Баклановский Дмитрий Сергеевич, доцент кафедры прикладной математики и экономико-математических методов, Санкт-Петербургский государственный экономический университет, г. Санкт-Петербург

Воробьёв Илья Тимофеевич, магистрант, Санкт-Петербургский государственный экономический университет, г. Санкт-Петербург

ОЦЕНКА ПРОИЗВОДИТЕЛЬНОСТИ КЛАССИЧЕСКИХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ТРАНСФОРМЕРНЫХ МОДЕЛЕЙ ПРИ РЕШЕНИИ ЗАДАЧ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

Аннотация. В данной статье представлен сравнительный анализ эффективности классических методов машинного обучения и современных трансформерных моделей для решения задач многоклассовой классификации текстов. Исследование особо актуально в условиях растущего объема текстовых данных и необходимости их быстрой и точной обработки. Авторы проводят детальную оценку производительности различных алгоритмов, включая логистическую регрессию, методы опорных векторов, ансамбли деревьев и нейросетевые архитектуры, используя стандартные метрики качества.

Ключевые слова: многоклассовая классификация, машинное обучение, трансформеры, BERT, логистическая регрессия, NLP.

Annotation. This paper presents a comparative analysis of the effectiveness of classical machine learning methods and modern transformational models for solving multiclass text classification problems. The study is particularly relevant in the context of the growing volume of text data and the need to process it quickly and accurately. The authors perform a detailed performance evaluation of various algorithms, including logistic regression, support vector methods, tree ensembles and neural network architectures, using standard quality metrics.

Keywords: Multiclass classification, machine learning, transformers, BERT, logistic regression, NLP.

Основной текст статьи

Современные задачи многоклассовой классификации текстов требуют высокоэффективных решений, способных обрабатывать большие объемы данных с минимальными временными затратами. В этом контексте сравнение классических методов машинного обучения и современных трансформерных моделей представляет значительный научный и практический интерес. Традиционные алгоритмы, такие как логистическая регрессия, методы опорных векторов и ансамбли решающих деревьев, долгое время оставались основой для автоматической классификации текстов. Однако с появлением трансформерных архитектур, таких как BERT, RoBERTa и GPT, возможности обработки естественного языка существенно расширились[1], что требует комплексного анализа их производительности в сравнении с классическими подходами.

Актуальность данного исследования обусловлена необходимостью выявления оптимальных методов для многоклассовой классификации, учитывающих как точность предсказаний, так и вычислительную эффективность. В работе проводится сравнительный анализ производительности традиционных алгоритмов машинного обучения и современных трансформерных моделей на различных наборах текстовых данных. Результаты исследования могут быть использованы для выбора наиболее подходящих методов в зависимости от специфики задачи, объема данных и доступных вычислительных ресурсов.

Целью данной работы является сравнительный анализ производительности классических методов машинного обучения и современных трансформерных моделей в задаче многоклассовой классификации текстовых данных. Исследование направлено на оценку эффективности различных алгоритмов по ключевым метрикам точности,

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention is all you need. [Электронный ресурс]. Режим доступа: https://arxiv.org/abs/1706.03762, свободный. – Загл. с экрана. – Яз. англ

скорости обработки и масштабируемости. Полученные результаты позволят определить оптимальные подходы для решения задач автоматической категоризации контента в условиях различных требований к вычислительным ресурсам и качеству классификации.

Практическая значимость исследования заключается в сравнительном анализе эффективности различных подходов к многоклассовой классификации, что позволяет обоснованно выбирать оптимальные методы для конкретных прикладных задач. Полученные результаты дают возможность:

1. Рационально подходить к выбору между классическими ML-методами и трансформерными моделями, учитывая требования к точности и вычислительным ресурсам
2. Оптимизировать процессы автоматической категоризации контента в медиаиндустрии
3. Повышать качество пользовательского опыта за счет более точной классификации данных

Проведенный анализ имеет универсальное значение и может быть применен:

- При разработке систем автоматической обработки текстов для медиаплатформ
- Для решения широкого круга задач многоклассовой классификации за пределами медиа-сферы
- В качестве методической основы для выбора архитектур моделей при работе с текстовыми данными

Исследование создает основу для дальнейшей работы по оптимизации и адаптации алгоритмов классификации под специфические требования различных предметных областей.

В рамках исследования применяются современные методы обработки естественного языка, включая предварительный анализ текстовых данных, алгоритмы многоклассовой классификации и нейросетевые архитектуры.

Эффективность моделей оценивалась по метрикам точности accuracy, F1-меры и другим показателям качества классификации[2]. Для выявления значимых признаков и интерпретации результатов использовались методы анализа важности признаков и ошибок классификации.

Проведенная оценка производительности классических методов машинного обучения и трансформерных моделей выявила существенные различия в их эффективности при решении задачи многоклассовой классификации. Наибольшую точность (accuracy = 96,65%) продемонстрировала классическая логистическая регрессия, что свидетельствует о высокой линейной разделимости классов в признаковом пространстве для рассматриваемой задачи. Данный результат превосходит показатели как ансамблевых методов, так и нейросетевых архитектур.

