УДК 004.89

ИССЛЕДОВАНИЕ ТЕХНОЛОГИЙ ДЕТЕКТИРОВАНИЯ ЭМОЦИЙ

Котов Д.О.

(Университет ИТМО, г. Санкт-Петербург)

Научный руководитель – к. т. н. Махныткина О.В.

(Университет ИТМО, г. Санкт-Петербург)

В данной работе были протестированы стандартные нейросетевые архитектуры (CNN, LSTM, BiLSTM) а также модель BERT на существующих русскоязычных датасетах для сентимент-анализа текста.

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».

Исследования в области сентимент анализа в настоящее время в основном основываются на нейросетевых подходах, что предполагает обучение и тестирование на специально размеченных датасетах. Англоязычные популярные датасеты для этой задачи включают в себя Sentiment Treebank datasets SST, IMDB dataset, Twitter sentiment datasets и многие другие. Для остальных языков существует гораздо меньше данных. Для русского языка несколько таких датасетов было создано, а именно: ROMIP2012-2013 и SentiRuEval2015-2016. Они включают в себя размеченные данные обзоров фильмов, книг и цифровых камер, вырезки из новостей и твиты. Так как конкурсы, для которых эти данные были агрегированы, проходили более пяти лет назад, лучшие результаты были получены такими классическими моделями машинного обучения, как SVM, рекуррентные нейронные сети и даже методы, основанные на правилах и лексиконе.

В данной работе предпринята попытка применения модели BERT, основанной на архитектуре трансформера, для решения задачи сентимент анализа на уровне аспектов на вышеперечисленных датасетах. Была проверена гипотеза, о том, что эта модель способна показать результат выше, чем классические алгоритмы машинного обучения. Также был проведен сравнительный анализ различных моделей, предобученных на русскоязычных данных, и апробированы современные популярные подходы применения модели BERT к различным задачам.