DOI 10.37882/2223–2966.2020.11.11

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ СООБЩЕНИЙ НА РУССКОМ ЯЗЫКЕ

COMPARATIVE ANALYSIS OF TONALITY METHODS FOR AUTOMATIC DETERMINATION OF MESSAGES IN RUSSIAN

A. Gurin

Summary. This article is about of automatic sentiment methods. It discusses in detail the methods of determining the tonality of their strengths and weaknesses. There is also a comparison of the popular software for determining the emotional coloring SentiStrenght.

Keywords: Sentiment analysis, machine learning, dictionary approaches for determining the sentiment of text, comparison of methods for automatic sentiment determination.

Гурин Анатолий Анатольевич
Аспирант, Российский Экономический Университет им. Г. В. Плеханова
Anatoly196674@gmail.com

Аннотация. Данная статья посвящена методам автоматического определения тональности текста. В ней подробно рассматриваются методы определения тональности их сильные и слабые стороны. Также происходит сравнение популярного ПО для определения эмоциональной окрашенности SentiStrenght с разрабатываемым алгоритмом в рамках НКР.

Ключевые слова: Сентимент анализ, машинное обучение, словарные подходы для определения тональности текста, сравнение методов автоматического определения тональности.

Сентимент-анализ — это набор технологий и методов, предназначенных для извлечения информации из текстов, дисциплина, базирующаяся на стыке поиска информации и вычислительной лингвистики, которая исследует не столько содержание текста, сколько его тональность.[1]

Составляющие, которые необходимо выявить при анализе тональности:

1. Субъект тональности — источник мнения, автор сообщения.
2. Объект тональности — о чем идет речь.
3. Аспект тональности — характеристика объекта.
4. Тональная оценка — тип мнения, оценочный компонент, отношение автора к описываемому предмету, конкретное сообщение об аспектах (свойствах) объекта.

Анализ тональности в первую очередь связан с задачей классификации. Простейший случай — бинарное представление: положительно или отрицательно окрашен данный текст. Можно ввести и более дробное деление на классы, однако, с увеличением количества классов, уменьшается точность классификации. [2]

Автоматический анализ тональности текста базируется на технологиях лингвистической интерпретации эмоций, машинного обучения, извлечения эмоционального смысла из информации и т.д. Технология может использоваться для автоматической оценки новостных событий, продуктов, персоналий, организаций, стран и т.д. К задачам относятся распознавание и интерпретация мнения, кластеризация текстов, исходя из позитивных или негативных мнений; сегментация текстов по разным мнениям; прогнозирование мнений, исходя из анализируемых текстов.

Основные подходы к автоматическому определению тональности:

1. Алгоритмы, основанные на правилах (rule-based). Чаще применяются для анализа тональности текста на русском языке.
2. Алгоритмы, использующие методы машинного обучения (machine learning). Чаще применяются для анализа тональности текста на английском языке, т.к. для английского языка имеется множество общедоступных коллекций, на которых можно тренировать модели машинного обучения и большое количество программных продуктов с открытым кодом.

Подход с использованием правил и словарей заключается в следующем:

- Правила для анализа тональности используют заранее разработанные шаблоны, которые описывают определенную предметную область. По этим шаблонам из текста извлекаются n-компонентные цепочки (n-граммы), тональность которых определяется и на основе правил, и на основе словарей.

- Правила могут строиться на основе модели «если…то…»: «Если цепочка содержит глагол из списка («любить», «нравиться», «обожать» и др.) и не содержит глагола из другого списка («ужасать», «отвращать» и др.) или отрицания, то ее тональность положительная». Таким образом собираются оценки различных цепочек документа. Для получения итоговой окраски тональности общую сумму всех весов можно рассчитать по формуле, составленной разработчиками решения, т.к. универсальная формула отсутствует.[3]

Помимо правил для оценки тональности зачастую используют словари оценочной лексики, которые хранят в себе слова и словосочетания, каждому из которых поставлен в соответствие уровень эмоциональной оценки по определенной шкале. Также существуют специальные тезаурусы, в которых размечена эмоциональная составляющая текста

Словари могут составляться как вручную, так и автоматически. Вручную можно создать словарь с нуля, использую помимо всего прочего толковые словари или корпусы текстов. После этого полученные словари можно автоматически расширять, включая в них новую лексику по разработанным правилам, которые используются для извлечения из текстов оценочных слов, не попавших в словарь. [4]

Подход с использованием машинного обучения делиться на два типа:

- метод, основанный на применении машинного обучения с учителем;
- метод, основанный на применении машинного обучения без учителя.

