НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2015 Том 15 № 1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2015 Vol. 15 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.4

ПРИМЕНЕНИЕ СЕНТИМЕНТ-АНАЛИЗА ТЕКСТОВ ДЛЯ ОЦЕНКИ ОБЩЕСТВЕННОГО МНЕНИЯ

Р.В. Посевкин[a], И.А. Бессмертный [a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: igor_bessmertny@hotmail.com

Информация о статье
Поступила в редакцию 28.04.14, принята к печати 03.07.14
doi: 10.17586/2226-1494-2015-15-1-169-171
Язык статьи – русский
Ссылка для цитирования: Посевкин Р.В., Бессмертный И.А. Применение сентимент-анализа текстов для оценки общественного мнения// Научно-технический вестник информационных технологий, механики и оптики. 2015. Том 15. № 1. С. 169–171

Аннотация. Описывается подход к оценке эмоциональной окрашенности естественно-языковых текстов на основе словарей тональности. Предложен метод автоматической оценки общественного мнения с помощью сентимент-анализа отзывов и обсуждений опубликованных документов в сети Интернет, базирующийся на статистике использованных слов. Разработан исследовательский прототип программной системы, производящей сентимент-анализ естественно-языкового текста на русском языке на основе линейной шкалы. Для более точного сопоставления каждого слова в предложении словарю выполняются синтаксический анализ и лемматизация. Словари тональности представлены в открытом и удобочитаемом виде, что позволяет его расширять и корректировать. Программная система сентимент-анализа русскоязычного текста, реализованная на открытых словарях тональности, разработана впервые.

Ключевые слова: анализ тональности текста, тональность, сентимент-анализ, естественно-языковой текст.

TEXTS SENTIMENT-ANALYSIS APPLICATION FOR PUBLIC OPINION ASSESSMENT

R.V. Posevkin [a], I.A. Bessmertny [a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: igor_bessmertny@hotmail.com

Article info
Received 28.04.14, accepted 03.07.14
doi: 10.17586/2226-1494-2015-15-1-169-171
Article in Russian
Reference for citation: Posevkin R.V., Bessmertny I.A. Texts sentiment-analysis application for public opinion assessment. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2015, vol. 15, no. 1, pp. 169–171 (in Russian)

Abstract. The paper describes an approach to the emotional tonality assessment of natural language texts based on special dictionaries. A method for an automatic assessment of public opinion by means of sentiment-analysis of reviews and discussions followed by published Web-documents is proposed. The method is based on statistics of words in the documents. A pilot model of the software system implementing the sentiment-analysis of natural language text in Russian based on a linear assessment scale is developed. A syntactic analysis and words lemmatization are used to identify terms more correctly. Tonality dictionaries are presented in editable format and are open for enhancing. The program system implementing a sentiment-analysis of the Russian texts based on open dictionaries of tonality is presented for the first time.

Keywords: text tonality processing, tonality, sentiment-analysis, natural language text.

Сентимент-анализ, или анализ тональности текста, представляет большой интерес для сфер и институтов общества, оперирующих с текстовыми документами. Особенно это относится к сферам образования, журналистики, культуры, издательской деятельности, эффективность которых обусловлена качеством текста, а умения и навыки работы с ним входят в состав профессиональных требований [1]. Эмоциональная окраска текста в общем случае является многомерной, и ее идентификация требует наличия мощных специально подготовленных словарей [2]. В настоящей работе решается частная задача анализа отзывов на публикации в сети Интернет на основе линейной шкалы позитивных или негативных оценок. Таким образом, с помощью сентимент-анализа отзывов и переписки людей на форумах предлагается автоматически оценивать общественное мнение относительно обсуждаемых событий.

Тональность текста в целом определяется лексической тональностью составляющих его единиц и правилами их сочетания [3]. Тональность текста определяется тремя факторами: субъект тональности, тональная оценка, объект тональности. Субъектом тональности является автор текста, объект тональности – то, о чем или о ком идет в тексте речь [4]. Тональная оценка может быть представлена в одном из

следующих видов: бинарный (положительный / отрицательный), тернарный (положительный / нейтральный / отрицательный), ранжированный [5].

Поиск лексической тональности в тексте предлагается осуществлять по заранее составленным тональным словарям (спискам паттернов) с применением лингвистического анализа. Данный метод позволяет не только показать цепочки тональной лексики, но и получить синтаксически корректные эмоциональные выражения [6].

Разработанный авторами исследовательский прототип анализатора тональности текста реализует многофазный процесс [7], состоящий из следующих этапов. На первом этапе текст разбивается на отдельные предложения, предложения – на отдельные слова. На втором этапе производятся морфологический анализ каждого слова, лемматизация и определение частей речи. Для лемматизации используется Томита-парсер. Перечисленные этапы анализа предложений необходимы для точного сопоставления найденных слов тональному словарю.

