Разработка и оценка эффективности автоматизированных методов сбора и анализа данных для выявления и противодействия дезинформационным кампаниям в социальных сетях

К.С. Корнеев, М.С. Спирин, Д.А. Куль

***Аннотация*—Противодействие дезинформационным кампаниям в условиях современной информационной войны являются актуальными. Особое внимание уделяется роли социальных сетей как ключевого инструмента распространения фейков и формирования общественного мнения, что представляет угрозу для социальной стабильности государства. Для решения данной задачи предлагается использование автоматизированных систем анализа открытых источников информации (OSINT) для выявления и нейтрализации дезинформационных кампаний на ранних этапах их развития.**

В статье рассмотрена, аналитическая платформа PolyAnalyst, в которой реализован принцип Low-code разработки, позволяющий автоматизировать сбор, обработку и визуализацию данных из социальной сети Telegram, включая использование методов машинного обучения и элементы нейролингвистического программирования (NLP) для анализа текстов и выявления фейков. Особенность системы заключается в возможности построения графовых моделей для отслеживания механизмов распространения дезинформации, что позволяет определить источники, пути и ключевых участников информационных атак.

***Ключевые слова*—Дезинформация, фейки, социальные сети, OSINT, Polyanalyst, NLP, графовые модели, Low-code, обработка данных.**

I. Введение

Происходящие в мире события, позволяют нам увидеть, как важен контроль за информацией, распространяемой в социальных сетях. Активное распространение фейковой информации – может подорвать основы социальной стабильности государства. Дезинформационные кампании, являются неотъемлемой частью информационной войны [2].

С развитием информационных технологий, информационные войны становятся все более изощренными. Если в XIX веке использовались газеты или листовки, направленные на очернение оппонента, то в современных реалиях, средства и формы, стали более сложными и изощренными.

В наше время представить человека, у которого бы не было того или иного, цифрового источника информации, социальной сети [9].

Уже длительное время Россия вовлечена в информационные войны, с так называемыми «Странами коллективного Запада». Наши противники, кроме традиционных СМИ, активно используют и социальные сети. Они, будучи владельцами различных платформ и ресурсов, могут беспрепятственно продвигать «рекомендательными» алгоритмами, нужную им информацию и формировать общественное мнение как внутри страны, так и за ее пределами.

В истории стран, окружающих Россию, достаточно примеров, когда государство не смогло или не захотело остановить иностранное влияние на свое информационное поле, итогом этого стали перевороты: 2003 год «Революция роз» в Грузии, 2004 год «Оранжевая революция» на Украине, 2004 год «Тюльпановая революция» в Киргизии, 2013-2014 года «Евромайдан» на Украине.

Главная задача наших оппонентов – ослабление России изнутри и на международном уровне, создавать образ «отсталой» страны, на международной арене. Единственным выходом из сложившейся ситуации – это грамотно проработанная стратегия противодействия.

Одним из механизмов противодействия может являться поиск и выявление источников дезинформационной кампании на ранней стадии, благодаря автоматизированным системам OSINT (open source intelligence), разведка по открытым данным [3].

Разработка отечественной автоматизированной системы OSINT для выявления дезинформационных кампаний, обусловлена необходимостью, своевременного поиска и выявления очага, вызывающего общественные бурления, направленные против государства [6].

Для решения данной задачи, выполнен проект, получивший свою реализацию на базе отечественной платформы «PolyAnalyst».

Статья получена 29 января 2026 г.
К.С. Корнеев младший научный сотрудник ВИТ «ЭРА» (email: era_op_open@mil.ru);
М.С. Спирин начальник отдела (научно-исследовательского) ВИТ «ЭРА» (email: era_op_open@mil.ru);
Д.А. Куль старший оператор ВИТ «ЭРА» (email: era_op_open@mil.ru);

II. Анализ социальных сетей

Проект представляет собой гибкую систему, созданную в аналитической среде визуального программирования PolyAnalyst, способную существенно ускорить поиск источника дезинформации [4]. В рамках проекта решались задачи сбора, обработки и визуализации информации, полученных из социальной сети Telegram [5].

