Метод построения и анализа корпуса коротких текстов для задачи классификации отзывов

© Ю. В. Рубцова
Институт систем информатики им. А.П. Ершова
СО РАН, Новосибирск
yu.rubtsova@gmail.com

Аннотация

В работе описывается метод построения русскоязычного корпуса коротких текстов, предназначенного для решения задачи классификации отзывов на три класса: «положительные», «отрицательные», «нейтральные». Проведен морфологический анализ корпуса с целью выделения характерных для каждого класса отзывов признаков и терминов. На основе предложенного метода был создан русскоязычный корпус коротких текстов, автоматически классифицированный на три группы: «заведомо положительные», «заведомо отрицательные» и «заведомо нейтральные». Корпус состоит из 50 000 коротких текстов.

Работа выполнена при частичной финансовой поддержке РФФИ (проект № 13-07-00422).

1. Введение

В настоящее время каждый пользователь сети Интернет имеет возможность высказать свое мнение относительно товара, услуги, явления, персоны или организации. Для этого не нужно создавать и поддерживать собственные web-ресурсы, достаточно зарегистрироваться в социальной сети, форуме или на тематическом сайте с обзорами и отзывами. Появилось много площадок, на которых пользователи Интернета могут свободно высказывать свое мнение, и число таких площадок увеличивается год от года.

Пользователи, высказывающие свое мнение или отношение к чему-либо, делают это с помощью оценочных слов, выражений и специальных символов, имеющих положительную или отрицательную окраску. Более того, пользователи в своих высказываниях предпочитают использовать определенные формы частей речи в зависимости от окраса отзыва.

Важными характеристиками для отзывов в Интернете является скорость получения отзыва заинтересованным лицом и авторитетность автора отзыва. От скорости получения отзыва зависит быстрота реагирования на него. Как правило, из-за длительного цикла производства товаров, для отзывов на товары, скорость реагирования не имеет большого значения. Напротив, скорость реагирования на отзывы на услуги, информационные продукты или новости крайне важна: через несколько дней актуальность отзыва может снизиться или вообще пропасть. Современные поисковые системы и имеющиеся в открытом доступе инструменты по сбору текстовых отзывов не удовлетворяют задаче сбора актуальных отзывов и оперативной работы с данными. В связи с этим на основе программного интерефейса API twitter был разработан программный инструмент для извлечения отзывов об интересующих товарах, услугах, событиях, персонах из микроблоггинг-платформы twitter, который позволяет учитывать время публикации сообщения и автора сообщения. Извлечение отзывов из других микроблоггинговых платформ, а также популярных социальных сетей относится к перспективам данной работы и выходит за рамки публикации. С помощью созданного программного инструмента был собран корпус коротких текстов, описываемый в статье.

В статье рассматривается подход к построению корпуса размеченных отзывов для тренировки тонового классификатора. Автоматическая классификация отзывов осуществляется по методу, предложенному Jonathon Read в [8].

После сбора корпуса, работа с корпусом состоит из двух этапов:

1. морфологическая разметка корпуса;
2. извлечение списков наиболее часто употребляемых слов с целью выделить оценочные слова для общетематических отзывов. При этом для каждого слова рассчитывается набор статистических характеристик и выделяется список общетематических оценочных слов, характерных для положительной или отрицательной окраски отзыва.

Анализ результатов морфологического анализа и наборы оценочных слов используются в дальнейшей работе для построения классификатора отзывов.

