Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики

© М. В. Клековкина © Е. В. Котельников
Вятский государственный гуманитарный университет,
Киров
klekovkina.mv@gmail.com kotelnikov.ev@gmail.com

Аннотация

В статье представлен метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики. Приводится описание процесса создания словаря: выделение оценочных слов, назначение им весов, определение влияния слов-модификаторов и слов, выражающих отрицание. Для оценки метода используются коллекции семинара РОМИП. Результаты сравниваются с результатами других методов.

1 Введение

Активное развитие в настоящее время социальных сетей, блогов и форумов привело к увеличению интереса, как со стороны научного сообщества, так и со стороны многих организаций к задаче автоматического анализа мнений пользователей Интернета по различным вопросам (отношение к товарам, услугам, событиям, высказываниям). Одной из основных проблем при анализе мнений является классификация текстов по тональности. Тональностью текста называется эмоциональная оценка, выраженная в тексте по отношению к некоторому объекту, и определяется тональностью составляющих его лексических единиц и правилами их сочетания. В простейшем случае классификация текстов по тональности осуществляется на два классса, обозначающие позитивные и негативные эмоциональные оценки.

В данной работе предлагается метод автоматической классификации текстов по тональности с использованием словаря эмоциональной лексики. Для экспериментов используются коллекции отзывов о фильмах Российского семинара по оценке методов информационного поиска (РОМИП) [20].

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

2 Подходы для автоматического определения тональности текста

Для автоматического определения тональности текста можно выделить следующие подходы [13]:

1) на основе правил с использованием шаблонов (rule-based with patterns) [9]. Подход заключается в генерации правил, на основе которых будет определяться тональность текста. Для этого текст разбивается на слова или последовательности слов (N-grams). Затем полученные данные используются для выделения часто встречающихся шаблонов, которым присваивается положительная или отрицательная оценка. Выделенные шаблоны применяются при создании правил вида «ЕСЛИ *условие*, ТО *заключение*»;

2) машинное обучение без учителя (unsupervised learning) [16]. Данный подход основан на идее, что наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции. Выделив данные термины и определив их тональность, можно сделать вывод о тональности всего текста;

3) машинное обучение с учителем (supervised learning) [4]. В этом подходе требуется наличие обучающей коллекции размеченных в рамках эмотивного пространства текстов, на базе которой строится статистический или вероятностный классификатор (например, байесовский);

4) гибридный метод (hybrid method) [6]. Данный подход сочетает все или несколько из рассмотренных выше принципов и заключается в применении классификаторов на их основе в определенной последовательности.

На конференции «Диалог-2012» были представлены доклады участников семинара РОМИП 2011 года [1], где впервые предложены дорожки для анализа и классификации отзывов пользователей по тональности. В докладах использованы различные подходы, например, в работе [17] применен метод классификации отзывов, основанный на правилах, сформированных экспертами и методы машинного обучения с учителем для уточнения построенных

правил. В работе [19] также представлен подход на основе правил с использованием шаблонов. Для выделения шаблонов использовался лингвистический метод. Изначально эксперт выделил наиболее значимые для пользователей атрибуты объекта оценки и расширил данный список синонимами и гипонимами. Также были составлены словари оценочной лексики. Далее использовались семантические шаблоны, описывающие возможные синтаксические связи в предложении между группами получившихся словарей. Выделенные пары «атрибут + оценка» использовались в качестве терминов для автоматического определения тональности текста, для классификации использовались методы машинного обучения с учителем. В семинаре РОМИП 2011 принимала участие система, основанная на методе SVM с моделями представления документа в виде традиционных *n-grams* и предложенных авторами *d-grams* [10]. D-grams представляют собой тройки, построенные на основе дерева зависимостей для текста. Два элемента тройки – слова, третий элемент – синтаксическая связь между словами. В работе [12] для классификации отзывов по тональности использовался модифицированный авторами наивный алгоритм Байеса.

