УДК 004.056.5:004.738.5:004.89

Давыдов Виталий Владимирович
Санкт-Петербургский государственный университет
телекоммуникаций им. проф. М. А. Бонч-Бруевича,
студент, 6 курс, Россия, г. Санкт-Петербург
Зверев Павел Сергеевич
Санкт-Петербургский государственный университет
телекоммуникаций им. проф. М. А. Бонч-Бруевича,
бакалавр, Россия, г. Санкт-Петербург

МЕТОДИКА РАНЖИРОВАНИЯ ИСТОЧНИКОВ РАСПРОСТРАНЕНИЯ AI-СГЕНЕРИРОВАННОЙ ВРЕДОНОСНОЙ ИНФОРМАЦИИ В СОЦИАЛЬНЫХ СЕТЯХ

Аннотация: *В статье рассмотрена задача ранжирования источников распространения AI-сгенерированной вредоносной информации в социальных сетях. Показано, что развитие генеративных моделей изменяет характер информационных угроз, поскольку такие модели позволяют создавать масштабируемые текстовые, графические и мультимедийные информационные объекты, применимые в сценариях дезинформации, репутационных атак и иных форм деструктивного воздействия. Цель работы состоит в разработке методики, позволяющей формировать приоритетный список источников для последующего анализа оператором системы мониторинга. Предложена модель информационного объекта, модель источника распространения и система показателей, учитывающая активность источника, вовлеченность аудитории, признаки AI-сгенерированности и признаки вредоносности. Для оценки применимости методики приведен модельный расчет интегрального показателя приоритета. Сделан вывод о возможности использования предложенного подхода в системах мониторинга социальных сетей и поддержки принятия решений при противодействии вредоносной информации.*

Ключевые слова: *социальные сети, вредоносная информация, AI-сгенерированный контент, генеративный искусственный интеллект, ранжирование источников, информационная безопасность, мониторинг социальных сетей*

Annotation: *The paper considers the problem of ranking sources of AI-generated malicious information dissemination in social networks. The relevance of the study is determined by the increasing*

availability of generative models that can produce textual, visual and multimedia information objects used for disinformation, reputation attacks and other forms of destructive influence. The purpose of the research is to develop a methodology for forming a priority list of sources for a monitoring system operator. The proposed approach includes an information object model, a source model and a set of indicators reflecting source activity, audience engagement, AI-generated content and harmfulness. A model calculation is presented to demonstrate the procedure for forming an integral priority indicator. The results can be used in social media monitoring systems and decision support tools for counteracting malicious information dissemination.

Key words: *social networks, malicious information, AI-generated content, generative artificial intelligence, source ranking, information security, social media monitoring*

Введение

Социальные сети и мессенджеры в настоящее время рассматриваются в исследованиях по информационной безопасности как среда распространения общественно значимой, недостоверной и вредоносной информации, где отдельные сообщения, комментарии, изображения, видеоматериалы и ссылки формируют динамичное информационное пространство [1, 2, 4]. В работах, посвященных противодействию вредоносной информации, подчеркивается, что рост объема сообщений, скорости их тиражирования и количества источников затрудняет ручной анализ оператором системы мониторинга [1, 3].

Одной из прикладных задач такого мониторинга является не только обнаружение отдельного сообщения, но и определение источника, требующего первоочередного анализа. В исследованиях Витковой Л.А. и соавторов предложены модели социальной сети, источника и вредоносной информации, а также подходы к анализу аудитории, каналов распространения и выбору мер противодействия [1, 2, 3, 5]. Эти работы показывают, что приоритет источника должен определяться с учетом опубликованных информационных объектов и реакции аудитории, а не только фактом наличия вредоносного сообщения [2, 6].

Развитие генеративного искусственного интеллекта добавляет к указанной задаче новый фактор риска. Обзоры по LLM-опосредованной дезинформации

показывают, что крупные языковые модели могут использоваться для подготовки текстовых и мультимедийных материалов, увеличивая скорость, вариативность и масштаб распространения недостоверных сообщений [9]. При этом вопросы злоупотребления AI-системами, атак на модели и нежелательных сценариев их применения рассматриваются как самостоятельное направление безопасности искусственного интеллекта [13]. Для визуального контента дополнительно отмечается, что AI-сгенерированные изображения могут влиять на восприятие достоверности сообщения пользователями [14].

