11. *Avdeeva Z.K., Kovriga S.V.* Analytical and instrumental framework for the analysis and resolution of stakeholder interest conflict using cognitive maps // Journal of Physics: Conference Series. – 2021. – Vol. 1828. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/1828/1/012108/pdf (дата обращения 14.10.2021).

Мамченко М.В., Рей А.С.

Оценка рисков распространения деструктивного контента в социальных сетях

Аннотация: В работе представлен обобщенный подход к оценке рисков в задаче выявления и блокирования деструктивного контента в социальных сетях, позволяющий вводить дополнительные метрики для оценки значений параметров, связанных с передаваемым сообщением. Представлены структура типовой системы выявления и блокирования деструктивного контента в социальных сетях, место в ней системы оценки рисков и обобщенный алгоритм ее работы. Задействование системы оценки рисков теоретически позволяет сократить количество обращений к модератору, используя значения показателей доверия для отправителя, осуществлять стратификацию и агрегацию категорий деструктивного контента в зависимости от уровня его угрозы, а также контролировать возраст целевой аудитории, являющейся адресатом направляемого сообщения.

Ключевые слова: социальные сети, оценка риска, критерии риска, социо-киберфизическая система, деструктивный контент

В настоящее время сеть Интернет уже возможно рассматривать в качестве социо-киберфизической системы (СКФС), которая оказывает сильное влияние на поведение и способы коммуникации людей. Использование ресурсов Интернета предоставляет большие возможности, но также несет большие риски, связанные с влиянием деструктивного контента на индивидуальное и групповое сознание пользователей. Среди всех пользователей Интернета несовершеннолетние дети и подростки считаются наиболее уязвимой группой риска, так как, будучи одной из наиболее

активных групп в сети Интернет, они начинают взаимодействовать с киберпространством еще с дошкольного возраста. В целом деструктивный контент в сети Интернет и других СКФС является одним из основных негативных факторов воздействия на пользователей (особенно – на молодое поколение), и задача выявления и блокирования подобного контента является актуальной [1]. Социальные сети предпринимают самостоятельные меры по выявлению и блокировке запрещенного контента. В частности, реализуются и совершенствуются алгоритмы фильтрации сообщений и потоковых трансляций пользователей и сообществ в режиме реального времени. Предложено большое количество соответствующих подходов, методов, алгоритмов и архитектур. Например, в работе [2] предложена архитектура мультиагентной системы выявления деструктивного информационного воздействия в социальных сетях, включающая в себя множество агентов, а в статье [1] представлена независимая и федеративная реализации архитектуры отдельной СКФС для выявления разнородного негативного контента в сети Интернет. Недостатком подобных систем является необходимость полной проверки всех сообщений вне зависимости от уровней доверия и охвата его отправителя (источника) и характеристик целевой аудитории, отсутствие разделения деструктивного контента в зависимости от степени его опасности, а также невозможности снижения уровня доверия источника сообщения при попытке передать в нем деструктивный контент. Таким образом, целью настоящей работы является разработка подхода к комплексной оценке риска распространения сообщения в контуре системы выявления и блокировки деструктивного контента.

В соответствии с [3], проведем детализацию задач системы комплексного оценивания риска распространения сообщений в виде дерева критериев. Комплексным показателем (критерием) будет «уровень риска распространения сообщения» (К1), который будет определяться «состоянием угрозы источника» (К21), «возрастным показателем целевой аудитории» (К22) и «уровнем деструктивности контента» (К23). В свою очередь, показатель «состояние источника» будет определяться «уровнем доверия источника» (К211) и «уровнем популярности источника» (К212) (рисунок 1).

Рисунок 1 – Дерево критериев

Введем дискретную шкалу с агрегированными оценками для каждого критерия, где значения уровня риска возрастают от 1 до 4. На основании сформированных критериев и их дискретных оценок сформируем матрицы свертки. Свертка оценок критериев К211 и К212 позволит сформировать оценки критерия К21. Затем возникает необходимость осуществить операцию свертки для оценок трех критериев (К21, К22, К23), чтобы получить значения интегрального критерия К1. В этом случае возможны три комбинации критериев, в результате которых получаются три матрицы свертки, среди которых выбирается матрица с наивысшими выходными значениями уровня риска. Следует отметить, что конкретные значения в матрицах свертки для сопоставления оценок критериев выбираются ответственными должностными лицами или экспертами [3]. Графическое описание формирования матриц свертки для оценок всех критериев представлено на рисунке 2.

