УДК 004.912

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПОДХОДОВ К ВЫЯВЛЕНИЮ ТОКСИЧНЫХ СООБЩЕНИЙ

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

В работе рассматривается возможность применения заранее выделенных дополнительных признаков для выявления токсичных сообщений и их влияние на точность классификации различными методами машинного обучения.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 619423 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе аргументационного подхода и глубокого машинного обучения»

В настоящее время высказать свое мнение касательно того или иного вопроса в сети Интернет достаточно просто. Однако, не всегда обмен мнениями происходит в культурной форме, нередко собеседники прибегают к нецензурной лексике и другим высказываниям, способным оскорбить оппонента. Такого рода комментарии чаще всего называются токсичными, и для их выявления нередко используются методы машинного обучения. Так, Google и Jigsaw уже два года проводят соревнование, в ходе которого участники предлагают различные модели, способные не только определить, является ли комментарий токсичным, но и отнести его к одному из классов.

Предоставленные данными компаниями датасеты используются в научных исследованиях. Например, в статье «A Multi-Task Deep Learning Approach» авторы используют набор «Toxic Comment Classification Challenge» 2018 года для построения многоуровневой архитектуры, включающей в себя FastText как метод извлечения признаков из текста, сверточные и рекурентные нейронные сети. Авторы статьи «Convolutional Neural Networks for Toxic Comment Classification» для обработки текста используют метод Мешка Слов (Bag Of Words) с мерой TF-IDF, а классификатора использована CNN. Статья «Challenges for toxic comment classification: An in-depth error analysis» рассматривает ансамбль, решающий, какой из классификаторов наиболее эффективен для конкретного вида комментариев. Ансамбль наблюдает особенности в комментариях, взвешивает и изучает оптимальный выбор классификатора для данной комбинации признаков. В качестве классификаторов используются Логистическая регрессия, RNN и CNN, для извлечения признаков из текста используются Glove и FastText. Помимо уже перечисленных алгоритмов классификации, для сравнения в статьях так же рассматриваются Наивный байес, Метод опорных векторов, метод k-ближайших соседей и линейный дискриминантный анализ.

В данной работе в качестве набора данных был выбран набор, предоставленный на платформе Kaggle в рамках соревнования «Jigsaw Unintended Bias in Toxicity Classification». Данный датасет включает комментарии, собранные со страниц обсуждения Википедии на английском языке. В качестве ключевых слов использовались значения, принадлежащие столбцам, в которых содержатся данные о гендерной принадлежности, расе и религии. Однако, данные представлены не для каждого комментария, поэтому из всего датасета были выбраны те сообщения, для которых хотя бы в одном из рассматриваемых 16 столбцов значение превышает 0,5.

Далее была проведена предварительная обработка текста, включающая в себя удаление знаков препинания и стоп-слов, токенизацию и лемматизацию. После было получено векторное представление текстов с помощью метода Word2vec. Для каждого предложения был построен вектор размерностью в 300 элементов. После чего для каждого вектора была проведена конкатенация, в ходе которой размер был увеличен до 316 элементов.

В качестве классификаторов были выбраны одни из наиболее простых и действенным методов, а именно метод опорных векторов, так же известный как SVM, и логистическая регрессия. Для оценки качества построенных моделей были использованы такие метрики как precision (точность), recall (полнота) и F-мера.

В целом, использование дополнительных признаков может иметь смысл, однако следует учитывать распространенность данных признаков и их наличие для сообщений в больших количествах, нежели при выполнении данной работы.