Цель статьи состоит в разработке методики ранжирования источников распространения AI-сгенерированной вредоносной информации в социальных сетях. Методика опирается на ранее разработанные подходы к выявлению каналов распространения информации и ранжированию источников [2, 6], но расширяет их за счет учета признаков AI-сгенерированности. Научная новизна работы состоит в том, что синтетичность контента рассматривается не как самостоятельное окончательное решение, а как один из факторов приоритизации источника наряду с активностью, вовлеченностью аудитории и вредоносностью сообщений.

Анализ существующих подходов

Задачи анализа вредоносной информации в социальных сетях обычно рассматриваются в нескольких взаимосвязанных направлениях: выявление нежелательного контента, анализ источников и каналов распространения, оценка аудитории, выбор контрмер и поддержка принятия решений [1, 3, 4]. В русскоязычных работах по данной тематике отдельно выделяются модели вредоносной информации, модели ее распространителя и архитектуры систем выявления и противодействия нежелательной информации [4, 5].

Для задач противодействия вредоносной информации важным является определение источника, на который целесообразно направить внимание оператора. В работе [2] предложена методика выявления каналов распространения

информации в социальных сетях, основанная на анализе информационных объектов и взаимодействий пользователей. В работе [6] рассматривается подход к ранжированию источников распространения информации, в котором используются показатели сообщений источника и вовлеченности аудитории. Данные подходы позволяют выделять значимые источники без обязательного построения полного графа социальной сети [2, 6].

Отдельный блок исследований связан с выбором мер противодействия. В работе Витковой Л.А., Чечулина А.А. и Сахарова Д.В. рассматривается задача выбора мер противодействия вредоносной информации в социальных сетях, что подтверждает необходимость связи между обнаружением угрозы, анализом источника и последующим решением оператора [3]. Поэтому ранжирование источников в настоящей работе рассматривается не как самостоятельная конечная процедура, а как этап поддержки принятия решений в системе мониторинга.

Социальные боты являются одним из объектов, влияющих на распространение недостоверной и вредоносной информации. В работе Коломейца М.В. и Чечулина А.А. предложены метрики вредоносных социальных ботов, включая доверие, живучесть, цену, скорость и экспертное качество, что показывает переход от бинарного обнаружения бота к качественному описанию нарушителя [7]. Экспериментальная работа Коломейца М.В., Тушкановой О.Н., Десницкого В.А., Витковой Л.А. и Чечулина А.А. показывает, что распознавание социальных ботов людьми затруднено и может ограничивать качество обучающих наборов данных [8].

Для AI-сгенерированного контента развиваются методы обнаружения синтетического текста и цифровых водяных знаков. Обзор Liu и соавторов, опубликованный в ACM Computing Surveys, систематизирует методы text watermarking, их обнаружимость, влияние на качество текста и устойчивость к атакам [10]. Работа Dathathri и соавторов демонстрирует возможность масштабируемого watermarking для выходов больших языковых моделей, но также

указывает на требования к качеству, detectability и вычислительной эффективности [11]. Следовательно, признаки AI-сгенерированности целесообразно использовать как вероятностный фактор анализа, а не как безусловное доказательство вредоносности объекта [10, 11, 13].

В исследованиях по обнаружению дезинформации применяются также графовые методы и графовые нейронные сети, позволяющие учитывать связи между пользователями, контентом и событиями [12]. Однако для прикладных систем мониторинга доступ к полному графу социальной сети может быть ограничен правилами платформ, настройками приватности и техническими ограничениями API. Поэтому сохраняется практическая потребность в методах предварительного ранжирования, использующих доступные признаки источника, опубликованных им информационных объектов и реакции аудитории [2, 6, 12].

Модель AI-сгенерированного вредоносного информационного объекта

В рамках исследования под информационным объектом понимается единица контента, опубликованная в социальной сети или мессенджере и доступная для анализа системой мониторинга. Такое понимание согласуется с подходами, в которых вредоносная информация рассматривается через совокупность информационных объектов, источников и признаков вредоносности [1, 5]. Информационный объект может представлять собой текстовое сообщение, комментарий, изображение, видео, аудиофрагмент, ссылку либо комбинированный мультимедийный материал.

