УДК 004.272.3; 519.217.2

РАЗРАБОТКА ГИБРИДНОГО МЕТОДА ОБНАРУЖЕНИЯ ФЕЙКОВЫХ НОВОСТЕЙ С ПОМОЩЬЮ МОДЕЛИ BERT

Лобов А.С., студент гр. 4411, IV курс
Научный руководитель: Сотников С.В., к.т.н., доцент
Казанский национальный исследовательский технический
университет им. А.Н. Туполева – КАИ,
г. Казань

В современном мире стремительного развития цифровых технологий наблюдается рост объёмов распространения информации, что приводит к увеличению количества недостоверного контента. Фейковые новости являются серьезной краткосрочной угрозой, так как они оказывают деструктивное влияние на общественное мнение, политическую и социальную стабильность [1]. В русскоязычном сегменте интернета фиксируется рост недостоверных публикаций на 30-50%, что особенно критично в период геополитической напряженности [2]. Традиционные подходы, такие как ручной фактчекинг или анализ контента по ключевым словам, демонстрируют низкую эффективность против современных, лингвистически сложных форм дизинформации.

Существуют иные способы обнаружения фейковых новостей, среди которых присутствуют как базовые алгоритмы классификации, так и с использованием современных методов машинного обучения и обработки естественного языка. Однако традиционные подходы имеют ряд ограничений, связанный с ручной инженерией признаков, неспособностью учитывать контекстные и семантические особенности текста, а технологии с применением NLP требуют колоссальных мощностей и используют закрытые датасеты, что делает данный подход малодоступным для массового пользователя.

В данной статье я хочу уделить особое внимание трансформерной модели BERT (Bidirectional Encoder Representations from Transformers), которая демонстрирует высокую точность в задачах классификации текста благодаря способности учитывать двунаправленный контекст и извлекать глубокие семантические представления [3 - 4]. На основе анализа научных публикаций показано, что модели на базе BERT превосходят традиционные подходы и являются перспективным направлением для решения задачи обнаружения дезинформации.

Суть гибридного подхода состоит в сочетании возможностей BERT для обработки текстовых данных с дополнительными структурированными признаками, извлеченными из метаданных и лингвистических характеристик. Такой подход позволяет объединить преимущества глубокого обучения и интерпретируемых признаков, повышая качество классификации и устойчивость модели.

В качестве экспериментальной базы используется открытый датасет LIAR [5] (Wang W., 2017), содержащий размеченные политические высказывания с различными уровнями достоверности. Рассматриваемый набор данных отличается наличием метаданных, включая информацию о спикерах и контексте высказываний, что делает его подходящим для построения гибридных моделей. В рамках исследования выполнена предобработка данных, инженерия признаков и обучение моделей, а также проведено сравнительное тестирование предложенного подхода с базовыми алгоритмами.

В рамках исследования был проведен анализ существующих подходов к обнаружению фейковых новостей, включающий как традиционные алгоритмы, так и современные модели глубокого обучения [6 - 9]. Основное внимание уделялось их способности учитывать контекст, семантическую структуру текста, а также вычислительным затратам и интерпретируемости. Сравнительные характеристики рассмотренных моделей представлены в таблице 1.

Модель	Тип	Преимущества	Недостатки
Logistic Regression	Классическая	Простота, интерпретируемость, низкие затраты	Не учитывает контекст, требует инженерии признаков
SVM	Классическая	Эффективна в высокоразмерных данных	Плохо масштабируется, нет учета контекста
Random Forest	Классическая	Устойчивость к шуму, важность признаков	Не понимает семантику текста
XGBoost	Классическая	Высокая точность, работа с нелинейностями	Зависимость от признаков
AdaBoost	Классическая	Хорошо на структурированных данных	Чувствителен к шуму
CNN	Глубокое обучение	Очень быстрая работа; хорошо находит специфические маркеры фейков	Игнорирует глобальный контекст предложения и порядок слов
LSTM	Глубокое обучение	Учитывает порядок слов и грамматическую структуру	Деградация информативности начальных токенов при увеличении длины входной

			последовательности; медленное обучение
BERT	Глубокое обучение	Лучшее понимание контекста и скрытого смысла	Требует больших вычислительных мощностей (GPU)

Таблица 1. Сравнение методов обнаружения фейковых новостей

Анализ представленных методов показывает, что традиционные алгоритмы уступают современным моделям глубокого обучения в способности учитывать контекст и семантические особенности текста. В то же время способы с применением глубокого обучения несовершенны и их результативность можно связать с качественно подготовленными датасетами, чьё содержимое не доступно базовому пользователю.

В исследовании применяется датасет LIAR в качестве экспериментальной базы ввиду его открытости, что обеспечивает прозрачность и воспроизводимость результатов, а также позволяет гибко модифицировать данные для более полного раскрытия возможностей моделей BERT в условиях ограниченных вычислительных ресурсов.

Ниже, в таблице 2, представлен итоговый набор данных для обучения модели на базе LIAR. Выбранные поля включают как текстовую информацию, используемую моделью BERT для извлечения контекстных представлений, так и дополнительные признаки, отражающие метаданные и статистические характеристики спикеров.

