Сравнительный анализ методов определения эмоциональной окраски сообщений в социальных сетях с применением обучения с учителем

Рязанова Н.Ю., к.т.н., доцент каф. ИУ7 МГТУ им. Н.Э. Баумана
ryaz_nu@mail.ru
Сперцян К.М., студент МГТУ им. Н.Э. Баумана
spertsiankamo@gmail.com

Аннотация

В данном документе рассматривается проблема анализа эмоциональной окраски сообщений в социальных сетях, способы представления текста в виде векторов признаков, анализируются и сравниваются существующие методы определения эмоциональной окраски текста: наивный байесовский классификатор, метод максимальной энтропии и метод опорных векторов.

1 Введение

Социальные сети (Twitter, Facebook, ВКонтакте и т.д.) являются одними из самых популярных площадок в сети Интернет, где люди могут делиться различного рода информацией: обмениваться сообщениями, публиковать свои мнения и отзывы относительно тех или иных событий, например, политических новостей, выхода киноновинок в прокат и других ([Kaplan, Haenlein, 2010]).

В информационном обмене через социальные сети участвуют миллионы людей. Передаваемая информация может быть как личного, так и публичного характера. Количество текстов публичной направленности регулярно растёт, что привлекает к ним внимание различных социальных, рекламных и маркетинговых служб [Дементьева, 2014]. По данным [Brand Analytics, 2017] в среднем летом 2017 года в самых популярных социальных сетях в России было опубликовано более 200 млн публичных сообщений в месяц. Зачастую эти тексты размещаются пользователями в открытом доступе с целью высказывания собственного отношения к тому или иному событию. Именно колоссальные размеры этой информации и её оценочная составляющая позволяют собирать статистику о событиях в виде мнений. В результате подобного анализа выявляется оценка социумом того или иного события.

Большой объём информации, используемой при анализе, обусловил применение компьютерных технологий для автоматизации этого процесса. Для использования компьютеров в процессе извлечения мнений из текста требуется применение формализованных методов оценки мнения. Основным подходом к выявлению выраженного в тексте отношения человека к событию, является определение эмоциональной окраски (тональности) этого текста. Эмоциональная окраска текста формируется из слов, выражающих положительное или отрицательное отношение автора. Под анализом эмоциональной окраски текста понимается классификация текстов на группы, положительно или отрицательно оценивающих событие, или текстов, которые невозможно однозначно классифицировать [Pang, Lee, 2008].

Применяемые методы обработки информации должны учитывать специфические особенности анализируемых текстов. Отношение человека может быть выражено в нескольких словах, коротком тексте или длинном повествовании, в литературном или разговорном стилях. Так, например, отзыв о фильме на специализированном интернет-ресурсе (например, «КиноПоиск») зачастую представляет из себя объёмную рецензию с соблюдением правил русского языка и использованием эмотивной лексики (лексики, имеющей ярко выраженную эмоциональную окраску [Пазельская, Соловьев, 2011]), в то время как публикации в социальных сетях обычно отличаются малым размером текста и наличием сокращений, а также в них иногда встречаются орфографические ошибки и сленговые выражения. При этом такие тексты могут не содержать прямую оценку события [Adedoyin-Olowe, 2014]. В настоящее время данная проблема привлекает к себе пристальное внимание экспертов, благодаря чему сформирован ряд подходов к решению задачи автоматического анализа мнения.

2 Формализация задачи автоматического анализа текстовой информации

Задача автоматического анализа текста требует формализованного подхода к представлению текста. В задачах классификации документ представляется в виде вектора *N-грамм*, то есть совокупности N последовательных слов текста. Чаще используются N-граммы малых порядков – униграммы и биграммы, – так как они дают лучшие результаты. Последовательность из трёх и более связанных по смыслу слов встречается значительно реже, поэтому N-граммы высших порядков являются избыточными и дают менее точные результаты [Осокин, Шегай, 2014]. Также иногда применяется комбинация униграмм и биграмм, так как в некоторых случаях такой подход позволяет получить более точные результаты, потому что учитывается не только частота использования конкретных слов, но и их последовательных пар [Будыльский, Подвесовский, 2015].

