УДК 004.89

А.А. Прокапович, А.А. Егошина
Донецкий национальный технический университет, г. Донецк
кафедра систем искусственного интеллекта

АНАЛИЗ МЕТОДОВ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ЕСТЕСТВЕННО - ЯЗЫКОВЫХ ТЕКСТОВ

Аннотация

Прокапович А.А., Егошина А.А. Анализ методов определения тональности естественно - языковых текстов. *Проведен анализ методов определения тональности естественно-языковых текстов, рассмотрены особенности существующих систем анализа тональности лексики русского языка. Показано, что наиболее популярным является лингвистический подход к анализу эмоциональной оценки сообщения, в связи с тем, что работа этих методов тесно связана с семантикой слов, в отличие от методов машинного обучения, оперирующих со статистикой и теорией вероятности*

Ключевые слова: *естественно-языковые тексты, анализ тональности, эмоциональная окраска лексики*.

Постановка проблемы

В последние годы происходит активное развитие Интернета, в том числе русскоязычного сегмента. Вместе с увеличением числа пользователей сети Интернет, возрастает и количество генерируемого ими контента. Люди оставляют сообщения на форумах, пишут посты в блогах, комментируют товары на страницах интернет-магазинов и пишут в социальных сетях.

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 класса, обозначающие позитивные и негативные эмоциональные оценки.

Методы определения тональности естественно языковых текстов

Рассмотрим наиболее популярные методы и алгоритмы оценки тональности текстовых документов.

Метод k ближайших соседей (k Nearest Neighbors) – еще один алгоритм классификации текстов. Для его реализации нужна обучающая выборка размеченных рецензий. Для определения класса рецензии из тестовой выборки, нужно определить расстояние от вектора этой рецензии до векторов из обучающей выборки. Определить k объектов обучающей выборки,

расстояние до которых минимально (k задается экспертом или выбирается согласно оценкам эффективности). Класс входного вектора – это класс, которому принадлежат больше половины из соседних k векторов. В качестве функции расстояния было выбрано Евклидово расстояние:

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^2}$$

Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики, определяет тональность на основе подсчета весов входящих в него оценочных слов, веса которых извлекаются из словаря эмоциональной лексики. Для каждого текста из обучающей коллекции подсчитывается его вес, равный среднему весу входящих в него оценочных слов:

$$W_T = \frac{\sum_{i=1}^{N} W_i}{N}$$

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 *Ti* помещаются в одномерное эмотивное пространство в соответствии со своим весом *Ti W* , 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:

$$AW_T^C = \frac{\sum_{i=1}^{N_C} W_T}{N_C}, T_i \in C,$$

Здесь AW_T – средний вес текстов класса тональности C;

N_C – количество текстов, принадлежащих классу тональности C.

Далее вычисляется граница d – середина отрезка $[AW_T^-; AW_T^-]$.

Текст T относится к положительному классу, если значение веса W_T находится справа от d, иначе – относится к отрицательному классу.

В работе [1] 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 работе [2] 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 [3] 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 [4] 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 [5], где с помощью толкований слов в словаре определяется их ориентация. Данный метод основывается на идее, что слова с одинаковой эмоциональной оценкой имеют схожие толкования.

Обзор систем анализа тональности текстовых документов

«SentiStrength» [3] — система, разработанная M. Thelwall, K. Buckley, G. Paltoglou и D. Cai. Начальное назначение было, для анализа коротких неструктурированных неформальных текстов на английском языке. Система может быть сконфигурирована для работы с текстом, также и для других языков, в том числе и для текста на русском языке.

Результат выдается в виде двух оценок – оценка позитивной составляющей текста (по шкале от +1 до +5) и оценка негативной составляющей (по шкале от -1 до -5). Также, возможно предоставления оценок в другом виде: бинарная оценка (позитивный/негативный текст); тернарная оценка (позитивный/негативный/нейтральный); оценка по единой шкале от -4 до +4.

Алгоритм основан на поиске максимального значения тональности в тексте для каждой шкалы (т.е. поиск слова с максимальной негативной оценкой и слова с максимальной позитивной оценкой). При работе алгоритма учитывается простейшее взаимодействие слов (например, слова-усилители усиливают значение тональности для слова, на которое они действуют – «очень злой» будет иметь более негативную оценку, нежели просто «злой») и идиоматические выражения.[4]

Недостатки системы: система может быть сконфигурирована для русского языка, реализованные в ней алгоритм не учитывают его специфику, в том числе русскую морфологию, что приводит к ряду проблем. Кроме того, система считает лишь общую тональность текста, не выделяя субъекты и объекты тональности.