2 Обзор предметной области

В последние годы ведется довольно много исследований в области определения тональности текстов, в частности отзывов на товары и услуги. При этом отзывы чаще всего разбиваются либо на два класса (положительные и отрицательные), либо на три класса (положительные, отрицательные и нейтральные). Много работ посвящено тоновой классификации отзывов на продукты и фильмы [1, 6], а также анализу блогов и новостей. Другими словами, изучались достаточно длинные тексты, принадлежащие некоторой заранее определенной предметной области. Иногда для работы использовались тексты отзывов, размеченные самим автором по пяти или десяти балльным шкалам [7]. Для предыдущих исследований [1, 6, 7, 11] разрабатывались тренировочные корпусы текстов, обладающие следующими параметрами:

- корпуса отзывов с вручную проставленными потребителями оценками;
- узкотематические корпуса отзывов (на фильмы, на книги).
- корпуса общезначимых новостей (тексты, состоящие из нескольких абзацев).

Существующие коллекции на русском языке, подготовленные для задачи автоматической классификации отзывов на два или три класса, представляют собой коллекции, объединенные одной тематикой, например, коллекция отзывов о фильмах с оценками пользователей (РОМИП 2011, [11]). Таким образом, все доступные коллекции являются коллекциями отзывов, принадлежащими определенной предметной области, а не общетематическими коллекциями микроблогов.

Микроблогинг сильно отличается от отзывов на специализированных площадках: в то время как отзыв является обдуманным, структурированным заключением автора о продукте или услуге, сообщения микроблога более спонтанны, менее продуманы и ограничены по длине В отзывах, как правило, преобладает конструктивная критика или похвала продукта, сообщения микроблога более эмоциональны и менее конструктивны. В отзывах на специализированных ресурсах можно выделить оценочные слова, характерные для определенной предметной области. Микроблоги являются общетематическими ресурсами, поэтому при их анализе стоит более сложная задача выделения ярких оценочных терминов, которые имеют положительную или отрицательную окраску во многих предметных областях, а не в одной, исследуемой. Как правило, это слова-жаргонизмы.

В работе [2] автор сравнивает результаты использования одного алгоритма тоновой классификации текстов для трех различных предметных областей: книги, кино и цифровые фотокамеры. А так же, использование одного алгоритма для классификации на два, три и пять классов. В основе алгоритма лежит подход основанный определении весов слов в коллекции; слов, выражающих мнения и слов, которые меняют полярность отзывов. Алгоритм показал хорошие результаты почти во всех предметных областях и при классификации на два, три или пять классов. Однако, автор отмечает, что наилучшие результаты достигаются при сходстве тренировочных и тестовых коллекций.

В работе [4] автор приводит алгоритм классификации текстов на русском языке на 2 класса (положительные и отрицательные) и 3 класса (положительные, отрицательные и нейтральные) без привязки к объекту или тематической области. Однако, чтобы повысить точность классификации текстов, автор рекомендует использовать свой словарь эмоционально-окрашенных терминов для каждой отдельно взятой предметной области.

При классификации коротких сообщений исследуется влияние микроблогинга на бренды [3]. В работе [3] авторы выяснили, что порядка 20% сообщений 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.

Классификацию на уровне коротких фраз и выражений, а не на уровне длинных текстовых пассажей или целых документов, проводили Wilson, Wiebe и Hoffmann [10]. Авторы статьи показали, что зачастую важно знать окраску (положительная или отрицательная) отдельно взятого предложения, а не всего текста целиком, так как в длинном документе мнение автора о рассматриваемом предмете может меняться с положительного на отрицательный и наоборот. Другими словами, не всегда длинный документ или отзыв однозначно можно классифицировать как положительный или отрицательно окрашенный.

Работа [10] состояла из двух этапов: сначала определялось, является ли фраза нейтральной или имеет положительную/отрицательную окраску и, если фраза имела окраску, только после этого она классифицировалась как положительно или отрицательно окрашенная.

В работе [5] авторы построили тоновый классификатор на англоязычном корпусе twitter-постов. Исследователи описали алгоритм автоматического извлечения корпуса текстов микроблогов методом Read [8] и использовали полученный корпус для последующей тренировки тонового классификатора, построенного по методу Байеса. В результате, натренированный на тестовой выборке классификатор, показал лучшие результаты, чем нетренированные классификаторы.