Например, представление предложения «Эта новость меня приятно удивила» в виде вектора униграмм будет иметь следующий вид: *(«эта», «новость», «меня», «приятно», «удивила»)*, в виде вектора биграмм: *(«эта новость», «новость меня», «меня приятно», «приятно удивила»)*, и, наконец, в виде комбинации униграмм и биграмм: *(«эта», «новость», «меня», «приятно», «удивила», «эта новость», «новость меня», «меня приятно», «приятно удивила»)*.

В качестве другого примера рассмотрим фразу «ужасно интересный фильм». Она имеет ярко выраженную положительную эмоциональную окраску, однако, если рассматривать её в виде вектора униграмм *(«ужасно», «интересный», «фильм»)*, то слово «ужасно» в отсутствие контекста будет иметь отрицательную тональность, что может привести к ошибке автоматического определения эмоциональной окраски всего выражения. Устранить эту проблему можно с помощью биграмм – *«ужасно интересный», «интересный фильм»*. Теперь оба элемента вектора признаков имеют положительную тональность и вероятность ошибки автоматического анализа значительно меньше.

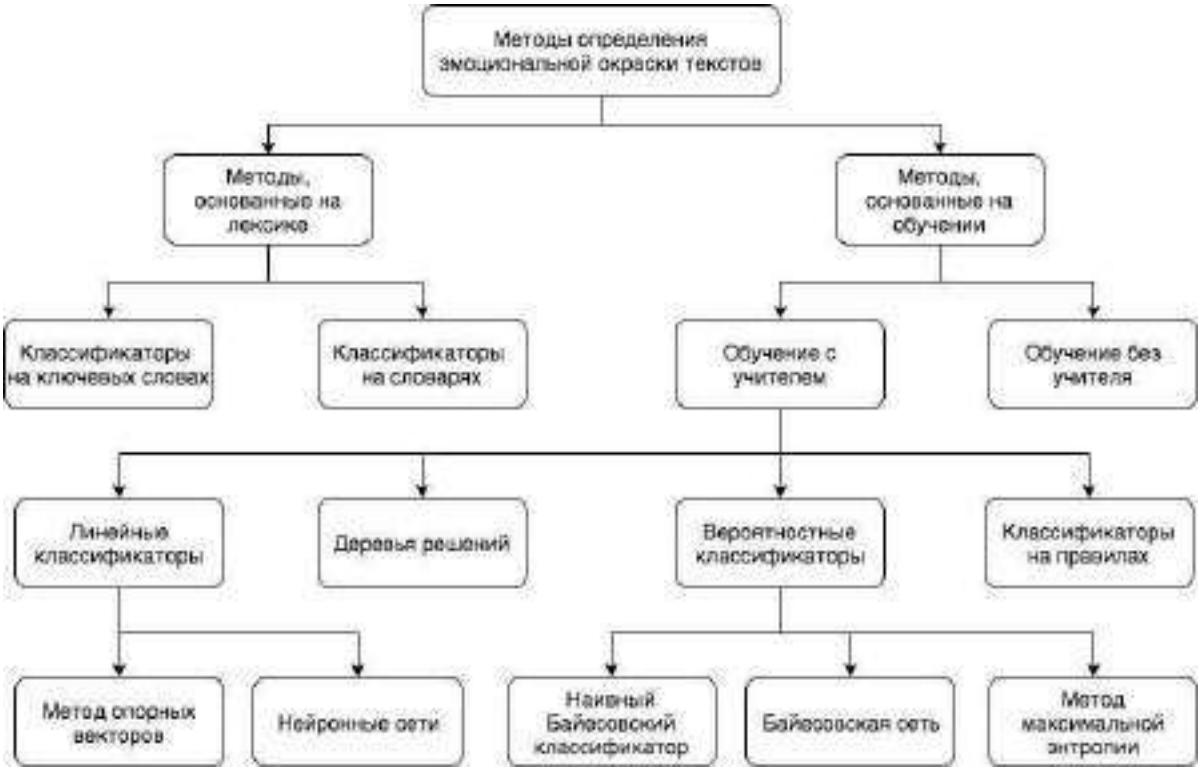

Рис. 1. Классификация методов определения эмоциональной окраски текстов

3 Классификация методов определения эмоциональной окраски текстов

Методы анализа эмоциональной окраски текста можно разделить на две большие группы: методы, основанные на лексике, и методы, основанные на машинном обучении. Последние в свою очередь разделяются на обучение с учителем и обучение без учителя (рис. 1) [Medhat, 2014].