Ансамблевые алгоритмы (LightGBM и XGBoost) показали сопоставимую производительность на уровне 93% точности, подтвердив свою надежность для задач классификации. Однако их неспособность превзойти по эффективности логистическую регрессию указывает на возможную избыточную сложность для данного конкретного случая. Особого внимания заслуживает низкая результативность трансформерной модели BERT (accuracy = 30,75%), что требует дополнительного исследования факторов, повлиявших на столь низкую производительность.

Ключевые выводы исследования:

1. Простые классические методы могут демонстрировать превосходство над сложными современными архитектурами в задачах с высокой линейной разделимостью классов[3]
2. Все рассмотренные подходы сталкиваются с проблемой классификации миноритарных классов[4]

---

[2] Соловьева Я.В., Некрасова А.С. ИССЛЕДОВАНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ// Труды Международной научно-технической конференции «Перспективные информационные технологии», 2022, стр. 225

[3] Mario Graff, Daniela Moctezuma, Eric S. Téllez. Bag-of-Word approach is not dead: A performance analysis on a myriad of text classification challenges// Natural Language Processing Journal 11 (2025) 100154, pp.1-13

[4] Талғатұлы, О. О., Каиргельдиевич, С. Е., & Талапқызы, Т. А. (2021). ПРОБЛЕМА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ в МАШИННОМ ОБУЧЕНИИ». [Электронный ресурс]. Режим

3. Выбор модели должен основываться на комплексном анализе характеристик данных, а не только на репутации алгоритмов[5]

Полученные результаты подчеркивают важность сравнительного анализа при выборе методов классификации.

Литература

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention is all you need. [Электронный ресурс]. Режим доступа: https://arxiv.org/abs/1706.03762, свободный. – Загл. с экрана. – Яз. англ.

2. Соловьева Я.В., Некрасова А.С. ИССЛЕДОВАНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ// Труды Международной научно-технической конференции «Перспективные информационные технологии», 2022, стр. 225

3. Талғатұлы, О. О., Каиргельдиевич, С. Е., & Талапқызы, Т. А. (2021). ПРОБЛЕМА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ в МАШИННОМ ОБУЧЕНИИ». [Электронный ресурс]. Режим доступа:https://cyberleninka.ru/article/n/problema-mnogoklassovoy-klassifikatsii-v-mashinnom-obuchenii, свободный. – Загл. с экрана. – Яз. рус.

доступа:https://cyberleninka.ru/article/n/problema-mnogoklassovoy-klassifikatsii-v-mashinnom-obuchenii, свободный. – Загл. с экрана. – Яз. рус.

[5] Шерстнев, П., Липинский, Л. (2021). ВЫЧИСЛЕНИЕ ВЕКТОРА ДОКУМЕНТА с ИСПОЛЬЗОВАНИЕМ МЕРЫ TF-IDF. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/vychislenie-vektora-dokumenta-s-ispolzovaniem-mery-tf-idf, свободный. – Загл. с экрана. – Яз. рус.

4. Шерстнев, П., Липинский, Л. (2021). ВЫЧИСЛЕНИЕ ВЕКТОРА ДОКУМЕНТА с ИСПОЛЬЗОВАНИЕМ МЕРЫ TF-IDF. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/vychislenie-vektora-dokumenta-s-ispolzovaniem-mery-tf-idf, свободный. – Загл. с экрана. – Яз. рус.

5. Mario Graff, Daniela Moctezuma, Eric S. Téllez. Bag-of-Word approach is not dead: A performance analysis on a myriad of text classification challenges// Natural Language Processing Journal 11 (2025) 100154, pp.1-13

Literature

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention is all you need. [Электронный ресурс]. Режим доступа: https://arxiv.org/abs/1706.03762, свободный. – Загл. с экрана. – Яз. англ.

2. Solovieva Y.V., Nekrasova A.S. ISSUE OF METHODS OF TEXT CLASSIFICATION IN NATURAL LANGUAGE// Proceedings of the International Scientific and Technical Conference “Perspective Information Technologies”, 2022, p. 225.

3. Talgatuly, O. O., Kairgeldievich, S. E., & Talapkyzy, T. A. (2021). PROBLEM OF MULTI-CLASS CLASSIFICATION IN MACHINE TRAINING.” [Electronic resource]. Access mode:https://cyberleninka.ru/article/n/problema-mnogoklassovoy-klassifikatsii-v-mashinnom-obuchenii, free. - Zagl. from the screen. - Russian language.

4. Sherstnev, P., Lipinskii, L. (2021). Computation of the DOCUMENT VECTOR using the TF-IDF measure. [Electronic resource]. Access mode: https://cyberleninka.ru/article/n/vychislenie-vektora-dokumenta-s-ispolzovaniem-mery-tf-idf, free. - Zagl. from the screen. - Russian language.

5. Mario Graff, Daniela Moctezuma, Eric S. Téllez. Bag-of-Word approach is not dead: A performance analysis on a myriad of text classification challenges// Natural Language Processing Journal 11 (2025) 100154, pp.1-13