УДК 004.85

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВЕКТОРИЗАЦИИ ТЕКСТА ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ В NLP

Скворцов Александр Александрович, к.п.н., доцент,
Анурьева Мария Сергеевна, к.п.н., доцент,
Солодовников Александр Николаевич, ведущий специалист IT-центра,
Тамбовский государственный университет имени Г.Р. Державина

COMPARATIVE ANALYSIS OF TEXT VECTORIZATION METHODS FOR CLASSIFICATION TASKS IN NLP

Skvorcov Aleksandr, candidate of Pedagogical Sciences, Associate Professor, skvor_88@mail.ru,
Anureva Mariya, candidate of Pedagogical Sciences, Associate Professor, anuryeva@mail.ru,
Solodovnikov Aleksandr, Senior Specialist at the IT Center, bearbearovich@gmail.com,
Tambov State University named after G.R. Derzhavin, Russian Federation

Проведен сравнительный анализ методов векторизации текста для задач классификации в NLP. Эксперименты показали преимущество моделей BERT по качеству предсказаний. Подчеркивается значение выбора подхода к векторизации для повышения точности.

Ключевые слова: *обработка текста, классификация, NLP, векторизация, TF-IDF, FastText, BERT, трансформеры.*

The abstract: A comparative analysis of text vectorization methods for NLP classification tasks is presented. Experiments show BERT models outperform others in prediction quality. The study emphasizes the importance of vectorization choice for improving accuracy.

Keywords: *text processing, classification, NLP, vectorization, TF-IDF, FastText, BERT, transformers*.

Одним из подходов к выявлению интересов пользователя является анализ его текстовой активности в социальных сетях. При этом интересы можно интерпретировать как категории, что позволяет сформулировать задачу в терминах автоматической классификации текстов. Для построения эффективной модели классификации необходимо выбрать подходящий способ представления текстов в числовом виде. С этой целью в настоящей работе проводится анализ современных методов векторизации текста, применяемых в задачах обработки естественного языка.

Курейчик В.В., Родзин С.И. и Бова В.В. в своей работе «Методы глубокого обучения для обработки текстов на естественном языке» рассматривают подходы на основе искусственных, сверточных и рекуррентных нейронных сетей [1]. Отдельное внимание уделено архитектурам CNN и их применимости к задачам семантического анализа и классификации текста. Подчеркивается значимость выбора эффективного представления текста, включая векторные модели, в достижении высокой точности. Также описан нейроэволюционный алгоритм для автоматического подбора архитектуры CNN, что подтверждает растущую роль оптимизации в построении адаптивных NLP-моделей. Работа актуальна в контексте выбора и сравнения методов векторизации текста при решении прикладных задач классификации.

Аспекты классификации текстов, извлечения информации и анализа тональности подробно рассматриваются в работе Пудаковой В.Е. и Кулакова П.А. «Методики использования искусственного интеллекта» [2]. Подчеркивается значение этапа подготовки текста – токенизации, лемматизации и синтаксического анализа – для построения эффективных моделей классификации. Приведен обзор метрик оценки качества, включая точность, полноту, F1-меру и коэффициент ошибок. Работа подчеркивает практическую значимость выбора подходящих методов представления и анализа текста для решения прикладных задач в NLP.

Хоружая А.Н., Козлов Д.В., Арзамасов К.М. и Кремнева Е.И. в своем исследовании [3] проводят сравнительный анализ моделей BERT и ансамбля алгоритмов машинного обучения для задачи бинарной классификации текста.Использованы три метода векторизации – bag-of-words, TF-IDF и Word2Vec — в сочетании с семью алгоритмами: логистической регрессией, деревом решений, случайным лесом, SVM, KNN, CatBoost и XGBoost. Лучшие модели объединены методом стекинга. Предобученная модель BERT (MedRuBertTiny2) дообучалась на том же корпусе. Результаты

показали более высокие метрики у BERT по сравнению с ансамблем, особенно по чувствительности. Работа демонстрирует преимущества современных трансформеров и значимость выбора подхода к векторному представлению текста в задачах классификации.

Джозеф Т. акцентирует внимание на применении NLP в задачах анализа тональности в социальных сетях [4]. Обзор показывает, что модели машинного и глубокого обучения (CNN, RNN) эффективно справляются с классификацией текста, учитывая контекст и сленг. Отмечена важность векторизации текста и развитие гибридных моделей для многоязычного анализа. Работа подчеркивает роль масштабируемых и этически устойчивых NLP-систем. Лейн Х. и Дишел М. уделяют особое внимание построению векторных представлений слов и предложений [5]. Подчеркивается, что векторизация является ключевым этапом в обработке текста, позволяющим преобразовать неструктурированные данные в формат, пригодный для машинного анализа. Это подчеркивает актуальность сравнительного анализа методов векторизации для повышения эффективности классификации текстов.

Классификация текста представляет собой одну из ключевых задач обработки естественного языка (Natural Language Processing, NLP), заключающуюся в автоматическом присвоении одной или нескольких категорий текстовым данным.

В зависимости от формата разметки выделяют следующие типы классификации:

1. Бинарная классификация – задача, в которой текст относится к одному из двух классов.
2. Мультиклассовая классификация – предполагает присвоение текста одному из нескольких возможных классов.
3. Мультилейбл классификация – допускает одновременное присвоение нескольким классам, что отражает многотематичность текста.

Сравнение текстовых данных в их исходной форме затруднено из-за лексической неоднородности, вариативности и отсутствия структуры. В связи с этим важнейшим этапом в задачах классификации является преобразование текста в числовое представление – векторизацию. Существует несколько распространенных подходов к векторному представлению текста, среди которых наибольшее распространение получили TF-IDF, FastText и BERT. Для выбора наилучшего метода в конкретной задаче необходим сравнительный анализ их эффективности.

