СИСТЕМА ВЫЯВЛЕНИЯ ИНФОРМАЦИОННЫХ АТАК

Я.Р. Мустакимова, А.Н. Рабчевский
Пермский государственный национальный исследовательский университет

Аннотация. Данная статья посвящена системе автоматического выявления информационных атак в социальных сетях. В системе анализируются следующие признаки информационных атак: использование ботов для распространения информации, распространение четких и нечетких дубликатов сообщений, использование независимых каналов распространения информации. Целью работы является создание комплексного подхода к мониторингу и анализу информационных потоков в социальных сетях, что поможет в своевременном обнаружении и противодействии информационным атакам.
Ключевые слова: *информационные атаки, методы обнаружения информационных атак, социальные сети, боты, четкие и нечеткие дубликаты, независимые каналы.*

Введение

В наши дни социальные сети играют важную роль, они стали одним из основных способов общения и взаимодействия людей. Популярность социальных сетей непрерывно растет, что объясняется простотой их использования и широкой доступностью. Социальные платформы предоставляют возможность быстро находить интересных собеседников, делиться опытом, а также получать актуальную информацию и развлекательный контент. Тем не менее, с увеличением популярности социальных сетей увеличиваются и риски, связанные с их использованием. К сожалению, эти платформы часто привлекают злоумышленников, которые используют социальные сети для деструктивного влияния на общественное мнение. Злоумышленники активно распространяют ложную информацию через социальные сети, провоцируют конфликты, что в свою очередь приводит к росту социальной напряженности в обществе. Таким образом злоумышленники используют социальные сети для проведения информационных атак. Чтобы предотвратить такие атаки, их сначала нужно обнаружить. Создание методов, алгоритмов и программных решений для этой цели является крайне важной и актуальной задачей.

Обзор литературы

Информационная атака – это спланированное, организованное, целенаправленное, массированное информационное воздействие на объекты с целью формирования общественного мнения и поведения в соответствии с задачами организаторов атаки [1].

Существует множество различных объектов, на которые оказывают влияние информационные атаки. Целью информационных атак может быть общество в целом. Здесь речь идет о гражданах, которые становятся жертвами дезинформации, распространяемой в социальных сетях. Чаще всего такие атаки нацелены на создание паники, внушение страха или же изменение общественного мнения по вопросам, касающимся политики, экологии, безопасности или других актуальных тем. Массовая дезинформация может приводить к протестам и серьезным общественным беспорядкам. Также объектами информационных атак являются государственные институты. Злоумышленники могут пытаться подорвать репутацию конкретных лидеров или партий, распространяя недостоверную информацию. Не следует забывать и о специфических группах и сообществах. Это могут быть этнические или религиозные меньшинства, а также любые другие группы с определенной социальной идентичностью. Атаки на них могут быть направлены на разжигание ненависти, подстрекательство к дискриминации и насилию, на создание ложных стереотипов.

Если информационная атака достигла своих первоначальных целей, будь то изменение общественного мнения, манипуляция выборами, подрыв доверия к определенным личностям и организациям, или создание беспорядка в обществе, то она считается успешно проведенной.

Типовая информационная атака в социальных сетях проводится в два этапа. Первый этап – это вброс информации. Под информационным вбросом понимается резкое заполнение сетевого пространства какой-либо короткой, вызывающей массу эмоций информацией [2]. В социальных сетях этап вброса происходит через публикацию постов на страницах пользователей и в сообществах, причем публикацию могут осуществлять как реальные аккаунты, так и автоматизированные боты. Второй этап – это разгон контента. Он подразумевает реакцию на уже размещенные посты, включая репосты или комментарии. Пользователь может выразить свое одобрение через лайки или просто ознакомиться с постом. Все эти действия имеют значительное влияние на распространение информации. Обнаружение информационной атаки на этапе вброса значительно ускоряет реакцию на неё и позволяет осуществить противодействие на самом начальном этапе, предотвращая достижение целей злоумышленника.

