УДК 004.056

ФИЛЬТРАЦИЯ ДЕСТРУКТИВНОГО КОНТЕНТА НА ОСНОВЕ АНАЛИЗА ЗАМАСКИРОВАННОГО ТЕКСТА И ОПРЕДЕЛЕНИЯ ИСТОЧНИКА ЕГО РАСПРОСТРАНЕНИЯ

Певзнер А.Д.[1]
Научный руководитель – научный сотрудник, Еритенко Н.А.[1]
[1]Университет ИТМО
wescom1324@gmail.com

Введение

Распространённость деструктивного контента является актуальной проблемой в цифровом пространстве. Фильтрация и удаление опасной информации осложняется из-за огромного количества контента и несовершенства существующих автоматизированных решений, которые не учитывают концептуальный сдвиг [1] в контексте текста и путей распространения контента в Интернет-пространстве. Также, злоумышленники используют обфускацию и различные методы маскировки, для обмана системы распознавания текста. Учитывая все вышеперечисленное, возникает необходимость в создании адаптивной системы выявления деструктивного контента с поиском источника, для предотвращения его дальнейшего распространения.

Основная часть

В рамках разработки адаптивной системы реализованы следующие задачи:

1. Определение распространенных форм деструктивного контента в России. Обзор российских и зарубежных решений по борьбе с деструктивным контентом и результаты их тестирования;
2. Представление схемы разрабатываемой системы, на основе NLP-модели [2]. Особенностями системы обнаружения деструктивного контента являются модули поиска источника распространения, с учетом неявных связей пользователей через публикации и группы [3]. Рассматривается выставление приоритетов поиска контента, его категоризация, оценка деструктивности и составление отчета, а также регуляция системы через элементы дообучения языковой модели и пополнение словаря статических методов;
3. Анализ и оптимизация алгоритмов, используемых для распознавания деструктивного контента, с учетом изменения информации в текстовом виде с течением времени или с целью маскировки противоправной информации.

Выводы

В результате проектировки системы были учтены недостатки существующих систем и алгоритмов в данной области и предложено решение выявления деструктивного контента при помощи распознавания замаскированного текста. Данные, полученные в результате анализа признаков, способствуют определению источника распространения [4]. В дальнейшем планируется комплексная реализация системы и тестирование в социальных сетях.

Литература

1. Understanding Data Drift and Why It Happens URL: https://www.dqlabs.ai/blog/understanding-data-drift-and-why-it-happens/
2. Цитульский Антон Максимович, Иванников Александр Владимирович, Рогов Илья Сергеевич NLP - обработка естественных языков // StudNet. 2020. №6. URL: https://cyberleninka.ru/article/n/nlp-obrabotka-estestvennyh-yazykov.
3. Кузин, М. А. Автоматизированная информационная система анализа и оценки активности пользователей в социальной сети Вконтакте: Социальные связи и

взаимодействия / М. А. Кузин, В. В. Родионов // Информатика и вычислительная техника : Сборник научных трудов XVI Всероссийской научно-технической конференции аспирантов, студентов и молодых ученых, Ульяновск, 13–14 июня 2024 года. – Ульяновск: Ульяновский государственный технический университет, 2024. – С. 125-131. – EDN QHCVXH.

4. Куртукова Анна Владимировна, Романов Александр Сергеевич, Федотова Анастасия Михайловна, Шелупанов Александр Александрович ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ОТБОРА ПРИЗНАКОВ НА ОСНОВЕ ГЕНЕТИЧЕСКОГО АЛГОРИТМА В РЕШЕНИИ ЗАДАЧИ ОПРЕДЕЛЕНИЯ АВТОРА РУССКОЯЗЫЧНОГО ТЕКСТА ДЛЯ КИБЕРБЕЗОПАСНОСТИ // Доклады ТУСУР. 2022. №1. URL: https://cyberleninka.ru/article/n/primenenie-metodov-mashinnogo-obucheniya-i-otbora-priznakov-na-osnove-geneticheskogo-algoritma-v-reshenii-zadachi-opredeleniya.