Emotion Detection in Written Sentences

Nina Bu | RA, FTN - Novi Sad

Uvod

Motivacija

Cilj projekta je klasifikacija rečenica na osnovu šest osnovnih emocija: ***sadness, anger, love, surprise, fear*** i ***joy***.
Ideja je da se svaka rečenica klafisikuje prema **tačno** jednoj od ovih emocija.
Izbor ovih šest emocija je zasnovan na njihovoj čestoj prisutnosti u pisanim tekstovima i njihovoj značajnoj ulozi u ljudskom izražavanju.

Skup podataka

Skup podataka se sastoji od **~100k rečenica** na engleskom jeziku, u formatu *text, emotion*, gde se svakoj rečenici pridružuje odgovarajuća emocionalna kategorija.
Skup je podeljen na tri skupa: trening, validacioni i test skup u odnosu 70/20/10

Pretprocesiranje

Pre obučavanja modela vrši se **pretprocesiranje** i **normalizacija** teksta. Proces podrazumeva uklanjanje emoji-a, znakova interpunkcije, kao i uklanjanje *stop-words*. Nakon toga, tekst se tokenizuje i tokeni se pretvaraju u mala slova. Na kraju se vrši lematizacija tokena tj. Reči se svode na njihov osnovni oblik.

Model

1. Naive Bayes

2. Random Forest

Pre treniranja modela izvršena je transformacija tekstualnih podataka u vektore brojeva pomoću **TF-IDF** *Vectorizer-a.* TF-IDF (Term Frequency-Inverse Document Frequency) je tehnika koja se koristi za pretvaranje tekstualnih podataka u numerički oblik. Ona meri koliko je određena reč važna za određeni dokument u kontekstu celokupne kolekcije dokumenata.
Kombinacija TF i IDF daje TF-IDF vrednost koja odražava značajnost reči u dokumentu u odnosu na celokupnu kolekciju. Reči koje se često pojavljuju u dokumentu, ali su retke u celokupnoj kolekciji, dobijaju visoku TF-IDF vrednost.
Što se tiče oba modela, primenjene su implementacije Naivnog Bajesa i Random Forest algoritma iz ***scikit-learn*** biblioteke. Kod Random Forest modela podešeni su hiperparametri **min_samples_leaf** i **min_samples_split**, koji koji određuje minimalan broj uzoraka koji moraju biti u listu čvora stabla i minimalan broj uzoraka potreban za razdvajanje čvora.

3. LSTM

Pre ubacivanja u neuronsku mrežu reči su tokenizovane pomoću ***Tokenizer-a*** iz ***Keras*** biblioteke. LSTM Model se sastoji iz sledećih slojeva:
Prvi sloj je **Embedding** sloj implementiran kao pretrenirani GloVe word embedding model. Drugi po redu je **Spatial Dropout** sloj koji nasumično postavlja određeni broj ulaznih elemenata na vrednost 0 tokom svake iteracije tokom treninga, kako bi se sprečio *overfitting*. Zatim slede 2 **LSTM** sloja, gde prvi ima 128, a drugi 64 neurona sa **tanh** aktivacionom funkcijom. Poslednji sloj je **Dense** output sloj, sa 6 neurona i **softmax** aktivacionom funkcijom. Kao loss funkcija korišćena je funkcija **sparse categorical crossentropy**, a za optimizaciju je korišćen **adam.**

Testiranje i rezultati

Naive Bayes	Random Forest	LSTM
84.98%	88.03%	90.53%

Tačnost modela

Glavna metrika za procenu performansi je sva tri modela je tačnost tj. ***accuracy***. Takođe su korišćenje i druge metrike kao što su ***precision*** i ***recall***, koje ovde nisu prikazane. Rezultati su očekivani, gde najjednostavniji model, Naive Bayes ima najmanju tačnost. Zatim sledi Random Forest klasifikator, ali ubedljivo najbolje celokupne performanse ima LSTM model.

Zaključak

Evaluacija projekta

U ovom istraživanju proučena je klasifikacija emocija u pisanim rečenicama.
Rezultati su pokazali da rekurentne neuronske mreže postižu visoku tačnost u klasifikaciji emocija, dok su Naive Bayes i Random Forest takođe dali zadovoljavajuće rezultate.
Važno je napomenuti, da pored samih modela, ogroman značaj imaju podaci. Klasifikacija emocija je kompleksan problem jer se ista rečenica može različito klasifikovati od strane različitih ljudi. Emocije su subjektivne i zavise od konteksta, individualnih iskustava i interpretacije. Imajući to u vidu, projekat bi mogao da se unapredi, korišćenjem sofisticiranijeg skupa podataka koji sadrži više nijasni emocija ili čak, preći u domen **multi-label** klasifikacije. Ovo je smisleno jer neke rečenice, mogu izražavati više nepovezanih emocija.

Reference

- https://scikit-learn.org/stable/index.html
- https://https://nlp.stanford.edu/projects/glove/
- https://www.kaggle.com/code/colearninglounge/nlp-data-preprocessing-and-cleaning/notebook#Visualizing-the--Vector-Space