Первая группа методов основана на эмотивной лексике в тексте. Тональность текста вычисляется исходя из тональности конкретных слов и их комбинаций по заранее составленным словарям тональностей и правилам вида «если – то» с применением лингвистического анализа. Ключевым моментом применения подобных техник является трудоёмкий процесс составления используемых словарей и правил. Готовые словари тональностей для русского языка на данный момент отсутствуют в открытом доступе, а процесс сопоставления каждому слову эмоциональной окраски не тривиален. К тому же данный подход неустойчив к орфографическим ошибкам или сокращениям, которые часто имеют место в публикациях в социальных сетях.

Методы второй группы отличаются между собой наличием или отсутствием обучающей выборки (корпуса) данных. Обучение без учителя – это раздел машинного обучения, в котором закономерности и взаимосвязи между объектами определяются из некоторой неразмеченной выборки (выборки, эмоциональная окраска текстов в которой заранее не определена экспертом (учителем)) [Воронина, Гончаров, 2015]. Данные из выборки разбиваются на классы, близкие по различным свойствам (например, позитивно окрашенные и негативно окрашенные тексты). Одной из разновидностей методов обучения без учителя является кластеризация. Любой алгоритм кластеризации требует наличия функции определения расстояния между двумя объектами выборки. Например, это могут быть расстояния между векторами признаков двух документов. Чаще всего для анализа используется расстояние Евклида (см. формулу 1), где $X = (x_1, x_2, \ldots, x_n)$ и $Y = (y_1, y_2, \ldots, y_n)$ – n-мерные вектора представления объектов исходной выборки [Zafarani, Ali Abbasi, Liu, 2014]. В задачах классификации текстов это могут быть рассмотренные вектора униграмм, биграмм и другие.

Идея подхода обучения без учителя в задаче анализа эмоциональной окраски текстов заключается в том, что больший вес при классификации конкретного текста имеют слова, которые часто встречаются нём, но реже – в остальных объектах выборки. Из весовых коэффициентов таких слов и определяется эмоциональная окраска всего текста [Стригулин, Журавлева, 2014]. Весовой коэффициент обычно принимает значение на отрезке [-1; 1] и показывает степень принадлежности слова к тому или иному классу. Например, для слова «отличный» весовой коэффициент стремится к единице, в то время как для слова «ненавижу» он будет близок к -1, а для слова «привет», не несущего никакой эмоциональной окраски, этот коэффициент будет находиться в некоторой окрестности нуля. Однако данный подход так же требует наличия словаря тональностей слов, который отсутствует в открытом доступе, как было сказано ранее, и как следствие не способен учитывать особенности текстов в социальных сетях.

Методы обучения с учителем основаны на заранее размеченной (обучающей) выборке (выборке, эмоциональная окраска текстов в которой заранее определена экспертом (учителем)), на которой происходит первоначальное обучение системы, и поэтому не требуют наличия тональных словарей. Формально эта выборка представляет из себя массив пар (x, y), где x – вектор характеристик конкретного объекта выборки, а y – класс, к которому учитель отнёс этот объект. Составление обучающей выборки избавляет от необходимости использования тональных словарей и при этом позволяет учесть особенности предметной области. Обученная модель затем используется для классификации неразмеченных текстов, также называемых тестовой выборкой, размер которой может быть в несколько раз больше размера обучающей выборки. Последняя в свою очередь представляется в виде массива пар (x, ?), то есть результирующий класс неизвестен [Воронина, Гончаров, 2015]. При этом точность классификатора напрямую зависит от размера обучающей выборки: чем большее

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \quad (1)$$

количество текстов разметит эксперт, тем выше будет точность. Ручная настройка обучающей выборки также позволяет учесть некоторые специфические особенности конкретной предметной области, например, включить в размеченный корпус тексты с применением узкоспециальных терминов, благодаря чему увеличить точность автоматического анализа.

