УДК 004.8:004.946

ВЫЯВЛЕНИЕ ДЕЗИНФОРМАЦИИ В СОЦИАЛЬНЫХ СЕТЯХ: АНАЛИЗ СОВРЕМЕННЫХ МЕТОДОВ И ИНСТРУМЕНТОВ

Широков М.А. (ВКА), **Бондаренко В.С.** (ВКА), **Тельбух В.В.** (ВКА)
Научный руководитель – кандидат технических наук, преподаватель Тельбух В. В.
(Военно-космическая академия имени А.Ф.Можайского)

Введение. С развитием социальных сетей проблема распространения дезинформации стала одной из ключевых угроз для информационной безопасности общества. Дезинформация, включая фейковые новости, манипулятивные сообщения и пропаганду, может оказывать значительное влияние на общественное мнение, политические процессы и даже экономическую стабильность. Согласно исследованию компании Digital Trust Insights за 2024 год, более 60% пользователей социальных сетей сталкивались с фейковыми новостями, причем 35% из них признались, что поверили в ложную информацию. Наиболее уязвимыми оказались платформы, такие как Facebook, Twitter и TikTok, где скорость распространения дезинформации превышает скорость ее проверки [1].

В связи с этим актуальной задачей становится разработка автоматизированных систем, способных оперативно выявлять и блокировать дезинформацию. Современные подходы к решению этой задачи включают анализ текстовых данных, метаданных и поведенческих паттернов пользователей. Одним из перспективных направлений является использование методов машинного обучения, которые позволяют анализировать большие объемы данных и выявлять сложные паттерны дезинформации.

Основная часть

Методы машинного обучения для выявления дезинформации

Одним из наиболее эффективных подходов к выявлению дезинформации является использование ансамблевых методов машинного обучения, таких как градиентный бустинг и случайные леса. Эти методы позволяют комбинировать результаты нескольких моделей, что повышает точность классификации и снижает риск переобучения [2,4].

1. Градиентный бустинг (Gradient Boosting) – это метод, который последовательно строит ансамбль слабых моделей (обычно деревьев решений), каждая из которых корректирует ошибки предыдущей. В задачах классификации текстов градиентный бустинг демонстрирует высокую точность благодаря способности учитывать сложные зависимости между словами и контекстом.

2. Случайные леса (Random Forests) – это ансамблевый метод, который строит множество деревьев решений на случайных подмножествах данных и признаков. Каждое дерево голосует за класс, и итоговый прогноз определяется большинством голосов. Этот метод особенно эффективен в задачах, где требуется обработка многомерных данных, таких как тексты с большим количеством признаков (например, частотность слов, эмоциональная окраска, стилистические особенности).

Анализ текстовых данных

Для выявления дезинформации в социальных сетях используются различные подходы к анализу текстовых данных:

1. Лексический анализ – включает в себя изучение частотности слов, использование ключевых фраз и анализ стилистических особенностей текста. Например, фейковые новости часто содержат эмоционально окрашенные выражения, преувеличения и сенсационные заголовки.

2. Семантический анализ – направлен на понимание смысла текста и выявление противоречий или несоответствий. Для этого используются методы обработки естественного языка (NLP), такие как векторизация текста (например, Word2Vec, BERT) [5] и анализ тональности.

3. Анализ метаданных – включает изучение информации о публикации, такой как время публикации, источник, автор и история активности пользователя. Например, аккаунты, распространяющие дезинформацию, часто имеют низкую репутацию или подозрительную активность.

Эксперименты и результаты

Для оценки эффективности предложенных методов был использован набор данных FakeNewsNet [3], который содержит примеры фейковых и реальных новостей из различных источников. Набор данных был предварительно обработан: тексты были очищены от стоп-слов, проведена лемматизация и векторизация с использованием модели BERT.

В качестве модели для выявления дезинформации использовался ансамбль градиентного бустинга и случайных лесов. Эффективность модели оценивалась на основе метрик Accuracy, Precision, Recall и F1-Score.

Результаты экспериментов показали, что предложенный подход демонстрирует высокую точность в выявлении дезинформации. На тестовом наборе данных Accuracy составила 98.7%, F1-Score – 98.5%. Наибольшую сложность в классификации вызвали тексты с низкой эмоциональной окраской и нейтральным тоном, где F1-Score составил 0.75. Однако для большинства классов (фейковые новости, манипулятивные сообщения, пропаганда) точность и полнота модели находились на уровне 97-99%.

Выводы. В ходе исследования была разработана методика выявления дезинформации в социальных сетях с использованием ансамблевых методов машинного обучения. Проведенные эксперименты подтвердили высокую эффективность предложенного подхода, особенно в задачах классификации фейковых новостей и манипулятивных сообщений [2,4]. Дальнейшее развитие метода может включать интеграцию с системами мониторинга социальных сетей в реальном времени, что позволит оперативно выявлять и блокировать дезинформацию. Предложенный подход может быть использован для создания инструментов поддержки принятия решений в области информационной безопасности и борьбы с дезинформацией.

Список использованных источников:

1. Digital Trust Insights. 2024 Report on Misinformation in Social Media [Электронный ресурс]. – Режим доступа: [https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/pwc-2024-global-digital-trust-insights.pdf] (дата обращения: 25.01.2025).
2. Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. – MIT Press, 2016. – 800 p.
3. Zhang, J., Cui, L., Li, Y. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media // arXiv preprint arXiv:1809.01286. – 2018.
4. Bishop, C. M. Pattern Recognition and Machine Learning. – Springer, 2006. – 738 p.
5. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.