Научная статья
УДК 004.9
doi: 10.34822/1999-7604-2022-2-14-19

О НЕКОТОРЫХ ПОДХОДАХ К РЕШЕНИЮ ЗАДАЧИ КЛАССИФИКАЦИИ ТЕКСТОВ ПО ТОНАЛЬНОСТИ НА ПРИМЕРЕ АНАЛИЗА АНГЛОЯЗЫЧНЫХ ОТЗЫВОВ

Екатерина Рафаэлевна Болтачева[1], Светлана Анатольевна Никитина[2]✉
[1, 2]Челябинский государственные университет, Челябинск, Россия
[1]katya473@mail.ru, https://orcid.org/0000-0001-5170-5738
[2]nikitina@csu.ru✉, https://orcid.org/0000-0002-1975-1256

Аннотация. Рассмотрены способы автоматического определения тональности текста. Проведены разбор и программная реализация методов определения эмоциональной окраски текстовой информации с последующим анализом полученных результатов. Для решения поставленных задач предложен метод, основанный на весовых коэффициентах, проанализирована его эффективность. Отдельно рассмотрены базовые алгоритмы для составления текстовой модели и проведена предобработка текстовой информации. В практической части работы методы протестированы на реальных данных: проведен тоновый анализ англоязычных отзывов о различных фильмах, взятых с сайта www.imdb.com. Точность классификации отзывов для всех реализованных методов составила более 80 %.

Ключевые слова: предобработка текстовой информации, «Мешок слов», технология TF-IDF, эмоциональная тональность текста, автоматизированный анализ текста, обработка естественного языка

Для цитирования: *Болтачева Е. Р., Никитина С. А. О некоторых подходах к решению задачи классификации текстов по тональности на примере анализа англоязычных отзывов // Вестник кибернетики. 2022. № 2 (46). С. 14–19. DOI 10.34822/1999-7604-2022-2-14-19.*

Original article

ON CERTAIN APPROACHES TO SOLVING THE PROBLEM OF SENTIMENT CLASSIFICATION OF A TEXT: AN ANALYSIS OF ENGLISH REVIEWS

Ekaterina R. Boltacheva[1], Svetlana A. Nikitina[2]✉
[1, 2]Chelyabinsk State University, Chelyabinsk, Russia
[1]katya473@mail.ru, https://orcid.org/0000-0001-5170-5738
[2]nikitina@csu.ru✉, https://orcid.org/0000-0002-1975-1256

Abstract. The article discusses methods for automatic text sentiment analysis. The study analyzes the results obtained via the analysis and software implementation of the methods for determining text emotionality. A method based on weight coefficients was proposed for solving the problems in question, and its efficiency was analyzed. Separately, basic algorithms for a text model were discussed and text preprocessing was carried out. The practice section of the article includes an application of the methods on real data, namely, a sentiment analysis of English-language reviews of various films published at www.imdb.com. All methods applied showed 80 % of accuracy in review classification.

Keywords: preprocessing of text information, Bag of Words, TF-IDF technology, text sentiment, automated text analysis, natural language processing

For citation: *Boltacheva E. R., Nikitina S. A. On Certain Approaches to Solving the Problem of Sentiment Classification of a Text: An Analysis of English Reviews // Proceedings in Cybernetics. 2022. No. 2 (46). P. 14–19. DOI 10.34822/1999-7604-2022-2-14-19.*

ВВЕДЕНИЕ

Развитию и исследованию методов обработки естественного языка посвящено достаточное количество работ, например [1–5]. В статье [1] разработан инструмент для сбора и анализа корпуса коротких русскоязычных текстов. В [2] был проведен анализ сообщений из новостной ленты в социальной сети «ВКонтакте».

В работе [3] рассмотрены методы кластеризации текстовых документов. Было исследовано применение нескольких алгоритмов к решению задачи кластеризации научных статей.

В [4] для разработки системы автоматической классификации текстов было проведено сравнение эффективности использования нескольких способов вычисления расстояния. В работе описан компьютерный эксперимент с использованием русскоязычных текстов.

В [5] показано применение автоматизированного анализа текстов вакансий с сайта HeadHunter для оценки резюме кандидата на соответствие требованиям вакансии.