Выполнено сравнение наиболее распространённых методов векторизации текста, применяемых в задачах обработки естественного языка, с представлением их ключевых характеристик в табличной форме (табл. 1).

Таблица 1. – Сравнение методов векторизации текста

Метод	Описание	Преимущества	Недостатки
TF-IDF	Статистическая оценка важности слов	Простой, быстрый, устраняет шум	Не учитывает порядок и смысл слов
FastText	Разбивает слова на n-граммы, учитывает морфологию	Хорошо работает с редкими словами, учитывает морфологию	Большой размер, теряет контекст в длинных текстах
BERT	Модель на основе трансформеров, учитывает контекст с двух сторон	Высокая точность, контекст, дообучение под задачу	Требует ресурсы и много размеченных данных

В начале каждого предложения BERT добавляет специальный токен [CLS] (classification). Он аккумулирует информацию обо всем предложении в одном векторе. Эти вектора можно использовать сравнивать между собой классификации текста. Например, можно сравнить вектор предложения про IT и программирование с вектором текста от пользователя, чтобы узнать, связано ли это предложение с IT или нет.

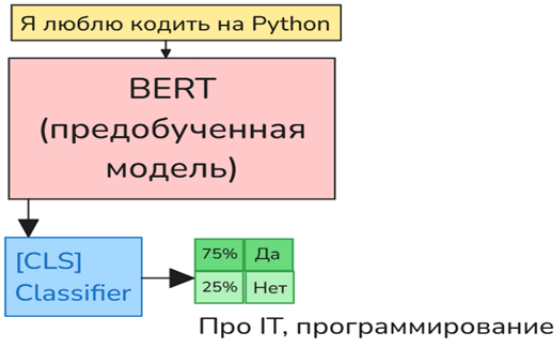

Рисунок – Классификация с помощью BERT

Для оценки эффективности различных языковых моделей, обученных на русском языке, были протестированы предобученные модели BERT с платформы HuggingFace:

1. cointegrated/rubert-tiny2;
2. DeepPavlov/rubert-base-cased;
3. cointegrated/LaBSE-en-ru;
4. ai-forever/ruRoberta-large.

Эксперимент проводился в рамках задачи бинарной классификации: определение принадлежности текста к тематике информационных технологий и программирования. Каждая модель была протестирована на одинаковом наборе текстов, для которых заранее задан бинарный признак – наличие или отсутствие связи с IT (табл.2).

Таблица 2. – Сравнительные метрики качества моделей классификации текста

Метод	Accuracy	ROC-AUC	F1-score
BERT cointegrated/rubert-tiny2	0.9160	0.9352	0.9182
BERT DeepPavlov/rubert-base-cased	0.7334	0.8086	0.7425
BERT cointegrated/LaBSE-en-ru	0.8013	0.8468	0.8092
BERT ai-forever/ruRoberta-large	0.4281	0.4304	0.4501
TF-IDF	0.6090	0.7193	0.6093
FastText	0.8433	0.8695	0.8488

Пример входных данных тестовой выборки представлен в Таблице 3.

Таблица 3. – Примеры входных данных тестовой выборки

Текст	IT-класс
Ты был на активном отдыхе?	0
Figma — инструмент для дизайна интерфейсов.	1
Tableau — инструмент для визуализации данных.	1
Я хочу отправиться на концерт классической музыки.	0
Мне нужно улучшить интерфейс нашего приложения...	1
Я люблю кататься на велосипеде.	0
MySQL — популярная реляционная база данных.	1

Количество предложений в тестовой выборки: 619 (табл. 4).

В ходе анализа рассмотрены различные подходы к векторному представлению текста, которые играют ключевую роль в задачах классификации в рамках обработки естественного языка. Выявлено, что эффективность классификационной модели во многом зависит от выбранного метода векторизации. Статистические (TF-IDF), морфологически ориентированные (FastText) и контекстные (BERT) подходы имеют свои сильные и слабые стороны.

Таблица 4. – Количество верных предсказаний классификационных моделей

Метод	Количество верных предсказаний
BERT cointegrated/rubert-tiny2	567
BERT DeepPavlov/rubert-base-cased	454
BERT cointegrated/LaBSE-en-ru	496
BERT ai-forever/ruRoberta-large	265
TF-IDF	377
FastText	522

Современные трансформерные модели, такие как BERT, демонстрируют наилучшие результаты по метрикам качества (F1, Accuracy, ROC-AUC), особенно в задачах, где требуется учитывать смысл и контекст. Исследование подтверждает актуальность использования компактных и специализированных моделей, таких как rubert-tiny2, для тематической классификации текста, а также подчеркивает перспективность комбинирования методов в рамках ансамблевых моделей.

Список использованных источников

1. Курейчик В.В., Родзин С.И., Бова В.В. Методы глубокого обучения для обработки текстов на естественном языке // Известия ЮФУ. Технические науки. 2022. №2 (226). С. 189-199.

2. Пудакова В.Е., Кулаков П.А. Методики использования искусственного интеллекта // Известия ТулГУ. Технические науки. 2023. №4. С. 303-306.

3. Хоружая А Н., Козлов Д В., Арзамасов К М., Кремнева Е И. Сравнение ансамбля алгоритмов машинного обучения и BERT для анализа текстовых описаний КТ головного мозга на предмет наличия внутричерепных кровоизлияний // Соврем. технол. мед.. 2024. №1. С. 27-36.

4. Joseph T. Natural Language Processing (NLP) for Sentiment Analysis in Social Media //International Journal of Computing and Engineering. – 2024. – Т. 6. – №. 2. – С. 35-48.

5. Lane H., Dyshel M. Natural language processing in action. – Simon and Schuster, 2025.