Высокая точность даже для коротких текстов является основным преимуществом данного подхода. Сложность применения обусловлена трудоёмким процессом составления обучающей выборки.

Среди методов определения эмоциональной окраски с применением обучения с учителем выделяются своей эффективностью следующие [Pang, Lee, 2008]:

- метод опорных векторов,
- наивный Байесовский классификатор,
- метод максимальной энтропии.

4 Краткое описание методов определения эмоциональной окраски текста с применением обучения с учителем

Идея метода опорных векторов (SVM) заключается построении разделяющей гиперплоскости. Для этого необходимо увеличить размерность пространства признаков текстов. Разделяющая гиперплоскость располагается между двумя параллельными гиперплоскостями, которые в свою очередь представляют группы текстов, схожие по свойствам (например, позитивные и негативные). Расстояния между этими гиперплоскостями зависит от того, насколько точно можно разбить выборку на группы. Например, если все тексты можно строго отнести либо к положительным, либо к отрицательным, то это расстояние будет максимальным. Таким образом, расстояние между гиперплоскостями характеризует точность классификатора – чем оно больше, тем меньше будет величина ошибки [Statnikov, Aliferis, Hardin, 2011].

Наивный байесовский классификатор (NB) является статистическим методом и основывается на теореме Байеса, описывающей соотношение между условными вероятностями:

$$P(A|B) = P(B|A) \times \frac{P(A)}{P(B)}, \quad (2)$$

где A и B – классы, к которым может принадлежать текст. Вероятность принадлежности документа d классу c_i определяется исходя из частоты встречаемости признаков документа d в документах класса c_i в обучающей выборке, а также из соотношения количества документов класса c_i к общему количеству документов обучающей выборки. В методе также делается предположение об условной независимости признаков [Bing, 2011; Barbosa, Feng, 2010].

Метод максимальной энтропии (ME) в отличие от наивного байесовского классификатора не исходит из предположения об условной независимости признаков. Для определения степени принадлежности документа к некоторому классу c_i вводится функция (3).

$$F_{i',c}(d,c) = \begin{cases} 1, w_i > 0 \wedge c = c_i \\ 0, w_i \leq 0 \vee c \neq c_i \end{cases} \quad (3)$$

где w_i – очередной признак вектора признаков корпуса документов. Далее оценка принадлежности документа d классу c вычисляется следующим образом [Аксёнов, 2016]:

$$P(c|d,\lambda) = \frac{1}{Z(d)} exp \sum_{i'} \lambda_{i'} F_{i',c} \quad (4)$$

где

- λ_i – вес i-го признака,
- Z(d) – функция нормировки, определяемая формулой (5):

$$Z(d) = \sum_{c' \in C} exp \sum_{i'} \lambda_{i'} F_{i',c} \quad (5)$$

5 Формальное сравнение методов определения эмоциональной окраски текста с применением обучения с учителем

Для сравнения рассмотренных методов воспользуемся метриками точности – precision (6), и полноты – recall (7):

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

где

- TP – истинно-положительное решение (количество документов, имеющих положи-

тельную окраску и отнесённых классификатором в эту группу),

- FP – ложно-положительное решение (количество документов, имеющих не положительную окраску, но отнесённых классификатором в эту группу),
- FN – ложно-отрицательное решение (количество документов, имеющих не отрицательную окраску, но отнесённых классификатором в эту группу).

Также для наглядности воспользуемся F_1-мерой, которая даёт гармоничную оценку, исходя из точности и полноты (8):

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}. \quad (8)$$

В таблице 1 приведены оценки метода максимальной энтропии и метода опорных векторов при представлении текста в виде вектора униграмм по данным [Аксёнов, 2016]. Так, метод максимальной энтропии показал незначительно лучший результат.

