УДК 004.91

СЕНТИМЕНТ-АНАЛІЗ ТЕКСТУ З ВИКОРИСТАННЯМ МОДЕЛЕЙ BERT ТА DISTILBERT

SENTIMENT ANALYSIS OF TEXT USING BERT AND DISTILBERT MODELS

Маріанна Притула, Ігор Оленич

Львівський національний університет імені Івана Франка, вул. Драгоманова, 50, 79005, Львів, Україна

Abstract. *The work describes the fine-tuning of BERT and DistilBERT pre-trained transformer models for the binary sentiment classification of user comments in the Ukrainian language. The models were evaluated according to accuracy, precision, recall, and F1-score metrics. An overall accuracy of 91% obtained for both models demonstrates the significant potential of incorporating these models into sentiment analysis tasks.*

Сентимент-аналіз став важливим напрямком досліджень у сфері обробки природної мови у зв'язку зі швидким збільшенням кількості текстових даних, таких як рецензії, твіти, реакції та коментарі. Автоматизоване визначення емоційного забарвлення надає можливість швидко реагувати та проводити аналіз рівня задоволення користувачів певними товарами чи послугами.

У роботі досліджено налаштування bert-base-multilingual-cased та distilbert-base-multilingual-cased моделей з Hugging Face для сентимент-аналізу текстів, оскільки попередньо-навчені моделі BERT показують високу ефективність для вирішення багатьох завдань обробки природної мови [1]. Для навчання та тестування моделей використано датасет з 11000 коментарями користувачів українською мовою про магазини, ресторани, готелі, медичні заклади, фітнес-клуби та отримання послуг з бінарним поділом на класи: з позитивним та негативним забарвленням. Після попередньої обробки, набір даних був розділений на навчальну та тестову вибірки у співвідношенні 80 і 20 %. Перетворення текстових даних у формат, який модель використовує для навчання та прогнозування, включає етапи: застосування Bert/Distilbert Tokenizer для конвертації у BERT-сумісні токени, числове представлення токенів, додавання спеціальних токенів, створення маски уваги та типів токенів. Оптимізатор AdamW був використаний з $2e^{-5}$ швидкістю навчання, розмір пакету даних встановлено рівним 8. Навчання моделей проводилось упродовж трьох епох.

Табл. 1. Результати ефективності BERT та DistilBERT моделей для сентимент-аналізу тексту

Модель	Клас	Кількість коментарів	Precision	Recall	F1-Score
BERT	Негативний	1191	92 %	90 %	91 %
	Позитивний	1009	89 %	91 %	90 %
DistilBERT	Негативний	1191	91 %	92 %	91 %
	Позитивний	1009	90 %	90 %	90 %

У підсумку, хоча DistilBERT вимагає значно менше обчислювальних ресурсів та є швидшою у порівнянні з BERT, загальна точність для обох моделей була однаковою – 91 %.

Бібліографія

1. Akpatsa S.K., Lei H., Li X., Obeng V.K.S., Martey E.M., Addo P.C., Fiawoo D.D. Online News Sentiment Classification Using DistilBERT. *Journal of Quantum Computing*. Vol. 4. P. 1–11.