

**ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ С ЦЕЛЬЮ
АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ КОНТР-НАРРАТИВОВ
В СИСТЕМЕ АНАЛИЗА И БЛОКИРОВКИ
ДЕСТАБИЛИЗИРУЮЩЕГО КОНТЕНТА В СОЦСЕТЯХ**

Демьяненко Г.В. (ВКА), Миронов И.В. (ВКА), Дудкин А.С. (ВКА)

**Научный руководитель – кандидат технических наук, доцент Дудкин А.С.
(Военно-космическая академия имени А.Ф.Можайского)**

Введение. Развитие технологий искусственного интеллекта предоставляет эффективные инструменты для анализа и обработки текстовых данных, включая обнаружение и нейтрализацию дискредитирующего контента. Особый интерес представляет применение моделей машинного обучения для автоматической генерации аргументированных контр-нарративов, направленных на опровержение враждебных высказываний.

В рамках создания программного комплекса, способного противодействовать деструктивной информации, методы машинного обучения играют ключевую роль, обеспечивая анализ содержания текстовых данных, выявление наиболее агрессивных сообщений и их классификацию, а также автоматическую генерацию контраргументов. Данный подход позволяет формировать конструктивный диалог, предоставляя пользователям альтернативные точки зрения, подкрепленные фактами и логикой [1].

Основная часть. К Ключевым компонентом разрабатываемого программного комплекса являются алгоритмы машинного обучения, обеспечивающие:

1. Сбор и подготовку данных. На данном этапе:
 - формируется корпус текстов, включающий дискредитирующие высказывания, аргументированные ответы и нейтральные сообщения;
 - используются технологии предобработки текстов, такие как очистка данных от шума, исправление ошибок и устранение избыточной информации [2].
2. Классификацию текстов:
 - анализ входящих сообщений на основе модели BERT или её адаптированной версии RuBERT-tiny2 [3];
 - модель классифицирует текст по степени агрессивности и типу позиции (проукраинская, пророссийская, нейтральная) [4].
3. Генерацию контраргументов:
 - используются большие языковые модели (GPT, T5, Llama), прошедшие fine-tuning на данных, содержащих примеры аргументированных опровержений [5];
 - модель анализирует текст с дискредитирующим содержанием и на его основе генерирует логичный и убедительный контраргумент [6].
4. Технические аспекты реализации:
 - генератор контраргументов реализуется в виде микросервиса, что обеспечивает модульность и гибкость системы [7];
 - управление нагрузкой и масштабируемостью с помощью Docker-контейнеров и платформы оркестрации Kubernetes [8].
5. Оптимизацию и оценку качества:
 - настройка параметров генерации (temperature, top_k, top_p) позволяет улучшить качество создаваемых ответов [9];
 - качество текста оценивается с помощью метрик BLEU, ROUGE и перплексии [10].

Применение алгоритмов машинного обучения в генерации контраргументов позволяет оперативно и эффективно формировать ответы на деструктивные сообщения, обеспечивая их точность, релевантность и убедительность [11].

Выводы. Современные технологии искусственного интеллекта предоставляют мощный инструмент для противодействия дискредитирующему интернет-контенту,

автоматически формируя аргументированные опровержения. Интеграция таких решений в разрабатываемый программный комплекс позволит не только выявлять деструктивные сообщения, но и обеспечивать объективность и конструктивность диалога в цифровом пространстве [12].

Реализация предложенного подхода повышает эффективность информационного противодействия, создавая условия для защиты национальных интересов государства и укрепления позиций в области информационной безопасности [13].

Список использованных источников:

1. Васильев, В. В., Иванов, Д. С. Применение машинного обучения для анализа текстовых данных в социальных сетях // Вестник информационной безопасности. – 2022. – № 3. – С. 45–58.
2. Brown, T., Mann, B., Ryder, N., et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems (NeurIPS). – 2020. – Vol. 33. – P. 1877–1901.
3. Devlin, J., Chang, M., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. – 2019.
4. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. Improving Language Understanding by Generative Pre-Training // OpenAI Technical Report. – 2018.
5. Lewis, M., Liu, Y., Goyal, N., et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension // arXiv preprint arXiv:1910.13461. – 2019.
6. Zhang, T., Kishore, V., Wu, F., et al. BERTScore: Evaluating Text Generation with BERT // arXiv preprint arXiv:1904.09675. – 2019.
7. Кудрявцев, С. Л. Автоматическое выявление и противодействие деструктивному контенту в социальных сетях // Компьютерные исследования и моделирование. – 2023. – Т. 15, № 2. – С. 95–110.
8. Xu, J., Sun, X. BERTRank: Learning to Rank with BERT for Query-focused Summarization // arXiv preprint arXiv:1910.03196. – 2019.
9. OpenAI. GPT-4 Technical Report // arXiv preprint arXiv:2303.08774. – 2023.
10. Vaswani, A., Shazeer, N., Parmar, N., et al. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). – 2017. – P. 5998–6008.
11. Чесноков, П. Е. Использование генеративных моделей для формирования контр-нарративов в онлайн-среде // Информационные технологии и безопасность. – 2022. – № 4. – С. 102–117.
12. King, G., Pan, J., Roberts, M. E. How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument // American Political Science Review. – 2017. – Vol. 111, No. 3. – P. 484–501.
13. Лапшин, А. В., Королёв, Д. П. Методы предобработки текстовых данных в системах анализа интернет-контента // Вестник цифровой безопасности. – 2021. – № 2. – С. 68–79.

Демьяненко Г.В. (автор)

Подпись

Миронов И.В. (автор)

Подпись

Дудкин А.С. (автор)

Подпись