УДК 004.912

ИССЛЕДОВАНИЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ КЛАССИФИКАЦИИ ТОКСИЧНЫХ СООБЩЕНИЙ

Богорадникова Д.А. (Университет ИТМО)
Научный руководитель – к.т.н. Махныткина О.В.
(Университет ИТМО)

В данной работе рассмотрено применение глубоких нейронных сетей для задачи классификации токсичных сообщений. Представлено описание архитектур сверточных, рекуррентных нейронных сетей, трансформеров и набора данных.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 620183 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе моделей-трансформаторов и понимания естественного и математического языка».

Публикация комментариев в осуждениях на различных онлайн-платформах является одним из основных способов выразить своё мнение в Интернете. Но, как и в реальной жизни, при обсуждении того или иного можно подвергнуться словесной атаке: враждебно настроенные пользователи нередко мешают культурному общению своими токсичными комментариями. Токсичный комментарий определяются как грубый, неуважительный или чрезмерно агрессивный комментарий, который может заставить других пользователей покинуть обсуждение.

Для исследований был выбран датасет, предоставленный на платформе Kaggle в рамках соревнования «Jigsaw Unintended Bias in Toxicity Classification», проводимого в 2019 году. Набор содержит 1804874 комментария, из них 144334 комментария (около 8%) токсичны и 1660540 нетоксичны.

В настоящее время для задач классификации текстов используются различные архитектуры глубоких нейронных сетей, точность классификации при этом зависит от выбранной архитектуры и имеющихся данных.

В работе рассматриваются:

- сверточные нейронные сети (CNN);
- рекуррентные нейронные сети (RNN), такие как LSTM и GRU;
- трансформеры.

Сверточные нейронные сети (CNN) получили свое название из-за использования в их архитектуре операции свертки, которая позволяет выявить местные закономерности рассматриваемых данных. Благодаря этому, CNN эффективны при решении задачи распознавания образов, анализа тональности и извлечения именованных сущностей.

Рекуррентные нейронные сети (RNN) являются мощным инструментом для последовательных данных, таких как текст, видео и речь, поскольку позволяют «запоминать» информацию о вычислениях для всей последовательности данных. Однако, нередко такие сети страдают от проблем исчезающего и взрывающегося градиентов. Слои LSTM (долгая краткосрочная память) и GRU (управляемые рекуррентные блоки) преодолевают проблему исчезающего градиента с помощью фильтров.

Трансформеры, как и RNN, направлены на обработку последовательностей данных, таких как текст, благодаря чему они активно используются при обработке естественного языка и связанных с ней задач машинного обучения. В общем случае, сеть трансформер состоит из кодировщика и декодировщика, однако они могут использоваться и по отдельности. Например, BERT, XLNet и XLM используют слои кодировщика, а GPT-2 – декодировщика.

Таким образом, исследование заключается в проведении сравнительного анализа использования рассмотренных архитектур глубоких нейронных сетей для решения задачи бинарной классификации токсичных сообщений.