УДК 004.853

Ю.В. Рубцова

Методы автоматического извлечения терминов в динамически обновляемых коллекциях для построения словаря эмоциональной лексики на основе микроблоговой платформы Twitter

Представлен подход к построению динамично обновляемого словаря эмоциональной лексики для задачи построения и тренировки тонового классификатора. Проведено сравнение методов извлечения оценочных слов, основанных на различных статистических мерах. Показана вычислительная сложность создания и пересчета весов терминов при добавлении новых документов в коллекцию в зависимости от выбранных весовых схем.

Ключевые слова: корпусная лингвистика, текст классификация и категоризация, анализ данных социальных сетей, весовые схемы.

Человеческая речь постоянно меняется и развивается: новые слова входят в активный словарь, старые перестают употребляться. Вместе с разговорной речью постоянно трансформируется и развивается естественный язык. Каждый день рождаются новые слова, и примерно половина из них – это сленг. Сленг быстрее остального языка реагирует на изменения во всех сферах. Сленг активно используют в разговорной речи и для письменного общения в социальных сетях, а также для выражения эмоционального отношения по отношению к тому или иному вопросу. Пользователи социальных сетей начинают использовать новые термины в повседневном общении одни из первых. В связи с этим необходимо учитывать сленг при разработке тоновых классификаторов, в частности, при создании словарей эмоциональной лексики. Более того, так как активный словарный запас регулярно пополняется новыми терминами, словари эмоциональной лексики также должны регулярно обновляться, а веса терминов в этих словарях – пересчитываться.

В данной работе представлен подход к извлечению терминов и назначению им весов для построения словаря эмоциональной лексики, который постоянно обновляется. Приведены сравнения методов, основанных на различных статистических мерах, и показана вычислительная сложность пересчета весов терминов словаря в зависимости от используемых методов.

Характеристики корпуса текстов. В предыдущей работе [1] автор описывал подход к построению корпуса коротких текстов на русском языке на основе сообщений социальной сети Twitter. Twitter – это социальная сеть и сервис микроблогинга, который позволяет пользователям писать сообщения в реальном времени. Зачастую сообщение пишется с мобильного устройства прямо с места событий, что добавляет сообщению эмоциональности. Из-за ограничения платформы длина twitter-сообщения не превышает 140 символов. В связи с этой особенностью сервиса (короткие сообщения, которые публикуются в реальном времени, возможно, с помощью мобильных устройств), люди используют аббревиатуры, сокращают слова, используют смайлики, пишут с орфографическими ошибками и опечатками. Так как Twitter имеет особенности социальной сети, пользователи Twitter имеют возможность выражать свое мнение относительно разнообразных вопросов: от качества телефонов до экономических и политических событий в мире. Поэтому площадка Twitter привлекает внимание исследователей.

С помощью API Twitter было собрана коллекция, состоящая из около 15 миллионов коротких сообщений, на основе которой с помощью метода [2] и предложенной автором фильтрации [1] был сформирован корпус, состоящий из следующих коллекций:

- коллекция положительных сообщений 114 991 записей;
- коллекция негативных сообщений 111 923 записей;
- коллекция нейтральных сообщений 107 990 записей.

Соотношение количества словоформ и уникальных словоформ в коллекциях представлено в (табл. 1). Корпус доступен для публичного скачивания и ознакомления по ссылке http://study.mokoron.com.

Таблица 1

Соотношение коллекций по их объемам в корпусе текстов, собранного на основе русскоязычных постов социальной сети Twitter

Тип коллекции	Количество словоформ в коллекции	Количество уникальных словоформ в коллекции
Положительные сообщения	1 559 176	150 720
Негативные сообщения	1 445 517	191 677

Чтобы использовать собранные коллекции для построения словаря эмоциональной лексики, необходимо, чтобы коллекции содержали достаточно большое количество лемм. Несмотря на то, что русский язык богатый и разнообразный, далеко не все слова используются в повседневной жизни и тем более для неформального общения в социальных сетях. Для построения словаря эмоциональной лексики необходим «достаточно представительный корпус». «Достаточно представительный корпус» означает, что добавление новых сообщений к коллекции повлечет за собой добавление очень небольшого числа новых терминов. Чтобы проверить, является ли корпус достаточно представительным, три коллекции были объединены в одну, после чего было произведено вычисление количества уникальных терминов в зависимости от размера коллекции. График (рис. 1) показывает, что при небольшом количестве твитов добавление к коллекции новых сообщений влечет за собой резкое увеличение числа уникальных терминов. С ростом числа сообщений рост числа терминов уменьшается. Так, если до 2 тыс. сообщений число терминов растет до 5 тыс. (прирост в среднем более 50 терм./собщ.), то в диапазоне 100–200 тыс. сообщений увеличение терминов составляет примерно 50 тыс. (0,5 терм./сообщ.).