Формально информационный объект oi предлагается описывать кортежем oi = <idi, typei, sourcei, timei, contenti, metai, ei, gi, hi>, где idi - идентификатор объекта, typei - тип объекта, sourcei - источник публикации, timei - время публикации, contenti - содержательная часть, metai - доступные метаданные, ei - показатель реакции аудитории на объект, gi - показатель AI-сгенерированности, hi - показатель вредоносности. Компоненты sourcei и ei включены в модель с учетом работ по

анализу аудитории и ранжированию источников распространения информации [2, 6].

Показатель AI-сгенерированности gi отражает наличие признаков синтетического происхождения информационного объекта. Он не интерпретируется как безусловное доказательство использования генеративной модели, поскольку современные методы обнаружения синтетического текста и watermarking имеют ограничения по устойчивости, качеству детектирования и зависимости от условий генерации [10, 11]. В рамках методики gi формируется на основе лингвистической шаблонности, повторяемости тезисов, отсутствия локального контекста, признаков синтетического изображения или видео, отсутствия исходного источника и резкого появления серии однотипных материалов [9, 10, 14].

Показатель вредоносности hi отражает степень потенциальной опасности информационного объекта для защищаемого субъекта, аудитории или информационного пространства. В состав признаков вредоносности могут входить недостоверные утверждения, призывы к распространению неподтвержденной информации, дискредитация лица или организации, манипулятивные эмоциональные конструкции, провокационные сообщения, фальсифицированные доказательства, а также признаки координированного информационного воздействия. Такой набор признаков соотносится с русскоязычными моделями вредоносной информации и задачами выбора контрмер [1, 3, 5].

Методика ранжирования источников

Под источником распространения информации Sj понимается объект социальной сети, от имени которого опубликовано множество информационных объектов за рассматриваемый период времени. Источником может быть пользовательский аккаунт, сообщество, канал, группа, страница организации или ботоподобный аккаунт. Такое понимание источника опирается на подходы к

выявлению каналов распространения информации и анализу источников в социальных сетях [2, 6, 7]. В рамках методики источник описывается кортежем Sj = <idj, urlj, Oj, audj, Aj, Ej, Gj, Hj, Pj>, где Oj - множество релевантных информационных объектов источника Sj, то есть объектов, прошедших предварительную тематическую фильтрацию и содержащих признаки вредоносности или AI-сгенерированности, а audj - размер аудитории источника (число подписчиков или участников).

Показатель активности Aj характеризует интенсивность публикации источником релевантных информационных объектов, то есть объектов множества Oj. В базовом варианте он рассчитывается как нормированное число таких объектов за заданный период: Aj = nj / max(n), где nj = |Oj| - число релевантных объектов источника Sj, а max(n) - максимальное число релевантных объектов среди рассматриваемых источников. Использование активности источника как фактора ранжирования согласуется с работами по анализу источников и каналов распространения информации [2, 6].

Показатель вовлеченности Ej отражает реакцию аудитории на сообщения источника. Он необходим, поскольку источник с небольшим числом сообщений может быть более значимым, если его публикации получают высокий охват и активно ретранслируются. Для отдельного объекта показатель реакции может быть представлен как ei = ln(1 + viewsi + alpha * repostsi + beta * commentsi + gamma * reactionsi). Показатель вовлеченности источника рассчитывается аналогично показателю активности: Ej = (1 / |Oj|) * sum(ei), нормированное относительно максимального значения среди рассматриваемых источников. При наличии данных о размере аудитории audj показатель ei может дополнительно нормироваться относительно audj, чтобы крупный источник не получал завышенную оценку вовлеченности только за счет размера аудитории. Учет обратной связи аудитории соответствует подходам к анализу каналов распространения и ранжированию источников [2, 6].

Показатель AI-сгенерированности источника Gj характеризует долю и выраженность синтетического контента в публикациях источника: Gj = (1 / |Oj|) * sum(gi). Включение этого показателя обосновано исследованиями, показывающими возможность генерации мультимедийной дезинформации с применением LLM и наличие отдельных методов идентификации AI-сгенерированного текста [9, 10, 11]. При необходимости расчет может учитывать только объекты, для которых значение вредоносности превышает заданный порог, чтобы нейтральный AI-контент не повышал приоритет источника.

