Секция «Лингвистика: Современные лингвистические исследования: фонетика, грамматика, лексика»

Автоматическая классификация направленности агрессии для сообщений анонимных форумов.

Гордеев Денис Игоревич

Аспирант

Московский государственный лингвистический университет, Москва, Россия

E-mail: achil92@mail.ru

Среди методов позволяющих изучать тональность текста и эксплицитное проявление эмоций наибольший интерес представляют автоматические методы, так как они позволяют используя ограниченные ресурсы анализировать многие сообщения и не требуют дорогостоящего времени квалифицированных лингвистов-модераторов. Кроме того, автоматические методы обычно предполагают формализацию использованных методов и широко описываются в научной литературе.

Однако работ посвященных именно анализу агресии и кибербулинга в сети не так много. Так M. Dadvar и другие анализировали кибербуллинг для сайта MySpace и использовали гендерный подход для анализа этого явления [3]. Они использовали метод опорных векторов и его имплементацию в библиотеке WEKA. С помощью MOB классификатора они проанализировали 381000 сообщений. В качестве критериев классификации они использовали самую употребимую обсценную лексику английского языка, наличие личных местоимений и наличие прочих местоимений, также они привели в качестве критерия вес каждого слова, использую TFIDF алгоритм. К сожалению, точность их алгоритма составляет 31% без учета пола пользователей. Гендерный подход позволяет улучшить результат до 43%.

Стоит правда отметить, что задача исследования тональности довольна близка к анализу агрессию, так как оба направления занимаются различными человеческими эмоциями, и с небольшими доработками методы анализа общей тональности могут использоваться для анализа вербальной агрессии. Изучение тональности в твиттере и других социальных сетях особенно близко теме нашего исследования, поскольку сообщения на анонимных форумах обычно короткие, так средняя длина сообщения на 4chan.org составляет 15 слов [7] и не более 140 символов на твиттере.

Огромное число работ было опубликовано по данным и смежным темам в последние годы. Correa и др. изучали влияние полной анонимности на поведение пользователей [2] по сравнению с частичной анонимностью твиттера. Они обнаружили, что пользователи более открыты и легче выражают негативные эмоции (не только агрессию) в анонимной среде. Однако они изучали сайт Whisper, специально созданный для того, чтобы делиться секретами и признаниями, это могло повлиять на результаты.

Martínez-Cámara провёл обзорное исследование различных методов анализа тональности в твиттере [4]. Другое исследование было проведено Dos Santos, который успешно (от 76% до 88% точности для различных выборок) устанавливал анализ тональности для сообщений в твиттере [8] без использования неавтоматических параметров, но он обладал большим объёмом аннотированных данных. Tang and Wei анализировали сообщения в твиттере, используя смайлики и нейронные сети [9]. Как мы видим, многие современные исследователи используют машинное обучение и нейронные сети для определения тональности. Как бы то ни было, Paltoglou [6] утверждает, что «неконтролируемые» методы, основанные на словарях, превосходят «передовые» методы машинного обучения. Нужно сказать, что он не упомянул методы, основанные на нейронных сетях или глубин-

ном обучении, а результаты сложно применить для других языков, кроме английского.

В нашей работе мы классифицировали сообщения, содержащие агрессию. Сначала мы аннотировали слова по категории агрессии. По направленности агрессии мы делили сообщения на психологофизиологические, этнические, социально-экономические, политические, геополитические, конфессиональные, культурологические, нравственно-этические. Сообщения, не попадающие в эти категории попадали в категорию «прочее». Затем мы выделили некоторые критерии, и на основе этих критериев тренировали Random Forest [1] классификатор. В качестве критериев была принята семантическая близость данного сообщения до наиболее популярных фраз их соответствующей категории. Семантическая близость вычислялась с помощью алгоритма word2vec [5].

Источники и литература

1) Breiman L. Random forests // Machine learning. 2001.
2) Correa D. [и др.]. The Many Shades of Anonymity: Characterizing Anonymous Social Media Content Oxford, UK:, 2015.
3) Dadvar M. [и др.]. Improved cyberbullying detection using gender information // 2012.
4) Martínez-Cámara E. [и др.]. Sentiment analysis in Twitter // Natural Language Engineering. 2014. № 01 (20). С. 1–28.
5) Mikolov T. [и др.]. Vector Space С. 1–12.
6) Paltoglou G., Thelwall M. Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media // ACM Transactions on Intelligent Systems and Technology. 2012. № 4 (3). С. 1–19.
7) Potapova R., Gordeev D. Determination of the Internet Anonymity influence on the level of aggression and usage of obscene lexis BT - Proceedings of the 17th International conference Speech and Computer (SPECOM 2015). Athens, Greece, September 20-24, 2015 University of Patras Press, Patras, 2015. 29–36 с.
8) Santos C.N. Dos Think Positive: Towards Twitter Sentiment Analysis from Scratch // Semeval-2014. 2014. С. 647–651.
9) Tang, D. and Wei, F. and Yang, N. and Zhou, M. and Liu, T. and Qin B. Learning sentiment-specific word embedding for twitter sentiment classification // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. С. 1555–1565.

Слова благодарности

Исследование частично финансировалось в рамках проекта № 14-18-01059 Российского научного Фонда (РНФ) на базе МГЛУ. Науч. рук. доктор фил. наук, профессор Р. К. Потапова.