В работе [18] авторами были проанализированы результаты тестирования нескольких методов машинного обучения с учителем. Данный подход для задачи автоматического определения тональности текста показал неплохие результаты. В настоящей работе рассматривается подход на основе правил с использованием шаблонов, учитывающий эмоциональную оценку отдельных слов, а не полагающийся лишь на частоту их употребления, как в методах машинного обучения. Для этого в тексте выделяются оценочные слова, для них вычисляется эмоциональный вес, затем эти веса объединяются при помощи некоторой функции (например, среднее арифметическое или сумма). Существует несколько подходов к извлечению оценочных слов и вычислению их эмоционального веса.

В работе Turney [16] изначально выбираются два эталонных множества оценочных слов: положительное и отрицательное. Далее из отзывов извлекаются наборы, состоящие из прилагательных в сочетании с существительными и наречия в сочетании с глаголами. Turney использует наборы, считая, что, хотя изолированное слово может указывать на субъективность, его может оказаться недостаточно для определения контекста эмоциональной оценки. Тональность отзыва рассчитывается как среднее эмоциональных оценок наборов, взятых из этого отзыва. Для расчета эмоциональной оценки для набора Turney использовал поисковую систему Altavista, которая для каждого набора вычисляет оценку путем определения совместной встречаемости со словами из эталонного множества.

Множества оценочных слов также создаются вручную экспертами. Для обогащения данных множеств могут использоваться словари. В работе [3] предложен метод, использующий тезаурус для пополнения заданного вручную множества оценочных слов. Идея метода в следующем: если слово оценочное, то его синонимы и гипонимы также будут оценочными и относятся к одной тональности, а антонимы – к противоположной тональности. Еще один подход представлен в работе [2], где с помощью толкований слов в словаре определяется их ориентация. Данный метод основывается на идее, что слова с одинаковой эмоциональной оценкой имеют схожие толкования.

В нашей работе используется следующий метод выделения оценочных слов. Первоначально в словарь эмоциональной лексики вручную заносятся оценочные слова, наиболее ярко характеризующие данную предметную область. Далее словарь пополняется оценочными словами из обучающей коллекции, имеющими наибольший вес, вычисленный по методу RF (Relevance Frequency – релевантная частота) [7], для каждого класса тональности.

3 Формирование словаря эмоциональной лексики

Для решения задачи автоматического определения тональности текста был создан словарь эмоциональной лексики, характерной для заданной предметной области. В нашей работе в качестве текстов для экспериментов использовались отзывы пользователей о фильмах, поскольку коллекция по этой предметной области была представлена на семинаре РОМИП 2011 года.

Изначально вручную было отобрано 60 оценочных слов и поставлены им веса по шкале от –5 до –1 для отрицательно ориентированных слов и от +1 до +5 для положительно ориентированных. Далее данный список дополнялся ключевыми словами для положительного и отрицательного классов тональности категории «отзывы о фильмах». С этой целью для каждого слова из словаря, созданного по обучающей коллекции текстов, вычислялся вес для каждого класса по методу RF.

В методе RF [7] для вычисления веса термина используется информация о распределении этого термина по текстам обучающей коллекции с учетом принадлежности текстов к классам.

Обозначим a – количество текстов, содержащих i-й термин и относящихся к классу C, b – количество текстов, содержащих термин и не относящихся к классу C. Тогда, значимость i-го термина для класса C будет выражаться формулой:

$$RF_i^C = \log_2\left(2 + \frac{a}{\max(1, b)}\right). \quad (1)$$

В список заносились оценочные слова из множества слов с наибольшим весом, пороговый вес определялся отдельно для каждого класса. Всего в словарь было добавлено 200 оценочных слов, большинство из которых положительного класса тональности. Веса данных слов определялись

вручную по шкале от –5 до –1 для отрицательно ориентированных слов и от +1 до +5 для положительно ориентированных. Общее количество оценочных слов в словаре – 260.

Кроме оценочных слов для предметной области «отзывы о фильмах», в словарь добавлялось множество слов-модификаторов, в зависимости от которых увеличивался, либо уменьшался вес следующего за ним оценочного слова. Также в словарь включались слова и частицы, использующиеся в текстах для отрицания следующего за ними высказывания. Веса для слов-модификаторов и отрицания определялись методом скользящего контроля по *Q* блокам на основе обучающей коллекции (Q-fold cross validation) [5].