3 Метод построения корпуса текстов

Так как на сегодняшний день существует очень мало публичных русскоязычных коллекций отзывов, которые можно было бы использовать для решения задачи классификации отзывов на три класса (положительные, отрицательные, нейтральные), и не обнаружено ни одной русскоязычной публичной коллекции постов микроблогов, было принято решение подготовить собственный корпус текстов.

В качестве источника текстов была выбрана платформа микроблогинга Twitter по следующим причинам:

- Twitter является достаточно популярным ресурсом, в нем уже содержится большое количество постов и их количество растет с каждым днем. Соответственно, из него можно получить достаточно репрезентативную выборку.
- Микроблоги используются для выражения субъективного мнения на различные темы, поэтому выборка постов из микроблогов интересна для исследования классификации сообщений на положительно, отрицательно и нейтрально окрашенные.
- 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.

Подготовка корпуса состоит из двух этапов:

1. Создание словаря специальных символов, характерных для выражения положительного и отрицательного отношения к чему-либо.
2. Создание фильтров для устранения дубликатов и неопределенностей.

3.1 Создание словаря специальных символов

Метод, описанный в [8], показал эффективность использования специальных символов иконок, обозначающих эмоции в письменных сообщениях, для классификации текстов на положительные и отрицательные. В отзывах с высокой точностью можно определить эмоцию, если автор указал символ обозначения эмоции на письме (смайлик). Поэтому, прежде всего, были составлены словари символов, обозначающих положительное и отрицательное отношение автора к предмету высказывания.

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

Пример положительных символов: “:)”, “=)”, “:))”, “=))”, “-)”, “-))”, “:-)”, “:-))”, “(:”, “(=”, “:D”, “:-D”, “))”, “)))”, “8)”, “8-)”, “8-))”

Пример отрицательных символов: “:(”, “=(”, “=((”, “:((”, “=((”, “-(”, “-((”, “:-(”, “:-((”, “):”, “)=”, “((”, “(((”, “:’(”, “:’-(”, “O_o”, “;(”, “:|”, “:-\”.

В соответствии с письменным обозначением эмоций был произведен поиск позитивно и негативно окрашенных сообщений и сформировано две выборки. Эти выборки будут использованы для последующего анализа позитивно и негативно окрашенных твитов и выявления характеристик структуры позитивного и негативного сообщения.

Для формирования выборки нейтральных сообщений были взяты сообщения новостных аккаунтов микроблогов.

3.2 Фильтрация

Тестовая выборка была отфильтрована, согласно следующим критериям:

- Удалялись все твиты, содержащие одновременно и положительные и отрицательные эмоции. Подобные тексты могут повлиять на качество анализа и последующих выводов, так как невозможно однозначно определить эмоцию сообщения.
- Удалялись все ретвиты. Retweet – это копирование чужого сообщения в свой аккаунт. Как правило, ретвиты сопровождаются аббревиатурой RT. Подобные сообщения удалялись, так как они могут придать дополнительный вес частям речи или оценочным словам при проведении анализа сообщения.
- Как выяснилось, API twitter отдает в результатах выдачи копии twitter-постов. В связи с этим необходимо было удалять одинаковые посты из тестовой выборки, чтобы не добавлять дополнительного веса частям речи или оценочным словам при проведении анализа.
- Удалялись малоинформативные твиты, длина которых составляла менее 40 символов.
- Удалялись реплаи. Реплай – это ответ на сообщение другого пользователя. Как правило, чтобы не показаться грубыми или выразить поддержку автору твита, пользователи отвечают на сообщения, добавляя смайлик. Таким образом, эмоция относится скорее к автору сообщения, а не к самому сообщению.

Следует заметить, что API twitter имеет ограничения, например, на каждый поисковый запрос с географической привязкой отдается не более 1000 twitter-постов. Географическая привязка

использовалась для того, чтобы собрать twitter-посты только на русском языке.

