УДК 004.934

АНАЛИЗ МЕТОДОВ ДЕТЕКТИРОВАНИЯ ТОКСИЧНЫХ СООБЩЕНИЙ

Мостицкий В.О. (Университет ИТМО)

Научный руководитель — к.т.н. Махныткина О.В. (Университет ИТМО)

Введение. В настоящее время, благодаря цифровым технологиям, мы имеем возможность общаться с миллионами людей по всему миру в режиме реального времени. Однако, среди всех преимуществ онлайн-коммуникаций, есть и негативные стороны, такие как распространение токсичных сообщений. В наши дни все больше людей использует социальные сети, форумы и мессенджеры для общения и обмена информацией. Но, к сожалению, такие платформы также могут стать местом для распространения оскорбительного контента, угроз и ненависти, которые могут нанести вред психическому здоровью людей и нарушить общественную гармонию. В связи с этим, обнаружение токсичных сообщений является одной из самых актуальных проблем в области регулирования онлайн-коммуникации.

Основная часть. Работа посвящена исследованию методов детектирования токсичных сообщений в цифровых коммуникациях. Основной задачей работы является анализ различных методов предварительной обработки текстовых данных, извлечения признаков и классификационных моделей, используемых для решения задачи детектирования токсичных сообщений. Особое внимание будет уделено сравнительному анализу различных методов, выявлению методов, показывающих наилучшее качество. Результаты данного исследования могут быть полезны для создания эффективных систем фильтрации токсичного контента и улучшения качества цифровых коммуникаций [1], [2]. Одной из главных проблем для создания систем детектирования токсичных сообщений на русском языке является недостаточное количество доступных наборов. В работе будет представлен обзор существующих наборов данных и предложены способы увеличения доступных данных.

Вывод. Целью данного исследования является анализ методов детектирования токсичных сообщений в цифровых коммуникациях и оценка их эффективности. В работе будут рассмотрены различные подходы к детектированию токсичности сообщений, включая методы машинного обучения. Основной задачей исследования является оценка эффективности каждого из методов и выявление наиболее эффективных подходов к детектированию токсичных сообщений.

Список использованных источников:

1. Risch, J., Krestel, R.: Toxic comment detection in online discussions. In: Deep learning-based approaches for sentiment analysis. Springer, 2020.
2. Kiritchenko, S. Automatic Identification of Hate Speech in Social Media Text. First Workshop on Trolling, Aggression and Cyberbullying, 2018.