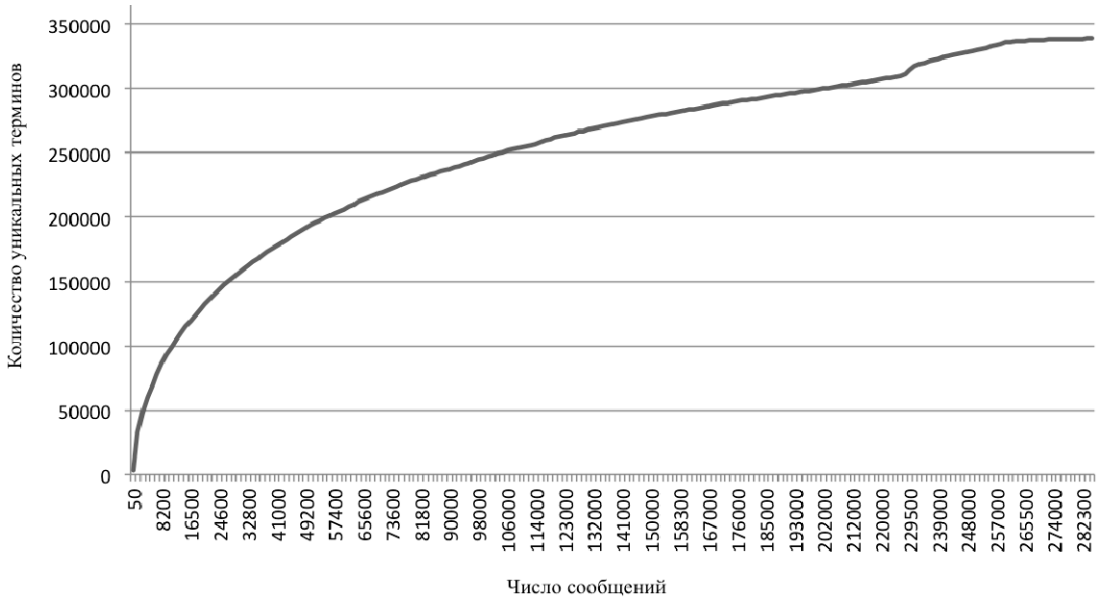

Рис. 1. Распределение числа уникальных терминов в зависимости от числа твитов

Построение словаря эмоциональной лексики. Существуют различные подходы к извлечению оценочных слов из текстов и определения их веса в коллекции. В работе [3] авторы используют тезаурус для расширения словаря оценочных слов, собранного вручную. В корпусной лингвистике широко применяются методы извлечения терминов, основанные на мере значимости этого термина для коллекции, например методы, основанные на мере TF-IDF [4]. Однако методы извлечения терминов, основанные на мере TF-IDF, показали на коллекции текстов, принадлежащих к разным классам, результаты хуже, чем методы, основанные на мере RF (Relevance Frequency – релевантная частота) [5–7].

Работа большинства из существующих методов автоматического и полуавтоматического извлечения слов из текстов строится на предположении, что все данные заранее известны, доступны и статичны. Например, для использования метода, основанного на мере TF-IDF [4], необходимо знать частоту встречаемости термина в документе, следовательно, набор данных не должен меняться во время расчета. Это существенно усложняет вычисления, если требуется провести обсчет данных в

реальном времени. Так, например, при добавлении нового текста в коллекцию требуется пересчитать веса для всех терминов коллекции. Вычислительная сложность перерасчета всех весов терминов в коллекции: $O(N^2)$.