Система состоит из более чем 20 функциональных узлов, частично ее структура представлена на рисунке 1. Первым и ключевым этапом системы является сбор данных. Поскольку Telegram не предоставляет открытого доступа к истории сообщений всех каналов, необходимо использовать программные методы для выгрузки (парсинга) контента. Это можно реализовать через Telegram API с помощью библиотек, таких как Telethon, библиотека языка Python или pyrogram, либо разработав собственного бота, который будет отслеживать заданные каналы и чаты [14].

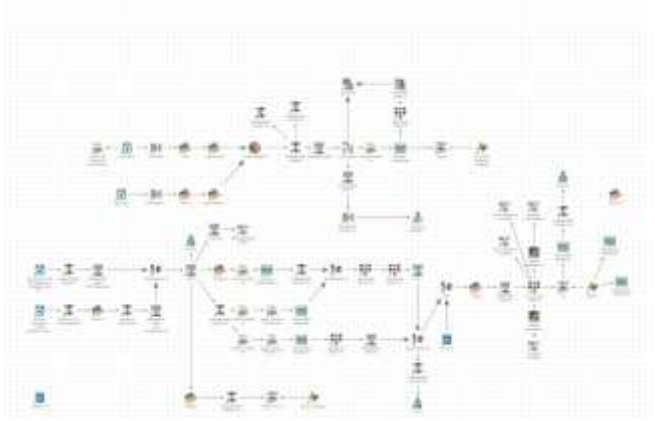

Рис. 1 Сценарий в программной среде PolyAnalyst

Не все каналы представляют интерес для анализа – некоторые могут быть малоактивными, или узкоспециализированными. Поэтому предварительно формируется список целевых источников, которые либо уже были замечены в распространении фейков, или обладают высокой аудиторией, что делает их потенциальными проводниками дезинформации [10]. Для организации выборки можно использовать несколько способов: использовать открытые базы данных или каталоги Telegram-каналов, так же можно использовать постепенное добавление связанных или указанных через репосты каналов, такой метод называется Snowball sampling.

Скрипт для выгрузки сообщений учитывает следующие аспекты:

А) Авторизация: работа с API требует аккаунт с номером телефона, а для больших объемов данных – несколько аккаунтов, для избегания блокировок.

Б) Фильтрация: чтобы не загружать весь контент подряд, можно сразу отсеивать сообщения, по ключевым словам, например рекламные посты [7].

Выгруженные сообщения сохраняются в структурированном виде в JSON файле вместе с метаданными: дата отправки, автор сообщения или поста, количество комментариев, количество просмотров, количество репостов и т.д.

После сбора сырых данных начинается этап обработки, для этого мы используем методы NLP (Natural Language Processing) – токенизация текста, лемматизация, анализируем последовательности слов в предложениях, и методы машинного обучения для выявления потенциальных вбросов и фейков.

В начале с помощью функционала PolyAnalyst мы проводим токенизацию – разбиваем текст на отдельные слова и предложения. Затем, очищаем от лишнего шума, удаляем предлоги, союзы, спецсимволы, эмодзи. Приводим слова к нормальной форме (бежал - бежать).

Следующим шагом мы начинаем процесс классификации текста и его тематическое моделирование [12]. Используя такие NLP-модели, как BERT, NATASHA, RoBERTA, мы определяем тональности «Нейтральная/негативная/положительная» и выделяем ключевые темы. Краткое описание моделей представлено в таблице 1.

Natasha – идеальный выбор, для нашей задачи. Эта модель не требует дополнительного дообучения в отличии от BERT и RoBERTA. Она уже подготовлена под специфику русского языка, а именно – корректно обрабатывает морфологию, распознает имена, даты и другие сущности, справляет опечатки и нормализует текст.

