Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора

Ю. Рубцова (mokoron@gmail.com)

Институт систем информатики им. А.П. Ершова СОРАН, Новосибирск

Abstract

В работе представлен первый этап по разработке классификатора корпуса текстов по тону (эмоциональному окрасу) на основе частей речи. Здесь и в дальнейшем в работе, термин “тон” обозначает эмоциональный окрас сообщения (положительное отношение к чему-либо, отрицательное отношение или нейтральное). Первый шаг разработки классификатора состоял из выделения паттернов (шаблонов) использования тех или иных частей речи в зависимости от тонового окраса сообщения. Далее на основе паттернов будет построен классификатор.

В начале работы был собран тренировочный корпус текстов. Корпус был автоматически классифицирован на положительно окрашенные сообщения, отрицательно и нейтрально. Для того, чтобы выделить шаблоны и зависимости эмоционально окрашенных текстов, выборки были размечены по частям речи. После чего был проведен анализ и выделены основные шаблоны использования тех или иных форм частей речи в положительно окрашенных, отрицательно окрашенных и нейтральных сообщениях. В последующем, на основе шаблонов строится алгоритм классификации. В последующих работах планируется проверить алгоритм на не размеченных корпусах текстов.

Введение

В наше время микроблоггинг стал очень популярным средством коммуникации среди интернет-пользователей. Ежедневно пользователи оставляют тысячи сообщений на популярных площадках вроде twitter, futubra, facebook. Они пишут о своей жизни, обсуждают новости, высказывают свое мнение относительно товаров и услуг. За последние годы интернет-пользователи стали активнее вести микроблоги. Перечислим основные свойства микроблогов:

- инструменты для микроблоггинга доступны и практически всегда под рукой;
- нет заведомо выбранного шаблона сообщений – можно писать в произвольной форме, не ограничивая себя;
- сообщение ограничено сверху по длине символов (нет необходимости долго расписывать мысли);
- чем больше друзей пользователя используют платформу, тем больше пользователей присоединяются к ведению микроблогов (эффект

снежного кома).

В микроблогах все больше и больше людей высказывают свое мнение о продуктах, которые они используют; об услугах, которые им оказывали, о политике и религии – тем самым делая микроблоги интересными для социологических и маркетинговых исследований. Для классификации текстовых сообщений на положительные, отрицательный и нейтральные классы, в работе используются сообщения микроблогов. Для исследований была выбрана платформа микроблоггинга Twitter по следующим причинам:

- Twitter является достаточно популярным, в нем уже содержится большое количество постов и их количество растет с каждым днем. Соответственно, можно получить достаточно репрезентативную выборку.
- Микроблоги используются для выражения субъективного мнения на различные темы; таким образом, выборка постов из микроблогов интересна для исследования тонового анализа.
- Социально-демографический срез аудитории twitter очень разнообразен: аккаунты имеют и студенты, и звезды шоу-бизнеса, и политики, и, даже президенты. Таким образом, twitter позволяет собирать данные различных групп людей с различными интересами.
- Аудитория twitter покрывает различные страны и континенты: в последующем можно расширить работу, включив в исследование блоги на различных языках.

Обзор предметной области

За последние годы было проведено достаточно большое количество исследований в области тонового анализа. Много внимания уделялось тоновому анализу отзывов о продуктах, отзывов на фильмы [4] и анализу блогов, другими словами, изучались достаточно длинные текстовые пассажи. Микроблоггинг в общем и twitter в частности достаточно сильно отличается от отзывов по назначению использования. В то время, как отзыв является обдуманным заключением автора о продукте или услуге, твиты более спонтанны, менее продуманы и ограничены по длине 140 символами. В отзывах, как правило, преобладает конструктивная критика или хвала продукта, твиты более эмоциональны и менее конструктивны. Тем не менее, twitter полезен для компаний, работающих с обратной связью.

Wilson, Wiebe и Hoffmann [5] проводили тоновую классификацию на уровне коротких фраз и выражений, а не на уровне длинных пассажей текстов или целых документов. Авторы статьи показали, что зачастую важно знать тоновый окрас отдельно взятого предложения, а не всего текста целиком. Так же, в своей работе авторы сначала определяли, является ли фраза тоново окрашенной или нет и, если является, – определяли ее к положительно окрашенным или отрицательно.

В области тонового анализа коротких сообщений существуют исследования влияния микроблоггинга на бренды [6]. В работе Micro-blogging as Online Word-of-Mouth Branding авторы выяснили, что порядка 20% твитов тестовой выборки содержат упоминания какого-либо бренда. Из них 50% упоминаний носит положительный характер, около 33% - отрицательный, остальные – высказывания нейтрального характера. Таким образом, пятая часть сообщений

микроблогов содержит упоминания какого-либо бренда, что является достаточно большой цифрой. Поэтому организациям, беспокоящимся о своем имидже, стоит уделять внимание микроблоггам в маркетинговых стратегиях.

