Wzór raportu z projektu

06-DUMAUI0 2023/SL

Cel projektu

Celem projektu było stworzenie modelu, który przewiduje, czy dany artykuł jest tzw. Fake Newsem, czy może jest prawdziwy, na podstawie analizy tekstu.

Dane

Dane pochodzą ze strony Kaggle.com [https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification]

Po odrzuceniu kolumn z wartościami NA, uzyskano 71537 przykładów, które podzielono na zbiór treningowy 80% (57229 przykładów) i testowy 20% (14308 przykładów).

Modele

W projekcie porównano działanie 3 modeli:

- Regresja logistyczna wielomianowa 1. stopnia. Brak regularyzacji.
- Naiwny klasyfikator bayesowski. Użyto modelu Multinomial. Porównano wyniki dla kolumny tekstu po usunięciu znaków specjalnych i stopwords oraz bez przetworzenia.
- Sieć neuronowa – model BI LSTM. Użyłam 7 warstw. Wykorzystane zostały funkcje aktywacji Relu i Sigmoid. Dodano także m.in. warstwę Dropout (w celu redukcji overfittingu).

Ewaluacja

Do ewaluacji wykorzystano metryki *accuracy, precision, recall* i *F1-score*. Wyniki ewaluacji przedstawia poniższa tabelka:

Model	Accuracy	Precision	Recall	F1-score
Regresja logistyczna bez regularyzacji	0.9672211350293543	0.9593901780895845	0.9769199065805743	0.9680756926009122
Naiwny klasyfikator bayesowski	- With stopwords 0.8952334358400894 - Without stopwords 0.9037601341906626	- With stopwords 0.9048147634244484 - Without stopwords 0.9094158075601375	- With stopwords 0.8891464412326152 - Without stopwords 0.9020998091082629	- With stopwords 0.8969121793549274 - Without stopwords 0.9057430351153399
BI LTSM	0.9733715403969807	0.9669576897246475	0.9815925824925007	0.9742201772785709

Wnioski

Najlepsze wyniki pod względem F1-score uzyskano przy pomocy sieci neuronowej BI LTSM. Warto natomiast nadmienić, iż występuje tutaj zjawisko overfittingu. Model regresji logistycznej bez regularyzacji poradził sobie dużo lepiej niż naiwny model Bayesowski. Usunięcie znaków i stopwords na przykładzie modelu Bayesa pokazuje, iż ta operacja pozwala podwyższyć accuracy i F1-score.