Дезинформация, фейковая информация, часто содержат эмоционально заряженные слова и кликбейтные заголовки, например: «катастрофа, заговор, ужас» и т.п.; «ШОК! Раскрыта правда об…»; также повторы одинаковых тезисов без подтверждения.

Часто бывает так, что фейковые новости начинают распространять не один источник или аккаунт. Чтобы выявлять такие кампании мы можем анализировать общие шаблоны сообщений – одинаковые эмоциональные окраски, одинаковые тезисы, одинаковые картинки.

Визуализировать распространение фейков можно с помощью графовых моделей. После того, как данные собраны и обработаны, наступает этап визуализации, который играет ключевую роль в понимании механизмов распространения дезинформации.

Графовые модели дают возможность увидеть скрытые связи и закономерности, которые невозможно обнаружить при беглом осмотре текстовых данных [11]. Основой анализа становится графовая модель, в которой узлами выступают пользователи, каналы или боты, а связями между ними – различные виды взаимодействий: репосты, упоминания, цитирования, ответы в обсуждениях и чатах. Чем активнее узел, тем больше связей он имеет, это позволяет выявить ключевых распространителей [15].

Таблица 1. Модели и описание

Модель	Описание
BERT	Модель на основе Transformer, предобученная на предсказании замаскированных слов и связи предложений. Она учитывает полный контекст, что даёт высокую точность в задачах вроде классификации текста или вопросно-ответных систем [8]. Плюсы: мощная и универсальная. Минусы: требует больших ресурсов и тонкой настройки. Несмотря на это, BERT остаётся эталоном в NLP.
NATASHA	Готовая NLP-библиотека для русского языка, основанная на BERT и Rule-Based подходах. Оптимизирована для быстрой работы с NER, лемматизацией, синтаксическим анализом и исправлением опечаток. Плюсы: простота использования и специализация под русский язык. Минусы: менее гибкая, чем чистые Transformer-модели.
RoBERTA	RoBERTa – улучшенный BERT от Facebook. Обучена на больших данных без NSP, лучше работает с контекстом и длинными текстами. Даёт более точные результаты для классификации, NER и QA, но требует больше ресурсов. Плюсы: мощнее BERT. Минусы: прожорлива к железу.

Используя алгоритмы RageRank или Betweenness Centrality, мы можем выделить тех, кто находится в важных точках информационного потока, используя эти методы мы можем вычислить небольшой канал, со скромной аудиторией, но он может выступать главным источником вброса, соответственно он будет играть роль важного узла, через который информация переходит в более крупные каналы.

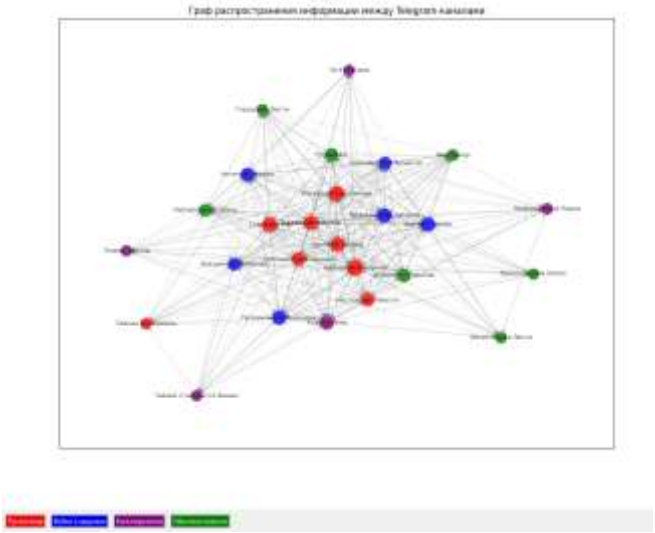

Рис. 2 Графовая модель распространения информации

III. Заключение

На текущий момент на рынке представлен ряд решений от российских и американских компаний. Предложенные программные продукты относятся к категории коммерческих решений, не предоставляющих доступ к исходному коду, что существенно ограничивает возможность обеспечения требуемого уровня информационной безопасности. Программы предлагаются по подписке, с минимальной стоимостью, при этом базовый функционал является неполным.