Первый тип является алгоритмом классификации и тренируется на основе обучающей выборки, которую необходимо собрать и разметить. После этого классификатор используется для определения тональности новой выборки. При втором типе подхода под названием кластеризация в используемой для тренировки обучающей выборке неизвестны присвоенные документам тональности, а наибольший вес получают термины, которые наиболее часто встречаются в данном тексте, но одновременно с этим присутствуют лишь в ограниченном количестве текстов всего множества, таким образом, данные слова отражают тональность определённых текстов и на их основе можно сделать вывод о тональности документа в целом.

Можно также совместить кластеризацию и классификацию выборки. При помощи инструментов анализа тональности русскоязычных текстов определить выборку текстов, для который тональность прослеживается особенно хорошо, а затем использовать данную, уже размеченную выборку в качестве обучающей и произвести классификацию для оставшейся части выборки диалогов. В случае с непрерывной функцией оценки тональности можно использовать модель линейной регрессии для обучения. [5]

Одним из самых известных методов анализа тональности текста на основе лексикона является Наивный Байесовский классификатор. В данном методе используются апостериорные вероятности связанности двух слов. В контексте данного классификатора рассматривается набор вероятностей типа: вероятность того, что слово A будет следовать за словом B при удалении из текста диалога самых часто используемых и самых редко используемых слов. Типичная сфера применения Наивного Байесовского классификатора — задача нахождения сообщений со спамом среди общей коллекции писем. Но в случае с задачей нахождения диалогов с негативной окраской также можно применить данный классификатор. [6]

Наиболее точными на сегодняшний день являются алгоритмы, основанные на правилах. В отдельных случаях точность определения тональности может достигать до 96%. Но данные алгоритмы имеют ряд минусов:

- Жесткая привязка к домену. При смене тематики сообщений, алгоритмы становятся полностью бесполезными и работают с большой долей ошибок.
- Данный подход не интересен для исследования. Т.к. необходимо описать грамматические правила и роль отдельных слов.
- Трудоемкость. Для достижения хороших результатов определения тональности, необходимо составить много правил, комбинацию правил, а также их приоритеты.

Данный подход не является популярным и нашел свое применение в узконаправленной области, например, ресторанной.

Подходы, основанные на словарях, используют так называемые тональные словари для анализа текста. Обычно, тональный словарь представляет из себя список слов со значением тональности для каждого слова. Данные словари могут составляться вручную, либо интегрироваться. На сегодняшний день существует много готовых, размеченных по тональности словарей. Например, переведенная база ANEW. Данные подходы не являются универсальными, но просты в применении. Как и в подходе с правилами, словарный подход дает очень точную оценку, при этом не сильно зависит от домена и тематики сообщений. Считается одним из популярных подхо-

Таблица. 1. Сравнение алгоритмов (Жирный шрифт = sig при 0,01, курсив = sig при 0,05 по сравнению с SentiStrength.)

Алгоритм	Оптимальные характеристики	Точность	Точность +/- 1 класс	Корр.	Абсолютное Значение Ошибки в%
SentiStrength (standard configuration, 30 runs)	-	60.6%	96.9%	.599	22.0%
Simple logistic regression	700	58.5%	96.1%	.557	23.2%
SVM (SMO)	800	57.6%	95.4%	.538	24.4%
J48 classification tree	700	55.2%	95.9%	.548	24.7%
JRip rule-based classifier	700	54.3%	96.4%	.476	28.2%
SVM regression (SMO)	100	54.1%	97.3%	.469	28.2%
AdaBoost	100	53.3%	97.5%	.464	28.5%
Decision table	200	53.3%	96.7%	.431	28.2%
Multilayer Perceptron	100	50.0%	94.1%	.422	30.2%
Naïve Bayes	100	49.1%	91.4%	.567	27.5%
Baseline	-	47.3%	94.0%	-	31.2%
Random	-	19.8%	56.9%	.016	82.5%

дов для определения тональности не только в русском, но и других языках, например: английский, немецкий, греческий, арабский и др. В доказательство тому проект Sentistrength — Программное обеспечение для анализа тональности, разработано проф. Майклом Фелловолоем, главой Statistical Cybernetics Research Group университета Вулверэмптона и ассоциированным научным сотрудником Oxford Internet Institute, Великобритания. [7]

Машинное обучение с учителем является наиболее распространенным методом, используемым в исследованиях. Его суть состоит в том, чтобы обучить машинный классификатор на коллекции заранее размеченных текстах, а затем использовать полученную модель для анализа новых документов. Именно про этот метод я расскажу далее.