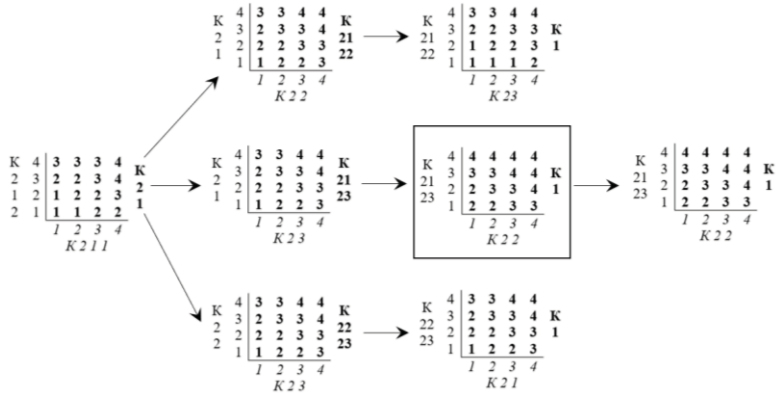

Рисунок 2 – Матрицы свертки для значений оценок критериев

Рассмотрим архитектуру СКФС обнаружения негативного контента в социальных сетях, представленную в [1]. Система состоит отдельно из парсера контента, модуля анализа, базы данных (БД) оценок деструктивных свойств контента и модуля принятия решения, с которым взаимодействует оператор (модератор контента).

Систему оценки рисков возможно внедрить в архитектуру данной системы в виде отдельного блока в составе модуля принятия решения. Кроме того, необходимо внедрение дополнительной БД, содержащей оценки надежности отправителей сообщений. Алгоритм функционирования блока системы оценки рисков (БСОР) включает в себя следующие шаги (действия):

На вход БСОР подаются сведения об отправителе и адресате контента, а также само сообщение и результат работы модуля анализа с оценкой критерия «уровень деструктивности контента» (К23).

Осуществляется обращение к БД, содержащей оценки надежности отправителей сообщений, соответствующие значениям критериев «уровень доверия источника» (К211) и «уровень охвата источника» (К212). Если значения для конкретного отправителя в БД отсутствуют, значения критериев получают наивысшее значение риска (К211 = 4; К212 = 4).

С помощью заданной матрицы свертки из значений критериев К211 и К212 получается оценка критерия К21.

Для формирования оценки критерия «возрастной показатель целевой аудитории» (К22) БСОР проверяет возраст получателей. В случае если получателей несколько (например, при публикации поста в сообществе), формируется массив возрастов получателей. Значение каждого элемента массива сравнивается с диапазонами критерия К22, сопоставляя им определенное значение риска.

Из значений оценок критериев К21, К22 и К23 с помощью трех матриц свертки формируется три значения интегрального критерия, из которых выбирается показатель с максимальным значением риска.

Далее БСОР по заранее установленному правилу принимает одно из следующих решений: переслать сообщение на дополнительную (усиленную) проверку на наличие деструктивного контента; разрешить отправку сообщения потребителям; заблокировать отправку сообщения; или направить сообщение модератору на дополнительную проверку. На усиленную проверку сообщения могут направляться, например, при наличии повышенного (но не максимального) уровня интегрального риска (например, К1 = 3). Если усиленная проверка не выявит деструктивного контента, а новое значение К1 не станет ниже предыдущего, сообщение может быть направлено на проверку модератору. Если отправка сообщения блокируется из-за наличия негативного контента, отправителю присваивается повышенное значение критерия «уровень риска» К211 (вплоть до К211 = 4 – не заслуживающий доверия источник).

Таким образом, оценка рисков в задаче выявления и блокирования деструктивного контента в социальных сетях позволяет вводить дополнительные метрики для оценки значений параметров, связанных с передаваемым сообщением (как между двумя лицами, так и в адрес неограниченного круга пользователей). Внедрение блока системы оценки рисков, алгоритм работы которого описан в настоящей статье, в состав типовой системы обнаружения деструктивного контента теоретически позволяет сократить количество обращений к модератору, используя значения показателей доверия для отправителя (в виде отдельной БД), осуществлять стратификацию и агрегацию категорий деструктивного контента в зависимости от уровня его угрозы, а также осуществлять контроль возраста целевой аудитории.

Исследование выполнено при частичной финансовой поддержке РФФИ в рамках научного проекта № 18-29-22104

Литература:

1. *Kulagina I., Iskhakov A.* Problems of Automation of the Aggression Analysis in Socio-Cyberphysical Environment / Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020). – Advances in Intelligent Systems Research. – 2020. – Volume 174. – P. 35-40.

2. *Охапкин В.П., Охапкина Е.П., Исхакова А.О., Исхаков А.Ю.* Деструктивное информационно-психологическое воздействие в социальных сетях // Моделирование, оптимизация и информационные технологии. – 2020. – №8(1). – С. 1-14.

3. *Новиков Д.А.* Теория управления организационными системами. – М.: Издательство физико-математической литературы, 2012. – 604 с.

Боресков Г.К.

Этические аспекты применения инструментов искусственного интеллекта для обеспечения пространства доверия в электронных СМИ

Аннотация: В работе рассматриваются специфические риски этического характера, связанные с применением инструментов искусственного интеллекта для обеспечения безопасного коммуникационного пространства доверия на площадках электронных СМИ.

Ключевые слова: социальные сети, электронные СМИ, инструменты искусственного интеллекта, безопасность, доверие

За последнее десятилетие рынок телевизионного вещания в Российской Федерации претерпел значительные изменения. Телевизионные информационные программы теряют привлекательность для современной аудитории, в то же время существенно возрастает популярность разного рода интернет-сервисов. Эту тенденцию отражают, например, результаты