В представленной статье исследована проблема определения тональности текста. Данная задача возникает, например, при анализе отзывов клиентов о товарах и услугах, чтобы определить их эмоциональную окраску (положительные, отрицательные, нейтральные). Выполнение подобного анализа вручную является трудоемким процессом, поэтому существует необходимость выполнять его автоматически. Указанные моменты определяют актуальность как разработки новых принципов и методов извлечения информации из текстовых данных на естественном языке, так и создания на их основе специальных информационных систем.

МАТЕРИАЛЫ И МЕТОДЫ

Перед тем как использовать любой из перечисленных далее методов обработки текстовой информации, необходимо провести с исходным текстом несколько процедур, которые представляют собой способы начальной обработки текста. К данным процедурам относятся: токенизация, изменение регистра, удаление стоп-слов и лишних знаков препинания, а также обработка отрицаний [6].

Токенизация представляет собой процесс разбиения текстового фрагмента на отдельные слова, которые называются токенами. Данная процедура нужна для удобства работы с отдельными словами.

Удаление стоп-слов используется для упрощения понимания семантики текста. В результате применения этой процедуры из исходного текста удаляются слова, не несущие в себе смысла. В основном это предлоги и союзы, например is, where, are и т. д.

Обработка отрицаний является объединением слова и частицы no или not в один токен, если частица not или no стоит перед данным словом. Обработка отрицаний поможет увеличить точность определения тональности текста.

Приведение слов к одному регистру нужно для того, чтобы система видела отличие между одним и тем же словом, написанным с заглавной буквы и с маленькой буквы. Также удаляются лишние пробелы, поскольку они не несут в себе никакой смысловой нагрузки при обработке текста.

Простейшая модель текста «Мешок слов» строится на основе словаря, содержащего слова всех используемых в исследовании документов [7]. Это называется «мешком», потому что всякая информация о порядке или структуре слов в документе отбрасывается. Модель заботится о том, встречаются ли известные слова в документе, а не об их положении.

Кратко опишем алгоритм построения модели:

1. Составляем матрицу, столбцы которой соответствуют входящим в текст словам, а строки – предложениям (документам). Если слово в документе есть, то записывается значение «единица», в противном случае – ноль. Это позволит создать матрицу размера $d \times n$, где d – это общее число различных слов, n – число документов.

2. Применяем технологию TF-IDF (Term Frequency (далее – TF, частота слова), Inverse Document Frequency (далее – IDF, обратная частота документа) [8].

Term Frequency – это отношение количества появлений слова в документе к числу всех слов в документе.

Если слово встречается во всех документах, оно не очень значимо для исследования.

Насколько редким является слово, можно определить по следующей формуле:

$$\mathrm{IDF}(w) = \ln\left(\frac{N}{N_w}\right), \quad (1)$$

где N – количество всех документов в наборе (корпусе);

N_w – число документов в корпусе, которые содержат слово w. Если слово w встречается часто, значение IDF уменьшается. Если слово w используется редко, то N_w снижается и тогда значение IDF возрастает.

TF-IDF – это результат произведения значений TF и IDF. Больший вес получат слова, которые встречаются в данном документе чаще, чем во всех остальных.

3. На основании технологии TF-IDF создаем матрицу, значения элементов которой используются в дальнейшем.

Наивный Байесовский классификатор. Данный метод основывается на теореме Байеса, а также использует технологию «Мешок слов» [9].

При этом считаем, что выполнены следующие условия:

1. Появление различных слов в тексте происходит независимо друг от друга (наивное предположение).

2. Взаимное расположение слов не имеет значения.

Принадлежность C к тому или иному классу c вычисляем по формуле:

$$C = \max_c \Pi_{i=1}^{|M|} P(x_i \mid c), \quad (2)$$

где C – максимальная апостериорная вероятность класса;

$|M|$ – количество слов в «Мешке слов»;

x_i – некоторое слово из «Мешка слов»;

$P(x_i \mid c)$ – вероятность принадлежности слова x_i классу c.

