Predlog diplomskog rada

1. **Definicija problema/cilja projekta.**

Klasifikacija kateogrije jela na osnovu naziva i opisa samog jela. Konkretno su podržane sledeće kategorije: Condiments & Snacks, Meals, Vegetables, Beverages, Fruit, Grain Products, Chicken, Roots & Tubers, Chees, Beef, Seafood, Turkey, Pork, Yogurt, Legumes, Milk & Dairy, Tree nuts & seeds, rice, eggs, Butter, Meat

1. **Motivacija problema rešavanog u projektu.**

Veoma čest zadatak u NLP oblasti vještačke inteligencije je klasifikacija teksta. Procjenjuje se da je oko 80% informacija nestrukturirano, pri tom tekst predstavlja najčešći oblik nestrukturiranih podataka. Zbog svoje neuredjenosti analiza, klasifikacija i druge operacije nad tekstom oduzimaju previše vremena, s toga klasifikacija teksta korištenjem nekih metoda iz oblasti mašinskog učenja je veoma popularno rešenje. Konkretno u ovom slučaju pokušavamo da klasifikujemo jela na kategorije u odnosu na naziv i opis jela. Jedna od mogućih primjena kateogrizacije jela su aplikacije koje se bave dostavom hrane iz restorana, gdje podatke o hrani dobijaju od nekog eksternog izvora, stoga je potrebno klasifikovati pristigla jela na osnovu teksta.

1. **Relevantna literatura (minimum 3 rada)**
   1. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. <https://doi.org/10.3390/info10040150>

Ovaj rad se bavi različitim načinima izdvajanja “feature”-a, različitim algoritmima i tehnikama za klasifikaciju teksta i evaluacionim metodama. U radu se prolazi kroz svaki korak pipeline-a koji se najčešće koristi u klasifikaciji teksta i razmatraju se različite metode. Za izdvajanje feature-a obrađene su dvije metode i to “Word embeddings” (gloVe, Word2Vec) i “Weighted word techniques”(TF-IDF, TF i sl.). Kao algoritmi za klasifikaciju detaljno su opisane prednosti i mane tradicionalnih algoritama kao što je Rocchio klasifikacija, zatim linearnih klasifikatora kao što su Logistic Regression, Naïve Bayes, SVM i drugi, kao i deep learning modeli kao što su RNN, LSTM, CNN. Za evaluaciju su korištene preciznost, F mjera, MCC I ROC. Korišteni su različiti skupovi podataka, sa različitim brojem klasa kao i različitom veličinom samog skupa podataka. Izbor modela kao i način preprocesiranja skupa podataka veoma zavisi od konkretnog primjera. Iako su deep learning modeli veoma moćni, često usljed malog obima skupa podataka tradiocionalne tehnike i linearni modeli mogu postići odlične rezultate.

* 1. C., E. (2020). Comparing BERT against traditional machine learning text classification. *ArXiv*. https://doi.org/10.47852/bonviewJCCE3202838

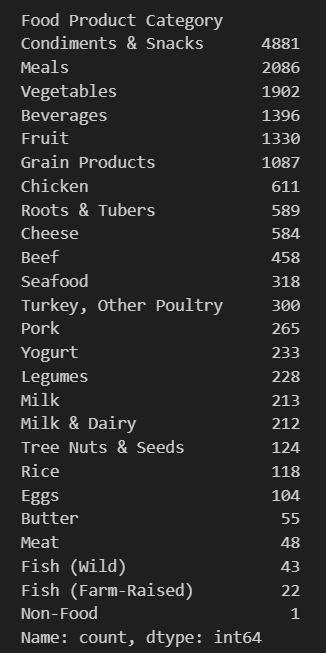
Zadatak ovog rada bio je da na različitim skupovima podataka uporedi rezultate klasifikacije teksta pomoću BERT modela i pomoću drugih modela mašinskog učenja (Logistic regression, Linear SVC, Voting Classifier, MNB, Ridge classifier, Passive Aggressive classifier) zajedno sa TF-IDF vektorizacijom. Za evaluaciju je korištena preciznost. U projektu su korištena četiri različita skupa podataka i to: IMDB skup podataka, RealOrNot tweets, Portuguese news i Chinese hotel reviews. Utvrđeno je da BERT postiže bolje rezultate nad svim skupovima podataka, kao i da je implementacija BERT modela jednostavnija od korištenja drugih modela mašinskog učenja, stoga se preporučuje korištenje BERT pretreniranog modela kao početnog u rešavanju NLP problema. Autori nisu uzeli u obzir i druge metode za evaluaciju koje bi bolje reprezentovale uspjehe modela na višeklasnim skupovima podataka.

* 1. Zaken, E. B., Ravfogel, S., & Goldberg, Y. (2021). BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. *ArXiv*. /abs/2106.10199

Kako su pretrenirani transformeri kao što je BERT veoma veliki, njihov trening i deployment su veoma skupi. Ovaj rad govori o fine-tuning metodi BERT transformera koja se naziva BitFit i služi da smanji broj parametara koji se obučavaju, a samim tim ubrza proces treniranja. Tačnije ovom metodoma se zaledi većina mreže, a fine-tuning se radi samo na “bias” komponentama. Čak je dokazano da uz mali pad performansi možemo raditi fine-tuning samo na dvije “bias” componente tako da se fine-tuning vrši na samo 0.04% parametara. Korišteni su javno dostupni BERT modeli i to: BERT base, BERT large i RoBERTa base. Evaluacija je vršena pomoću GLUE benchmark-a. Pokazano je da na malim i srednjim skupovima podataka BitFit metoda ostvaruje približne rezultate, a u nekim slučajevima i bolje od fine-tuning-a cijelog modela.

