УДК 004.912

АЛГОРИТМЫ ДОВЕРИЯ: ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ БОРЬБЫ С ФАЛЬШИВЫМИ НОВОСТЯМИ О COVID-19

Хохлачева Д.С. (Делийский университет, г. Дели, Индия)
Научный руководитель - Афанасьева Т.С., кандидат медицинских наук, ассистент кафедры пропедевтики внутренних болезней **(**ФГБОУ ВО Ижевская государственная медицинская академия МЗ РФ, г. Ижевск, РФ)

Введение: Цифровизация общества и активное использование социальных сетей привели к новым вызовам в научной коммуникации. Одной из наиболее острых проблем стало распространение фальшивых новостей (*fake news*), особенно во время кризисных ситуаций, таких как пандемия COVID-19. По данным ряда исследований [1], некорректная информация о COVID-19 не только искажала общественное восприятие пандемии, но и приводила к необоснованным решениям со стороны населения — от отказа от вакцинации до использования сомнительных методов лечения [2]. Несмотря на то, что научная коммуникация традиционно опирается на рецензируемые публикации, официальные доклады и научно-популярные издания, с развитием социальных сетей информация стала распространяться быстрее, а контроль за ее достоверностью — ослабевать. Из вышесказанного следует, что современные методы научной коммуникации нуждаются в надежных инструментах фильтрации и проверки информации. В этой связи искусственный интеллект (AI) и глубокие нейронные сети представляют собой мощные средства для выявления фейковых новостей и поддержания достоверности научных данных в цифровой среде [3, 4].

Целью данной работы является исследование возможностей глубокого обучения в контексте научной коммуникации и борьбы с дезинформацией в сфере общественного здравоохранения.

Основная часть: В рамках исследования были протестированы различные методы машинного обучения для детекции фейковых новостей. Был использован размеченный датасет CONSTRAINT@AAAI2021, содержащий посты из социальных сетей, которые классифицированы как «фейковые» или «реальные» [5].

Результаты проведенных экспериментов продемонстрировали высокий потенциал современных методов глубокого обучения для решения задачи классификации постов в социальных сетях. Модель Bidirectional LSTM показала точность 85.09%, эффективно справляясь с анализом длинных и контекстно сложных текстов за счет двунаправленного учета последовательности слов. Облегченная версия BERT, модель DistilBERT, достигла точности 94.72%, сохраняя преимущество BERT — двунаправленное понимание текста - при относительно низких вычислительных затратах. Наивысший показатель точности — 95.23% — был достигнут полноформатной моделью BERT, что обусловлено глубокой архитектурой модели и её способностью учитывать широкий контекст сообщений. Для сравнения был использован традиционный метод SVM, который после оптимизации гиперпараметров продемонстрировал точность 92.76%. Несмотря на конкурентоспособность, метод SVM уступает в результативности современным нейронным сетям, что подчеркивает преимущество использования методов глубокого обучения в задачах детекции фейковых новостей [6].

Для дальнейшего практического применения подобных решений предлагается несколько направлений развития:

1. Интеграция AI в социальные сети – автоматическая проверка постов и маркировка недостоверной информации (например, Google Fact Check API [7]).
2. Фильтрация новостей – использование AI-алгоритмов ранжирования, учитывающих надежность источников и репутацию авторов.
3. Поддержка медицинских организаций – AI-боты на основе BERT для проверки медицинских публикаций, быстрый анализ достоверности исследований.
4. Фактчекинговые платформы – разработка онлайн-сервисов для анализа публичных дискуссий (например, EARS от ВОЗ [8]) и интеграция AI в фактчекинговые базы (Meedan, Snopes, и др).

Выводы:

1. AI, основанный на глубоких нейросетях, является перспективным инструментом для выявления фейковых новостей в научной коммуникации.
2. В данном исследовании модели BERT и DistilBERT показали высокую точность (более 94.3%), превосходя традиционные ML-методы.
3. Для эффективной борьбы с дезинформацией требуется интеграция AI в существующие платформы научной коммуникации, включая соцсети и новостные агрегаторы.
4. Будущее научной коммуникации лежит в синергии AI и фактчекинга: автоматизация анализа совместно с экспертным контролем.

Список использованных источников

[1] Nelson T., Kagan N., Critchlow C., Hillard A., Hsu A. The Danger of Misinformation in the COVID-19 Crisis. Mo Med. 2020 Nov-Dec; 117(6):510-512. PMID: 33311767; PMCID: PMC7721433.

[2] Nyilasy G. Fake news in the age of COVID-19. The University of Melbourne. 10 апреля 2020 г. Доступно по ссылке: https://pursuit.unimelb.edu.au/articles/fake-news-in-the-age-of-covid-19 (дата обращения: 07.02.2025).

[3] Shu K., Sliva A., Wang S., Tang J., Liu H. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter. 2017; 19(1). Доступно по ссылке: https://doi.org/10.48550/arXiv.1708.01967 (дата обращения: 07.02.2025).

[4] Felber T. Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task. arXiv preprint, 2021. Доступно по ссылке: https://arxiv.org/abs/2101.03717 (дата обращения: 07.02.2025).

[5] Constraint@AAAI2021 – COVID-19 Fake News Detection in English. Организовано пользователем parthpatwa на платформе CodaLab. Доступно по ссылке: https://competitions.codalab.org/competitions/26655 (дата обращения: 07.02.2025).

[6] GitHub – Fake News Classification. Программная реализация исследования доступна по ссылке:https://github.com/Pineappledeydre/fake_news_classification (дата обращения: 07.02.2025).

[7] Google Fact Check Tools API. Инструмент для автоматизированного поиска и обработки проверок фактов с использованием ClaimReview и FactCheck Claim Search API. Доступно по ссылке: https://developers.google.com/fact-check/tools (дата обращения: 07.02.2025).

[8] Всемирная организация здравоохранения (ВОЗ). WHO launches pilot of AI-powered public-access social listening tool [Электронный ресурс]. 29 января 2021 г. Доступно по ссылке:https://www.who.int/news/item/29-01-2021-who-launches-pilot-of-ai-powered-public-access-social-listening-tool (дата обращения: 07.02.2025).