Машинное обучение без учителя представляет собой, наверное, наиболее интересный и в то же время наименее точный метод анализа тональности. Одним из примеров данного метода может быть автоматическая кластеризация документов.

Машинное обучение с учителем

Процесс создания системы анализа тональности очень похож на процесс создания других систем с применением машинного обучения:

- необходимо собрать коллекцию документов для обучения классификатора
- каждый документ из обучающей коллекции нужно представить в виде вектора признаков
- для каждого документа нужно указать «правильный ответ», т.е. тип тональности (например, положительная или отрицательная), по этим ответам и будет обучаться классификатор
- выбор алгоритма классификации и обучение классификатора
- использование полученной модели.

На данный момент, обучение с учителем является самым популярным методом определения тональности сообщений. Данный подход гарантирует точность определения тональности от 72% и выше.

Согласно исследованиям, опубликованным в статье Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology [8]. Было проведено тестирование различных алгоритмов определения силы положительного настроения для 1041 комментария с расширенным набором функций и 10-кратной перекрестной проверкой (в порядке убывания показателей силы положительного настроения). За исключением SentiStrength, результаты представляют собой средние результаты по 4 запускам различных случайных тестовых / обучающих разделов и для оптимального количества функций. Результаты представлены в таблице 1.

Касаемо силы негативных эмоций, большинство методов дают очень похожие результаты, а некоторые дают лучшие результаты, чем SentiStrength. Хотя, точность SentiStrength составляет 72,8%, это только на 2,9% лучше, чем базовый уровень, некоторые другие методы имеют аналогичные уровни точности, а SVM значительно точнее. SentiStrength является наиболее точным из методов, если допускается ошибка одного класса, и имеет наиболее высокую корреляцию с результата-

```
Очки тональности текста: 1.0779765908785147
Текст имеет положительную окраску
Содержание эмоции радости в тексте: 0.4669900369436992
Содержание эмоции горя в тексте: 0.1960093578661175
Содержание эмоции гнева в тексте: 0.2736904858683576
Содержание эмоции страха в тексте: 0.213171467541031
Содержание эмоции надежды в тексте: 0.18607339963327282
Содержание эмоции возмущения в тексте: 0.2718839480078404
Содержание эмоции жалости в тексте: 0.3089179741484432
Содержание эмоции облегчения в тексте: 0.16258840744654907
Содержание эмоции удовлетворения в тексте: 0.09574650660741223
Содержание эмоции разочарования в тексте: 0.0009032689302586059
Содержание эмоции радости за другого в тексте: 0.4669900369436992
Содержание эмоции гордости в тексте: 0.16258840744654907
Содержание эмоции восхищения в тексте: 0.807856633155392
Содержание эмоции стыда в тексте: 0.12916745702698065
Содержание эмоции подтвердившегося страха в тексте: 0.213171467541031
Содержание эмоции упрека в тексте: 0.05238959795499914
Содержание эмоции любви в тексте: 0.31253104986947766
Содержание эмоции ненависти в тексте: 0.38132762467369416
Содержание эмоции злорадства в тексте: 0.27289199613400822
Содержание эмоции благодарности в тексте: 0.82748466700990896
Содержание эмоции вознаграждения в тексте: 0.62957844439024831
Содержание эмоции раскаяния в тексте: 0.3251768148930981
Всего слов в тексте: 110709
Всего предложений в тексте:7719
Время работы алгоритма составляет: 75.0640115737915 секунд
```

Рис. 1. Вывод результата оценки тональности текста первого тома.

ми кодирования, полученными человеком. Теоретически ни один из методов не должен быть хуже базового, но это может произойти из-за оптимизации обучающего набора, а не набора оценки. В целом, может показаться, что SentiStrength не очень хорошо распознает негативные эмоции, но это сложная задача для анализируемых здесь коротких текстов. Также средняя процентная абсолютная ошибка для случайной категории превышает 100% из-за преобладания «1» как правильной категории для отрицательного настроения.

Таким образом, системы, основанные на словарном подходе не хуже систем, использующих машинное обучение, а иногда и лучше, все зависит от конкретной задачи.