Для того чтобы решить проблему поиска и расчёта весов терминов в режиме реального времени, была предложена [8] мера Term Frequency – Inverse Corpus Frequency (TF-ICF). Для расчета TF-ICF не требуется информация о частоте использования термина в других документах коллекции, таким образом, вычислительная сложность линейна. Чтобы оценить эффективность метода, основанного на мере TF-ICF для задачи извлечения оценочных терминов для словаря эмоциональной лексики, автор сравнил меру TF-ICF и ее модификацию icf-based [9] с тремя широко используемыми мерами: TF-IDF [4], TF-RF [5,6], prob-based [10].

Формула для вычисления меры TF-IDF:

$$\text{tf.idf} = \text{tf} * \log \frac{T}{T(t_i)}, \quad (1)$$

где tf – это частота встречаемости термина в коллекции (положительных или отрицательных твитов); T – общее число сообщений в коллекциях положительных и отрицательных; $T(t_i)$ – число сообщений в положительной и отрицательной коллекциях, содержащих термин.

Суть меры RF состоит в том, что вес слова вычисляется на основе информации о распределении этого термина в текстах коллекции и учитывает принадлежность текстов коллекции к определенным классам (положительные, отрицательные, нейтральные). В работе [7] показано, что методы, основанные на мере RF, показывают лучшие результаты при вычислении веса слова с учетом принадлежности слова к разным классам, чем методы, основанные на мере TF-IDF.

Формула для вычисления меры TF-RF:

$$\text{tf.rf} = \text{tf} * \log(2 + \frac{a}{\max(1,c)}), \quad (2)$$

где a – количество сообщений (положительной) коллекции, содержащих термин; C – количество сообщений (отрицательной) коллекции, содержащих взвешиваемый термин.

Формула для вычисления Prob-based:

$$\text{prob-based} = \text{tf} * \log(1 + \frac{a}{c} * \frac{a}{b}), \quad (3)$$

где a и C – аналогично формуле (2), b – число сообщений (положительной) коллекции, которые не содержат взвешиваемый термин.

Формулы TF-ICF и ICF-based:

$$\text{tf.icf} = \text{tf} * \log(1 + \frac{|C|}{cf(t_i)}), \quad (4)$$

$$\text{icf-based} = \text{tf} * \log(2 + \frac{a}{\max(1,c)} * \frac{|C|}{cf(t_i)}), \quad (5)$$

где C – это число категорий; Cf – число категорий, в которых встречается взвешиваемый термин.

Чтобы проверить эффективность подходов, основанных на выбранных мерах, поступим, как в работе [7], – возьмем 5 терминов из реального корпуса и оценим расчет весов термина в зависимости от принадлежности термина к коллекции (обидно, плохо, люблю, конечно, время). Первые два термина относятся к классу негативных сообщений, следующие два – к классу позитивных; последний термин нейтральный, встречается одинаково часто как в коллекции позитивных твитов, так и в коллекции негативных. В табл. 2 и 3 указан вес термина, рассчитанный одним из пяти вышеописанных методов. Число в скобках указывает на частоту встречаемости термина в объединенной коллекции позитивных и негативных твитов.

Таблица 2

Практический пример использования пяти методов для категории позитивных твитов

Термин	RF	Prob-based	idf	tf.icf	Icf-based
Обидно (899)	1,0463	0,0007	2,4019	16,5566	0,3149
Плохо (1872)	1,1971	0,1996	2,0834	127,6367	0,3603
Люблю (3908)	1,9296	43,4220	1,7638	757,6925	0,5807
Конечно (1735)	1,8516	6,9751	2,1164	322,1021	0,5571
Время (2690)	1,5624	6,2502	1,9260	395,2524	0,4702

Таблица 3

Практический пример использования пяти методов для категории негативных твитов

Термин	RF	Prob-based	idf	tf.icf	Icf-based
Обидно (899)	4,1165	40,4249	2,4019	254,0693	1,2323
Плохо (1872)	2,4370	27,5370	2,0834	435,8914	0,7330
Люблю (3908)	1,3520	4,18686	1,7638	418,7327	0,4070
Конечно (1735)	1,3904	1,07085	2,1164	200,1849	0,4185
Время (2690)	1,6082	7,76170	1,9260	414,5183	0,4840

Так как метод, основанный на мере IDF, не учитывает при расчете позитивные и негативные категории, то веса для положительных и негативных терминов в столбце IDF одинаковые. Остальные методы корректно определяют термины между двумя категориями. Это означает, что негативно окрашенные слова определяются как негативные, а позитивно окрашенные слова как позитивные и не наоборот. Из тестовой выборки видно, что применение метода, основанного на мере Prob-based, дает контрастные результаты на статических коллекциях при определении веса эмоционально окрашенного слова. Аналогичный эксперимент, показавший схожие результаты, был проведен для вычисления веса биграмм.

