

Использование системы AutoMOD в противодействии кибербуллингу

Научный руководитель – Макаримов Эдуард Аликович

Костина Варвара Сергеевна

Студент (бакалавр)

Московский государственный университет имени М.В.Ломоносова, Юридический факультет, Кафедра криминалистики, Москва, Россия

E-mail: varya1979kost@gmail.com

В своем докладе за 2024 год уполномоченный по правам человека в РФ Татьяна Москалькова отметила важность возвращения к обсуждению вопроса о введении нового административного наказания за травлю в Интернете или «кибербуллинг». [7]

Кибербуллинг (травля в Интернете [1]) – это нестандартная форма буллинга (травли, издевательств [6]), которая включает и моральную и психологическую агрессию, наличие власти и принуждения, бойкот или изоляцию, террор и шантаж, угрозы, реализуемые по средствам электронной коммуникации. [5]

Е.О.Мазурчук и Н.И.Мазурчук отмечают следующие признаки кибербуллинга:

1. Моральное и психологическое насилие;
2. Доминирование одного участника над другим;
3. Принуждение жертвы кибербуллинга к совершению каких-либо действий его инициатором;
4. Социальная изоляция и запугивание жертвы кибербуллинга его инициатором;
5. Систематический характер. [4]

Согласно статистике мониторингового центра «Безопасность 2.0», число проявлений кибербуллинга возросло на 15% за 2024 год. Также центр отмечает, что около 23% всех выявленных конфликтов в сети Интернет являются проявлениями кибербуллинга. [9]

Генеральная прокуратура РФ в 2024 году отмечала низкую раскрываемость киберпреступлений. Для оптимизации выявления, расследования и профилактики кибербуллинга возможно использование искусственного интеллекта. [10]

На данный момент активное использование искусственного интеллекта для борьбы с кибербуллингом осуществляется в социальных сетях при модерации контента. VK Play провели исследование, в котором выяснили, что около 70% ПК-геймеров сталкивались с кибербуллингом, таким образом не только у социальных сетей, но и стриминговых сервисов (непрерывный поток онлайн-видео [3]) возникает потребность в модерации контента. [13]

Стриминговая платформа Twitch использует систему AutoMOD. Twitch – это видеостриминговый сервис, специализирующийся на тематике компьютерных игр, в том числе геймплея и киберспортивных турниров. Поэтому данная платформа подвержена повышенному уровню интернет-агрессии.

AutoMOD – инструмент модерации на основе ИИ, который блокирует неприемлемый или оскорбительный контент в чате с помощью контроля модератора. Когда пользователь отправляет сообщение, программа отмечает его как потенциально неприемлемое. Отмеченное сообщение направляется модератору чата с пометкой «Разрешить» или «Запретить». Таким образом, программа самостоятельно не блокирует сообщение, а лишь «задерживает» его перед отправкой в чат. [11]

Программа распознает язык идентичности; языковые средства, содержащие сексуальную лексику; агрессивную лексику и нецензурную лексику.

Систему AutoMOD можно настраивать «по уровням». В зависимости от «уровня» программа будет распознавать разный объем информации: от часто блокируемой лексики до всех возможных вариаций неприемлемой лексики. Также программа обрабатывает информацию на большом количестве языков, например, русском, китайском, польском и других. [12]

Система быстро подстраивается под изменения языка и на появление сленга (разговорный или фамильярный стиль речи в особой среде [2]), поскольку позволяет удалять не только контент, который отметила сама, но и контент, который модератор самостоятельно обнаружит в чате, тем самым система постоянно обучается, обрабатывая большие массивы данных.

Также AutoMOD не только самостоятельно отбирает информацию в зависимости от установленного «уровня» работы, но и позволяет автору канала обозначить список нежелательных слов и словосочетаний. Данный список может быть «личным» или «публичным». Личность или публичность списка означает, может или не может модератор чата увидеть данный список. Пользователи платформы Twitch не видят перечень слов и словосочетаний. [8]

Система AutoMOD предназначена в первую очередь для публичных платформ общения, но не подходит для частной переписки, поскольку в ней отсутствует «модератор». В личной переписке возможно использование системы распознавания лексики, а также списка слов и словосочетаний, неприемлемых для человека. В таких случаях система не сможет уберечь от получения жертвой информации, но позволит успешнее расследовать преступление, поскольку сообщения могут быть направлены службе безопасности платформы.