Параметры для классификатора находим так:

$$P(x_i \mid c) = \frac{\varphi(x_i, c)}{\sum_c \varphi(x_i, c)}. \quad (3)$$

В (3) $\varphi(x_i, c)$ – вспомогательная функция, значения которой определяем следующим образом: составляем набор часто употребляемых слов в отзывах (учитываем отрицания) и составляем таблицу принадлежности данного слова к типам эмоциональной окраски (табл. 1).

Таблица 1

Таблица принадлежности слова к типам эмоциональной окраски

	Не нравится	Отвратительно	Прекрасно	Нейтрально
Положительный	1	1	3	1
Нейтральный	2	2	2	4
Отрицательный	3	3	1	1
Сумма	6	6	6	6

Примечание: составлено авторами на основании данных, полученных в исследовании.

Если слово x_i не встречается в данной таблице, то принимаем значение $P(x_i \mid c)$ за единицу и будем считать, что это слово не имеет отношения к определению эмоциональной окраски отзыва или является очень редким.

Если слово x_i встречается в данной таблице, то значение функции $\varphi(x_i, c)$ будем брать из приведенной таблицы. Например, нам попадается слово «прекрасно», тогда для положительного класса вышеупомянутая функция примет значение 3, для нейтрального – 2, для отрицательного – 1. Знаменателем предыдущей формулы будет являться сумма значений каждого слова для всех классов, что продемонстрировано в табл. 1 в последней строке.

Если вероятности принадлежности высказывания одновременно равны нулю или все равны между собой, то принимаем текст за нейтрально окрашенный. Если максимальная вероятность будет одинаковой у положительно и нейтрально окрашенной фразы, то будем считать, что фраза положительно окрашенная.

Если максимальная вероятность будет одинаковой у отрицательно и нейтрально окрашенной фразы, то будем считать, что фраза отрицательно окрашенная.

Метод, основанный на весовых коэффициентах. Данный метод основан на описанном выше алгоритме «Мешок слов» с технологией TF-IDF и является экспериментальным. Для начала необходимо составить набор слов, которые чаще всего используются в отзывах, и поделить данный набор на две группы: положительно окрашенные слова и отрицательно окрашенные слова.

Одной из задач представленного в работе исследования является формирование достаточно полного словаря эмоциональной лексики для заданной предметной области. В качестве текстов для выполнения экспериментов использовались англоязычные отзывы о фильмах. Оказалось, что, несмотря на многообразие языка, далеко не все слова используются для написания отзывов.

Первоначально вручную было отобрано несколько характерных слов, которые были отнесены к коллекции отрицательно или положительно ориентированных. Далее эмоциональный словарь дополнялся оценочными словами для положительного и отрицательного классов тональности в категории «отзывы о фильмах». В словарь были также включены некоторые слова, использующиеся в сообщениях для отрицания последующего высказывания.

Отметим, что наполнение характерными словами такого набора является достаточно трудоемким процессом. На начальном этапе возможно составление минимального словаря-набора при дальнейшем его пополнении, что помогает увеличить точность классификации.

Для определения эмоциональной принадлежности отзыва прогоняем его через метод «Мешок слов» с технологией TF-IDF, получаем в итоге смысловую значимость в виде весового коэффициента каждого слова в отзыве. После чего считаем две суммы:

1) P – сумма весов слов из текста, которые совпали со словами из набора положительно окрашенных слов;

2) N – сумма весов слов из текста, которые совпали со словами из набора отрицательно окрашенных слов.

Затем сравниваем значения P и N:

1. Если оба равны нулю, значит отзыв нейтральный.

2. Если $P > N$, $P > 0$, то считаем значение $\frac{N}{P}$. Если оно попадает в диапазон [0; 0,55], то отзыв имеет положительную окраску, если в диапазон (0,55; 1], то он имеет нейтральную окраску.

3. Если $P < N$, $N > 0$, то считаем значение $\frac{P}{N}$. Если оно попадает в диапазон [0; 0,55], то отзыв имеет отрицательную окраску, если в диапазон (0,55; 1], то он имеет нейтральную окраску.