Несмотря на ограничения, с помощью API twitter, был собран корпус русскоязычных twitter-постов, автоматически размеченных на два класса (положительные и отрицательные). Корпус нейтральных постов собирается отдельно. Каждый текст в корпусе имеет следующие атрибуты:

- дата публикации;
- имя автора;
- текст твита;
- класс, к которому принадлежит текст (положительный, отрицательный, нейтральный);
- количество реплаев (ответов других пользователей на это сообщение);
- количество ретвитов (количество копирований этого сообщения другими пользователями).

В результате был получен тренировочный корпус, состоящий из 34 235 положительных, 34 225 отрицательных и 32 065 нейтральных twitter-постов. Собранный корпус используется в дальнейшей работе.

4 Морфологический анализ тренировочного корпуса

Тренировочный корпус текстов был размечен с помощью TreeTagger для русского языка (Schmid, 1994 [9]). TreeTagger – это инструмент для аннотирования текста по частям речи. Задача исследования состояла из двух подзадач:

1. Выявить закономерности распределения частей речи между коллекциями, заведомо состоящими или не состоящими из эмоционально окрашенных высказываний.
2. Выявить закономерности распределения частей речи между «положительной» и «отрицательной» коллекциями.

В результате анализа выяснилось, что в зависимости от того, выражает ли автор положительный/отрицательный настрой или нет, он склонен использовать различные части и формы речи для построения предложений. Так, например, авторы твитов, содержащих эмоции, чаще используют наречия, частицы, существительные в именительном падеже мужского и женского родов (График 1, столбец R, Q, Ncmsnn, Ncfsnn соответственно). Примеры используемых наречий: “прекрасно”, “беспощадно”, “стыдно”, “интересно”, “дико”. Авторы эмоционально окрашенных сообщений часто описывают себя и свой опыт, поэтому в них преобладают местоимения первого и второго лица (График 1, столбец, P-1-snn). Количество использований местоимений первого и второго лица в различных формах в тестовой выборке составляет 15 249, в то время как использование местоимений третьего лица всего 3794.

Для выражения эмоций, как положительных, так и отрицательных, пользователи используют схожие речевые шаблоны. Поэтому, чтобы выделить закономерности использования частей речи в сообщениях пользователей в зависимости от положительного окраса или отрицательного, выборка была нормирована. Далее было произведено сравнение использования той или ной части речи в процентном соотношении от количества использования всех частей речи в положительной или отрицательной выборке.

На графике 2 видно, что в негативно окрашенных твитах сравнительно чаще, чем в позитивных используются глаголы настоящего времени, обозначающие продолжительность действия в третьем лице единственного числа несовершенного вида (столбец vmi3s-me) – пользователи описывают неприятную ситуацию, которая еще не разрешилась. В негативно окрашенных сообщениях сравнительно чаще используются имена собственные в единственном числе в винительном падеже (столбец Npmsny, График 2), чем в положительных.

Для выражения позитивных эмоций используются глаголы активного залога настоящего времени первого и третьего лица единственного числа (vmip1s-a-e и vmip3s-a-e, График 2), глаголы в активном залоге в прошедшем времени женского и мужского родов единственного числа (vmis-sfa-p и vmis-ama-p, График 2) – пользователи делятся своими достижениями, вспоминают, что с ними произошло.

В положительно окрашенных твитах чаще встречаются притяжательные местоимения мужского и женского родов в единственном числе в родительном падеже – тем самым пользователь указывает на положительный опыт обладания чем-либо или делится собственными достижениями. В негативно окрашенных твитах реже используются притяжательные местоимения.