Заключение. В работе показан подход к автоматическому построению словарей эмоциональной лексики, в которые входят как отдельные термины, так и биграммы. Словарь строится на основе размеченных коллекций и является общетематическим, т.е. не принадлежит никакой заранее определенной предметной области. Веса в словарях вычисляются с помощью методов, основанных на пяти статистических мерах. Для методов определена их вычислительная сложность при обновлении коллекции – добавлении новых сообщений. В отличие от методов, основанных на перерасчете всех весов терминов коллекции, вычислительная сложность метода, основанного на мере icf, линейна. Полученный в результате работы программный модуль позволяет динамически обновлять словарь эмотивной лексики, отслеживать и учитывать во времени лексические изменения речи, ввод в активный словарный запас новых терминов и пересчитывать веса этих терминов в зависимости от принадлежности к коллекции.

Использование полученных результатов для тренировки и оценки качества тонового классификатора, основанного на словарях эмоциональной лексики, и относится к перспективам дальнейшей работы.

Литература

1. Рубцова Ю.В. Метод построения и анализа корпуса коротких текстов для задачи классификации отзывов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды XV Всерос. науч. конф. RCDL’2013. Ярославль, Россия. – Ярославль, 2013. – С. 269–275.

2. Read J. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification // Proceedings of the Student Research Workshop at the 2005 Annual Meeting of the Association for Computational Linguistics. – USA, Ann Arbor, 2005. – P. 43–48.

3. Hu M. Mining and Summarizing Customer Reviews / M. Hu, B. Liu // Proceedings of Knowledge Discovery and Data Mining. – USA, Seattle, 2004. – P. 168–177.

4. Salton G. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Journal of Information Processing and management. – 1988. – Vol. 24, № 5. – P. 513–523.

5. Jones K.S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval // J. Documentation. – 1972. – Vol. 28, № 1. – P. 11–21.

6. Jones K.S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval // J. Documentation. – 2004. – Vol. 60, № 5. – P.493–502.

7. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization / M. Lan, C.L. Tan, J. Su, Y. Lu // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2009. – Vol. 31, № 4. – P. 721–735.

8. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams / J.W. Reed, Y. Jiao, T.E. Potoket al. // Proceedings of Machine Learning and Applications ICMLA. – USA, Orlando, 2006. – P. 258–263.

9. Inverse Category Frequency based supervised term weighting scheme for text categorization / D. Wang, H. Zhang, W. Wu, M. Lin // ArXiv e-prints. – 2010. – P. 1–12 [Электронный ресурс]. – Режим

доступа: http://arxiv-web3.library.cornell.edu/pdf/1012.2609v2.pdf, свободный (дата обращения: 02.06.2014).

10. Liu Y. Imbalanced text classification: A term weighting approach / Y. Liu, H.T. Loh, A. Sun // Expert Systems with Applications. – 2009. – Vol. 36, № 1. – P. 690–701.

Рубцова Юлия Владимировна
Аспирант лаборатории искусственного интеллекта
Института систем информатики им. А.П. Ершова СО РАН, Новосибирск
Тел.: 8-905-951-67-57
Эл. почта: yu.rubtsova@gmail.com

Rubtsova Yu.V.

Automatic term extraction approach in dynamic text collection for building Word-Emotion Dictionary for Twitter

This paper presents an approach of extraction and term weighting scheme for building sentiment vocabulary, which is constantly updated for the task of sentiment classification. The author compares the methods based on various statistical schemes and shows the computational complexity of generating representations for N dynamic documents depending on weighting schemes.

Keywords: corpus linguistics, text classification and categorization, social networks data analysis, weighting schemes.