Показатель вредоносности источника Hj определяется на основе оценок вредоносности опубликованных им информационных объектов: Hj = (1 / |Oj|) * sum(hi). Для более строгого варианта расчета может использоваться не среднее значение, а доля объектов, значение hi которых превышает экспертно заданный порог. Такой вариант соответствует логике перехода от анализа отдельных вредоносно-информационных объектов к анализу источника и выбору мер противодействия [1, 3, 5].

Интегральный показатель приоритета рассчитывается как взвешенная сумма нормированных показателей: Pj = wA * Aj + wE * Ej + wG * Gj + wH * Hj, где wA, wE, wG и wH - весовые коэффициенты, сумма которых равна единице. Для содержательной интерпретации значения Pj предлагается выделять три уровня приоритета: высокий (Pj >= 0,7), средний (0,4 <= Pj < 0,7) и низкий (Pj < 0,4); границы уровней могут уточняться экспертно в зависимости от специфики системы мониторинга. Данная формула является авторским методическим предложением настоящей статьи, а выбор групп показателей основан на работах по ранжированию источников, анализу вредоносной информации, социальным ботам и AI-сгенерированному контенту [1, 3, 6, 7, 9].

Алгоритм применения методики включает сбор информационных объектов за заданный период, предварительную фильтрацию по тематическим признакам, расчет признаков AI-сгенерированности и вредоносности, группировку объектов по

источникам публикации, расчет частных показателей, формирование интегрального показателя и передачу ранжированного списка оператору, который использует его для приоритетной проверки источников и принятия решения о дальнейших мерах противодействия (подробнее рассмотрено в разделе «Модельная апробация методики»). Выделение такого этапа согласуется с архитектурной логикой систем выявления и противодействия нежелательной информации, где мониторинг, анализ и поддержка принятия решений рассматриваются как связанные компоненты [3, 4].

Модельная апробация методики

Для демонстрации применимости предложенной методики рассмотрен модельный набор данных, имитирующий результаты мониторинга социальной сети за период 7 дней. Набор включает 200 информационных объектов, опубликованных 26 источниками. В состав объектов включены текстовые сообщения, комментарии, изображения и ссылки. Для каждого объекта заданы показатели реакции аудитории, а также экспертные оценки AI-сгенерированности и вредоносности в диапазоне от 0 до 1. Приведенный расчет является иллюстративным и не подменяет эксперимент на реальной выгрузке социальной сети.

Для наглядности приведены два примера информационных объектов из модельного набора данных, иллюстрирующие формирование показателей gi и hi. Объект o1 представляет собой текстовое сообщение одного из источников выборки: шаблонная тревожная формулировка без ссылки на первоисточник, за короткий промежуток времени опубликованная в нескольких почти идентичных вариациях; по совокупности признаков (шаблонность формулировки, отсутствие исходного источника, серийность публикации, провокационный характер утверждения) объекту присвоены значения gi = 0,86 и hi = 0,80. Объект o2, относящийся к другому источнику выборки, представляет собой изображение с признаками AI-генерации (искусственная детализация, отсутствие метаданных съемки), однако его

содержание нейтрально и не содержит недостоверных или провокационных утверждений, поэтому gi = 0,72 при hi = 0,15. Сопоставление этих двух объектов иллюстрирует общий тезис методики: признаки AI-сгенерированности сами по себе не определяют вредоносность объекта, а учитываются наряду с показателем hi при расчете итоговых показателей источника.

В демонстрационном расчете использовались равные веса показателей: wA = wE = wG = wH = 0,25. Для иллюстрации расчета из 26 источников выборки рассмотрены три источника, отражающие разные сочетания показателей. Источник S1 имел значения A = 0,90, E = 0,82, G = 0,75, H = 0,70 и получил интегральный показатель P = 0,79, что соответствует высокому приоритету. Источник S2 при значениях A = 0,75, E = 0,55, G = 0,62, H = 0,68 получил P = 0,65 и был отнесен к среднему приоритету. Источник S3 при значениях A = 0,55, E = 0,35, G = 0,80, H = 0,50 получил P = 0,55: высокий показатель AI-сгенерированности не привел к высокому приоритету из-за ограниченной вовлеченности аудитории и умеренной вредоносности, и источник также был отнесен к среднему приоритету.