5 Извлечение оценочных слов из тренировочного корпуса

Сообщения микроблогов зачастую отражают отношение общественности к текущим мировым событиям, новостям, анонсам, релизам. Несмотря на то, что корпус собирался в течение небольшого времени, в него попали сообщения о ярких новостных событиях. Поэтому для чистоты выбора оценочных терминов из корпуса был построен словарь «стоп-слов», состоящий из фамилий и названий продуктов. Например, часто встречается фамилия Навальный или упоминается футбольный клуб «Зенит», но ни положительной, ни отрицательной окраски они не имеют, поэтому подобные термины были отфильтрованы. Предлоги и союзы также были отфильтрованы.

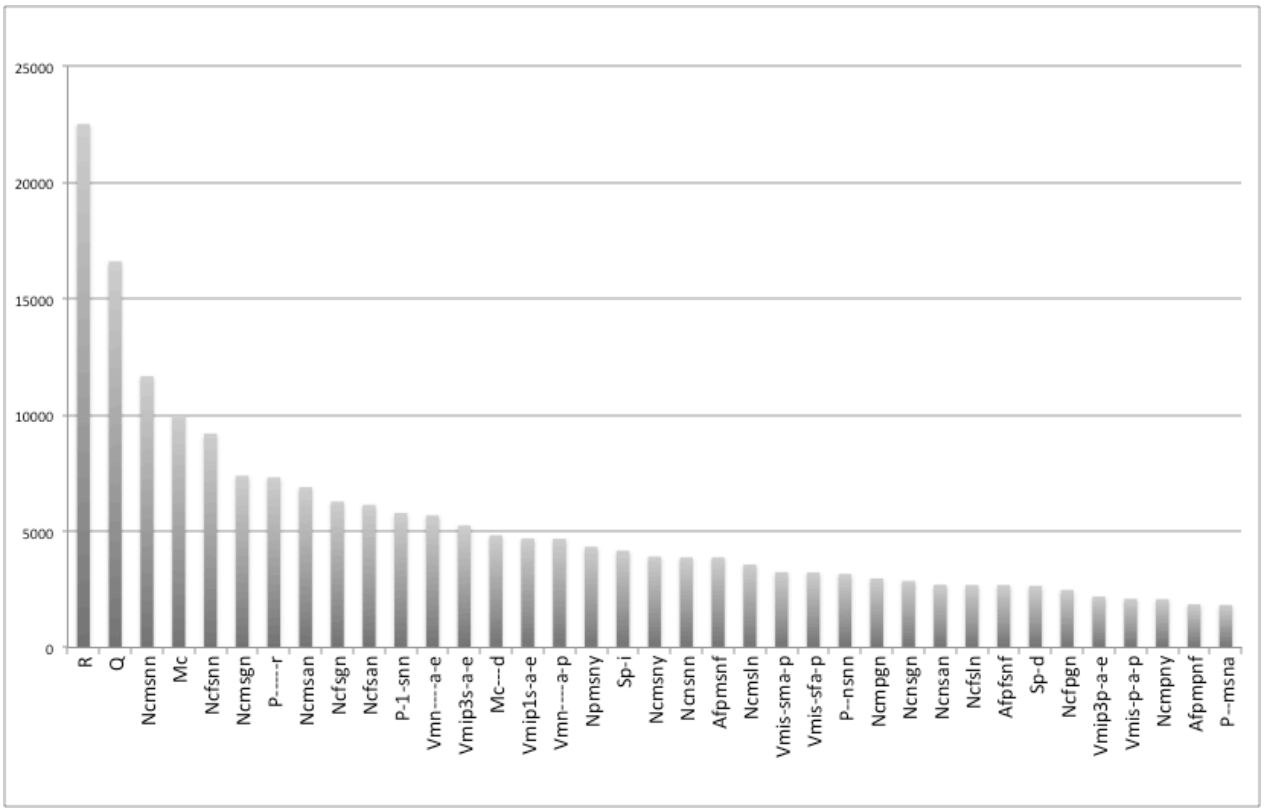

График 1. Распределение по частоте использования частей речи в эмоционально окрашенных твитах (союзы, предлоги и знаки препинания отфильтрованы)

