Исследование и применение методов машинного обучения для решения обратной задачи определения дезинформации

Турарбек А., Нарбаева С.М., Нургали А.А., Алиев Р.Ш., Арапова Ж.Е.
КазНУ им. аль-Фараби, Алматы, Казахстан
turarbek_ asem@mail.ru

В исследовании рассматривается распространение дезинформации в социальных сетях, которая в последние десятилетия стала одной из наиболее актуальных угроз информационной безопасности общества. Особенно остро эта проблема проявляется в условиях чрезвычайных ситуаций (ЧС), таких как землетрясения, наводнения, пандемии и техногенные катастрофы. Во время ЧС резко возрастает потребность населения в достоверной информации, однако одновременно усиливается распространение слухов, панических сообщений и преднамеренной дезинформации, что затрудняет принятие оперативных решений и подрывает доверие к официальным источникам.

В научных исследованиях выявления дезинформации в социальных сетях прошло несколько стадий развития. На ранних этапах применялись простые эвристические методы, основанные на ключевых словах и статистических признаках сообщений. Позднее начали использоваться методы краудсорсинга и полуавтоматические системы аннотации, такие как AIDR и MicroMappers [1]. С развитием машинного обучения (ML) и глубоких нейронных сетей акцент сместился в сторону алгоритмов классификации текстов, анализа сетевых графов распространения информации, а также применения трансформеров (BERT, RoBERTa, XLM-R) для выявления скрытых семантических связей [2], [3].

Обратная задача заключается в оценке достоверности сообщений и выявлении источников дезинформации с применением методов машинного обучения и статистики.

Определение дезинформации можно рассматривать как обратную задачу: по наблюдаемым признакам текста и взаимодействий пользователей требуется установить вероятность того, что сообщение является ложным. Пусть множество сообщений задано как $D = d_1, \ldots, d_n$, где каждое d_i описывается вектором признаков x_i. Необходимо построить отображение $f : X \to Y$, где $Y = 0, 1$ (0 - достоверное сообщение, 1 - дезинформация), и оценить вероятность $P(y = 1|x)$.

Исследование предполагает сбор и аннотацию данных из социальных сетей (Twitter, Telegram, Facebook) и локальных новостных источников, связанных с ЧС. Тексты подвергаются предобработке (токенизация, лемматизация, удаление стоп-слов), после чего используются различные алгоритмы машинного обучения, такие как логистическая регрессия, SVM и случайный лес. На уровне глубокого обучения - рекуррентные сети (LSTM, BiLSTM) и трансформерные архитектуры (BERT, RoBERTa, XLM-R). Классификация осуществляется по принципу выбора наиболее вероятного класса: $\widehat{y} = \arg\max_{c \in C} P(c|x; \theta)$. Для восстановления скрытых параметров источников дезинформации и закономерностей её распространения применяются графовые модели и эпидемиологические подходы (SIR-модели).

Решение обратной задачи выявления дезинформации требует комплексного подхода, объединяющего методы машинного обучения, NLP и анализа социальных сетей. Практическая значимость заключается в возможности интеграции таких систем в мониторинговые сервисы, что особенно важно в условиях ЧС для фильтрации панических слухов и поддержки принятия решений.

Работа выполнена при поддержке Комитета науки Министерства науки и высшего образования Республики Казахстан (грантовый проект AP 26197729).

Список литературы

1. *Imran, M., Castillo, C., Diaz, F., Vieweg, S.* Processing social media messages in mass emergency // A survey. ACM Computing Surveys. 2015.
2. *Castillo, C., Mendoza, M., Poblete, B.* Information credibility on Twitter// WWW '11. 2011.
3. *Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S.* AIDR: Artificial intelligence for disaster response // In Proceedings of the 23rd international conference on world wide web. 2014. Т. 1. № 2. С. 159–162.