УДК 004.85

СРАВНЕНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ НЕПРИЕМЛЕМОГО ОБЩЕНИЯ В ИНТЕРНЕТ-КОММЕНТАРИЯХ НА РУССКОМ ЯЗЫКЕ

Галиулина В.А. (Университет ИТМО)
Научный руководитель – к.ф.-м.н., доцент, Михайлова Е.Г.
(**Университет ИТМО**)

Аннотация. В работе представлено сравнение методов распознавания неприемлемого общения среди пользователей сети на русском языке. Реализация подходов к классификации текстов осуществляется в несколько этапов: предварительной обработки естественного языка, получения векторных форм слов и построения моделей машинного обучения. В итоге по моделям получены оценки их качества и интерпретированы результаты по наилучшей из них.

Введение. Общение в сети Интернет давно стало новой реальностью в жизни современного общества. Объем информации в виде комментариев, оставляемых пользователями сети, многократно увеличивается каждую секунду, что создает проблему невозможности вручную контролировать весь поток информации. Также, одной из важнейших проблем онлайн общения является высокая склонность к конфликтам ввиду присущих ему признаков: неформальности, личностного и бытового ориентирования, опосредованности и анонимности. Согласно опросу АО «Тинькофф Банк» 73% респондентов сталкивались с токсичным общением в сети в свой адрес.

Токсичные комментарии – это комментарии пользователей сети, содержащие неприемлемый контент: оскорбления, нецензурную лексику, угрозы и агрессию. Последствия такого общения могут быть еще более негативными: деградация культуры общения, конфликты между разными социальными группами, рост социальной напряженности в обществе.

В западном обществе проблема токсичных комментариев уже давно стоит достаточно остро, и задача детектирования токсичного общения при помощи методов машинного обучения является актуальной в крупных компаниях в последние три года. Уже существуют успешные результаты разработки моделей распознавания токсичных комментариев на английском языке, в том числе, с применением нейронных сетей. Отечественных работ, посвященных задаче распознавания неприемлемых комментариев в сети на данный момент мало, так как спрос на их разработку стал появляться намного позже, чем в западных странах. Более того, в российском Интернет-пространстве реализовать такую задачу технически сложнее ввиду особенностей русского языка и процесса его машинной обработки. Также только относительно недавно в свободном доступе появились качественные данные (комментарии пользователей, размеченные по принципу наличия токсичности).

Основная часть. Распознавание неприемлемых комментариев – это задача классификации, поскольку классы заранее определены. Исходные данные содержат 14 412 размеченных комментариев пользователей с оценкой их токсичности (1 - токсичный, 0 - приемлемый). Решение задачи распознавания токсичных комментариев осуществляется при помощи построения моделей машинного обучения на предварительно преобразованных текстовых данных на языке Python.

Методы машинного обучения невозможно применить непосредственно к тексту на естественном языке, поэтому перед этапом моделирования необходимо провести тщательное преобразование текста, которое включает в себя исправление опечаток, орфографических ошибок, удаление лишних символов и пробелов, цифр, а также знаков пунктуации. Затем осуществляется токенизация – разбиение текста на отдельные слова. Все слова приводятся к нижнему регистру. После очистки текста осуществляется процесс лемматизации – нормализации слова, который заключается в приведении слова к начальной форме. Принцип лемматизации индивидуален для каждого языка, поэтому существует несколько инструментов на языке Python для реализации алгоритма лемматизации. Для русскоязычных текстов

рекомендованы инструменты PyMorphy2 и PyMyStem3. В данной работе используется PyMorphy2, так как он быстрее и эффективнее работает с текстами, разбитыми на предложения, более того, его алгоритм способен обрабатывать неизвестные ему слова (что и требуется для решения задачи обработки неформальных текстов).

После лемматизации производится преобразование текстовых данных в векторные формы в виде числовых массивов. Способов векторизации текстов существует несколько, необходимо подобрать оптимальный, так как от этого зависит качество итогового результата. Для русского текста подходят инструменты Word2vec и fastText. Принцип векторизации с помощью Word2vec следующий: каждому слову сопоставляется вектор, создается словарь, а затем вычисляются векторные представления слов таким образом, что слова со схожим смыслом имеют высокое значение косинусного сходства. То есть, данный метод векторизации позволяет учесть контекст отдельного слова, но его недостаток в том, что он не распознает слова, отсутствующие в обучающей выборке. Метод fastText наоборот хорошо работает с редко встречающимися словами, но не учитывает контекст отдельных слов. Это объясняется принципом его работы: слова разбиваются на N-граммы (части слова), создаются векторные представления N-грамм, которые суммируются для создания векторного представления слова, что позволяет получить векторное представление незнакомого слова, суммируя векторные представления уже известных N-грамм.

После того как получены векторные представления слов, можно приступить к моделированию. В данной работе рассмотрены методы машинного обучения, использовавшиеся зарубежными исследователями для классификации англоязычных текстов: метод логистической регрессии и случайного леса. Модель логистической регрессии отлично подходит для задачи классификации текста на предмет наличия токсичности, поскольку существуют только 2 класса комментариев: токсичный и приемлемый. Модель случайного леса имеет высокую точность предсказания, редко переобучается и предполагает возможность сбалансировать вес каждого класса, что важно при их несбалансированности.

Для оценки качества моделей используются метрики:

1) Accuracy – доля верно определенных алгоритмом объектов;
2) Precision (точность) – доля объектов, названных классификатором положительными и при этом действительно являющимися положительными;
3) Recall (полнота) – доля найденных алгоритмом объектов положительного класса из всех объектов положительного класса;
4) F1 – агрегированный показатель из Precision и Recall.

Выводы. В данной работе были применены методы машинного обучения к предобработанным комментариям на русском языке и получены следующие результаты: F1 = 81% по модели логистической регрессии и F1 = 76% по модели случайного леса. Наилучшей оказалась модель логистической регрессии: доля верно распознанных токсичных комментариев составила почти 80%, при этом accuracy модели составляет 90%, то есть, верно определено 90% всех комментариев.

Таким образом, применяя несложные методы машинного обучения к русскоязычным комментариям, удалось добиться неплохих результатов распознавания токсичности, не прибегая к разработке нейросетей.

Галиулина В.А. (автор) Подпись

Михайлова Е.Г. (научный руководитель) Подпись