Было решено сравнить программное обеспечение SentiStrength с разрабатываемой программой по определению тональности текста, в рамках диссертационного исследования. SentiStrength представляет особый интерес в связи с известностью, а также апробации не только в английском и русском языках, но и других. К тому же существует надстройка под русский язык, выполненная в рамках проекта при поддержке Фонда академического развития НИУ ВШЭ — Санкт-Петербург в 2012 году.[9] Само программное обеспечение доступно для скачивания и тестирования с официального сайта, также существует коммерческая версия программы, которая позволяет интегрировать алгоритмы в уже существующее решение, либо использовать ПО с большей скоростью обработки.

Разрабатываемая программа работает следующим образом:

- Подготовка данных. На этом этапе весь текст проходит очистку от знаков препинания и прочих символов. Также все слова приводятся к одному регистру.
- Все слова приводятся к начальной форме, использую OpenCorpora.
- После, используя тональные словари, а также словари выражений и идиом, словам присваиваются веса.
- Присвоение весов идет всем словам, если слова нет в тональном словаре, то оно записывается в отдельный массив и учитывается позже при оценке текста в целом.
- Кроме подсчёта негативной, нейтральной и положительной тональностей, происходит вычисление эмоций, используя модель OCC.
- После определения тональности и вычисления значений 16 эмоций, программа дает оценку тональности текста.

Вычисление эмоций один из важных процессов, т.к. их выявление улучшает работу алгоритма определения тональности, а заодно и проверяет его на ошибки.

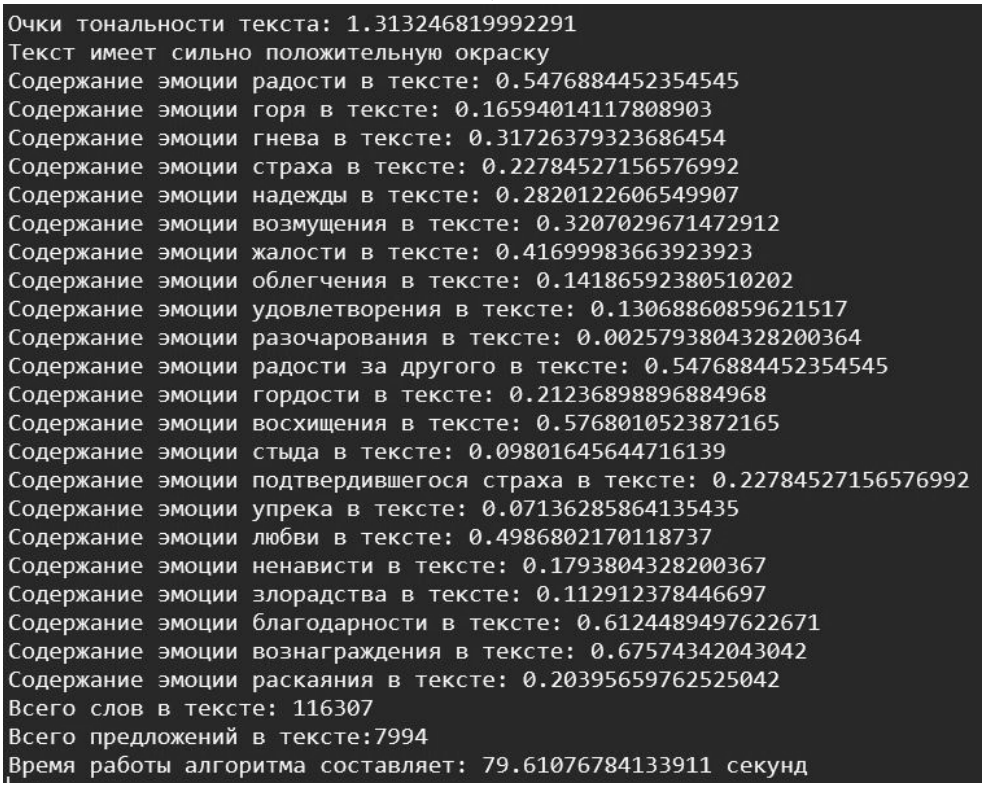

Рис. 2. Вывод результата оценки тональности текста второго тома.

Для сравнения работы алгоритмов было взято произведение Л. Н. Толстого «Война и Мир» Том I. Интерес представляет общая тональность и какие эмоции преобладают в данном произведении. Ниже на рисунке 1 представлен снимок экрана-результата тональности после обработки текста первого тома.