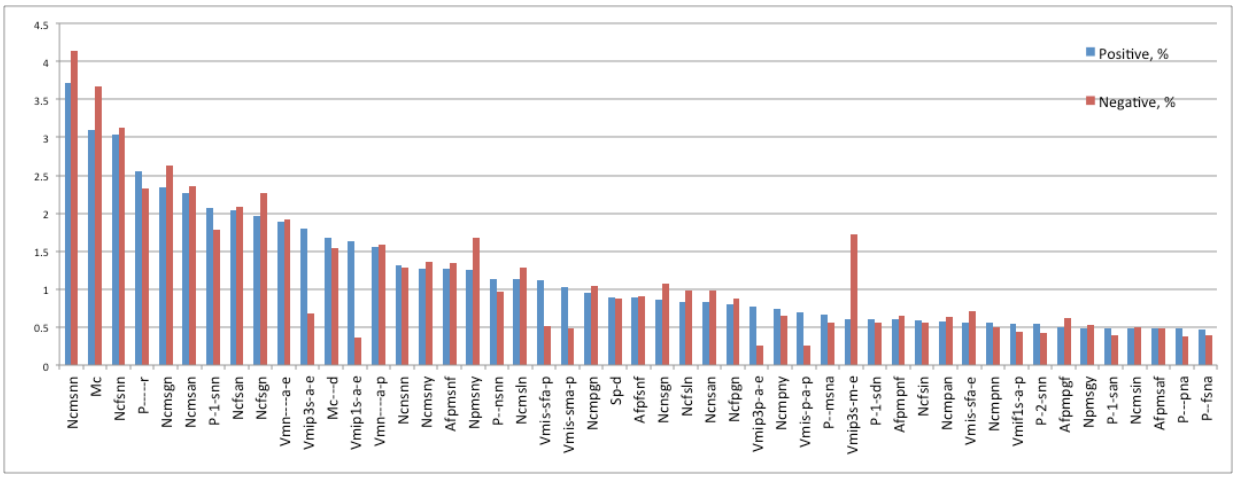

График 2. График использования частей речи в положительных твитах в сравнении с отрицательными. Данные нормированы и приведены в процентах.

В предыдущей работе автора [12] оценочные термины играли вспомогательную роль. Единственной характеристикой, на основе которой они извлекались, была частота. После извлечения из текста такие термины проходили ручную проверку на принадлежность к строго положительным или строго отрицательным высказываниям.

Один и тот же термин может использоваться одинаково часто как для выражения положительного мнения, так и отрицательного. Это может быть термин, служащий для усиления эмоции, а не для ее выражения. Поэтому некорректно строить предположения об оценочных терминах, основываясь исключительно на частотности его употребления. Поэтому в настоящей работе помимо частотности употребления термина в положительных или отрицательных высказываниях используется параметр обратный частоте использования этого термина в отрицательных или положительных высказываниях.

В этом случае вес каждого термина из положительной (отрицательной) категории состоит из двух множителей: частотность употребления термина в коллекции положительных (отрицательных) текстов, и обратная величина частотности употребления этого термина в коллекции отрицательных (положительных) текстов. (Таблицы 1,2)

Вес слова (отношение частоты встречаемости слова в отрицательных твитах к частоте встречаемости в положительных)	**Слово**	**Частота встречаемости слова в коллекции отрицательных твитов**
13.01650495	продажа	56
10.22725389	утро	484
6.5610792	встать	69
6.136352334	проснуться	88
5.835406825	вставать	53
4.881189357	вчера	119
4.685941782	погибнуть	56
4.623486813	школа	305
4.532532974	рано	52
4.434909187	жаль	53
4.408012837	дождь	59
4.183876591	погода	70
3.893329606	рабочий	67
3.892594044	из-за	147
3.521870802	сон	133
3.388940039	болеть	81
3.096068678	ненавидеть	74
3.053099134	получить	108
2.836526503	рубль	80
2.811042085	машина	86

Таблица 1 Веса и частота использования слов в негативных твитах.

