Эмоциональная окраска кратких высказываний по материалам SetiRuEval_2016

Дмитрий Чубаров

13.03.2017

Аннотация

Представлены первые результаты работы с корпусом `bank_train_2016`. Вычислены базовые статистические характеристики корпуса, предложено правило для выделения эмоционально окрашенных высказываний, качество распознавания по мере $F_1 = 0.19$. Тем самым обоснована необходимость применения более сложных моделей. Определены направления для дальнейшей работы.

Характеристика задачи

В социологии, философии науки и бизнесе известна концепция Zeitgeist – в каждый момент времени в истории человеческого общества существует набор настроений, представлений о мире и убеждений, некоторым образом влияющий на действия каждого человека. Регулярный мониторинг социальных сетей, состоящий в определении преобладающих тем, а также отношения к ним, может быть использован для того, чтобы измерять "дух времени".

Каждое сообщение имеет ряд характеристик: автор сообщения – *субъект*, предмет, к которому относится сообщение – *объект*, объект может обладать несколькими *аспектами*, по отношению к которым субъектом дается эмоциональная оценка, и собственно *тональность* или *сентимент* – эмоциональная оценка, определенная в некоторой шкале. Постановка задачи подробно изложена в [Лукашевич, 2015].

Для разработки алгоритма определения эмоциональной окраски сообщения предоставлены два размеченных корпуса кратких сообщений: `banks_` и `ttk_`. В последующем тексте представлены результаты работы с корпусом `bank_train_2016`.

Характеристика входных данных

Входные данные представляют собой XML–дамп базы данных СУБД MySQL, содержащий выгруженные из Twitter сообщения на русском языке с экспертной разметкой. Каждая запись содержит дату, текст сообщения, а также

оценки эмоциональной окраски сообщения по отношению к 8-ми объектам — известным банкам, действующим в нескольких странах.

Экспертные оценки зачастую несогласованы. Обзор первых 100 сообщений показал, что в четырех случаях оценки не согласованы.

```
5. @sawik_shuster @YevhenS Главное чтоб банки СБЕР и ВТБ!!!
```

Экспертом даны положительные оценки, но из контекста ясно, что речь идёт о санкциях по отношению к упомянутым банкам. Очевидно, что субъект относится к объекту негативно.

```
56. Сравнить кредитные ставки в сбербанке газпромбанке ираффайзенбанке
```

Экспертом сообщение не отнесено к райффайзенбанку

В следующих двух сообщениях эксперт определил нейтральное отношение, хотя во втором случае несомненна негативная эмоциональная окраска, а в первом можно дать положительную оценку.

```
77. #СМИ Сбербанк поможет своему должнику построить Керченский
    мост: @dailysmi_net Глава Сбербанка Герман...
```

```
92. @sberbank Но и когда уже Вы созреете господа банкиры.
```

Обзор 100 сообщений не позволяет дать статистически значимую нижнюю оценку доли неверно классифицированных экспертом сообщений, но для наших целей достаточно знать, что она вероятно не превышает 15%.

Статистика

Для дальнейшей работы с данными они были загружены в СУБД PostgreSQL с помощью XML–парсера и AWK. Типы данных были преведены к соответствующим типам данных СУБД.

Загрузка в СУБД позволила быстро получить некоторые статистические характеристики корпуса. Корпус `bank_train_2016` содержит 9392 сообщений,среди которых 704 содержат положительные оценки, 1734 негативных оценки и 6977 нейтральных. Небольшое число сообщений содержит высказывания различной окраски по отношению к разным объектам: 2 сообщения содержат как положительную, так и отрицательную оценку, 2 положительную и нейтральную и 19 отрицательную и нейтральную. Таким образом, на первом этапе работы можно пренебречь случаями смешения оценок.

Выборка смещена в сторону нейтральных оценок. Таким образом, первая задача состоит в отделении нейтральных оценок от эмоционально окрашенных.

При загрузке обнаружилось, что некоторые сообщения разбиты на несколько строк. Корректная обработка таких сообщений потребовала дополнительной настройки XML–парсера.

Частотный словарь

Многие известные технологии автоматизированного измерения эмоциональной окраски рассматривают сообщение как мультимножество слов и оценивают тональность по соотношению положительно и отрицательно окрашенных слов. Наличие размеченного корпуса позволяет построить соответствующий словарь. Этот подход обладает своими ограничениями. Даже исключая сравнения "Сбербанк лучше ВТБ", этот подход не позволяет учитывать грамматическую структуру языка.

```
569. У @sberbank на вопросы в личном кабинете Сберонлайна
     все равно отвечают "обратитесь в отделение".
     В чем тогда смысл улучшеший?
```

Тем не менее, ключевые слова являются значимым признаком эмоционально окрашенных слов. Был получен частотный словарь для заданного корпуса. Для этого были удалены знаки препинания: `.‘‘,‘‘!,#,"`; также были удалены гиперссылки, а прописные буквы заменены строчными. Размер полученного таким образом частотного словаря составил 14670 записей. Лемматизация и дополнительный отсев слов позволили бы сильнее сократить размер словаря. Большой по сравнению с размером обучающей выборки размер словаря потребует дополнительных усилий, чтобы избежать переобучения при применении методов машинного обучения.

Возможности оценки эмоциональной окраски по ключевым словам были проверены на нескольких примерах.

Слово "*санкции*" хорошо коррелирует с негативной оценкой: из 781 сообщения, содержащих формы этого слова, 652 имеют негативную оценку, 99 нейтральную и 33 положительную.

Слово "лучше" (см. пример выше), коррелирует с положительной оценкой: из 59 сообщений 33 имеют положительную оценку, 10 отрицательную и 17 нейтральную.

"Смайлики" также коррелируют с соответствующими эмоциями: `))` и `:)` – 27 положительных, 33 нейтральных, 27 негативных; `((` и `:(` — 2 нейтральных, 24 негативных.

Feature Engineering

Многие сообщения в Twitter содержат ссылки на внешние источники или онлайн-приложения. По наличию ссылки был определен признак `has_link`. Сообщения также содержат признак ретвита – буквы `RT` в начале сообщения – признак `is_retweet`.

Правило `not has_link or is_retweet` было проверено для выделения эмоционально окрашенных сообщений. Это правило даёт precision 14% и recall 30%, что соответствует значению 0.19 по мере F_1. Этот результат может быть принят за базовую величину для оценки более сложных методов.

Результаты

На первом этапе работы выполнена загрузка корпуса в СУБД, определены некоторые статистические характеристики корпуса. Предлжен метод, который может задавать базовый уровень для последующей работы.

Рабочие материалы размещены в GitHub https://github.com/d1mach/my_sentirueval_2016

Последующие этапы работы

В задачах, где требуется классифицировать сообщения с небольшим числом классов s распространены методы, основанные на поиске отображения W слов сообщений в векторное пространство размерности d, которое может быть продолжено некоторым образом на сочетания слов, и линейного отображения $W_s : R^d \to R^s$, таких что для каждого сообщения x из обучающей выборки значения $softmax(W_s(W(x))$ минимально отклоняются от определенных экспертом классов [Socher et al., 2013].

Библиография

1. Н.В. Лукашевич, Автоматический анализ тональности текстов по отношению к заданному объекту и его характеристикам, Russian Digital Libraries Journal. 2015. V. 18. No 3-4
2. Socher, R, et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)