Табл. 1. Результаты сравнения методов ME и SVM

Метод	Precision	Recall	F_1
ME	**0,81**	0,735	**0,77**
SVM	0,77	**0,75**	0,76

Как показано в работе [Pang, Lee, 2002] для униграмм и комбинированного подхода самым точным методом среди рассмотренных (NB, SVM, ME) является метод опорных векторов (точность 82-83%), а наивный байесовский классификатор и метод максимальной энтропии показывают приблизительно одинаковые результаты (точность 80-81%). Для биграмм точность всех трёх методов оказалась приблизительно одинакова – 77% (таблица 2).

Табл. 2. Результаты сравнения точности методов NB, ME и SVM в зависимости от типа признаков

Метод	Unigram	Bigram	Unigram + Bigram
NB	0,81	0,773	0,806
ME	0,804	**0,774**	0,808
SVM	**0,829**	0,771	**0,827**

На рисунке 2 показаны результаты сравнения наивного байесовского классификатора и метода опорных векторов на основании данных из [Barhan, Shakhomirov, 2011]. Точность последнего также оказалась выше как для униграмм и биграмм, так и для комбинированного подхода. Аналогичные результаты были получены в [Стригулин, 2016] (таблица 3). Показатели полноты и F_1-меры также выше у метода опорных векторов.

В то же время по данным из [Лебедева, 2014] результаты точности, полноты и F1-меры для всех рассмотренных методов приблизительно одинаковые (таблица 4). Тексты были представлены в виде векторов униграм.

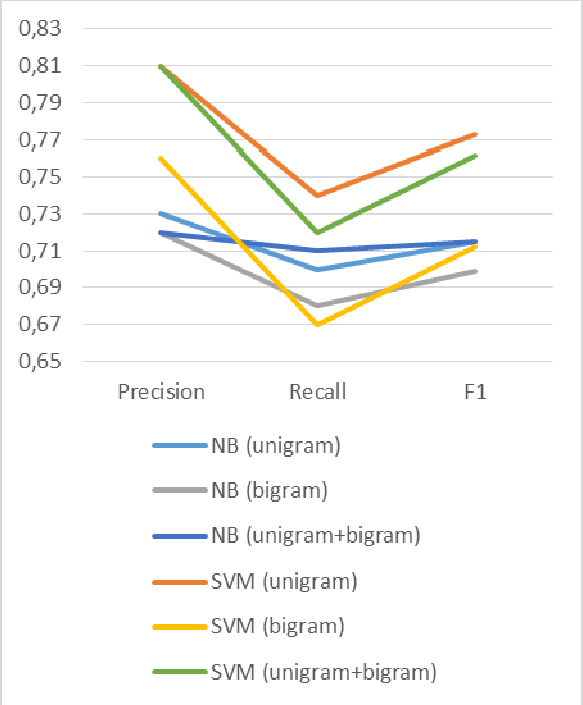

Рис. 2. График результатов сравнения показателей методов NB и SVM

Табл. 3. Результаты сравнения точности методов NB и SVM

Метод	Unigram	Bigram
NB	0,746	0,764
SVM	**0,767**	**0,777**

Табл. 4. Результаты сравнения методов NB, ME и SVM

Метод	Precision	Recall	F_1
NB	0,79	0,78	0,78
ME	0,8	**0,8**	0,79
SVM	0,8	0,79	0,79

Таким образом проведённый сравнительный анализ показывает, что наилучшие результаты получаются при применении метода опорных векторов. Точность и полнота наивного байесовского классификатора и метода максимальной энтропии отличаются незначительно, что означает, что главное преимущество метода максимальной энтропии – отсутствие предположения об условной независимости признаков – не оказывает существенного влияния на результаты. Так как предположение об условной независимости призна-

ков текста существенно упрощает затраты на проведение анализа, то использование наивного байесовского классификатора более обосновано в рамках задачи определения эмоциональной окраски текстов.

