DOI 10.36622/1682-7813.2025.28.2.003
УДК 004.62

КАТЕГОРИЗАЦИЯ ДЕЗИНФОРМАЦИИ ДЛЯ ПОСТРОЕНИЯ СИСТЕМЫ ЗАЩИТЫ

В.П. Лось, А.К. Шувалов

В сети Интернет затруднен как контроль достоверности информации, так и своевременное выявление информации, способной склонить большие группы пользователей к несанкционированным акциям. Одновременно с этим, множественные каналы распространения и площадки копирования контента затрудняют своевременный мониторинг. До сих пор не разработано эффективной автономной системы, выявляющей факт наличия и распространения дезинформации в режиме реального времени. Такая система может быть разработана на основе комплексного подхода к обеспечению ИБ, который включает адаптацию СЗИ для различных по типу и интенсивности информационных атак. В статье предлагается разработка начального этапа: категоризация внедряемой информации. Эта категоризация основана на реконструкции цели атаки, ее жизненного цикла и адаптации существующих методов категоризации в ИБ, разработанных для защиты от компьютерных атак, для задачи защиты от применения дезинформации.

Ключевые слова: дезинформация, категоризация, DISARM, матрица атак, жизненный цикл атаки.

Введение

Широкое распространение информационных технологий привело к тому, что теперь почти у каждого человека есть доступ к неограниченному объему информации, которая может использоваться не только на пользу, но и во вред, например, побудить к противоправным действиям. Такое влияние могут оказывать не только достоверные, но и ложные сведения, например, дезинформация.

Чтобы избежать разночтений в понимании данного термина, в работе будет использоваться определение, которое представлено в рамках плана ЕС по защите от ложной информации [1]. В документе дезинформация определяется как – создание, представление и распространение проверяемой ложной или вводящей в заблуждение информации с целью получения экономической выгоды или намеренного обмана общественности. Данное определение не только показывает негативные последствия от ложных сведений, но и подчеркивает основные этапы жизненного цикла атаки с использованием дезинформации. Актуальность работ в данной области подтверждается как тяжестью возможных последствий, так и статистикой. По данным представленным на конференции «Противодействие фейкам в 2024 году», 64% жителей России каждую неделю сталкиваются с недостоверной информаций [2]. Также на рис. 1 показано число новостных ресурсов, на которых было зафиксировано использование ИИ для генерации ложных и/или вводящих в заблуждение новостей за последнее время [3].

Из графика, представленного на рис. 1, видно, что число таких ресурсов постоянно увеличивается, что может привести к еще большему распространению ложной информации среди населения.

Международный уровень признания проблемы и запрос на однозначное определение свидетельствует, что необходимость в эффективных способах защиты от ложных сведений как никогда высока. Как отмечается в пособиях по информационной безопасности [4], эффективная система защиты возможна, только с использованием комплексного подхода. Под комплексным подходом понимается множество различных решений ИБ, каждое из которых специализируется на защите от конкретной категории угроз, но вместе они составляют широкий спектр СЗИ, который покрывает большинство возможных атак на ИС. Это утверждение позволяет заключить, что для полноценной защиты от дезинформации, необходимо составить набор

категорий, который мог бы использоваться на каждом из этапов жизненного цикла атаки.