В работе [8] исследователи строили тоновый классификатор и анализировали работу полученного классификатора на англоязычном корпусе твиттер-постов. Авторы описали алгоритм автоматического извлечения корпуса текстов микроблогов для последующей тренировки тонового классификатора, построенного по методу Байеса. В результате, натренированный на тестовой выборке классификатор, давал лучшие результаты, чем нетренированные классификаторы.

В моей работе используется аналогичная схема извлечения и разметки текстовой выборки по частям речи, которая была описана в работе [8], но работа по извлечению и разметке по частям речи ведется для русскоязычных текстов.

Несмотря на то, что Pang, Lee и Vaithyanathan проводили исследования применения различных тоновых классификаторов на коллекции отзывов о кино, результаты, описанные в работе [7], легли в основу других исследований, посвященных автоматическому тоновому анализу на текстовых выборках, отличных от отзывов о кино. Pang, Lee и Vaithyanathan исследовали точность следующих алгоритмов: метода Байеса, метода максимальной энтропии, метода опорных векторов в зависимости от изменений разметки тестовой выборки:

- Тестовая выборка размечена по частям речи;
- Выделение униграмм;
- Выделение биграмм;
- Определение позиции слов в тексте.

Read [1] показал эффективность использования иконок, обозначающих эмоции на письме для тоновой классификации текстов и снижения зависимости от машинного обучения. Предложенная техника использовалась для подготовки тренировочного корпуса текстов в этой работе.

Тренировочный корпус текстов

Существующие коллекции на русском языке, подготовленные для автоматической тоновой классификации, представляют собой коллекции объединенные одной тематикой, например, коллекция отзывов о фильмах с оценками пользователей (РОМИП 2011, [2]). Таким образом, все доступные коллекции являются коллекциями отзывов, а не коллекциями микроблогов.

Одна из целей работы – выявление зависимости эмоционального сообщения в микроблоге от частоты использования и формы используемых частей речи. Так как на сегодняшний день существует не много публичных тестовых коллекций на русском языке и не было найдено ни одной русскоязычной публичной коллекции постов микроблогов, было принято решение подготовить собственный корпус текстов.

С помощью API twitter, был собран корпус из 31 000 twitter постов. API twitter имеет ограничения, например, на каждый поисковый запрос с географической привязкой выдается не более 1000 twitter постов. Географическая привязка использовалась для того, чтобы собрать twitter посты только на русском языке.

Корпус был разделен на три класса: позитивно окрашенные, негативно окрашенные и нейтральные. Так как люди не ограничены ни форматом, ни формой написания сообщений в микроблогах, нельзя выделить общий паттерн или построить единый словарь для определения эмоциональной принадлежности любого абстрактного сообщения. Поэтому для определения положительно окрашенных и отрицательно окрашенных сообщений использовался подход, предложенный в (Read, 2005 [1]). Для сбора корпуса был выполнен поиск по запросам, характерным для выражения эмоционального отношения к чему-либо. В пользовательских текстах с высокой точностью можно определить эмоцию, если автор указал символ обозначения эмоции на письме (смайлик):

- Позитивные: “:-)”, “:)”, “-)”, “=)”, “:0”, “))”, “хаха” (в различных вариациях).
- Негативные: “:-(”, “:(”, “=(”, “;(” , “:’-(” и т.д.

В соответствии с письменным обозначением эмоций был произведен поиск позитивных и негативных сообщений и сформировано две выборки. Эти две выборки будут использованы для последующего анализа позитивно и негативно окрашенных твитов, выявления общих тенденций и построения структуры позитивного, негативного или нейтрального сообщения.

Тестовая выборка была отфильтрована, согласно следующим критериям:

- Все твиты, содержащие и положительные и отрицательные эмоции удалялись. Подобные тексты могут повлиять на качество анализа и последующих выводов, так как невозможно однозначно определить эмоцию сообщения.
- Удалялись также все ретвиты. Retweet – это процес копирования чужого сообщения в свой аккаунт. Как правило, ретвиты сопровождаются аббревиатурой RT. Подобные сообщения удалялись, так как они могут придать дополнительный вес частям речи при проведении анализа.
- Как выяснилось, API twitter отдает в результатах выдачи копии twitter постов. Одинаковые посты удалялись из тестовой выборки, чтобы не добавлять дополнительного веса частям речи при проведении анализа.

Так как длина твитта ограничена 140 символами, было сделано допущение, что выражение эмоции в виде смайлика относится ко всему сообщению, а не к отдельной его части.

Для выборки нейтральных текстов были использованы аккаунты СМИ, которые, как правило, публикуют объективные сообщения, новости, а не выражают субъективные эмоции.