Результаты модельного расчета показывают, что высокий приоритет получает не любой источник с признаками AI-сгенерированности, а источник, для которого одновременно наблюдаются высокая активность, значимая реакция аудитории и выраженные признаки вредоносности. Такой вывод методически связан с ранее предложенной логикой многокритериального анализа источников и выбора мер противодействия [3, 6]. В то же время сам факт использования генеративной модели не рассматривается как достаточное основание для вывода об опасности источника, поскольку AI-сгенерированный контент может быть как вредоносным, так и нейтральным [9, 10, 11].

Практическое значение методики состоит в том, что она позволяет предварительно упорядочить множество наблюдаемых источников и выделить объекты, требующие первоочередной проверки. Такой список может использоваться для экспертной оценки, выбора мер противодействия, уточнения

правил мониторинга и формирования отчетов об информационных инцидентах. Данный сценарий соответствует направлению исследований, где анализ источников рассматривается как этап повышения оперативности и обоснованности противодействия вредоносной информации [1, 3].

Обсуждение результатов

Отличие предложенной методики от подходов, ориентированных только на выявление AI-сгенерированного контента, состоит в том, что синтетичность рассматривается как один из факторов риска. Такое решение связано с ограничениями методов watermarking и детектирования AI-текста, которые оцениваются по detectability, влиянию на качество и устойчивости к атакам [10, 11]. Поэтому в методике показатель AI-сгенерированности объединяется с показателями активности, вовлеченности аудитории и вредоносности.

Преимуществом методики является возможность применения без построения полного графа социальной сети. Это важно для прикладного мониторинга, поскольку графовые методы и графовые нейронные сети обладают аналитическими преимуществами, но требуют данных о связях между пользователями, событиями и контентом [12]. В случаях, когда такие данные недоступны, ранжирование по опубликованным объектам и реакции аудитории может использоваться как предварительный фильтр перед более сложным социально-сетевым анализом [2, 6].

Ограничения методики связаны с качеством исходных данных и надежностью признаков AI-сгенерированности. Современные методы обнаружения синтетического текста и мультимедиа имеют вероятностный характер и могут быть чувствительны к перефразированию, редактированию человеком, изменению формата или отсутствию метаданных [10, 11, 13]. Для изображений дополнительную сложность представляет влияние визуальной правдоподобности AI-сгенерированного материала на пользовательское восприятие достоверности [14].

Еще одно ограничение связано с социальными ботами и качеством разметки. Эксперименты показывают, что люди не всегда надежно распознают ботов, особенно более сложные их разновидности, а это влияет на качество обучающих наборов и последующих моделей обнаружения [8]. Поэтому результаты ранжирования должны использоваться как инструмент предварительной приоритизации, а не как автоматическое основание для окончательного вывода о вредоносности источника.

Возможным направлением развития методики является адаптивная настройка весовых коэффициентов. В базовом варианте веса задаются экспертно, однако при накоплении данных о подтвержденных инцидентах может быть использована статистическая настройка весов или обучение модели ранжирования. Такой переход соответствует общей тенденции исследований, где анализ источника, аудитории и свойств нарушителя используется для более точного выбора мер противодействия [1, 3, 7].

Заключение

В статье предложена методика ранжирования источников распространения AI-сгенерированной вредоносной информации в социальных сетях. Разработана модель информационного объекта, учитывающая тип контента, источник публикации, реакцию аудитории, признаки AI-сгенерированности и признаки вредоносности. Сформирована модель источника распространения информации, включающая показатели активности, вовлеченности аудитории, AI-сгенерированности и вредоносности.

На основе указанных показателей предложен интегральный показатель приоритета, позволяющий формировать ранжированный список источников для анализа оператором системы мониторинга. Приведенный модельный расчет демонстрирует порядок применения методики и показывает, что приоритет источника должен определяться не одним признаком, а совокупностью факторов.

