06-DUMAUI0 2025/SL

Cel projektu

Celem projektu było porównanie skuteczności różnych metod klasyfikacji tekstu w zadaniu wieloklasowym na zbiorze wiadomości z Newsgroups. W szczególności oceniono następujące podejścia: 1. TF–IDF + regresja logistyczna 2. Word2Vec (średnia wektorów) + regresja logistyczna 3. RNN (LSTM) 4. Seq2Seq (encoder–decoder jako generacja etykiety) 5. Transformer (BERT)

Dane

Dane wykorzystane w projekcie pochodzą z publicznego zbioru 20 Newsgroups, ograniczonego do czterech kategorii: - alt.atheism - comp.graphics - sci.med - soc.religion.christian

Ostatecznie zestaw składał się z ~18 846 dokumentów, które po wstępnej obróbce (usunięcie nagłówków, stop-words) podzielono losowo na zbiór uczący (80%, ok. 15 077 próbek) i testowy (20%, ok. 3 769 próbek).

Modele

W projekcie zaimplementowano i przetestowano pięć różnych podejść do klasyfikacji:

1. **TF–IDF + regresja logistyczna**
    - wektory TF–IDF o maksymalnej liczbie cech 5 000
    - regresja logistyczna z domyślnymi parametrami (L2, solver liblinear)
2. **Word2Vec + regresja logistyczna**
    - trenowany lokalnie model Word2Vec (vector_size=100, window=5)
    - reprezentacja dokumentu jako średnia wektorów słów
    - regresja logistyczna analogicznie do podejścia TF–IDF
3. **RNN (LSTM)**
    - tokenizacja ograniczona do 10 000 najczęstszych słów
    - pad_sequences długością 200 tokenów
    - warstwa embedding (rozmiar 128), pojedynczy LSTM(128)
    - klasyfikacja przez Dense+softmax
4. **Seq2Seq**
    - architektura encoder–decoder oparta na LSTM
    - etykiety klas traktowane jako pojedyncze „znaki” (0–3)
    - model uczony do generacji sekwencji o długości 1

5. **Transformer (BERT)**
    - wstępnie wytrenowany `bert-base-uncased`
    - tokenizacja z maksymalną długością 128 tokenów
    - fine-tuning na całym modelu BERT z warstwą klasyfikacyjną

Ewaluacja

Do oceny jakości klasyfikacji zastosowano metryki: - **Accuracy** (dokładność) - **Precision** (precyzja) - **Recall** (czułość) - **F1-score** (miara F1)

Model	Accuracy	Precision	Recall	F1-score
TF–IDF + regresja logistyczna	0.85	0.85	0.85	0.85
Word2Vec + regresja logistyczna	0.80	0.80	0.80	0.80
RNN (LSTM)	0.88	0.88	0.88	0.88
Seq2Seq (encoder–decoder)	0.35	0.34	0.35	0.34
Transformer (BERT)	0.92	0.92	0.92	0.92

Wnioski

1. Najlepsze wyniki osiągnął model Transformer (BERT), uzyskując dokładność ~92% oraz najwyższe wartości wszystkich metryk.

2. Model RNN (LSTM) ustępuje nieznacznie BERT-owi, ale nadal przekracza 88% we wszystkich miarach, co potwierdza siłę sieci rekurencyjnych w zadaniach sekwencyjnych.

3. Klasyczne podejścia TF–IDF oraz Word2Vec w połączeniu z regresją logistyczną zapewniły solidne baseline'y (85% i 80%), jednak odstawały od metod głębokich.

4. Podejście Seq2Seq jako generacja etykiet okazało się nieskuteczne (Accuracy ~35%) i niezalecane do prostych zadań klasyfikacyjnych.

Podsumowując, dla zadania wieloklasowej klasyfikacji tekstu w zbiorze Newsgroups rekomenduje użycie fine-tuningu modelu BERT, ewentualnie LSTM, w zależności od dostępnych zasobów obliczeniowych.