruncatedSVD može generisati negativne vrednosti, dok MultinomialNB zahteva nenegativne ulazne atribute.

4. Korišćeni modeli

U projektu je korišćeno pet klasifikacionih algoritama:

1. Multinomial Naive Bayes
2. Logistic Regression
3. Linear Support Vector Machine (LinearSVC)
4. Decision Tree
5. Random Forest

Linearni modeli (Logistic Regression i LinearSVC) posebno su pogodni za visoko-dimenzionalne tekstualne podatke.
Decision Tree i Random Forest predstavljaju modele zasnovane na stablu i ensemble pristupu.

5. Rezultati

Modeli su trenirani na tri reprezentacije podataka:

- **FULL TF-IDF** (54877 atributa)
- **CHI2 top-20000** atributa

- **SVD 300** komponenti

Metrike koje posmatramo:

- **Accuracy (train/test)** – procenat tačno klasifikovanih primera
- **Precision (train/test)** – od svih predikcija „jeste Farm Ads“, koliko je zaista tačno (kontroliše **FP**)
- **Recall (train/test)** – od svih stvarnih „jeste Farm Ads“, koliko je model pronašao (kontroliše **FN**)
- **F1 (train/test)** – balans precision i recall
- **ROC-AUC (test)** – koliko dobro model razdvaja klase nezavisno od praga
- **Confusion matrix** – broj TN, FP, FN, TP grešaka

5.1 FULL TF-IDF (54877 atributa)

Najbolji model: LinearSVC (FULL TF-IDF)

Iz tabele (LinearSVC, FULL TF-IDF):

- **acc_train = 0.9394**, **acc_test = 0.9011**
- **precision_train = 0.9400**, **precision_test = 0.9030**
- **recall_train = 0.9394**, **recall_test = 0.9011**
- **f1_train = 0.9394**, **f1_test = 0.9007**
- **roc_auc_test = 0.9728**

Train vs Test

Razlika train - test postoji (≈ 0.94 -> 0.90), ali nije ekstremna, što znači da model generalizuje solidno. FULL TF-IDF ima jako mnogo atributa, pa je očekivano da trening bude malo bolji.

Confusion matrix (TEST) – FULL TF-IDF (LinearSVC)

- TN = 330, FP = 57
- FN = 25, TP = 417

Ukupan broj grešaka = FP + FN = 57 + 25 = 82 (od 829 test primera).

- FP=57: model ponekad „pretera“ i označi “nije Farm Ads” kao “jeste Farm Ads”
- FN=25: ređe promaši “jeste Farm Ads” i označi ga kao “nije Farm Ads”

Pošto je FN manji od FP, model je malo skloniji da “pusti” neke “nije Farm Ads” kao “jeste Farm Ads”, nego da propusti “jeste Farm Ads”.

Precision / Recall / F1

- **Precision_test ≈ 0.903** -> kada model kaže “jeste Farm Ads”, ~90% puta je u pravu (FP nisu preveliki).
- **Recall_test ≈ 0.901** -> model pronalazi ~90% stvarnih “jeste Farm Ads” oglasa (FN relativno mali).
- **F1_test ≈ 0.901** -> dobar balans između precision i recall.

ROC-AUC (test)

ROC-AUC ≈ 0.973 je vrlo visok -> model ima odličnu sposobnost razdvajanja klasa

Ostali modeli (FULL TF-IDF)

- **LogisticRegression**: malo slabiji od LinearSVC (F1_test ≈ 0.894)
- **RandomForest**: solidan ali slabiji (F1_test ≈ 0.867), i vidi se da na train radi znatno bolje nego na test -> znaci overfitting
- **DecisionTree**: još izraženije overfitting ponašanje (train ≈ 0.97, test ≈ 0.86)
- **MultinomialNB**: stabilan i brz, ali slabiji od LinearSVC/LogReg (F1_test ≈ 0.864)

Zaključak: Linearni modeli (LinearSVC, LogReg) su najbolji na sparse tekstualnim podacima visoke dimenzionalnosti.

5.2 CHI2 top-20000 atributa

CHI2 selekcija zadržava samo najinformativnije reči (atribute) u odnosu na target, čime se uklanja šum i smanjuje dimenzionalnost.

Najbolji model: LinearSVC (CHI2 top-20000)

Iz tabele (LinearSVC, CHI2 top-20000):

- **acc_train = 0.9900**, **acc_test = 0.9047**
- **precision_train = 0.9901**, **precision_test = 0.9074**
- **recall_train = 0.9900**, **recall_test = 0.9047**
- **f1_train = 0.9900**, **f1_test = 0.9042**
- **roc_auc_test = 0.9728**

Train vs Test

Train je skoro savršen (~0.99), a test je ~0.905.
To znači da CHI2 daje jasan znak, ali i dalje postoji razlika - model i dalje uči veoma snažne obrasce u treningu, ali generalizacija je i dalje veoma dobra (test ~0.905).

Confusion matrix (TEST) – CHI2 (LinearSVC)

- TN = 329, FP = 58
- FN = 21, TP = 421

Greške = 58 + 21 = 79 (manje nego FULL: 82).

- FN je još manji (21) nego u FULL (25) -> model još ređe propušta "jeste Farm Ads".
- FP je sličan (58 vs 57) -> “lažni alarmi” su približno isti.