В итоге определяем тональность сообщения на основе соотношения значений сумм весов характерных слов, содержащихся в предварительно подготовленном словаре эмоциональной лексики.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Данные для анализа взяты с сайта www.imdb.com. Эти данные представляют собой набор англоязычных отзывов о различных фильмах. Тональность текстовых сообщений определялась по трем категориям: положительная, негативная и нейтральная.

Для компьютерной реализации [10] представленных методов был использован язык программирования C++, среда разработки Visual Studio 2017.

В табл. 2 приведены примеры выполнения предобработки исходных сообщений.

Таблица 2

Примеры предобработки отзывов

Отзыв	Результат предобработки
Nothing what made Matrix great is here. To be honest this is more of a Dragon Ball Z wannabe than a Matrix sequel	nothing what made matrix great is here to be honest this is more of a dragon ball z wannabe than a matrix sequel

Окончание табл. 2

Отзыв	Результат предобработки
This I literally one of the worst movies I have ever seen. I was so close of leaving from the cinema. Good actors but absolutely no point and awful scenario. Don't waste your time and money to see this	this i literally one of the worst movies i have ever seen i was so close of leaving from the cinema good actors but absolutely no point and awful scenario don't waste your time and money to see this

Примечание: составлено авторами на основании данных, полученных в исследовании.

В таблице 3 приведено несколько примеров отзывов о фильмах на английском языке, а также соответствующие результаты выполнения программы.

Таблица 3

Примеры работы тонового классификатора

Сообщение	Тональность, определенная наивным Байесовским классификатором	Тональность, определенная методом, основанным на весовых коэффициентах	Реальная тональность отзыва
Without giving away too much, Benedict Cumberbatch just kills it. He is one class actor. Cumberbatch rocks as Strange and pulls it off with ease. The impressions his portrayal gives us is that he was born to play the role of the ex-neurosurgeon turned Sorcerer Supreme. Such is the talent of the lead cast, that it represents just how good the movie was overall. Every normal movie watcher will love the action, humour, adventure and fantasy throughout the film. A solid new character introduced to the Marvel Cinematic Universe with loads of potential. Cumberbatch solidifies and anchors his role in the MCU as Stephen Strange in a similar fashion to RDJ doing so with Tony Stark. A must watch! Enjoy!	Положительная	Положительная	Положительная
Nothing what made Matrix great is here. To be honest this is more of a Dragon Ball Z wannabe than a Matrix sequel.	Нейтральная	Нейтральная	Нейтральная
This I literally one of the worst movies I have ever seen. I was so close of leaving from the cinema. Good actors but absolutely no point and awful scenario. Don't waste your time and money to see this.	Негативная	Негативная	Негативная

Примечание: составлено авторами на основании данных, полученных в исследовании.

Сравнение результатов использованных методов. Исходя из результатов, полученных в исследовании, можно сделать вывод, что Байесовский метод и метод, основанный на весовых коэффициентах, работают достаточно точно. Метод весовых коэффициентов имеет точность 80 %, а наивный метод Байеса – 91 %. Отметим, что обычно ошибка возникает, когда отзыв имеет «пограничную» окраску. Например, когда в отзыве высказывается легкое недовольство или человек в целом доволен фильмом, но не испытывает сильных положительных эмоций, поэтому программа может принять такой отзыв за нейтральный.

ЗАКЛЮЧЕНИЕ

В статье рассмотрена задача классификации текстов по тональности, которая относится к актуальным задачам анализа текстовых данных. Для ее решения предложен метод, основанный на весовых коэффициентах.

Отметим, что задача определения тональности является весьма сложной, кроме того, ее решение зависит от естественного языка, на котором написаны сообщения.

Полученные результаты могут быть применены для дальнейших исследований по данной проблематике, а именно для анализа эмоциональной окраски текстов на русском языке.