В статье рассматривается пример использования отечественного программного обеспечения PolyAnalyst, основанного на принципах Low-code разработки. Это решение демонстрирует высокую эффективность для повышения оперативности и информированности пользователей в правительственных и коммерческих организациях. В масштабах государства это способствует улучшению управления, а на локальном уровне позволяет получать и анализировать данные в режиме реального времени. Одним из ключевых преимуществ PolyAnalyst перед конкурентами является то, что заказчику требуется приобрести только среду разработки. Все созданные с её помощью продукты становятся полной собственностью организаций и государственных структур. Это не только снижает долгосрочные затраты, но и значительно повышает уровень информационной безопасности.

Библиография

[1] Z. Zhang, A. Poguda [Research on the development of data augmentation techniques in the field of machine translation]. International Journal of Open Information Technologies, 2023, no. 5. Available: https://cyberleninka.ru/article/n/research-on-the-development-of-data-augmentation-techniques-in-the-field-of-machine-translation (accessed January 17, 2025).

[2] D. N. Volozhanina, M. N. Zhukova [Review of the procedure for assessing security using OSINT methods in the conditions of an unstable international political environment]. Aktualnye problemy aviatsii i kosmonavtiki, 2022, no. [номер выпуска]. Available: https://cyberleninka.ru/article/n/peresmotr-protsedury-otsenki-zaschischennosti-metodami-osint-v-usloviyah-nestabilnoy-mezhdunarodnoy-politicheskoy-obstanovki (accessed January 17, 2025).

[3] A. E. Lebid, V. V. Stepanov, M. S. Nazarov [Use of OSINT technologies for civil society institutions]. International Journal of Media and Information Literacy, 2023, no. 1. Available: https://cyberleninka.ru/article/n/use-of-the-osint-technologies-for-civil-society-institutions (accessed January 17, 2025).

[4] G. S. Yakovlev, F. F. Ivanov [Use of low-code platforms in the transition to a process approach in the creation of automated systems]. Vestnik KRaUNC. Fiz.-mat. nauki, 2020, no. 1. Available: https://cyberleninka.ru/article/n/ispolzovanie-low-code-platform-pri-perehode-na-protsessnyy-podhod-v-sozdanii-avtomatizirovannyh-sistem (accessed January 17, 2025).

[5] V. S. Magomadov [Low-code and no-code platforms as a way to make programming more accessible to the wider public]. MNIJ, 2021, no. 6-1 (108). Available: https://cyberleninka.ru/article/n/platformy-low-code-i-no-code-kak-sposob-sdelat-programmirovanie-bolee-dostupnym-dlya-shirokoy-obschestvennosti (accessed January 17, 2025).

[6] A. S. Rusakovich [Intelligent data analysis as a decision support tool]. Sovremennye innovatsii, 2022, no. 1 (41). Available: https://cyberleninka.ru/article/n/intellektualnyy-analiz-dannyh-kak-instrument-podderzhki-prinyatiya-resheniy (accessed January 17, 2025).