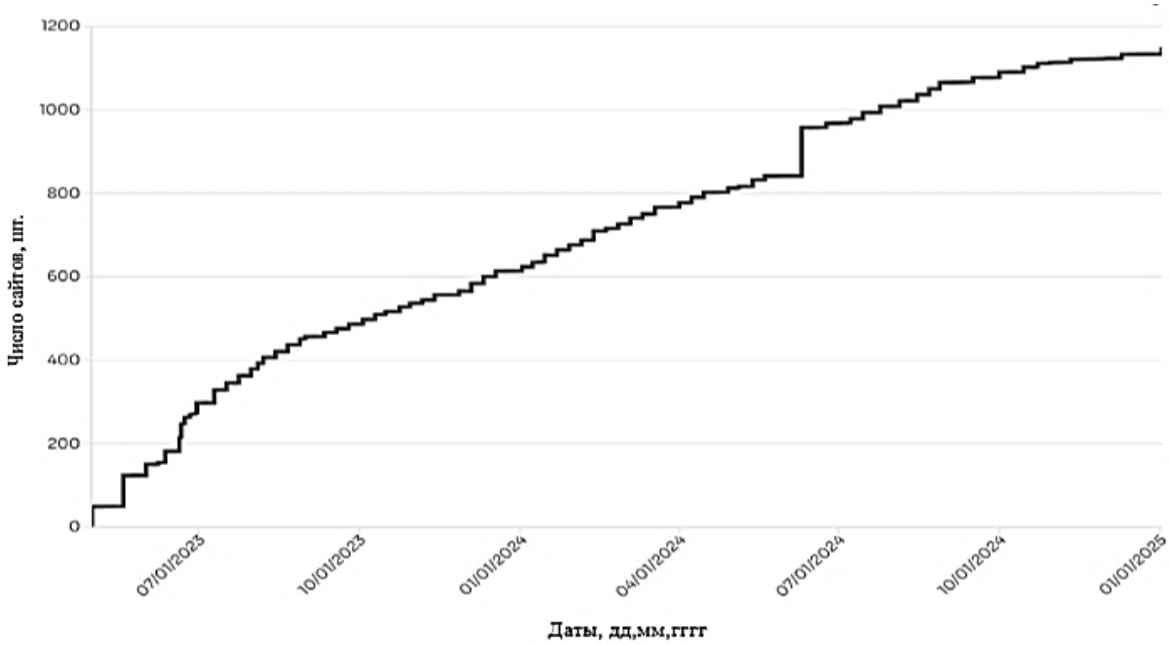

Рис. 1. График числа новостных ресурсов, которые используют сгенерированные новости

Применение категоризации в информационной безопасности

При рассмотрении задач, так или иначе, связанных с информационной безопасностью, прежде всего определяется целесообразность каждой из процедур последовательного обеспечения безопасности. Причина этого в том, что сами по себе решения ИБ не приносят прибыль для компании, а только требуют ресурсов. Из этого следует, что почти любое действие, которое занимает время для построения процессов ИБ должно быть обусловлено достижением некоторой цели.

Основная цель ИБ – заранее намеченный результат защиты информации, это может быть предотвращение ущерба собственнику, владельцу, пользователю информации в результате возможной утечки информации и/или несанкционированного и непреднамеренного воздействия на информацию. Дополнительно, стоит отметить, что необходимо не только обеспечить защиту, но и сделать это с использованием минимального количества ресурсов.

Получается, что вопрос категоризации дезинформации, так как это процесс, требующий затрат финансовых, человеческих, должен начинаться с обоснования необходимости применения такого подхода. Доказать его эффективность для построения системы защиты от ложной информации, можно на примере уже существующих категориальных матриц, которые активно применяются как отечественными, так и зарубежными организациями по ИБ.

Например, база знаний Mitre Att&ck (Adversarial Tactics, Techniques & Common Knowledge — «тактики, техники и общеизвестные факты о злоумышленниках»). Этот проект был создан в 2013 году и как говорится в его описании – он служит для составление структурированной матрицы используемых киберпреступниками приемов, чтобы упростить задачу реагирования на киберинциденты. Действительно, защититься от различных атак и/или реагировать на них можно и без представленной базы знаний. Однако это будет менее эффективно в сравнении с ее применением. Так, в табл. 1 показаны основные результаты исследования [5], которые свидетельствуют, что более 80% опрошенных компаний, которые используют данную базу знаний, подтверждают ее применимость для оценки и выявления киберинцидентов, также отмечается, что эта

технология позволяет эффективнее использовать средства защиты и находить несоответствия в их конфигурации [6].

Таблица 1

Оценка применимости матрицы Mittre Att&ck для выявления киберинцидентов

Характеристика	Процент компаний, которые подтвердили этот факт
Используют данную матрицу	63%
Подтвердили полноту информации для выявления киберинцидентов	83%

Категорированию подвергаются не только различные типы атак, но и сами объекты защиты. Так, например, постановлением Правительства [7] введены критерии для КИИ. Это позволяет различным организациям провести оценку своих информационных ресурсов и сопоставить их с законодательными требованиями по ИБ.