Вес слова (отношение частоты встречаемости слова в положительных твитах к частоте встречаемости в отрицательных)	**Слово**	**Частота встречаемости слова в коллекции положительных твитов**
4.097362044	клип	60
2.230786002	сериал	56
2.22893319	бл*	401
2.192768754	х**	2211
1.657155316	приятный	52
1.6101914	рад	64
1.434076715	зато	60
1.414158983	крутой	71
1.354405787	смеяться	51
1.284693724	фотка	86
1.283121272	наконец	51
1.23280279	счастье	98
1.212136271	именно	71
1.19506393	миллион	70
1.119279388	счастливый	96
1.079412582	гулять	70
1.068527749	интересно	76
1.029593539	милый	56

Таблица 2 Веса и частота использования слов в положительных твитах.

6 Заключение

В результате работы был построен корпус текстов, автоматически размеченный на тексты, содержащие положительную, отрицательную или нейтральную окраску. Каждый текст в корпусе содержит атрибуты, которые помогут сделать выводы об актуальности высказывания и силе его воздействия, важности. Была произведена морфологическая разметка корпуса. На основе морфологической разметки были выявлены закономерности о зависимости окраски сообщения от используемых в нем частей речи. Также была проведена работа по извлечению оценочных терминов, не относящихся к одной заранее определенной предметной области.

Планируется, что тоновый классификатор, обученный на собранном по предложенному в статье методу корпусе текстов, будет использоваться для автоматический оценки отзывов на интернет-ресурсы, найденные в качестве кандидатов на включение в интеллектуальные научные интернет-ресурсы (ИНИР) по заданной тематике. Каждый такой ИНИР представляет собой информационную систему, обеспечивающую систематизацию и интеграцию научных знаний и информационных ресурсов определенной области знаний, содержательный эффективный доступ к ним, а также поддерживающую их использование при решении различных научных и производственных задач.

Литература

[1] Chetviorkin I. I.,Loukachevitch N. V. Cross-domain opinion word extraction model //

Proceedings of 6-th Russian Yong Schientists Conference in Information Retrieval, Yaroslavl, 2012. - p.5-15.

[2] Chetviorkin I. I. testing the sentiment classification aproach in various domains — ROMIP 2011. Proceedings of the International Conference “Dialog 2012” Issue 11 Volume 2 p15-27.

[3] 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.

[4] Kan D. Rule-based approach to sentiment analysis at ROMIP 2011. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2012” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2012”]. Bekasovo, 2012.

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

[6] 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.

[7] Pang B., Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect of rating scales // Proc. of ACL, 43rd Meeting of the Association for Computational Linguistics. Ann Arbor: ACM, 2005. 115–124.

[8] Read, J (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In: Proceedings of the Student Research Workshop at the 2005 Annual Meeting of the Association for Computational Linguistics. Ann Arbor, Michigan, pp.43-48.

[9] 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.

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

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

[12] Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // Сборник трудов конференции «Инженерия знаний и технологии семантического веба – 2012». – СПб.: НИУ ИТМО, 2012. – С. 109–115.

A Method for development and analysis of short text corpus for the review classification task

Yuliya Rubtsova

The paper describes an approach to constructing the corpus of short texts in Russian. This corpus will be used to train tone classifier for the problem of classification reviews into three classes: "positive", "negative", "neutral". In order to extract features and characteristic for each set of texts (positive, negative, neutral), corpus was tagged by part of speech. Also frequently used terms was extracted for each of the sets and proceeded. Using proposed method the corpus of short Russian texts was created. This coprpus was automatically classified into three groups: "a priori positive", "a priori negative" and "a priori neutral." The corpus consists of 100,525 short texts.