3.1 Модификаторы

Все слова-модификаторы можно разделить на две группы в зависимости от их направленности. К первой группе относятся слова-модификаторы, которые увеличивают эмоциональный вес соседнего слова (например, «очень»), ко второй – те, которые уменьшают ее (например, «несколько»).

Для изменения веса следующего за модификатором слова можно использовать метод простого сложения и вычитания. Если модификатор увеличивает эмоциональный вес слова, то к его оценке добавляется 1, иначе – вычитается 1. Одной из проблем данного подхода является то, что он не учитывает широкий диапазон модификаторов в пределах группы. Например, *абсолютный* сильнее изменяет эмоциональный вес слова, чем *значительный*. Также при усилении слова с уже большим весом увеличение его эмоционального веса должно быть больше по сравнению со словом, обладающим меньшим весом. Например, *действительно восхитительный* и *действительно хороший*.

В нашей работе используется подход, который в зависимости от слова-модификатора изменяет вес соседнего слова на некоторый процент. Например, если слово *хорошо* имеет вес 3, то *действительно хорошо* будет иметь вес $3 \cdot (100\% + 15\%) = 3.45$. Модификаторы применяются рекурсивно, начиная от ближайшего слова-модификатора к оценочному слову. В качестве модификаторов используются наречия и прилагательные. Модификаторы-прилагательные изменяют вес только существительных. Например, *полный успех* лучше, чем просто *успех*. Этот подход был рассмотрен в работе [15], в которой процентные значения для слов-модификаторов фиксировались. В нашей работе процентные значения для слов-модификаторов представлялись в виде параметров, настройка которых осуществлялась методом скользящего контроля по *Q* блокам (Q = 5) на основе обучающей коллекции. Для категории «отзывы о фильмах» было выделено 19 слов-модификаторов:

Слово-модификатор	Изменение оценки
слова-модификаторы, уменьшающие оценку	
довольно	-15%
весьма	-15%
несколько	-10%
немного	-10%
менее	-10%
незначительный	-20%
слова-модификаторы, увеличивающие оценку	
очень	50%
совсем	50%
особенно	35%
действительно	20%
намного	35%
более	10%
полный	50%
огромный	40%
самый	50%
явный	20%
абсолютный	50%
значительный	30%
весомый	40%

3.2 Отрицание

Очевидным подходом к учету слов, выражающих отрицание, является инвертирование эмоционального веса оценочного слова, находящегося после слова, выражающего отрицание. Например, если слово *хорошо* имеет вес +3, то *не хорошо* будет иметь вес –3. Однако данный подход хорошо работает лишь в некоторых случаях. Например, «великолепно» имеет вес +5. При отрицании вес изменится на –5, но *не великолепно* более позитивно, чем *ужасно*, которое также имеет вес –5. Кроме того, при данном подходе сочетание отрицания с модификаторами приведет к нежелательному результату. Например, *не очень хорошо* будет более отрицательно, чем *плохо*. В работе [15] вместо смены знака, значение эмоционального веса сдвигается к противоположной полярности на фиксированную величину. В нашей работе сдвиг отрицания настраивается методом скользящего контроля по *Q* блокам на основе обучающей коллекции. В качестве слов, выражающих отрицание, используются частицы *не*, *ни*, а также местоимение *ничего*.

4 Метод определения тональности

В предлагаемом методе тональность текста определяется на основе подсчета весов входящих в него оценочных слов. Веса оценочных слов извлекаются из словаря эмоциональной лексики, формирование которого было описано в предыдущем параграфе.

Для каждого текста из обучающей коллекции подсчитывается его вес, равный среднему весу входящих в него оценочных слов:

$$W_T = \frac{\sum_{i=1}^{N} W_i}{N}, \quad (2)$$

где W_T – вес текста; W_i – вес оценочного слова i; N – количество оценочных слов в тексте T.

Слова-модификаторы и слова, выражающие отрицание, не являются оценочными словами, а изменяют вес оценочных слов при вычислении веса текста.