В результате был получен корпус текстов на русском языке автоматически размеченный на три класса в соответствии с тональностью.

Анализ корпуса

Первое, что было сделано с корпусом текстов – он был размечен с помощью TreeTagger для русского языка (Schmid, 1994 [3]). Для исследования необходимо было понять, как распределяются части речи между выборками,

заведомо содержащими и заведомо не содержащими эмоции.

В результате анализа выяснилось, что в зависимости от того, выражает ли автор эмоции или нет, он склонен использовать чаще ту или иную часть речи.

Так, например, авторы твитов, содержащих эмоции, чаще используют личные местоимения (*я, ты, он, она, оно* …). А в нейтральных твитах чаще употребляют имена собственные. Авторы эмоционально окрашенных сообщений часто описывают себя и свой опыт, поэтому в них преобладают местоимения первого и второго лица. Если местоимение встречается в нейтральных сообщениях, то, как правило, оно имеет форму третьего лица.

В нейтральных сообщениях активно используются существительные в именительном и родительном падежах. Также отмечено наличие большого количества числительных в нейтральных сообщениях.

Прилагательные в превосходной форме чаще встречаются в эмоционально окрашенных сообщениях, а прилагательные в сравнительной форме в нейтральных сообщениях.

Наречия, как правило, употребляются в эмоционально окрашенных сообщениях для усиления смысла глаголов, реже в неэмоциональных текстах.

Таким образом, удалось выделить основные закономерности и отличия текстов с эмоциональным окрасом и без него — выделены шаблоны использования частей речи, на основе которых строится алгоритм классификации.

Также было проведено сравнение эмоционально-окрашенных постов между собой. В результате было установлено, что в негативно окрашенных твитах чаще используется прошедшее время – люди выражают свое негодование или разочарование относительно событий в прошлом, в то время как радостные (позитивные) события происходят в настоящем времени. По сравнению с положительно окрашенными сообщениями в негативных встречается много глаголов (см. график 1).

График 1. Разница в значениях частей речи между положительными сообщениями и отрицательными.

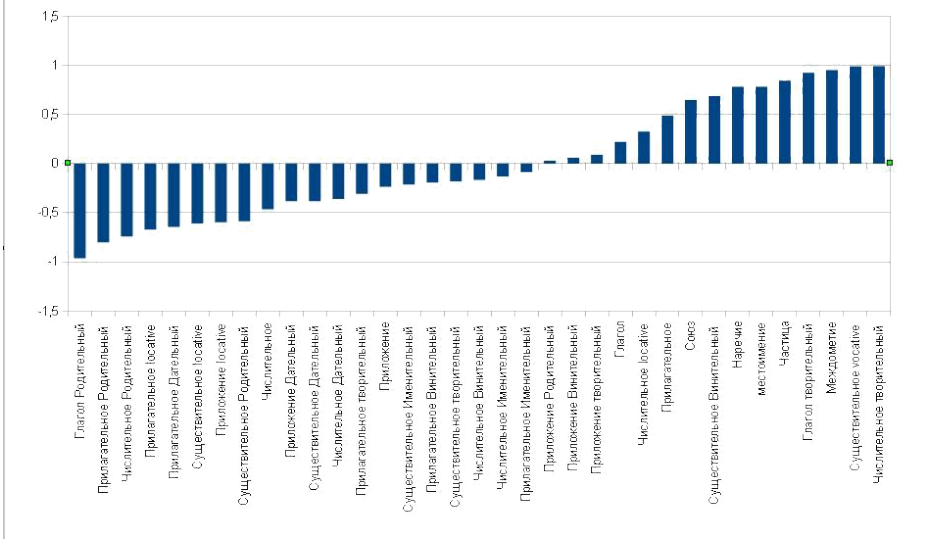
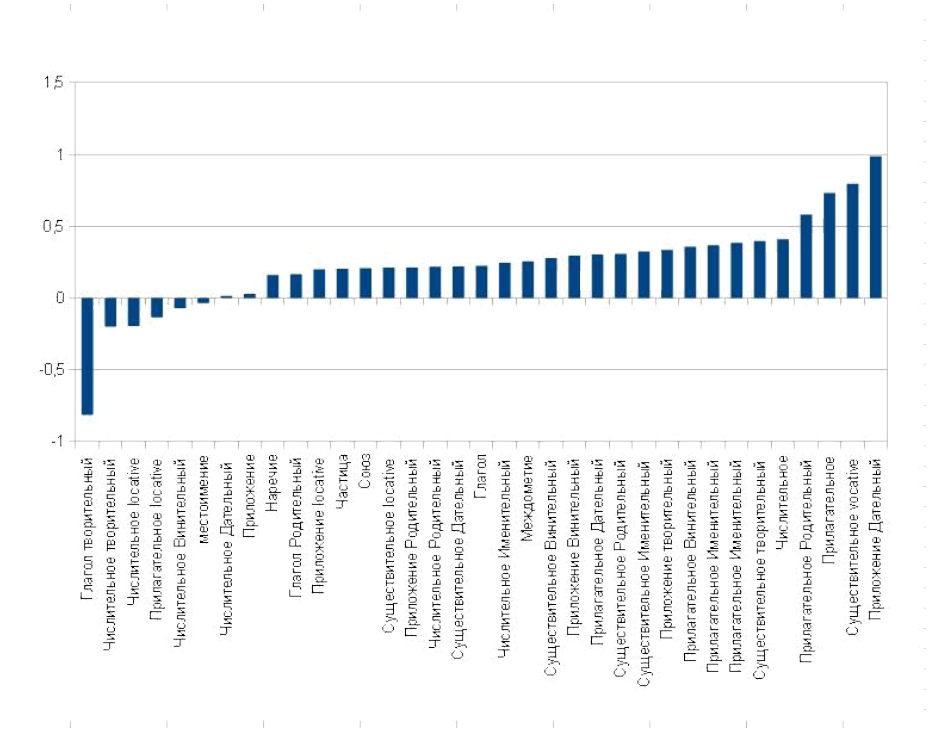


