УДК 004.89

ИССЛЕДОВАНИЕ И ПРИМЕНЕНИЕ СОВРЕМЕННЫХ НЕЙРОСЕТЕВЫХ АРХИТЕКТУР В ЗАДАЧЕ ДЕТЕКТИРОВАНИЯ ТОКСИЧНОСТИ ПИСЬМЕННОЙ РЕЧИ НА РУССКОМ ЯЗЫКЕ

Кореневская А.М. (ИТМО)

Научный руководитель – кандидат технических наук, доцент Махныткина О.В. (ИТМО)

Введение. При человеко-машинном взаимодействии встречается проявление агрессии и неуважения как со стороны пользователей, так и диалогового агента. В связи с этим возникает необходимость в разработке систем, способных автоматически определять и фильтровать токсичные реплики. В данной работе проведен анализ существующих подходов к обучению детектора токсичной речи, а также предлагается вариант нейросетевой архитектуры для решения задачи детектирования токсичности письменной речи на русском языке.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 623088 «Разработка русскоязычного персонифицированного эмоционального диалогового агента».

Основная часть. Задача детектирования токсичной речи включает построение скрытых представлений слов или других синтаксических единиц. Многие нейросетевые архитектуры, такие как Word2Vec, GloVe или BERT, порождают высокоуровневые эмбеддинги слов и показывают хорошие результаты для задачи классификации токсичных и нетоксичных реплик. Авторы статьи [1] предлагают использовать для классификации механизм внимания [2] применительно к n-граммам разной длины. Такой подход позволяет детектировать токсичные сообщения, принимая в расчет контекст предложения.

В данной работе проведено исследование существующих методов детектирования токсичной речи, а также предлагается нейросетевая архитектура, представляющая собой усреднение архитектуры BERT с моделью, предложенной в статье [1].

Выводы. Проведено исследование подходов к детектированию токсичности письменной речи нейросетевыми методами. По результатам исследования предложена архитектура для решения задачи детектирования токсичности в русскоязычном тексте.

Список использованных источников:

1. Jarquín-Vásquez, H.J., Montes-y-Gómez M.,Villaseñor-Pineda. Not All Swear Words Are Used Equal: Attention over Word n-grams for Abusive Language Identification // Lecture Notes in Computer Science, v. 12088. – 2020. – p. 282–292.

2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. – 2017. – URL: https://arxiv.org/pdf/1706.03762.pdf (дата обращения 29.02.2024).