Идея о выявлении информационных атак не нова. Существует два основных подхода к ее реализации [3]:

1. Классический метод. Этот метод заключается в анализе количества сообщений определенной направленности в пределах заданного времени. Превышение количества сообщений над пороговым значением является свидетельством проведения информационной атаки.

2. Метод, основанный на анализе динамики информационных потоков. Этот метод включает в себя построение графика изменения объема публикуемых сообщений и его сопоставление с заранее заданным шаблоном. Если графики совпадают, то это указывает на наличие информационной атаки.

Недостатком данных методов является то, что они не способны оперативно выявлять информационные атаки, поскольку требуется анализ публикуемых сообщений за достаточно длительный срок.

Анализируя общедоступные источники, можно заметить, что исследователи применяют различные сочетания признаков для выявления информационных атак. Среди этих признаков можно выделить: публикация схожего по содержанию целевого контента и высокая частота таких публикаций [3], небольшой временной интервал между публикациями [4], значимость источника и стремление к максимальному охвату аудитории [5] и др. Также уже существует патент «Система и способ выявления информационной атаки», автор – Нежданов Игорь Юрьевич [6]. В данном патенте представлено изобретение, которое позволяет автоматически выявлять факты проведения информационных атак и незамедлительно уведомлять ответственных лиц о таких атаках. Описанная в патенте система направлена на анализ различных интернет-ресурсов, таких как новостные сайты, блоги, форумы, видеохостинги, стриминговые платформы и сервисы вопросов и ответов. Для определения факта информационной атаки исследуются следующие параметры: общее число публикаций; количество сообщений, размещенных ботами, и количество управляемых ботами аккаунтов; количество публикаций, отправленных группами связанных источников; общее число дублирующихся публикаций.

Нашей же задачей является разработка системы автоматического выявления информационных атак в социальных сетях.

Система выявления информационных атак

В предлагаемой нами системе планируется использовать следующие признаки информационных атак:

1. Использование ботов для распространения информации. До сих пор не выработано однозначное определение термина «бот социальной сети». Мы будем рассматривать бот как специализированную страницу (аккаунт) в социальной сети, которая имитирует обычного пользователя и автоматически выполняет действия по публикации и продвижению контента для достижения целей злоумышленников.

2. Распространение четких и нечетких дубликатов сообщений. Четкий дубликат – это сообщение, которое точно совпадает с исходным сообщением по содержанию, формату и другим критериям. Например, если два сообщения полностью идентичны (одинаковый текст, изображения, ссылки), они считаются четкими дубликатами. Обычно такие дубликаты легко обнаруживаются с помощью алгоритмов сравнения текстов. А нечеткий дубликат – это сообщение, которое не является идентичным исходному сообщению, но имеет значительное совпадение по содержанию или смыслу. Нечеткие дубликаты могут включать небольшие переработки текста, изменения в формулировках или в порядке представления информации. Например, публикации, говорящие об одном и том же событии, но с разной формулировкой, могут считаться нечеткими дубликатами. Обнару-

жить нечеткие дубликаты намного сложнее, поэтому злоумышленники могут использовать их при проведении информационных атак.

3. Использование независимых каналов распространения информации. Под независимыми каналами в данном контексте подразумеваются аккаунты в социальных сетях, которые не имеют связей друг с другом. Вероятность того, что множество незнакомых людей станет одновременно публиковать одинаковый контент без общей цели, крайне мала. Такое поведение свидетельствует об организованной деятельности тех, кто инициирует информационную атаку.

Мы считаем, что именно эти признаки позволят нам обнаружить информационные атаки на ранней стадии. Архитектура нашей системы представлена на рисунке.