График 2. Разница в значениях частей речи между эмоционально окрашенными сообщениями и нейтральными.



С целью выделить положительно и отрицательно окрашенные термины, характерные для различных предметных областей, а не только для определенной узкоспециализированной области, были проанализированы наиболее часто встречающиеся слова и словосочетания в каждой из выборок. Так, оказалось, достаточно сложно выбрать слова, характеризующие положительную эмоцию. Например, слово “хорошо” в одном контексте может иметь положительное значение, в другом – нейтральное или даже отрицательное *(“хорошо, я сделаю это” и “хорошо, что хоть часть денег вернули”*). Термины, которые можно отнести к отрицательно окрашенным более конкретны и однозначны (*скучно, гадко, чушь* …). Выделенные термины также были включены в шаблоны, на основе которых строится алгоритм классификации постов микроблогов.

Для построения алгоритма классификации, требуется выделить шаблоны совместного использования различных частей речи. Например, в первом приближении было выявлено, что в отрицательно окрашенных сообщениях достаточно часто используется связка «личное местоимение + глагол в прошедшем времени». Выявленные шаблоны совместного использования частей речи будут включены в разрабатываемый алгоритм тоновой классификации.

Результаты

Для начала работы была подготовлена тестовая коллекция twitter-постов с использованием метода автоматического сбора корпуса текстов [1]. Коллекция текстов прошла постфильтрацию по предложенным автором работы фильтрам, чтобы снизить возможный шум. В результате исследования полученного корпуса текстов выявлено, что в зависимости от тона сообщения, авторы сообщений склонны чаще использовать ту или иную часть и форму речи. Некоторые части речевой разметки могут служить индикатором эмоциональности текстового сообщения.

Основываясь на анализе коллекции, были выявлены шаблоны обособленного использования частей речи в положительно окрашенных сообщениях, отрицательно окрашенных и нейтральных. Так же были выделены положительно и отрицательно окрашенные термины. На основе шаблонов использования частей речи и эмоционально окрашенных терминов будет построен алгоритм тоновой классификации twitter-постов.

Дальнейшая работа

В последующем планируется проверить гипотезу, полученную в этой работе, на выборке текстовых сообщений микроблогов, которые не содержат смайлики – то есть заранее не определен тоновый окрас сообщения.

Так же будет проведен анализ совместного использования различных частей речи и выделены основные шаблоны. Алгоритм классификации будет дополнен этими шаблонами

В данный момент ведется работа по реализации алгоритма тоновой классификации.

Список литературы

[1] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. InACL. The Association for Computer Linguistics.

[2] Российский семинар по оценке методов информационного поиска (РОМИП). URL http://romip.ru/ru/collections/imhonet-films.html

[3] Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, pages 44–49.

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania. Pages 79-86, 2002.

[5] T.Wilson, J.Wiebe and p.Hoffmann. Recognizing contextual polarity in phrese-level sentiment analysis. In proceedings of Human Languages Technologies Conference/ Conference on Emperical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA, 2005.

[6] B.J. Jansen, M. Zhang, K.Sobel and A. Chowdury. Micro-blogging as Online Word-of-Mouth Branding. In CHI EA '09: Proceedings of the 27th international conference exended abstract on Human factors in computing sestems, pages 3859-3864, New York, NY, USA, 2009. ACM.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP, pp. 79--86, 2002

[8] Alexander Pak, Patrick Paroubek: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC 2010