Raport z projektu

06-DUMAUI0 2025/SL

# Cel projektu

Celem projektu było stworzenie modelu, który przewiduje sentyment tekstu (czy dany tekst jest pozytywny czy negatywny).

# Dane

Dane pochodzą z gotowego zbioru przygotowanego przez Hugging Face, zawierającego zbalansowaną liczbę 25 000 tekstów o pozytywnym lub negatywnym wydźwięku. Dane zostały podzielone na zbiór uczący (25 000 przykładów) oraz zbiór testowy (również 25 000 przykładów). (link: <https://huggingface.co/datasets/stanfordnlp/imdb>).

# Modele

W projekcie porównano działanie 3 modeli:

* Naiwny klasyfikator bayesowski: wykorzystano rozkład wielomianowy na danych po transformacji TF-IDF.
* Model LSTM oparty na embeddingach GloVe: dane wejściowe przekształcono w embeddingi 300-wymiarowe, a następnie przetworzono przez jednokierunkową warstwę LSTM i klasyfikator.
* Model transformerowy BERT: użyto wstępnie wytrenowanego modelu distilbert-base-uncased-finetuned-sst-2-english z biblioteki Hugging Face, dostrojonego na zadaniu klasyfikacji sentymentu.

# Ewaluacja

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score |
| Naiwny klasyfikator bayesowski | 0,86 | 0,86 | 0,86 | 0,86 |
| LSTM + GloVe | 0,8064 | 0,8065 | 0,8064 | 0,8064 |
| Transformer (BERT) | 0,93 | 0,93 | 0,93 | 0,93 |

# Wnioski

Najlepsze wyniki pod względem wszystkich metryk osiągnął model typu transformer (BERT). Wysoka skuteczność wynika z faktu, że BERT jest modelem głębokim, wstępnie wytrenowanym na dużej liczbie danych i uwzględnia kontekst słów. Model LSTM oparty na embeddingach GloVe poradził sobie znacznie gorzej niż transformer, co może być związane z brakiem kontekstu i trudnością w trenowaniu sieci od zera na ograniczonym zbiorze danych. Naiwny klasyfikator bayesowski, mimo swojej prostoty, uzyskał zaskakująco dobre wyniki – lepsze niż LSTM – co potwierdza skuteczność klasycznych metod na zadaniach z dobrze dobranymi cechami (TF-IDF). Może to świadczyć o przetrenowaniu LSTM lub niedostatecznym czasie trenowania.