[7] A. A. Solomonov [Optimization of ETL processes for big data]. Vestnik nauki, 2024, no. 9 (78). Available: https://cyberleninka.ru/article/n/optimizatsiya-etl-protsessov-dlya-bolshih-dannyh (accessed January 17, 2025).
[8] R. Gruetzemacher, (2022). The Power of Natural Language Processing. Harvard Business Review. Retrieved February 1, 2022, from https://hbr.org/2022/04/the-power-of-natural-language-processing
[9] V. Yu. Statyev, V.A. Dokuchaev, V. V. Maklachkova, (2022). [Information security in the "big data" space]. T-Comm, 4. Retrieved January 20, 2025, from https://cyberleninka.ru/article/n/informatsionnaya-bezopasnost-na-prostranstve-bolshih-dannyh.
[10] P. A. Basina, D. O. Dunayeva, A. Yu. Sarkisova, (2022). [Validation of machine learning models for automated determination of the tonality of Russian-language texts]. Vestnik Tomskogo gosudarstvennogo universiteta, 485. Retrieved January 20, 2025, from https://cyberleninka.ru/article/n/validatsiya-modeley-mashinnogo-obucheniya-dlya-avtomatizirovannogo-opredeleniya-tonalnosti-russkoyazychnyh-tekstov.
[11] V. I. Smirnov, O. V. Novoselova, (2024). [Review of modern methods of big data analysis for various subject areas]. Vestnik nauki, 6(75). Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/obzor-sovremennyh-metodov-analiza-bolshih-dannyh-dlya-razlichnyh-predmetnyh-oblastey
[12] T. S. Kuchkarov, (2023). [On the methods and tools of big data analysis]. Ekonomika i sotsium, 12(115)-2. Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/o-metodah-i-instrumentah-analiza-bolshih-dannyh
[13] A. A. Arlanova, A. M. Nobatov, (2023). [Intelligent data analysis: Types and methods]. Vestnik nauki, 1(58). Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/intellektualnyy-analiz-dannyh-vidy-i-metody
[14] D. K. Karpov, (2021). [Big data processing using Python]. StudNet, 6. Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/obrabotka-bolshih-dannyh-s-ispolzovaniem-sredstv-yazyka-python
[15] D. I. Safikanov, A. A. Artamonov, Yu. E. Fomina, A. I. Cherkassky, (2024). [Statistical model for searching target objects in a social network]. International Journal of Open Information Technologies, 10. Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/statisticheskaya-model-poiska-tselevyh-obektov-v-sotsialnoy-seti

Development and evaluation of the effectiveness of automated OSINT data collection and analysis methods for detecting and countering disinformation campaigns in social networks

K.S. Korneev, M.S. Spirin, D.A. Kul

***Abstract*—Countering disinformation campaigns in the context of modern information warfare is a pressing issue. Particular attention is paid to the role of social networks as a key tool for spreading fake news and shaping public opinion, which poses a threat to a state's social stability. To address this challenge, the use of automated open-source intelligence (OSINT) analysis systems is proposed to detect and neutralize disinformation campaigns in their early stages.**

This article examines the PolyAnalyst analytical platform, based on Low-code development principles, which automates the collection, processing, and visualization of data from the Telegram social network. The system incorporates machine learning methods and neuro-linguistic programming (NLP) techniques to analyze texts and identify fake news. A key feature of the system is its ability to construct graph models to track the mechanisms of disinformation dissemination, enabling the identification of sources, pathways, and key participants in information attacks.

***Keywords*—Disinformation, fake news, social networks, OSINT, PolyAnalyst, NLP, graph models, Low-code, data processing.**

REFERENCES

[1] Z. Zhang, A. Poguda [Research on the development of data augmentation techniques in the field of machine translation]. International Journal of Open Information Technologies, 2023, no. 5. Available: https://cyberleninka.ru/article/n/research-on-the-development-of-data-augmentation-techniques-in-the-field-of-machine-translation (accessed January 17, 2025).

[2] D. N. Volozhanina, M. N. Zhukova [Review of the procedure for assessing security using OSINT methods in the conditions of an unstable international political environment]. Aktualnye problemy aviatsii i kosmonavtiki, 2022, no. [номер выпуска]. Available: https://cyberleninka.ru/article/n/peresmotr-protsedury-otsenki-zaschischennosti-metodami-osint-v-usloviyah-nestabilnoy-mezhdunarodnoy-politicheskoy-obstanovki (accessed January 17, 2025).

