УДК 004.8

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ BERT ДЛЯ КЛАССИФИКАЦИИ НЕСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Бадретдинов Д.В., студент 3 курса инженерно-экономического факультета
Научный руководитель – Зайченко Е.А.,старший преподаватель
Белорусско-Российский университет, Могилёв, Беларусь

Ключевые слова: *Нейросети, обработка естественного языка, модели BERT, классификация текста*

Работа посвящена сравнительному анализу различных моделей BERT для задачи классификации текстов. Рассматриваются производительность и точность предобученныхтрансформерных моделей. Охватываются ключевые аспекты обработки естественного языка (NLP), методы оценки качества классификации и влияние архитектурных особенностей моделей на их результаты.

В последние годы задачи обработки естественного языка (NLP) активно развиваются благодаря появлению трансформерных моделей[1], среди которых особенно выделяется архитектура BERT (BidirectionalEncoderRepresentationsfromTransformers). Эти модели демонстрируют высокую эффективность в классификации текстов, однако существуют различные модификации BERT, отличающиеся по структуре, объёму предобученных данных и вычислительной сложности. Оптимальный выбор модели остаётся актуальной задачей, особенно при работе с неструктурированной текстовой информацией.

Целью данной работы является сравнительный анализ различных предобученных моделей BERT в задаче классификации неструктурированного текста [2]. Исследование направлено на оценку точности, производительности и эффективности различных версий BERT. Особое внимание уделяется влиянию архитектурных различий моделей на их результаты и вычислительные затраты.

В рамках исследования проведены эксперименты с несколькими вариантами BERT, анализируются их показатели на одном и том же наборе данных, а также рассматриваются преимущества и недостатки каждой модели в контексте реального применения.

Для экспериментов использовался датасет, содержащий неструктурированные текстовые данные, размеченные по категориям. В ходе предварительной обработки выполнены следующие операции.

1) Все лейблы классов были пронумерованы, что позволило преобразовать категориальные метки в числовой формат, необходимый для работы моделей.

2) Данные были разделены на обучающую и тестовую выборки. Чтобы улучшить баланс классов и снизить переобучение, был выбран равномерный поднабор записей различных классов.

3) Текст был очищен и токенизирован с использованием встроенного токенизатора BERT, что позволило привести входные данные к формату, необходимому для работы модели (добавление специальных токенов [CLS]и [SEP], приведение к фиксированной длине, создание масок внимания).

Модель и процесс обучения. Для классификации использовалась предобученная модель на основе архитектуры BERT. В ходе экспериментов выполнены следующие операции.

1) Модель была загружена из библиотеки HuggingFaceTransformers, что позволило использовать предобученные веса и дообучать её на подготовленном заранее датасете, состоящим из 900 предложений, разделенных на три разных класса [3].Количество объектов в классах одинаково.

2) В качестве оптимизатора применялся алгоритм оптимизацииAdam, который адаптирован для работы с трансформерными моделями.

3) Обучение проходило в несколько эпох с использованием кросс-энтропийной функции потерь, так как задача является многоклассовой.

4) Для ускорения вычислений использовались GPU-ускорения (если доступны) или вычисления на TPU/CPU в облачной среде (например, GoogleColab).

Оценка точности и анализ результатов. После обучения модель тестировалась на заранее отложенной тестовой выборке. Для оценки производительности использовалась стандартная метрика –accuracy– общая доля правильных предсказаний.

Данные были замерены на пятой эпохе. Объем обучающей выборки всегда равен 750 объектов. Объем тестовой выборки всегда равен 150 объектов. Количество классов равно 3. Количество объектов в классах одинаково. Размер батча (подмножество данных, которое обрабатывается моделью за один шаг/итерацию) равен 4. Используемая GPU - RTX 6000Ada. Потери тренировки – среднее значение функции потерь для последних 100 батчей при обучении. Потери тестирования– среднее значение функции потерь для последних 100 батчей при тестировании.

Эксперименты показали, что модели google/rembert, FacebookAI/xlm-roberta-large и sentence-transformers/LaBSE продемонстрировали высокую точность классификации, превышающую 95%. Лидером среди них стала google/rembert с показателем 97%, что свидетельствует о её высокой эффективности в задаче обработки неструктурированного текста. FacebookAI/xlm-roberta-large и sentence-transformers/LaBSE также показали достойные результаты (по 95%), подтверждая свою пригодность для решения задач текстовой классификации. При этом можно обратить внимание, что сложные модели (xml-roberta-base) не всегда справляются хорошо, а крупные компании (microsoft) не всегда способны сделать пригодный продукт.

Полученные данные указывают на высокую производительность многоязычных моделей и подчёркивают значимость выбора архитектуры в зависимости от конкретных требований к задаче.

Библиографический список:

1. Нежников, Р. И. Сравнительный анализ моделей трансформера для классификации неструктурированной текстовой информации / Р. И. Нежников, А. Н. Марьенков // Прикаспийский журнал: управление и высокие технологии. – 2024. – № 2(66). – С. 32-38. – EDN LREEXX.

2. Краснов, Ф. В. Использование языковых моделей на основании архитектуры трансформеров для понимания поисковых запросов на

электронных торговых площадках / Ф. В. Краснов // InternationalJournalofOpenInformationTechnologies. – 2023. – Т. 11, № 9. – С. 33-40. – EDN YJMLDT.

3. Lepekhin, M. Experiments with adversarial attacks on text genres / M. Lepekhin, S. Sharoff // Computational Linguistics and Intellectual Technologies, 15–18 июня 2022 года. Vol. Выпуск 21 S, 2022. – P. 1106-1117. – DOI 10.28995/2075-7182-2022-21-1106-1117. – EDN BKMFNW.

COMPARATIVE ANALYSIS OF BERT MODELS FOR CLASSIFICATION OF UNSTRUCTURED TEXT INFORMATION

Badretdinov D.V.
Scientific supervisor – Zaichenko E.A.
Belarusian-Russian University, Mogilev, Belarus

Keywords: *Neural networks, natural language processing, BERT models, text classification.*

The paper is devoted to a comparative analysis of various BERT models for the task of text classification. The performance and accuracy of pre-trained transformer models are considered. The key aspects of natural language processing (NLP), classification quality assessment methods, and the impact of architectural features of models on their results are covered.