Precision / Recall / F1

- **Precision_test ≈ 0.907** -> malo bolja nego FULL (0.903)
- **Recall_test ≈ 0.905** -> malo bolja nego FULL (0.901)
- **F1_test ≈ 0.904** -> najbolja od sva tri pristupa

CHI2 je dao blago poboljšanje uz smanjenje dimenzionalnosti (54877 -> 20000). To je odličan rezultat: manje atributa, a performanse iste ili bolje.

ROC-AUC

ROC-AUC ≈ 0.973 ostaje na istom nivou kao FULL -> separabilnost klasa je zadržana.

Ostali modeli (CHI2)

- **LogisticRegression**: slabije od LinearSVC (F1_test ≈ 0.886)
- **RandomForest**: solidan (F1_test ≈ 0.880), ali i dalje overfit (train ≈ 0.98, test ≈ 0.88)
- **DecisionTree**: test ≈ 0.86, train ≈ 0.97 -> overfitting
- **MultinomialNB**: slično FULL, stabilan ali slabiji od LinearSVC

Zaključak: Ovo je najbolji kompromis – manja dimenzionalnost i najbolji test rezultat (LinearSVC).

5.3 SVD 300 komponenti

SVD pravi novu reprezentaciju od 300 "latentnih" komponenti (kombinacije reči). Ovo najviše smanjuje dimenzionalnost.

Najbolji model: LinearSVC (SVD 300)

Iz tabele (LinearSVC, SVD 300):

- **acc_train = 0.9258**, **acc_test = 0.8951**
- **precision_train = 0.9265**, **precision_test = 0.8983**
- **recall_train = 0.9258**, **recall_test = 0.8951**
- **f1_train = 0.9256**, **f1_test = 0.8944**
- **roc_auc_test = 0.9647**

Train vs Test

Razlika train - test je mala (~0.926 -> 0.895), što znači da model generalizuje stabilno, ali ukupni maksimum performansi je niži jer smo kompresovali podatke na 300 dimenzija.

Confusion matrix (TEST) – SVD (LinearSVC)

- TN = 323, FP = 64

- FN = 23, TP = 419

Greške = 64 + 23 = 87 (više nego FULL 82 i CHI2 79).
Zato je i accuracy malo niži (~0.895).

- FP raste (64) -> više “lažnih alarma” nego FULL/CHI2
- FN je sličan (23) -> promašaji “jeste Farm Ads” nisu mnogo porasli

Precision / Recall / F1

- **Precision_test ≈ 0.898** (niži nego FULL/CHI2) – više FP
- **Recall_test ≈ 0.895** (niži nego FULL/CHI2)
- **F1_test ≈ 0.894** – očekivan pad zbog kompresije informacija

ROC-AUC

ROC-AUC ≈ 0.965 je i dalje visok, ali niži od FULL/CHI2 (≈0.973).
To znači da SVD reprezentacija i dalje dobro razdvaja klase, ali je izgubila deo separabilnosti.

Izostavljanje MultinomialNB u SVD

MultinomialNB zahteva **nenegativne** ulaze (brojevi kao učestalosti/težine).
SVD komponente mogu biti **negativne**, pa je legitimno preskočiti MultinomialNB za SVD.

Zaključak za SVD: Ogromno smanjenje dimenzionalnosti (54877 -> 300) uz umeren pad performansi. Dobro kada su performanse “dovoljno dobre”, a cilj je brzina/memorija.

5.4 Ukupno poređenje (sve reprezentacije zajedno)

Ako gledamo najbolje rezultate na test skupu (po F1_test i acc_test), dobija se:

- **Najbolje: LinearSVC + CHI2 top-20000**
  - acc_test ≈ **0.9047**
  - f1_test ≈ **0.9042**

    - roc_auc_test ≈ **0.9728**
    - greške: **79/829**
- **Odmah iza: LinearSVC + FULL TF-IDF**
    - acc_test ≈ **0.9011**
    - f1_test ≈ **0.9007**
    - roc_auc_test ≈ **0.9728**
    - greške: **82/829**
- **Najefikasnije (najmanje dimenzije): LinearSVC + SVD 300**
    - acc_test ≈ **0.8951**
    - f1_test ≈ **0.8944**
    - roc_auc_test ≈ **0.9647**
    - greške: **87/829**

Zaključci iz poređenja:

1. **LinearSVC je najbolji model** u svim reprezentacijama - linearni modeli su najpogodniji za sparse tekstualne podatke velike dimenzionalnosti.
2. **CHI2 top-20000 daje najbolji rezultat**, uz smanjenje dimenzionalnosti - selekcija atributa uklanja šum i zadržava najinformativnije reči.
3. **SVD donosi najveću kompresiju**, ali uz očekivan pad performansi - odličan kompromis kada je prioritet brzina ili memorija.
4. **DecisionTree i RandomForest** imaju visoke train rezultate, ali znatno slabije test rezultate - kod njih je vidljiv **overfitting**, posebno kod DecisionTree.
5. **ROC-AUC je vrlo visok** u sva tri slučaja (≈0.965–0.973), što znači da modeli generalno dobro razdvajaju klase i da rezultati nisu slučajni.