[3] A. E. Lebid, V. V. Stepanov, M. S. Nazarov [Use of OSINT technologies for civil society institutions]. International Journal of Media and Information Literacy, 2023, no. 1. Available: https://cyberleninka.ru/article/n/use-of-the-osint-technologies-for-civil-society-institutions (accessed January 17, 2025).

[4] G. S. Yakovlev, F. F. Ivanov [Use of low-code platforms in the transition to a process approach in the creation of automated systems]. Vestnik KRaUNC. Fiz.-mat. nauki, 2020, no. 1. Available: https://cyberleninka.ru/article/n/ispolzovanie-low-code-platform-pri-perehode-na-protsessnyy-podhod-v-sozdanii-avtomatizirovannyh-sistem (accessed January 17, 2025).

[5] V. S. Magomadov [Low-code and no-code platforms as a way to make programming more accessible to the wider public]. MNIJ, 2021, no. 6-1 (108). Available: https://cyberleninka.ru/article/n/platformy-low-code-i-no-code-kak-sposob-sdelat-programmirovanie-bolee-dostupnym-dlya-shirokoy-obschestvennosti (accessed January 17, 2025).

[6] A. S. Rusakovich [Intelligent data analysis as a decision support tool]. Sovremennye innovatsii, 2022, no. 1 (41). Available: https://cyberleninka.ru/article/n/intellektualnyy-analiz-dannyh-kak-instrument-podderzhki-prinyatiya-resheniy (accessed January 17, 2025).

[7] A. A. Solomonov [Optimization of ETL processes for big data]. Vestnik nauki, 2024, no. 9 (78). Available: https://cyberleninka.ru/article/n/optimizatsiya-etl-protsessov-dlya-bolshih-dannyh (accessed January 17, 2025).

[8] R. Gruetzemacher, (2022). The Power of Natural Language Processing. Harvard Business Review. Retrieved February 1, 2022, from https://hbr.org/2022/04/the-power-of-natural-language-processing

[9] V. Yu. Statyev, V. A. Dokuchaev, V. V. Maklachkova, (2022). [Information security in the "big data" space]. T-Comm, 4. Retrieved January 20, 2025, from https://cyberleninka.ru/article/n/informatsionnaya-bezopasnost-na-prostranstve-bolshih-dannyh.

[10] P. A. Basina, D. O. Dunayeva, A. Yu. Sarkisova, (2022). [Validation of machine learning models for automated determination of the tonality of Russian-language texts]. Vestnik Tomskogo gosudarstvennogo universiteta, 485. Retrieved January 20, 2025, from https://cyberleninka.ru/article/n/validatsiya-modeley-mashinnogo-obucheniya-dlya-avtomatizirovannogo-opredeleniya-tonalnosti-russkoyazychnyh-tekstov.

[11] V. I. Smirnov, O. V. Novoselova, (2024). [Review of modern methods of big data analysis for various subject areas]. Vestnik nauki, 6(75). Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/obzor-sovremennyh-metodov-analiza-bolshih-dannyh-dlya-razlichnyh-predmetnyh-oblastey

[12] T. S. Kuchkarov, (2023). [On the methods and tools of big data analysis]. Ekonomika i sotsium, 12(115)-2. Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/o-metodah-i-instrumentah-analiza-bolshih-dannyh

[13] A. A. Arlanova, A. M. Nobatov, (2023). [Intelligent data analysis: Types and methods]. Vestnik nauki, 1(58). Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/intellektualnyy-analiz-dannyh-vidy-i-metody

[14] D. K. Karpov, (2021). [Big data processing using Python]. StudNet, 6. Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/obrabotka-bolshih-dannyh-s-ispolzovaniem-sredstv-yazyka-python

[15] D. I. Safikanov, A. A. Artamonov, Yu. E. Fomina, A. I. Cherkassky, (2024). [Statistical model for searching target objects in a social network]. International Journal of Open Information Technologies, 10. Retrieved January 18, 2025, from https://cyberleninka.ru/article/n/statisticheskaya-model-poiska-tselevyh-obektov-v-sotsialnoy-seti