Все тексты T_i помещаются в одномерное эмотивное пространство в соответствии со своим весом W_{T_i}, причем тексты положительной тональности занимают преимущественно положение справа, а тексты отрицательной тональности – слева. Для повышения уверенности классификации из рассмотрения исключаются тексты положительной тональности, которые расположены существенно левее большинства положительных текстов, ближе к отрицательным текстам, и наоборот, исключаются тексты отрицательной тональности, которые расположены существенно правее большинства отрицательных текстов, ближе к положительным. Для определения процента исключаемых текстов применяется метод скользящего контроля. Для обучающей коллекции «отзывы о фильмах» из рассмотрения были исключены 35% отрицательных текстов и 40% положительных.

После исключения текстов вычисляется среднее значение весов текстов положительного класса тональности и среднее значение весов текстов отрицательного класса тональности:

$$AW_T^C = \frac{\sum_{i=1}^{N_C} W_{T_i}}{N_C}, \; T_i \in C. \quad (3)$$

Здесь AW_T – средний вес текстов класса тональности C; N_C – количество текстов, принадлежащих классу тональности C.

Далее вычисляется граница d – середина отрезка $\left[AW_T^-; AW_T^+\right]$ (рис. 1). Решение об отнесении текста T к одному из классов принимается следующим образом: текст T относится к положительному классу, если значение веса W_T находится справа от d, иначе – относится к отрицательному классу.

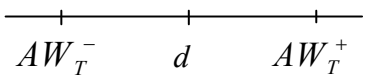

Рис. 1. *Расположение средних весов текстов положительной и отрицательной тональностей и границы d между текстами разных классов*

5 Эксперименты и результаты

В этом параграфе представлены результаты автоматического определения тональности текстов для метода, основанного на словаре эмоциональной лексики, которые сравниваются с результатами метода опорных векторов (Support Vector Machine, SVM) [4, 11] и простейшего способа классификации (baseline), в котором все тексты относятся к наиболее часто встречаемому классу тональности текстов для данной коллекции. Также приводится лучший результат классификации отзывов о фильмах, полученный на РОМИП-2011 (к сожалению, неизвестен метод, на основе которого получен данный результат). В качестве реализации метода опорных векторов использовалась библиотека LIBSVM [8]. Проводился выбор ядра и подбор оптимальных параметров. Наилучшие результаты показало линейное ядро с регулирующим параметром $C = 1$.

Для составления словаря и тестирования методов использовались соответственно обучающая и тестовая коллекции текстов, предоставленные в рамках семинара РОМИП. Коллекция обучающих текстов содержит отзывы пользователей по фильмам. Каждый отзыв, помимо текстового комментария, включает оценку данного фильма по шкале от 1 до 10. Объем обучающей коллекции составляет 15718 отзывов. В качестве тестовых данных использовались отобранные и оцененные экспертами РОМИП отзывы, их количество 312.

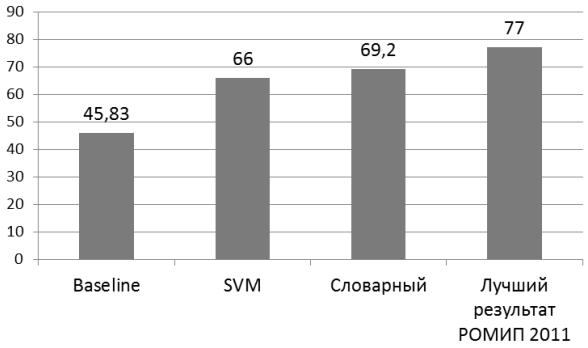

Рис. 2. *Результаты классификации (macro F1) отзывов пользователей о фильмах (%)*

Результаты классификации текстов по классам тональности представлены на рис. 2. Оценка качества классификации текстов производилась с помощью метрики *macro* F_1, вычисляемой по формуле [14]:

$$F_1 = \frac{2 \cdot R \cdot P}{R + P}. \quad (4)$$

Метрика F_1 используется в качестве характеристики, объединяющей метрики полноты R (recall) и точности P (precision).

