Программа для анализа тональности сообщений в социальных сетях

Сметанин Сергей Игоревич

Студент

Факультет компьютерных наук Национального исследовательского университета Высшая школа экономики, Москва, Россия

E-mail: sismetanin@gmail.com

Одной из наиболее актуальных задач в области анализа текстовых сообщений в социальных сетях является задача распознавания эмоциональной окраски текста, которая позволяет извлечь из текстовой информации мнение человека об объекте и его характеристики. Согласно [1], на 30 сентября 2015 года аудитория социальной сети Twitter составляла 320 миллионов активных пользователей в месяц на более чем 35 разных языках. Нет сомнений, что объемы генерируемых сообщений делают невозможным обработку этих данных человеческими силами.

В докладе рассматривается реализации программы для автоматического анализа тональности сообщений из русскоязычного сегмента социальной сети Twitter. Для бинарной классификации эмоциональной окраски сообщений [2] была разработана программа на языке Python, обрабатывающая сообщения в несколько этапов.

На первом этапе текст проверяется на наличие эмотиконов (пиктограмма либо последовательность типографских знаков, изображающая эмоцию). Эмоциональная окраска каждого эмотикона задавалась согласно экспертной оценки автора работы. Если сообщение содержит эмотикон, то тональность сообщения определяется тональностью эмотикона. В обратном случае, либо если сообщение содержит положительный и отрицательный эмотиконы, программа переходит на следующий этап.

Как правило, нормы общения в социальных сетях отличаются от норм литературного языка. Сообщениям в социальных сетях свойственны орфографические и пунктуационные ошибки, опечатки, сленг, использование эмотиконов и авторская пунктуация, что значительно затрудняет автоматический анализ. Для решения данной проблемы был предложен метод автоматической предварительной обработки текста: сначала удаляются символы, не являющиеся буквами. Далее строка приводится к нижнему регистру. Последовательности из трех одинаковых символов заменяются на последовательности из двух таких же символов. Каждое слово приводится

в начальную форму с последующим извлечением леммы (каноническая форма слова) с помощью библиотеки PyMorphy.

На последнем этапе используется наивный байесовский классификатор [2] с мультиномиальной моделью распределения, обученный на корпусе коротких текстов Юлии Рубцовой [3], который содержит 114911 положительных записей и 111923 отрицательных. Для решения проблемы неизвестных слов (у слов, не встретившихся в обучающей выборке, вероятность принадлежности к какому либо из классов равна нулю) применялось аддитивное сглаживание (сглаживание по Лапласу) [4]. Оценка эффективности алгоритма осуществлялась в критериях точности и полноты. Для усреднения показания метрик качества классификации применялся скользящий контроль (10-fold cross-validation [5]) с реализацией из библиотеки Scikit-Learn. Так же для контроля использовалось подмножество данных из 500 сообщений, размеченных автором работы.

Таким образом, была разработана программа для автоматического анализа тональности сообщений в русскоязычном сегменте социальной сети Twitter на основе методов машинного обучения. Был реализован наивный байесовский классификатор с мультиномиальной моделью распределения, выбраны метрики и произведены расчеты эффективности классификации, проведено тестирование методом кросс-валидации. Полученная точность классификации сопоставима с точностью современных аналогов.

Литература

1. О компании Твиттер: `https://about.twitter.com/ru/company`
2. Котельников Е. В., Клековкина М .В. Автоматический анализ тональности текстов на основе методов машинного обучения. Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции Диалог. том 2, 2012. С. 27–36.
3. Корпус коротких текстов на русском языке на основе постов Twitter: `http://study.mokoron.com/`
4. Additive smoothing: `https://en.wikipedia.org/wiki/Additive_smoothing`
5. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, No. 2 (12). (1995), P. 1137–1143.