Приведенные примеры доказывают, что категоризация различных объектов и подходов к защите информации, позволяет повысить эффективность их применения, найти ошибки в работе СЗИ и провести сопоставления с требованиями регуляторов. Из этого можно сделать вывод, что выделение у дезинформации различных категорий может способствовать построению комплексной системы защиты, которая будет показывать свою работоспособность на каждом из этапов жизненного цикла атаки.

Формирование критериев для определения категорий дезинформации

Опираясь на определение дезинформации и понимании того, что для злоумышленника важен не сам факт ее проведения, а конечный результат, можно сделать вывод о жизненном цикле атак с использованием ложной информации. Полученный цикл представлен на рис. 2.

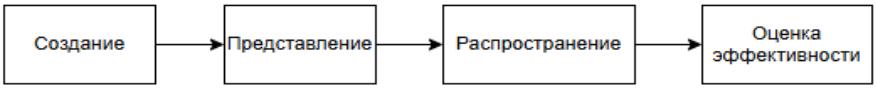

Рис. 2. Жизненный цикл атаки с использованием дезинформации

Так, в создании определяется стратегия и целевая аудитория атаки. В представлении происходит разработка контента и формирование каналов передачи ложной информации. При распространении осуществляется доставка сформированной дезинформации до цели, определенной на первом этапе. Оценка эффективности – дополнительный этап, который позволяет определить качество атаки и при необходимости доработать ее.

При этом в материалах по теме дезинформации нет четкой категоризации, как, например, для компьютерных атак. Многие исследования концентрируют внимание на одном из этапов жизненного цикла и выстраивают категории основываясь именно на различных признаках характерных конкретному этапу [8]. Такой подход позволяет разобрать возможные сценарии и способы защиты от них, однако, не дает полной картины, что может затруднить определение набора действий для пресечения комплексных угроз, что, как отмечалось ранее, и является основной целью защиты информации. Если рассматривать атаку через дезинформацию, как способ воздействия на некоторый информационный объект, которым, судя по определению, чаще всего является человек, то можно проследить зависимости между этапами жизненного цикла компьютерной атаки и через ложную информацию.

Представленный ранее список основных этапов атак показывает, что стадиями они схожи с жизненным циклом компьютерных атак, которые состоят из четырех пунктов, представленных на рис. 3 [9].

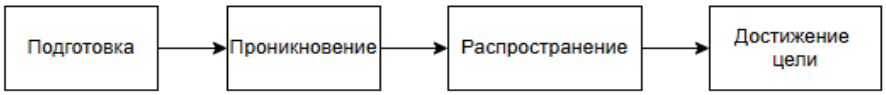

Рис. 3. Жизненный цикл компьютерной атаки

Данное сопоставление позволяет выдвинуть гипотезу, что если для компьютерных атак уже существует рабочая структура для их категорирования, то она будет применима и для защиты от дезинформации.

Так, матрица Mitre приводит обширный набор категорий, техник и тактик совершения атак, и возможных способов противодействия им. Если рассматривать дезинформацию с такой категориальной позиции, то внедрение подобной базы знаний позволит выстроить надежную систему защиты, которая поспособствует эффективному противодействию распространению ложных сведений на различных этапах.

В итоге, можно составить набор критериев, которые ограничивают поиск существующих решений по категоризации дезинформации на основе жизненного цикла атаки. Они представлены в табл. 2.

Таблица 2

Критерии для поиска категоризации дезинформации

Критерии
Открытый доступ
Наличие техник, тактик и категорий как для атаки, так и для противодействия ложной информации
Достоверность источников
Схожесть структуры с уже зарекомендовавшими себя решениями ИБ

По этим четырем критериям подходит фреймворк DISARM [10].

Анализ фреймворка DISARM

DISARM – это фреймворк, который распространяется в открытом доступе и предлагает базу знаний для определения и противодействия атакам с использованием дезинформации. По свой структуре он схож с матрицей Mitre [11], что упрощает работу для аналитиков безопасности.

Как говорится на официальном сайте компании, фреймворк зародился в 2018 году и уже в 2019 начал сотрудничество с различными правительственными и коммерческими организациями ЕС и США. Авторы говорят, что цель данного решения, дать возможность множеству людей отслеживать появление дезинформации и иметь способ защититься от нее.