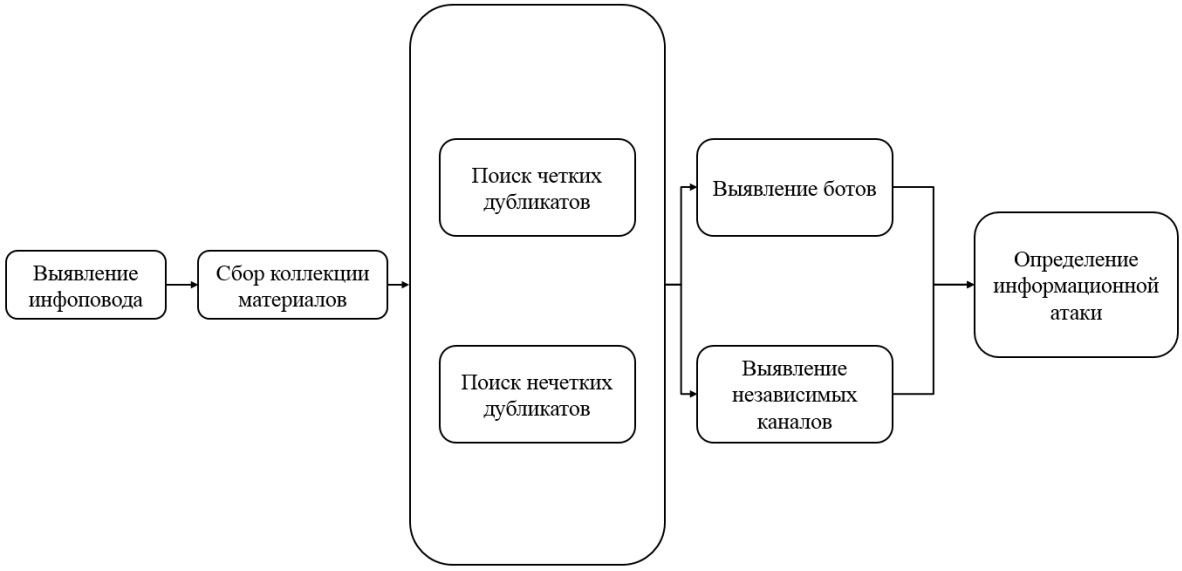

Рис. Архитектура системы автоматического выявления информационных атак в социальных сетях

Процесс работы нашей системы можно представить в виде следующей последовательности действий:

1. Выявление инфоповода. Инфоповод – это серия сообщений, публикуемых пользователями в социальных сетях, объединенных определенной тематикой. Инфоповод может быть связан с обсуждением какого-нибудь события, новости или темы, которая интересует большое количество людей в данное время. Инфоповоды могут быть как положительными, когда, например, обсуждается значимое событие или достижение, вдохновляющее людей и вызывающее одобрение, так и негативными. Такие темы, как коррупционные скандалы или социальные проблемы, тоже часто привлекают внимание людей и становятся предметом активных обсуждений в социальных сетях.

2. Сбор коллекции материалов. На данном шаге происходит сбор сообщений, которые относятся к выбранному информационному поводу. При этом собираются все сообщения, а не только те, которые имеют нега-

тивный характер. Потому что сообщения нейтрального и позитивного характера также могут быть частью информационных атак.

3. Формирование информационных треков. Информационный трек можно рассматривать как поток сообщений, созданных пользователями социальных сетей, состоящий из повторяющихся записей, упорядоченных по времени. Информационная атака, как правило, включает множество таких треков. В начале запуска трека контент распространяется быстро и в большом объеме от разных аккаунтов, что приводит к массовому вбросу информации. На данном шаге отдельно формируются треки с четкими и нечеткими дубликатами. Для работы с четкими дубликатами разработана и зарегистрирована программа «Информационный трек-детектор» [7], которая выявляет в общем потоке сообщений четкие дубликаты. Для выявление нечетких дубликатов также разработана и зарегистрирована программа «Программа выявления нечетких текстовых дубликатов» [8]. Процесс выявления нечетких дубликатов состоит из двух этапов: векторизация и кластеризации полученных векторов. Для векторизации была выбрана модель Transformers [9] – это архитектура нейронной сети, которая получила широкое применение в области обработки естественного языка (Natural Language Processing, NLP). Для кластеризации был выбран метод DBSCAN (Density-based spatial clustering of applications with noise, плотностной алгоритм пространственной кластеризации с присутствием шума) [10], который как следует из названия, оперирует плотностью данных.