Первый том был определен, как положительный. В нем присутствуют все эмоции, особенно выражены радость, гнев, возмущение, жалость, восхищение, любви, ненависти, благодарности и вознаграждения. Действия первого тома описывают события войны, однако в данном тексте содержится большое количество прилагательных и наречий, которые положительным образом сказываются на общей тональности.

Текущая реализация алгоритма позволяет классифицировать текст, определяя общую оценку, кроме того предоставляется возможность детализировать её и посмотреть вектор эмоций, которые преобладают в тексте.

Исследовав аналогичный текст в SentiStrength, были получены следующие результаты:

- Рейтинг позитивных настроений: 4 по шкале от 1 (нейтральный) до 5 (сильно + ve)
- Рейтинг негативных настроений составляет –4 по шкале от –1 (нейтральный) до –5 (сильно — ve)
- Средняя тональность предложения (без округления) положительное 1500 и отрицательное –1 458

По полученным данным можно сделать вывод, что текст имеет нейтральную окраску, что не совсем верно. В связи с этим было решено провести аналогичный опыт, но уже со 2 томом.

Определение эмоциональной тональности текста у второго тома. Данный том освещает события общественной жизни в томе автор описывает личные отношения героев и их переживания, затрагивает темы отцов и детей, дружбы, любви и поиска смысла жизни. На рисунке 2 представлен снимок экрана с определенной тональностью второго тома.

Данный том, алгоритм классифицировал как сильно положительный. Здесь также преобладают эмоции радости, гнева, возмущения, любви, благодарности и вознаграждения.

Результат в SentiStrength определил следующее:

- Рейтинг позитивных настроений: 2 по шкале от 1 (нейтральный) до 5 (сильно + ve)
- Рейтинг негативных настроений составляет –3 по шкале от –1 (нейтральный) до –5 (сильно — ve)

- Средняя тональность предложения (без округления) положительное 1 091 и отрицательное –1 636

Из полученных значений, видно, что текст скорее негативный, чем положительный.

Таким образом, можно сделать вывод, что все методы определения тональности хороши. В настоящий момент не существует универсального решения и все зависит от задачи и нужд заказчика. Так, например, подход, основанный на правилах, дает отличные результаты, но он ограничен в применении, а также в выбранном домене (тематике). Резкая смена домена сильно отражается на определении тональности, да и не очень хорошо начинает работать, когда происходит пересечение совершенно разных тем. Подходы, связанные с машинным обучением, тоже не совершены, но пользуются популярностью в виду технологического уклада и развития информационного общества. В данных подходах много всяких надстроек, которые при грамотном применении дают хорошие результаты. Словарные подходы тоже имеют свои недостатки и в некоторых задачах показывают хорошие результаты, в других не очень. Считается, что словарные подходы не универсальны, но данные подходы, комбинированные с другими, например машинным обучением, вполне могу обеспечивать очень точное определение тональности, независимо от тематики и структурированности данных.

ЛИТЕРАТУРА

1. Гурин А.А., Основные методы и инструменты анализа тональности текста. // Вестник Российского Экономического Университета Имени Г. В. Плеханова. Вступление. Путь В Науку, № 3 (27) стр. 29–38–2019.
2. O'Keefe T., Koprinska I., Feature selection and weighing methods. I sentiment analysis // Australasian Document Computing Symposium. 2009
3. Prabowo R., Thelwall M., Sentiment Analysis: A combined approach // Journal of Informatics. 2009
4. Cliche M. BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs //Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). — 2017. — С. 573–580
5. Zhang Y., Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification //arXiv preprint arXiv:1510.03820. — 2015.
6. Rosenthal S., Farra N., Nakov P. SemEval-2017 task 4: Sentiment Analysis in Twitter //Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). — 2017. — С. 502–518.
7. Mikolov T. et al. Distributed Representations of Words and Phrases and Their Compositionality //Advances in Neural Information Processing Systems. — 2013. — С. 3111–3119
8. Sentiment Strength Detection in Short Informal Text. Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai. Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK.
9. Павлова Ю. Метод автоматического анализа тональности текста в применении к социологическим задачам: на примере анализа комментариев к постам Живого Журнала / Избранные тезисы докладов IV Студенческой социологической межвузовской конференции / Отв. ред.: М. Р. Демин. СПб.: НИУ ВШЭ (Санкт-Петербург), 2013.

Журнал «Современная наука: актуальные проблемы теории и практики»