На текущей момент в открытом доступе фактически нет аналогов данному решению. Хотя такие попытки и были, но дальше, чем первичные исследования в открытом доступе публикаций больше не было [12].

Как отмечалось ранее, сам фреймворк сходен по структуре с матрицей Mitre. Такой подход позволил выделить определенные наборы групп, которые фактически и являются категориями дезинформации причем сразу с упором на дальнейшую защиту от атак.

В DISARM определены четыре основных направления, которые характеризуют этапы жизненного цикла атаки [13]:

1) планирование,
2) подготовка,
3) исполнение,
4) оценка эффективности.

Каждое из направлений разбивается на подкатегории, которые называются тактиками и обозначаются с префиксом TA. Тактики определяют общее направление атаки, целевые задачи, которые необходимо выполнить, и возможные техники, которые могут способствовать успешному достижению поставленной цели. Так как, необходимо не только понимать, что проводится атака, но и защититься от нее, то в фреймворке вводится специальный набор контрмер, которые соотносятся с различными техниками и тактиками.

Итоговый процесс анализа атаки представлен на рис. 4.

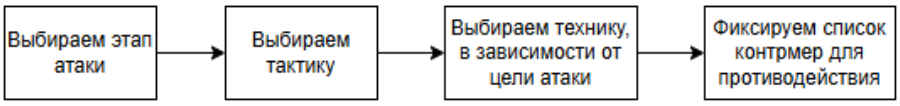

Рис. 4. Процесс анализа атаки с использование фреймворка DISARM

Однако, не смотря на все преимущества, которые могут быть получены, при использовании данного решения, оно не лишено недостатков. Основным, из которых является – наполненность информацией.

Действительно, сложно отрицать пользу DISARM с учетом отсутствия аналогов, удобства интеграции с другими системами для построения аналитики [14] и широким спектром тактик и техник, которые могут быть использованы для построения эффективной системы защиты. Но если отойти от способов проведения атак и обратить больше внимания именно на защиту от них, то станет заметно, что большое число техник не содержит в себе информацию по противодействию их реализации. Особенно в разрезе новых технологий, которые приобретают все большую популярность, например:

1) генерация текста,
2) использования DeepFake,
3) генерация речи,
4) создание мемов.

В табл. 3 представлена сводная статистика числа контрмер для различных тактик в DISARM.

Таблица 3

Число контрмер против тактик в DISARM

Этап жизненного цикла атаки	Категория. Тактики	Общее число техник	Число техник, покрытых контрмерами
Планирование	Планирование стратегии	6	0
	Планирование целей	41	1
	Анализ целевой аудитории	21	0
Подготовка	Подготовка нарративов	9	4
	Разработка контента	33	1
	Создание социальных активов	74	2
	Установление легитимности контента	38	0
	Микроцель	7	2
	Выбор каналов и возможностей распространения	58	1
Исполнение	Тестовая загрузка контента	5	4
	Доставка контента	10	0
	Увеличение спектра распространения	21	2
	Активные действия онлайн	16	1
	Автономная активность	12	2
	Закрепление в информационном поле	27	1
Оценка эффективности	Оценка эффективности	13	0

Данные из таблицы подтверждают факт того, что в рамках DISARM представлено малое число контрмер для защиты, что усложняет его применимость.

На рис. 5 представлена статистика, агрегированная статистическим сервисом Яндекса для анализа популярности нейронных сетей в обществе [15].