Список источников

1. Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1 (109). С. 72–78.
2. Гаршина В. В., Калабухов К. С., Степанцов В. А., Смотров С. В. Разработка системы анализа тональности текстовой информации // Вестн. ВГУ. Сер.: Систем. анализ и информ. технологии. 2017. № 3. С. 185–194.
3. Пархоменко П. А., Григорьев А. А, Астраханцев Н. А. Обзор и экспериментальное сравнение методов кластеризации текстов // Тр. ИСП РАН. 2017. Т. 29, Вып. 2. С. 161–200.
4. Глазкова А. В. Оценка результативности применения расстояний Евклида и Махаланобиса для решения одной из задач классификации текстов // Вестн. Дагестан. гос. технич. ун-та. Технич. науки. 2017. № 44 (1). С. 86–93.
5. Валиев А. И., Лысенкова С. А. Применение методов машинного обучения для автоматизации процесса анализа содержания текста // Вестник кибернетики. 2021. № 4 (44). С. 12–15.
6. Клековкина М. В., Котельников Е. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Электрон. библиотеки: перспектив. методы и технологии, электрон. коллекции : труды 14-й Всерос. науч. конф. (RCDL-2012). Переславль-Залесский, 2012. С. 81–86.
7. Потапенко А. А. Семантические векторные представления текста на основе вероятностного тематического моделирования : дис. … д-ра ф.-м. наук. М., 2018. 147 с.
8. Гомзин А. Г., Коршунов А. В. Тематическое моделирование текстов на естественном языке // Тр. ИСП РАН. 2012. № 23. С. 215–244.
9. Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Докл. РАН. 2014. Т. 455, № 3. С. 268–271.
10. Никитина С. А., Болтачева Е. Р. Preprocessing Text Information. Свидетельство о регистрации программы для ЭВМ RU 2022614365 от 21.03.2022. Заявка № 2022611703 от 09.02.2022.

References

1. Rubtsova Yu. V. Constructing a Corpus for Sentiment Classification Training // Software & Systems. 2015. No. 1 (109). P. 72–78. (In Russian).
2. Garshina V. V., Kalabukhov K. S., Stepantsov V. A., Smotrov S. V. Development of the System of Sentiment Analysis of the Text // Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies. 2017. No. 3. P. 185–194. (In Russian).
3. Parkhomenko P. A., Grigoryev A. A., Astrakhantsev N. A. A Survey and an Experimental Comparison of Methods for Text Clustering: Application to Scientific Articles // Proceedings of ISP RAS. 2017. Vol. 29, Is. 2. P. 161–200. (In Russian).
4. Glazkova A. V. Efficiency Assessment of Euclidean and Makhalanobis Distances for Solving a Major Text Classification Problem // Herald of Dagestan State Technical University. Technical Sciences. 2017. No. 44 (1). P. 86–93. (In Russian).
5. Valiev A. I., Lysenkova S. A. Application of Machine Learning Methods for Automation of the Process of the Text Contents Analysis // Proceedings in Cybernetics. 2021. No. 4 (44). P. 12–15. (In Russian).
6. Klekovkina M. V., Kotelnikov E. V. The Automatic Sentiment Text Classification Method Based on Emotional Vocabulary // Digital Libraries: Advanced Methods and Technologies : Proceedings of the 14th All-Russian Research Conference (RCDL-2012). Pereslavl-Zalessky, 2012. P. 81–86. (In Russian).
7. Potapenko A. A. Semanticheskie vektornye predstavleniia teksta na osnove veroiatnostnogo tematicheskogo modelirovaniia : Doctoral Dissertation (Physics and Mathematics). Moscow, 2018. 147 p. (In Russian).
8. Gomzin A. G., Korshunov A. V. Topic Modeling in Natural Language Texts // Proceedings of IST RAS. 2012. No. 23. P. 215–244. (In Russian).
9. Vorontsov K. V. Additivnaia reguliatsiia tematicheskikh modelei kollektsii testovykh dokumentov // Doklady RAN. 2014. Vol. 455, No. 3. P. 268–271. (In Russian).
10. Nikitina S. A., Boltacheva E. R. Preprocessing Text Information. Certificate for registering a program for a computer RU 2022614365 of 21.03.2022. Claim No. 2022611703 of 09.02.2022. (In Russian).

Информация об авторах

Е. Р. Болтачева – магистрант.
С. А. Никитина – кандидат физико-математических наук, доцент.

Information about the authors

E. R. Boltacheva – Master’s Degree Student.
S. A. Nikitina – Candidate of Sciences (Physics and Mathematics), Associate Professor.