

АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТІВ З ВИКОРИСТАННЯМ МОДЕЛЕЙ ГЛИБОКОГО НАВЧАННЯ

Ахмедов Р. Р., Безкоровайний В. С., Дербенцев В. Д.

Київський національний економічний університет імені Вадима Гетьмана, м. Київ, Україна

Анотація. Робота присвячена розробці моделі аналізу тональності текстів із використанням глибоких нейронних мереж. З цією метою в роботі було запропоновано архітектуру, що складається із перед-навченої моделі BERT (Bidirectional Encoder Representations from Transformers), що здійснює попередню обробку тексту, токенизацію та векторне представлення слів, та моделі класифікатора на базі логістичної регресії. Об'єктом дослідження є аналіз тональності текстів (сентимент-аналіз). Предметом дослідження є моделі обробки природної мови на основі глибоких нейронних мереж. Згідно з одержаними результатами наша модель показала на тестовій вибірці точність 82% в термінах метрики Ассурасу. Проведене експериментальне дослідження свідчить про перспективність запропонованого підходу здійснення сентимент-аналізу з використанням перед-навченої моделі BERT. В подальших дослідженнях можна використовувати більш складні моделі класифікаторів, здійснювати багатокласову класифікацію, аспектний сентимент-аналіз.

Ключові слова: моделі обробки природної мови, сентимент-аналіз, глибокі нейронні мережі, модель BERT.

TEXT SENTIMENT ANALYSIS USING DEEP LEARNING MODELS

Akhmedov R., Bezkorovainyi V., Derbentsev V.

Kyiv National Economic University named after Vadym Hetman, Kyiv, Ukraine

Abstract. This paper is devoted to the development of Sentiment Analysis model using Deep Neural Networks (DNNs). For this purpose, pre-trained BERT-based (Bidirectional Encoder Representations from Transformers) architecture has been proposed. BERT unit provides data preprocessing, tokenization and word embedding, and the Logistic Regression unit performs final classification. Sentiment Analysis (SA) is the object of research. As modelling tool Natural Language Processing models based on DNNs (subject) have been used. According to the obtained results, our model has shown an accuracy of 82% on the test dataset in terms of the Accuracy metric. Experimental setup has shown availability of proposed approach of applying pre-trained BERT model for SA task. Further research can be focused on the use of more complex classification models, providing multiclass classification, implementation of aspect-based SA.

Keywords: Natural Language Processing, Sentiment Analysis, Deep Neural Networks, BERT.

Вступ. Стрімкий розвиток електронних ЗМІ та соціальних мереж дав новий поштовх розвитку автоматизованих систем обробки природної мови (Natural Language Processing, NLP). Одним із важливих завдань NLP є аналіз тональності текстів, або сентимент аналіз (Sentiment Analysis, SA) [1]. SA – це спроба «витягти» з тексту суб'єктивні характеристики: емоції, сарказм, розгубленість, підозрілість тощо. Основна мета SA полягає у здійсненні класифікації полярності даного документа (висловлювання, речення, твіта тощо) та визначенні, чи є думка, висловлена в ньому, позитивною, негативною чи нейтральною. Прогрес у NLP був досягнутий завдяки використанню моделей штучного інтелекту (ШІ), зокрема моделей на основі глибоких нейронних мереж [2], [3]. З точки зору ШІ SA є проблемою бінарної або багатокласової класифікації.

Метою роботи є розроблення автоматизованої системи аналізу тональності текстів на основі глибокої мережі BERT [4].

Модель BERT, що запропонована дослідниками компанії Google [4], є перед-навченою моделлю для вирішення різних задач NLP. Замість рекурентних нейронних мереж модель використовує архітектуру Transformer, засновану не на зворотному зв'язку, а на механізмі

«уваги» та «самоуваги» (Attention and Self-attention Mechanism). Ця модель була навчена на величезному корпусі нерозміченого тексту, включаючи Вікіпедію.

Безпосереднє навчання цієї моделі вимагає великих обчислювальних ресурсів, оскільки навіть базова версія містить більше 100 млн. параметрів. Але, хоча ця модель не була призначена для класифікації текстів, моделі типу BERT є універсальними та здатними «витягувати» з тексту ознаки, корисні для вирішення багатьох задач текстового аналізу. Тому попередньо навчену BERT можна переналаштувати додаванням кількох додаткових вихідних повнозв'язних шарів, щоб створити нові моделі для різноманітних завдань NLP, включаючи SA. При цьому для навчання моделі необхідно, здебільшого, лише навчити класифікатор з мінімальними змінами самої моделі BERT під час фази навчання (можна лише «донавчити» додані кілька шарів з використанням бібліотеки TensorFlow [5], або через оболонку Keras BERT [6]). Такий процес навчання називається тонким налаштуванням для конкретної задачі. Для вирішення задачі SA електронних текстів нами було запропоновано таку архітектуру: модель BERT обробляє речення (здійснює попередню обробку тексту, токенизацію та створює векторне подання слів – ембедінги) та передає отриману інформацію у наступну модель-класифікатор, яка і здійснює sentiment аналіз.

В якості моделі BERT було використано базовий варіант моделі для англійської мови BERT-Base, Uncased (12-layer, 768-hidden, 12-heads, 110M parameters), яка знаходиться у вільному доступі у репозитарії GitHub [7]. Для токенизації було використано wordpiece-токенізатор для перетворення вхідних рядків у BERT-сумісний формат.

Як модель-класифікатор було обрано модель звичайної логістичної регресії з функцією активації Softmax. Оскільки sentiment-аналіз зводиться до задачі класифікації, то перевагами моделі логістичної регресії є те, що на виході вона дає ймовірності прогнозованих класів. Модель BERT-Base-Uncased була донавчена з використанням оптимізатора Adam протягом 5 епох. Як функцію похибки було використано категоріальну крос-ентропію.

Для реалізації моделі було використано датасет Large Movie Review Dataset [8], який містить близько 50 тис. розмічених текстів. Він призначений для бінарної класифікації настроїв і містить набагато більше даних, ніж будь-які попередні набори даних у цій галузі. Згідно з одержаними результатами наша модель показала на тестовій вибірці точність 82% в термінах метрики Ассигасу (відношення кількості вірно класифікованих прикладів до їх загального обсягу).

Висновки. Проведене експериментальне дослідження свідчить про перспективність запропонованого підходу здійснення SA з використанням перед-навченої моделі BERT. В подальших дослідженнях можна використовувати більш складні моделі класифікаторів (наприклад, інші моделі нейронних мереж), здійснювати багатокласову класифікацію, аспектний SA. Доцільним також є тестування на доменах інших предметних галузей, зокрема, сайтах новин.

Список використаних джерел

1. Liu B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*. 2012. Vol. 5 (1). P. 1–167.
2. Trisnaa K. W., Jiea H. J. Deep learning approach for aspect-based sentiment classification: A comparative review. *Applied Artificial Intelligence*. 2022. P. 1–37. doi: 10.1080/ 08839514.2021.2014186.
3. Drus Z., Khalid H. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*. 2019. Vol. 161. P. 707–714.
4. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*. 2019. P. 4171–4186.
5. Tensorflow. URL: <https://www.tensorflow.org/> (дата звернення: 11.05.2022).
6. Keras Documentation. URL: <http://keras.io> (дата звернення: 11.05.2022).
7. Git Hub. URL: <https://github.com/google-research/bert> (дата звернення: 11.05.2022).
8. Large Movie Review Dataset. URL: <http://ai.stanford.edu/~amaas/data/sentiment/> (дата звернення: 11.05.2022).