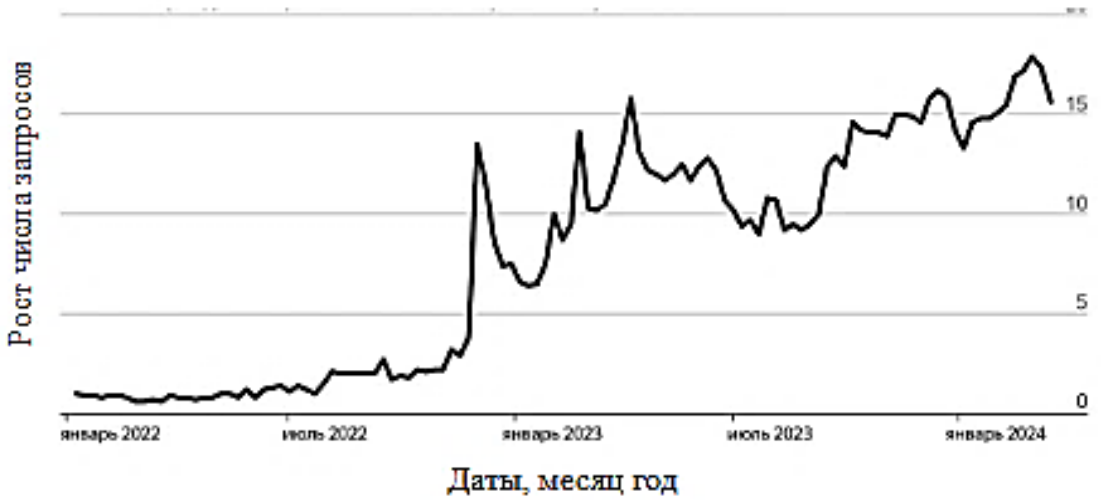

Рис. 5. Динамика интереса к нейронным сетям

Из графика на рис. 5 видно, что с каждым годом интерес общества все увеличивается. Также, фиксируется увеличение числа пользователей, которые применяют нейронные сети, например, для генерации текста. Так с февраля по июнь 2024 года, число таких лиц выросло с 26% до 33%.

Еще один недостаток, который выявляется исходя из сравнения с матрицей Mitre. Отсутствие ссылок на исследования. В матрице по компьютерным атакам, каждой технике в сопоставление идет набор работ, которые позволяют полноценно оценить и понять атаку, что может улучшить качество противодействия ей. В DISARM таких ссылок нет.

Основная задача исследования была именно провести категоризацию дезинформации для получения возможности построить более эффективную систему защиты от атак. Те группы/категории, которые представлены в фреймворке DISARM позволяют рассмотреть атаки с множества различных углов и подходов, что будет способствовать выбору корректного способа защиты. Однако, отсутствие большого числа контрмер затруднит интеграцию данного решения для построения аналитики и противодействию атак с использованием ложной информации.

Вывод

В работе были рассмотрены различные варианты категоризации дезинформации. Определены основные понятия, которые присущи данному типу атак. Было показано, что большинство существующих подходов к категоризации дезинформации, рассматривают ее только в рамках одного из этапов жизненного цикла, что не дает полного представления и не позволяет построить эффективную систему защиты. Данный тезис был выдвинут из-за оценки эффективности влияний таких категориальных матриц как Mitre на способность выявления атак. В исследованиях было доказано, что такой подход позволяет повысить показатели защищенности информационной системы и применяется в многих международных организациях, которые занимаются внедрением решений по информационной безопасности.

Из данного утверждения был сделан вывод, что подход с использование матрицы подобной Mitre, но адаптированной к анализу атак с использованием дезинформации,

может повысить эффективность системы защиты от подобных угроз. Было показано, что в открытом доступе существует всего одно такое решение, которое активно развивается сообществом и сотрудничает с различными правительственными организациями западных стран, что позволяет оценить авторитетность рассмотренного фреймворка. Однако, отмечается, что не смотря на удобство и эффективность использования подобного решения, оно не лишено недостатков, основным из которых является полнота информации о способах защиты от атак. При рассмотрении большинства современных подходов, которые также зафиксированы в рамках DISARM, не отмечаются способы и/или контрмеры для защиты, что с учетом активного развития генеративных моделей, которые показывают все большую результативность при создании контента, может привести к трудностям в использовании данного решения.

Список литературы

1. Antonios Kouroutakis EU Action Plan Against Disinformation // The International Lawyer. 2020. №2. С. 277-290.

2. 64% россиян сталкиваются с фейками в интернете раз в неделю и чаще // ДИАЛОГ URL: https://dialog.info/64-rossiyan-stalkivajutsya-s-fejkami-v-internete-raz-v-nedelju-i-chashhe/ (дата обращения: 02.12.2024).