Представление текста в виде вектора униграмм даёт лучшие результаты, чем остальные способы. Это может быть связано как с особенностями предметной области, так и с размерами тестовых выборок. Из этого также можно сделать вывод, что связи между двумя соседними словами не представляют значительного интереса в задаче определения эмоциональной окраски текста, а заключение о его тональности можно сделать исходя только из конкретных слов.

6 Заключение

В работе рассмотрена задача определения эмоциональной окраски текстов в социальных сетях и приведен сравнительный анализ методов определения тональности, основанный на обучении с учителем:

- метода опорных векторов,
- наивного байесовского классификатора,
- метода максимальной энтропии.

Анализа показателей точности, полноты и F1-меры привёл к заключению, что лучшие результаты вне зависимости от способов представления векторов признаков даёт метод опорных векторов. Результаты метода максимальной энтропии и наивного байесовского классификатора отличаются незначительно, при этом реализация последнего значительно проще, в связи с чем его применение более приоритетно. При этом для получения наибольшей точности в общем случае следует представлять исходные тексты в виде векторов униграмм.

Список литературы

Аксёнов А. В. 2016. *Анализ тональности текстовых сообщений социальной сети Twitter*. Научно-технический журнал МИФИ «Теория. Практика. Инновации».

Будыльский Д. В., Подвесовский А. Г. 2015. *Векторное представление текстовой информации на русском языке*. XIX Международная научно-техническая конференция «Информационно-вычислительные технологии и их приложения».

Воронина И. Е., Гончаров В. А. 2015. *Анализ эмоциональной окраски сообщений в социальных сетях (на примере сети «ВКонтакте»)*. Вестник ВГУ, серия: Системный анализ и информационные технологии.

Дементьева И. Н. 2014. *Мониторинг общественного мнения как инструмент исследования социальной ситуации в регионе*. IV международная социологическая конференция «Продолжая Грушина».

Лебедева Е. А. 2014. *Анализ эмоциональной окраски сообщений в микроблогах с помощью вероятностных моделей*. Санкт-Петербургский государственный университет.

Осокин В. В., Шегай М. В. 2017. *Анализ тональности русскоязычного текста*. Интеллектуальные системы. Теория и приложения.

Пазельская А., Соловьев А. 2011. *Метод определения эмоций в текстах на русском языке*. The international conference on computational linguistics and intellectual technologies “Dialogue 2011”.

Стригулин К. А., Журавлева Л. В. 2016. *Анализ тональности высказываний в Twitter*. Молодой ученый.

Стригулин К. А. 2016. *Анализ тональности высказываний в Twitter*. Санкт-Петербургский государственный университет.

Adedoyin-Olowe, A. 2014. *A Survey of Data Mining Techniques for Social Network Analysis*. Journal of Data Mining & Digital Humanities.

Barbosa L., Feng J. 2010. *Robust sentiment detection on twitter from biased and noisy data*. 23rd International Conference on Computational Linguistics.

Barhan A., Shakhomirov A. 2011. *Methods for Sentiment Analysis of Twitter Messages*. Proceeding of the 12th conference of fruct association.

Bing Liu. 2011. *Sentiment Analysis Tutorial*. AAAI-2011.

Brand Analytics. 2017. *Социальные сети в России, лето 2017: цифры и тренды*.

Kaplan, A.M., Haenlein, M. 2010. *Users of the world unite! The challenges and opportunities of social media.* Science direct.

Medhat, W. 2014. *Sentiment analysis algorithms and applications: A survey.* Ain Shams Engineering Journal.

Mewari R., Singh A., Srivastava A. 2015. *Opinion Mining Techniques on Social Media Data*. International Journal of Computer Applications.

Pang B., Lee L. 2008. *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval.

Pang B., Lee L. 2002. *Thumbs up? Sentiment Classification using Machine Learning Techniques.* Proceedings of EMNLP.

Statnikov A., Aliferis C. F., Hardin D. P. 2011. *A Gentle Introduction to Support Vector Machines in Biomedicine: Theory and methods.* World Scientific.

Zafarani R., Ali Abbasi M., Liu H. 2014. *Social Media Mining. An Introduction*. Cambridge University Press.