1. **Skup podataka.**

Skup podataka koji će se koristiti predstavlja podatke o nabavci hrane raznih kompanija iz 2019. godine. Skup podataka se može preuzeti na sledećem [linku](https://www.kaggle.com/datasets/anoopjohny/gourmet-food-procurement-data). Iz datog skupa podataka koristiće se naziv jela („Product Name“) i opis jela („Product Type“) koji će predstavljati ulazne podatke u model, a kao labela će se koristiti kategorija jela („Food Product Category“). Konkretno su podržane sledeće kategorije: Condiments & Snacks, Meals, Vegetables, Beverages, Fruit, Grain Products, Chicken, Roots & Tubers, Chees, Beef, Seafood, Turkey, Pork, Yogurt, Legumes, Milk & Dairy, Tree nuts & seeds, rice, eggs, Butter, Meat. Skup podataka je sačinjen od 17208 redova.



Na slici je prikazana trenutna raspodjela podataka po klasama. Klase “Fish(Wild)” i “Fish(Farm-Raised)” će se pripojiti klasi “Seafood” prilikom procesiranja skupa podataka, takođe je planirano da dođe do spajanja klasa “Milk” i “Milk & Dairy”.

1. **Metodologija**.

Ulaz u sistem će biti naziv i opis jela. Klasifikacija će biti vršena pomoću linearnog modela i BERT transformatora. Kao linearni klasifikator će se probati SVM model iz „Scikit learn“ biblioteke, Random Forest, XGBoost i Bagging (sa SVM modelom kao osnovnim). Zajedno sa linearnim klasifikatorom koristiće se TF-IDF način za izdvajanje feature-a, a zatim „gloVe“ i „Word2Vec“ word embedding kako bi se poredili rezultati linearnog klasifikatora sa različitim načinom vektorizacije. Prije prosleđivanja podataka u linearni klasifikator vršiće se pretprocesiranje gdje će se naziv i opis spojiti, zatim će se izbaciti sve što ne pripada engleskom alfabetu kao i „stopwords“ iz teksta i na kraju će se izvršiti lematizacija ulaznih podataka. Pored linearnih klasifikatora radiće se fine-tuning BERT transformatora. Arhitektura BERT modela biće proširena potpuno povezanim slojem od 512 neurona. Fine-tuning će se vršiti pomoću BitFit metode kojom će se vršiti treniranje samo „bias“ komponenti BERT modela. Na kraju će se vršiti poređenje rada najboljeg linearnog klasifikatora sa BERT modelom, kao i razlike izmedju linearnih klasifikatora u zavisnosti od načina izdvajanja feature-a. Projekat je moguće proširiti i treniranjem svih parametara BERT modela, ali zbog ograničenja resursa na kojima će se vršiti fine-tuning trenutno je plan da se vrši samo BitFit.

1. **Metod evaluacije.**

Kao mjeru performanse za klasifikaciju jela koristiće se accuracy, precision, recall i F-mjera, gdje će F-mjera biti glavna metrika za evaluaciju zbog nebalansiranosti skupa podataka.

Skup podataka će biti podijeljen na trening, validacioni i test skup i to u odnosu 70% trening, 20% validacioni i 10% test podaci. Važno je istaći da će se raditi stratifikovana podjela skupa podataka.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kategorija | Ukupno podataka | Trening podaci (70%) | Validacioni podaci (20%) | Test podaci (10%) |
| Condiments & Snacks | 4881 | 3416 | 976 | 488 |
| Meals | 2086 | 1461 | 417 | 208 |
| Vegetables | 1902 | 1331 | 380 | 190 |
| Beverages | 1396 | 977 | 279 | 139 |
| Fruit | 1330 | 931 | 266 | 133 |
| Grain Products | 1087 | 760 | 217 | 109 |
| Chicken | 611 | 427 | 122 | 62 |
| Roots & Tubers | 589 | 412 | 117 | 59 |
| Cheese | 584 | 408 | 116 | 58 |
| Beef | 458 | 320 | 92 | 45 |
| Seafood | 382 | 229 | 76 | 38 |
| Turkey, other Poultry | 300 | 210 | 60 | 30 |
| Pork | 265 | 185 | 53 | 27 |
| Yogurt | 233 | 163 | 46 | 23 |
| Legumes | 228 | 159 | 45 | 23 |
| Milk & Dairy | 425 | 297 | 85 | 42 |
| Tree Nuts & Seeds | 124 | 86 | 24 | 12 |
| Rice | 118 | 82 | 23 | 12 |
| Eggs | 104 | 72 | 20 | 10 |
| Butter | 55 | 39 | 11 | 5 |
| Meat | 48 | 33 | 10 | 5 |