Учет AI-сгенерированности позволяет адаптировать подходы к ранжированию источников вредоносной информации к условиям распространения синтетического контента, создаваемого с использованием генеративных моделей.

В дальнейшем предполагается провести экспериментальную проверку методики на реальных публичных данных социальных сетей и мессенджеров, уточнить состав признаков AI-сгенерированности для текстовых и мультимедийных объектов, а также разработать процедуру автоматической настройки весовых коэффициентов с учетом подтвержденных информационных инцидентов.

Использованные источники:

1. Виткова Л. А. Модели, алгоритмы и методика противодействия вредоносной информации в социальных сетях: дис. ... канд. техн. наук: 05.13.19. Санкт-Петербург: СПб ФИЦ РАН, 2021.

2. Проноза А. А., Виткова Л. А., Чечулин А. А., Котенко И. В., Сахаров Д. В. Методика выявления каналов распространения информации в социальных сетях // Вестник СПбГУ. Прикладная математика. Информатика. Процессы управления. 2018. Т. 14. Вып. 4. С. 362-377. DOI: 10.21638/11702/spbu10.2018.409.

3. Виткова Л. А., Чечулин А. А., Сахаров Д. В. Выбор мер противодействия вредоносной информации в социальных сетях // Вестник Воронежского института ФСИН России. 2020. № 3. С. 20-29.

4. Виткова Л. А., Саенко И. Б. Архитектура системы выявления и противодействия нежелательной информации в социальных сетях // Вестник Санкт-Петербургского государственного университета технологии и дизайна. Серия 1: Естественные и технические науки. 2020. № 3. С. 33-39.

5. Виткова Л. А., Сахаров Д. В., Голузина Д. Р. Модель вредоносной информации и ее распространителя в социальных сетях // Защита информации. Инсайд. 2020. № 3 (93). С. 66-72.

6. Vitkova L., Kotenko I., Chechulin A. An Approach to Ranking the Sources of Information Dissemination in Social Networks // Information. 2021. Vol. 12, No. 10. Article 416. DOI: 10.3390/info12100416.

7. Kolomeets M. V., Chechulin A. A. Properties of Malicious Social Bots // Proceedings of Telecommunication Universities. 2023. Vol. 9, No. 1. P. 94-104. DOI: 10.31854/1813-324X-2023-9-1-94-104.

8. Kolomeets M., Tushkanova O., Desnitsky V., Vitkova L., Chechulin A. Experimental Evaluation: Can Humans Recognise Social Media Bots? // Big Data and Cognitive Computing. 2024. Vol. 8, No. 3. Article 24. DOI: 10.3390/bdcc8030024.

9. Barman D., Guo Z., Conlan O. The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination // Machine Learning with Applications. 2024. Vol. 16. Article 100545. DOI: 10.1016/j.mlwa.2024.100545.

10. Liu A., Pan L., Lu Y., Li J., Hu X., Zhang X., Wen L., King I., Xiong H., Yu P. S. A Survey of Text Watermarking in the Era of Large Language Models // ACM Computing Surveys. 2024. Vol. 57, No. 2. Article 47. DOI: 10.1145/3691626.

11. Dathathri S., See A., Ghaisas S. et al. Scalable Watermarking for Identifying Large Language Model Outputs // Nature. 2024. Vol. 634. P. 818-823. DOI: 10.1038/s41586-024-08025-4.

12. Lakzaei B., Haghir Chehreghani M., Bagheri A. Disinformation Detection Using Graph Neural Networks: A Survey // Artificial Intelligence Review. 2024. Vol. 57. Article 52. DOI: 10.1007/s10462-024-10702-9.

13. Vassilev A., Oprea A., Fordyce A., Andersen H. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST Trustworthy and Responsible AI Report. 2024. DOI: 10.6028/NIST.AI.100-2e2023.

14. Farooq A., de Vreese C. Deciphering Authenticity in the Age of AI: How AI-Generated Disinformation Images and AI Detection Tools Influence Judgements of Authenticity // AI & Society. 2025. DOI: 10.1007/s00146-025-02416-5.

Информация о себе:
89522007310, vdavydov.science@gmail.com
89939661232, playximik29@gmail.com