6 Заключение

В ходе экспериментов метод, основанный на словаре эмоциональной лексики, при решении задачи автоматической классификации текстов по

тональности показал результаты, несколько превосходящие результаты метода опорных векторов. В рамках данной работы не удалось достичь результатов лучшего метода на РОМИП-2011, однако, стоит отметить, что в предложенном методе использовался словарь эмоциональной лексики общим объемом всего 260 слов. В дальнейшем планируется увеличить размер словаря эмоциональной лексики, а также предпринять шаги по повышению его универсальности, что возможно позволит улучшить качество классификации.

Исследование показало, что метод на основе словаря показывает достаточно неплохие результаты при классификации текстов на два класса. В будущем планируется протестировать работу данного метода для большего числа классов, а также на других предметных областях.

Литература

[1] Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011 // Компьютерная лингвистика и интеллектуальные технологии. – Вып. 11(18). – М.: Изд-во РГГУ, 2012. – С. 1–14.

[2] Esuli A., Sebastiani F. Determining the Semantic Orientation of Terms through Gloss Classification // Conference of Information and Knowledge Management (Bremen). ACM, New York, NY, 2005, pp. 617–624.

[3] Hu M., Liu B. Mining and Summarizing Customer Reviews // KDD, Seattle, 2004, pp. 168–177.

[4] Joachims T. Making large-scale SVM learning practical // In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), Advances in kernel methods: support vector learning, 1999. The MIT Press.

[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), pp. 1137–1143.

[6] König A. C. & Brill, E. Reducing the human overhead in text categorization // In Proceedings of the 12th ACM SIGKDD conference on knowledge discovery and data mining, August 20–23, 2006, pp. 598–603.

[7] Lan M., Tan C.L., Su J., Lu Y. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization // IEEE Transactions on Pattern Analysis and Machine Intelligence, (2009), Vol. 31, no. 4, pp. 721-735.

[8] LIBSVM – A Library for Support Vector Machines, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[9] Liu H. MontyLingua: An end-to-end natural language processor with common sense, 2004. Available at <http://web.media.mit.edu/hugo/montylingua> (accessed 1 February 2005).

[10] Pak A., Paroubek P. Language independent approach to sentiment analysis (LIMSI Participation in ROMIP'11) // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Вып. 11 (18), М.: Изд-во РГГУ, 2012, С. 37–50.

[11] Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques // Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Philadelphia, US, 2002, pp. 79–86.

[12] Poroshin V. Proof of concept statistical sentiment classification at ROMIP 2011 // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Вып. 11 (18), М.: Изд-во РГГУ, 2012, С. 60–65.

[13] Prabowo R., Thelwall M. Sentiment analysis: A combined approach // Journal of Informetrics, Vol. 3, No. 2. (April 2009), pp. 143-157.

[14] Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys, 34(1): 1–47, 2002.

[15] Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for sentiment analysis // Computational Linguistics, 37(2): 267–307, 2011.

[16] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[17] Васильев В. Г., Худякова М. В., Давыдов С. Классификация отзывов пользователей с использованием фрагментных правил // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Вып. 11 (18), М.: Изд-во РГГУ, 2012, С. 66–76.

[18] 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». Вып. 11 (18), М.: Изд-во РГГУ, 2012, С. 27–36.

[19] Поляков П. Ю., Калинина М. В., Плешко В. В. Исследование применимости методов тематической классификации в задаче классификации отзывов о книгах // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Вып. 11 (18), М.: Изд-во РГГУ, 2012, С. 51–59.

[20] Российский семинар по оценке методов информационного поиска (РОМИП). URL: http://romip.ru.

The Automatic Sentiment Text Classification Method based on Emotional Vocabulary

Evgeny Kotelnikov, Mariya Klekovkina

The method of automatic sentimental classification of texts based on the emotive vocabulary is presented in the article. The description of the process of vocabulary creating is given: the selection of the evaluation words, assigning the weights for them, defining the influence of modifying words and words, which show negation on the weights of evaluation words. The collection of ROMIP seminar is used for the appraisal of the method. The results are compared with the results of other methods.