Raport z projektu

06-DUMAUI0 2021/SL

Cel Projektu

Celem projektu było stworzenie modelu, który przewiduje, czy dany tekst pochodzi ze Skeptic subreddita czy jednego z subredditów “paranormalnych” (Paranormal, UFOs, TheTruthIsHere, Ghosts, ,Glitch-in-the-Matrix, conspiracytheories).

Dane

Dane pochodzą ze strony: https://archive.org/details/2015_reddit_comments_corpus
Zbiór treningowy składa się z 289579 przykładów, a zbiór testowy z 5272.

Modele

W projekcie porównano działanie 2 modeli:

- Naiwny klasyfikator Bayesowski. Do preprocessingu został użyty Label Encoder oraz TFIDF Vectorizer. Wykorzystany został wielomianowy wariant naiwnego klasyfikatora Bayesowski (Multinomial Naive Bayes)
- Klasyfikator bazujący na prostej sieci neuronowej feed forward w pytorchu. Zamiast tfidf została użyta reprezentacja gęsta word2vec. Jako optymalizatora użyto stochastic gradient descent z wielkością batcha równą 15.

Ewaluacja

Do ewaluacji wykorzystano metryki *accuracy, precision, recall* i *F1-score*. Do ewaluacji wykorzystano narzędzie geval(./geval -t dev-0). Wyniki ewaluacji przedstawia poniższa tabelka:

Model	Accuracy	Precision	Recall	F1-score
Naiwny klasyfikator Bayesowski	0.7367	0.8997	0.2883	0.4367
Sieć neuronowa w PyTorch	0.7523	0.6842	0.5573	0.6143
Support Vector Machines - SVM	0.8249	0.7905	0.6876	0.7355

Wnioski

Naiwny klasyfikator Bayesowski i sieć neuronowa w PyTorch dają podobne wyniki Accuracy z lekka przewaga sieci neuronowej, która również daje lepszy wynik F1-score. Najlepsze wyniki pod względem Accuracy i jednocześnie F1-score daje metoda SVM. Jest to jednak metoda bardzo czasochłonna, stworzenie modelu na tym zbiorze danych trwało ponad 12 godzin, a przy użyciu pozostałych metod trwało to około 4 razy krócej.