3. Tracking AI-enabled Misinformation: 1,150 ‘Unreliable AI-Generated News’ Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools // NewsGuard URL: https://www.newsguardtech.com/special-reports/ai-tracking-center/ (дата обращения: 05.12.2024).

4. Основы информационной безопасности: учебное пособие для студентов вузов / Е.В. Вострецова. Екатеринбург: Изд-во Урал. ун-та, 2019.— 204 с.

5. Basra J., Kaushik T. MITRE ATT&CK® as a Framework for Cloud Threat Investigation // Center for Long-Term Cybersecurity (CLTC): Berkeley, Italy. 2020.

6. Study: MITRE ATT&CK Improves Cloud Security, Yet Many Enterprises Struggle to Implement It // AITHORITY URL: https://aithority.com/it-and-devops/cloud/study-mitre-attck-improves-cloud-security-yet-many-enterprises-struggle-to-implement-it/ (дата обращения: 10.12.2024).

7. Постановление Правительства Российской Федерации от 8 февраля 2018 г. N 127 // ФСТЭК России URL: https://fstec.ru/dokumenty/vse-dokumenty/postanovleniya/postanovlenie-pravitelstva-rossijskoj-federatsii-ot-8-fevralya-2018-g-n-127 (дата обращения: 11.12.2024).

8. Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 1. С. 119–129.

9. Анатомия таргетированной атаки // InformationSecurity URL: https://lib.itsec.ru/articles2/Oborandteh/anatomiya-targetirovannoy-ataki (дата обращения: 14.12.2024).

10. DISARM is an open framework for those cooperating in the fight against disinformation // DISARM Foundation URL: https://www.disarm.foundation/ (дата обращения: 16.12.2024).

11. DISARM Framework Explorer // DISARM Foundation URL: https://disarmframework.herokuapp.com/ (дата обращения: 17.12.2024).

12. The ABCDE Framework // JSTOR URL: https://www.jstor.org/stable/pdf/resrep26180.6.pdf (дата обращения: 19.12.2024).

13. Newman H. Foreign information manipulation and interference defence standards: Test for rapid adoption of the common language and framework ‘DISARM’ //Nato Strategic Communications. Centre of Excellence. 2022.

14. AMITT Frameworks: User Guide // GitHub URL: https://github.com/DISARMFoundation/DISARMframeworks/blob/main/DISARM_DOCUMENTATION/05_AMITT_User_Guide.pdf (дата обращения: 24.12.2024).

15. Нейростат // Яндекс URL: https://ya.ru/ai/stat (дата обращения: 28.12.2024).

Российский государственный гуманитарный университет
Russian State University for the Humanities

Поступила в редакцию 9.04.25

Информация об авторах

Лось Владимир Павлович – д-р воен. наук, профессор, главный научный сотрудник (ИИНиТБ), Российский государственный гуманитарный университет, e-mail: los-vladimir@yandex.ru

Шувалов Александр Константинович – аспирант, Российский государственный гуманитарный университет, e-mail: mr.sasha.shuv@gmail.com

CATEGORIZATION OF MISINFORMATION FOR BUILDING A PROTECTION SYSTEM

V.P. Los, A.K. Shuvalov

The Internet makes it difficult both to control the reliability of information and to identify in a timely manner the information that can incline large groups of users to unauthorized actions. At the same time, multiple distribution channels and content copying sites make timely monitoring difficult. So far, no effective autonomous system has been developed to detect the presence and dissemination of misinformation in real time. Such a system can be developed on the basis of a comprehensive approach to ensuring IS, which includes the adaptation of protection systems for different types and intensity of information attacks. The paper proposes the development of the initial stage: categorization of the introduced information. This categorization is based on the reconstruction of the purpose of the attack, its life cycle and adaptation of existing categorization methods in IS, developed for protection against computer attacks, for the task of protection against the use of misinformation.

Keywords: disinformation, categorization, DISARM, attack matrix, attack life cycle.

Submitted 9.04.25

Information about the authors

Vladimir P. Los – Dr. Sc. (Military), Professor, Chief Scientific Officer (IINiTB), Russian State University for the Humanities, e-mail: los-vladimir@yandex.ru

Alexander K. Shuvalov – postgraduate student, Russian State University for the Humanities, e-mail: mr.sasha.shuv@gmail.com