4. Выявление ботов. Для обнаружения ботов разработана и зарегистрирована программа «Детектор ботов» [11]. На основе обученных нейросетевых моделей программа классифицирует профили пользователей социальной сети на две категории – «Боты» и «Не боты». В процессе исследования рассматриваются как открытые, так и закрытые профили пользователей.

5. Выявление независимых каналов. Для выявления независимых каналов разработана программа, которая определяет связи между аккаунтами, то есть ищет общих друзей, и на основе полученной информации строит граф связей.

Таким образом, анализируя динамику распространения дубликатов вместе с выявлением ботов и независимых каналов, можно установить факт проведения информационной атаки.

Заключение

Для разработки надежной и статистически обоснованной системы необходимо провести углубленное исследование разнообразных информационных атак. Ключевую роль в системе играют пороговые значения, такие как количество необходимых сообщений для формирования информационных треков или минимальный процент ботов среди участвующих аккаунтов. Эти пороговые значения помогут в принятии обоснованных решений относительно факта проведения информационной атаки. Важно понимать, что различные сценарии атак могут требовать различных порого-

вых значений, что подчеркивает необходимость дальнейших исследований в этой области. Установление и уточнение этих значений станет основой для более эффективного мониторинга и реагирования на угрозы, исходящие из информационного пространства.

Список литературы

1. Коцюбинская Л.В. Информационная атака: понятие и онтологические свойства // Политическая лингвистика. 2017. №6. С. 106-111. EDN: YMJXMA
2. Кочешев П.А. Информационные «вбросы». Методы минимизации последствий // Проблемы науки. 2016. № 6(7). С. 24-25. EDN: WDFTFZ
3. Потемкин А.В. Распознавание информационных операций средств массовой информации сети Интернет // Интернет-журнал «Науковедение». 2015. Т. 7. № 3. С. 1-11. DOI: 10.15862/139TVN315 EDN: UMFXXB
4. Еременко В.Т. Информационное противоборство в социотехнических системах / В. Т. Еременко, П. Н. Рязанцев. Орел: ОГУ имени И.С. Тургенева, 2016. 209 с. EDN: ZBYMMP
5. Расторгуев С.П. Информационные операции в сети Интернет / С. П. Расторгуев, М. В. Литвиненко, под. общ. ред. А. Б. Михайловского. Москва: АНО ЦСОиП, 2014. 128 с. ISBN: 978-5-906661-05-0 EDN: YNDAPK
6. Нежданов И.Ю. Система и способ выявления информационной атаки [Электронный ресурс] URL: https://www.fips.ru/registers-doc-view/fips_servlet?DB=RUPAT&DocNumber=2789629&TypeFile=html (дата обращения: 07.12.2024)
7. Программа «Информационный трек-детектор» свидетельство о регистрации программы для ЭВМ регистрационный № 2022668598 от 10.10.2022.
8. «Программа выявления нечетких текстовых дубликатов» свидетельство о регистрации программы для ЭВМ регистрационный № 2024665993 от 09.07.2024.
9. Sentence transformers. Distiluse-base-multilingual-cased-v. [Электронный ресурс] URL: https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1(дата обращения: 07.12.2024)
10. DBSCAN [Электронный ресурс] URL: https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.DBSCAN.html (дата обращения: 07.12.2024)
11. Программа «Детектор ботов» свидетельство о регистрации программы для ЭВМ регистрационный № 2024617353 от 01.04.2024.

INFORMATION ATTACK DETECTION SYSTEM

Y.R. Mustakimova, A.N. Rabchevsky
Perm State University

Abstract. This paper is devoted to the system of automatic detection of information attacks in social networks. The system analyzes the following attributes of information attacks: the use of bots to disseminate information, the dissemination of clear and fuzzy duplicates of messages, the use of independent channels of information dissemination. The purpose of the work is to create a comprehensive approach to monitoring and analyzing information flows in social networks, which will help in the timely detection and counteraction to information attacks.
Keywords: *information attacks, information attack detection methods, social networks, bots, clear and fuzzy duplicates, independent channels.*