Используются тональные словари для русскоязычного текста объемом порядка 35000 слов. Если слово присутствует в словаре, записывается его тональность. В словаре каждому слову соответствует тональная оценка. Такой показатель представляет собой набор из пяти значений. Каждое значение определяет степень принадлежности слова к одному из классов: крайне отрицательный, отрицательный, нейтральный, положительный, крайне положительный. Сумма всех значений для конкретного слова равна единице. В случае отсутствия слова в словаре его тональность считается нейтральной. Последним шагом вычисляется общая тональность предложения на основе тональностей составляющих слов. Для расчета общей тональности предложения применяется метод, основанный на теоретико-графовых моделях [8]. В основе этого метода используется предположение о том, что не все слова в текстовом корпусе документа равнозначны. Какие-то слова имеют больший вес и сильнее влияют на тональность текста. Для получения конечного результата нужно вычислить значения двух оценок: положительной и отрицательной составляющей. Чтобы найти положительную составляющую предложения, необходимо найти сумму тональностей всех положительных компонентов слов предложения. Значение отрицательной составляющей текста находится аналогичным образом. Для итоговой оценки тональности всего текста нужно вычислить отношение этих составляющих по формуле.

$$S^* = \frac{\sum_{i=1}^{N}(S4_i + S5_i)}{\sum_{i=1}^{N}(S1_i + S2_i)},$$

где S^* – тональная оценка предложения; $S = [S1, S2, S3, S4, S5]$ – тональная оценка слова; $S1$, $S2$ – отрицательная составляющая тональной оценки слова; $S4$, $S5$ – положительная составляющая тональной оценки слова; N – количество слов в предложении. Значение S^* сравнивается с некоторым значением T, которое вычисляется экспериментально. Текст, в котором значение S^* близко к значению T, будет считаться нейтральным, если превосходит T – положительным, меньше значения T – отрицательным.

Узкая специализация анализатора – оценка текстов с помощью линейной шкалы – позволяет обойтись словарем небольшого объема.

В ходе исследования проведены эксперименты по анализу работоспособности разработанных алгоритмов оценки тональности текста. В качестве условия проведения эксперимента выступали отзывы пользователей электроники и бытовой техники, оставленные в специализированном сервисе «Яндекс.Маркет» и отзывы зрителей, оставленные на кинофильмы в сервисе «Кинопоиск». На данных сервисах пользователь при публикации отзыва определяет тип рецензии: положительная / отрицательная / нейтральная.

Для экспериментов были отобраны 100 предложений. Таким образом, в эксперименте сравнивается сентимент, полученный в результате анализа разработанным программным средством, и сентимент, который определил пользователь при публикации отзыва. На основании полученных результатов всех экспериментов рассчитывается точность определения сентимента:

$$Prec = \frac{N_{Cor}}{N_{All}},$$

где $Prec$ – точность определения тональности текста; N_{cor} – количество экспериментов с верно определенной тональностью предложения; N_{All} – общее количество экспериментов. По результатам экспериментов точность определения тональности текста, разработанного программного прототипа, составила 78%.

Разработана программная система сентимент-анализа естественно-языкового текста на русском языке. Словари тональности представлены в открытом и удобочитаемом виде, что обеспечивает возможность для расширения и правки уже существующей информации. Программная система сентимент-анализа русскоязычного текста, реализованная на открытых словарях тональности, разработана впервые.

1. Бессмертный И.А., Джалиашвили З.О., Максимов В.В., Маркин Д.А. Лингвооценочное управление текстом // Тезисы докладов X Международной конференции «Применение новых технологий в образовании». Троицк: Фонд новых технологий в образовании «Байтик», 1999.

2. Nugumanova A., Bessmertnyi I. Applying the latent semantic analysis to the issue of automatic extraction of collocations from the domain texts // Communications in Computer and Information Science. 2013. V. 394. P. 92–101. doi: 10.1007/978-3-642-41360-5_8
3. Позельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Тезисы докладов Международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2011». Москва, РГГУ, 2011. С. 510–522.
4. Cruz F.L., Troyano J.A., Pontes B., Ortega F.J. Building layered, multilingual sentiment lexicons at synset and lemma levels // Expert Systems with Applications. 2014. V. 41. N 13. P. 5984–5994. doi: 10.1016/j.eswa.2014.04.005
5. Ермаков С.А., Ермакова Л.М. Методы оценки эмоциональной окраски текста // Вестник Пермского университета. Серия: математика, механика, информатика. 2012. № 1. С. 85–90.
6. Parau P., Stef A., Lemnaru C., Dinsoreanu M., Potolea R. Using community detection for sentiment analysis // Proc. IEEE 9th Int. Conf. on Intelligent Computer Communication and Processing (ICCP 2013). 2013. P. 51–54. doi: 10.1109/ICCP.2013.6646080
7. Chiru C.-G., Hadgu A.T. Sentiment-based text segmentation // Proc. 2nd Int. Conf. on Systems and Computer Science (ICSCS 2013). 2013. P. 234–239. doi: 10.1109/IcConSCS.2013.6632053
8. Минаков И.А. Анализ эмоциональной тональности текста и его применение для повышения качества переходов по релевантным объявлениям // Вестник Самарского государственного технического университета. Серия: технические науки. 2013. № 1 (37). С. 58–63.

Посевкин Руслан Владимирович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, rus_posevkin@mail.ru

Бессмертный Игорь Александрович – кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, igor_bessmertny@hotmail.com

Ruslan V. Posevkin – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, rus_posevkin@mail.ru

Igor A. Bessmertny – PhD, Associate professor, Associate professor, ITMO University, Saint Petersburg, 197101, Russian Federation, igor_bessmertny@hotmail.com