№	Имя	Описание
1	label	Бинарная классификация для определения достоверности новости.
2	combined_text	Очищенное и улучшенное текстовое представление каждого образца.
3	fake_ratio	Доля слов, характерных для фейковых новостей, относительно общего количества слов в тексте.
4	is_disappointed	Бинарный признак: присутствие слов, выражающих разочарование.
5	is_conspiracy	Бинарный признак: присутствие слов, связанных с теориями заговора.
6	is_political	Бинарный признак: присутствие политически окрашенной лексики.
7	is_shocking	Бинарный признак: присутствие слов, вызывающих эмоциональный шок.
8	is_propaganda	Бинарный признак: присутствие слов, характерных для пропагандистских приёмов.

9	num_entities	Количество именованных сущностей (люди, организации, локации), извлечённых из текста.
10	text_length	Общая длина текста.
11	word_count	Количество слов в объединённом тексте.
12	exclamation_count	Количество восклицательных знаков (!) в тексте.
13	question_count	Количество вопросительных знаков (?) в тексте.
14	quote_count	Количество кавычек («) в тексте.

Таблица 2. Поля датасета после предобработки и инженерии признаков

На основе сформированного набора данных был реализован гибридный алгоритм обнаружения фейковых новостей, сочетающий использование языковой модели BERT и дополнительных структурированных признаков (Рисунок 1). Архитектура модели включает два входных канала: текстовый, обрабатываемый BERT для извлечения контекстных и семантических представлений, и табличный, содержащий метаданные и статистические характеристики спикера.

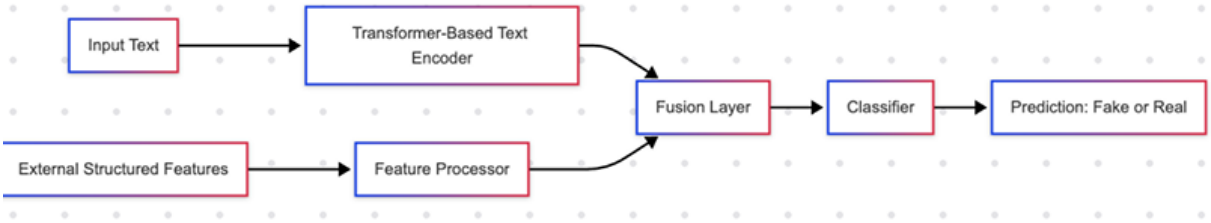

Рисунок 1. Схема работы алгоритма

Дополнительные признаки включают информацию о достоверности спикера, тематике высказывания и контексте его публикации, а также агрегированные показатели предыдущих утверждений. Это позволяет учитывать не только содержание текста, но и внешние факторы, влияющие на достоверность информации.

Объединение признаков осуществляется на этапе классификации, что обеспечивает интеграцию глубинных текстовых представлений и интерпретируемых характеристик. Такой подход позволяет повысить устойчивость модели к различным типам входных данных и улучшить обобщающую способность.

В результате применения данного алгоритма и традиционных способов на структурированном датасете гибридный подход продемонстрировал повышенную точность (Таблица 3), что подтверждает эффективность использования данной архитектуры для задачи обнаружения фейковых новостей.

Модель	Accuracy	Precision	Recall	F1 Score	AUC-ROC
XLM-RoBERTa	0.7443	0.6880	0.7577	0.7212	0.8305

XGBoost	0.7451	0.7386	0.6438	0.6879	0.8186
Logistic Regressio	0.7459	0.7362	0.6510	0.6910	0.8281
Random Forest	0.6093	0.7636	0.1519	0.2534	0.7665
SVM	0.7459	0.7584	0.6130	0.6780	0.8188
GradientBoosting	0.7482	0.7448	0.6438	0.6906	0.8217
MLP	0.7522	0.7357	0.6745	0.7038	0.8266

Таблица 3. Результаты тестирования BERT и эталонных алгоритмов

Список литературы

1. Global Risks Report 2024 [Электронный ресурс] / World Economic Forum. – Geneva, 2024. – 124 p. – URL: https://www.weforum.org/publications/global-risks-report-2024/ (дата обращения: 20.03.2026).
2. АНО «Диалог» : раздел исследований [Электронный ресурс]. – URL: https://dialog.info/ (дата обращения: 20.03.2026).
3. BERT (language model) [Электронный ресурс] // Wikipedia. – URL: https://en.wikipedia.org/wiki/BERT_(language_model) (дата обращения: 20.03.2026).
4. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // arXiv. – 2018. – 24 May. – URL: https://arxiv.org/abs/1810.04805 (дата обращения: 20.03.2026).
5. LIAR Dataset [Электронный ресурс] // GitHub. – URL: https://github.com/tfs4/liar_dataset (дата обращения: 21.03.2026).
6. Chen, T. XGBoost: A Scalable Tree Boosting System / T. Chen, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – San Francisco, 2016. – P. 785–794. – DOI: 10.1145/2939672.2939785.
7. Breiman, L. Random Forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, no. 1. – P. 5–32. – DOI: 10.1023/A:1010933404324.
8. FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection / [authors] // PMC. – 2023. – Article ID PMC10691701. – URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10691701/ (дата обращения: 21.03.2026).
9. An Empirical Comparison of Machine Learning and Deep Learning Models for Automated Fake News Detection / [authors] // MDPI. – 2025. – URL: https://www.mdpi.com/journal/ (дата обращения: 21.03.2026).