Поскольку система самостоятельно не блокирует и не удаляет контент, существует возможность сбора данных для формирования доказательственной базы, например, система на основе ИИ может хранить выделенные сообщения для дальнейшего изучения правоохранительными органами.

Каким образом возможно применение системы AutoMOD для выявления, расследования и профилактики кибербуллинга? При внедрении данной системы на различные платформы и в дальнейшем подключения ее к единой базе появляется возможность оперативного отслеживания информации. Чтобы избежать указанных ряда проблем, например, искаженность или неполнота информации, на основе которой искусственный интеллект принимает решение, нам кажется нужным внедрение системы проверки результатов работы программы. Поскольку система упрощает процесс обнаружения преступления, нельзя говорить о том, что программа выполняет работу вместо правоохранительных органов. Использование AutoMOD возможно в качестве архива, поскольку, как мы отмечали ранее, она не блокирует контент самостоятельно, а отслеживает и задерживает его, поэтому возможно сохранение выделенных сообщений или публикаций. Также использование системы возможно не только для фиксации травли в комментариях или в личных сообщениях, но и для обнаружения сообществ, созданных для травли жертвы.

Кибербуллинг – быстроразвивающееся и изменяющееся явление, требующее постоянного мониторинга. К сожалению, «человеческого ресурса» не хватает для данной задачи, поэтому использование искусственного интеллекта может позволить упростить процесс выявления преступлений, а также повысить их раскрываемость. В России есть успешные проекты применения ИИ в работе правоохранительных органов – система «Сейф», позволяющая систематизировать информацию о кражах из хранилищ. Таким образом, возможность использования искусственного интеллекта для целей выявления, расследования и профилактики кибербуллинга является важной и необходимой мерой противодействия преступности.

Источники и литература

- 1) Васильева А.Н. Герасимова М.Е. Влияние кибербуллинга на формирование личности подростка // Проблемы современного педагогического образования. 2023. №80 – 1. С. 92-94
- 2) Гончаров В.В. Возникновение и определение понятия «сленг» // Теория и практика современной науки. 2019. №6 С. 181-182
- 3) Ивлева А.А. Исследование актуальности использования стриминговых сервисов как канала продвижения lifestyle-брендов // Скиф. 2021. №9 С. 52-55
- 4) Мазурчук Е.О. Мазурчук Н.И. Девиантное поведение молодежи в виртуальном пространстве: кибербуллинг // Педагогическое образование в России. 2020. № 6. С. 224—229.
- 5) Макарова Е.А. Психологические особенности кибербуллинга как формы интернет-преступления / Е.А. Макарова, Е.Л.Макарова, Е.А.Махрина // Российский психологический журнал.2016. №3. С. 293-311
- 6) Якунина Е.О. Причины школьного буллинга // Вестник науки. 2021. -№5-1. С. 25-40
- 7) Не просто слова: в РФ предложили отключать от интернета за кибербуллинг // Известия [сайт. URL: <https://iz.ru/1846811/valentina-averanova/ne-prosto-slova-v-rf-p-redlozili-otklucat-ot-interneta-za-kiberbulling> (дата обращения: 02.03.2025)
- 8) Как пользоваться функцией AutoMOD // [сайт. URL: <https://help.twitch.tv/s/article/how-to-use-automod?language=ru#BlockedandPermitted> (дата обращения 02.03.2025)
- 9) В 2024 году в России на 15% выросло количество случаев кибербуллинга// ANTI-MALWARE [сайт. URL: <https://www.anti-malware.ru/news/2025-02-11-121598/45246> (дата обращения 02.03.2025)
- 10) В Генпрокуратуре заявили о низкой раскрываемости киберпреступлений в России// РИА Новости [сайт. URL: <https://ria.ru/20241010/kiberprestuplenie-1977431676.html> (дата обращения 02.03.2025)
- 11) Как настроить AutoMOD //сайт. URL: <https://help.twitch.tv/s/article/how-to-use-automod?language=ru> (дата обращения 02.03.2025)
- 12) How to Use AutoMOD //сайт. URL: https://twitch-help.fandom.com/wiki/How_to_Use_AutoMod (дата обращения 02.03.2025)
- 13) Исследование VK Play: российские геймеры осуждают кибербуллинг //сайт. URL: <https://vkplay.ru/media/news/issledovanie-vk-play-rossijskie-gejmery-osuzhdayut-kiberbulling/> (дата обращения 02.03.2025)