Компонент анализа тональности текста в составе систем «Аналитический курьер» и «X-files» — разработан компанией «Ай-Теко». Компонент определения тональности текста реализует метод, основанный на словарях и правилах.

Данная система выдает пользователю массив размеченных предложений. В предложениях размечаются объекты тональности (при наличии таковых) и цепочка слов, несущая в себе тональность по отношению к ним. Кроме того, на основании найденных цепочек слов подсчитывается общая тональность для каждого предложения. Для подсчета общей тональности используется ряд специальных правил. Например (для предложения «Доктор Смит вылечил больного гриппом»), есть правило, которое говорит, что сочетание

позитивного глагола «вылечить» с негативной цепочкой (в данном случай «больной гриппом») приписывает позитив подлежащему глагола (в нашем примере — «доктору Смиту»). Тональность оценивается по тернарной шкале (позитивный/негативный/нейтральный).

Система работает в несколько этапов: предварительная обработка текста, выделение и классификация найденных слов; объединение найденных слов в связанные друг с другом цепочки; выделение объектов тональности. Недостатки системы: отсутствие количественной оценки текста.

«Ваал» – система, разработанная Шалак Владимиром. Данная система предназначена для оценки «неосознаваемого эмоционального воздействия фонетической структуры текста и отдельных слов на подсознание человека». Работа системы основана на превращении текста в частотный словарь и отнесении некоторых слов к определенным психолингвистическим категориям. Результат анализа выдается пользователю в виде набора оценок по ряду критериев, относящихся к данному тексту/слову («гладкий – шероховатый», «могучий – хилый») и т.д. Недостатки системы: система не производит анализ семантики текста, что ведет к сильной ограниченности применимости продукта. Кроме того, использование данного продукта людьми, не являющимися специалистами в области психолингвистики, не представляется возможным.

Компонент анализа тональности в составе системы RCO Fact Extractor – система, разработанная компанией RCO. Для анализа тональности текста система использует подход, основанный на правилах. Данная система учитывает синтаксическую структуру текста и взаимодействие различных типов слов.

Работа компонента происходит в пять этапов:

1) распознавание всех упоминаний об объекте во всех формах, включая полные, краткие и другие формы упоминаний;

2) отсев и полный синтаксический разбор конструкций, в которых отражаются все события и признаки, связанные с целевым объектом;

3) выделение и классификация тех позиций, в которых явно выражается тональность, и тех пропозиций, которые описывают эмоционально-коннотативные ситуации;

4) для каждой пропозиции принятие решения о тональности «позитив-негатив» с учетом тех мест, которые занимают в её составе эмоционально-коннотативные, тональные и нейтральные слова, средства выражения отрицания;

5) оценка общей тональности текста на основе тональностей всех входящих в него пропозиций.

Для своей работы компонент использует модули синтаксического анализа текста и отождествления наименований, разработанные также в компании RCO. Недостатки системы: отсутствие количественной оценки текста.

Заключение. На основе проведенного анализ можно сделать вывод, что все методы анализа можно отнести к классу обучение с учителем. Результаты их работы отличаются от используемой метрики эффективности.

Работа этих методов обычно достигает более 70% точности. Исследователи часто комбинируют подходы для достижения наилучших результатов. Например, научная работа Васильева В.Г., Давыдова С. и Худяковой М.В. [6] использует лингвистический подход, дополненный методами машинного обучения для коррекции отдельных правил классификации путем обучения.

Более популярным является лингвистический подход, так как алгоритмы, основанные на правилах, дают более точные результаты, в связи с тем, что работа этих методов тесно связана с семантикой слов, в отличие от методов машинного обучения, оперирующих со статистикой и теорией вероятности.

Список литературы

1. 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.
2. 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.
3. 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.
4. 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.
5. Hu M., Liu B. Mining and Summarizing Customer Reviews // KDD, Seattle, 2004, pp. 168–177.
6. Худякова М.В., Давыдов С., Васильев В.Г. Классификация отзывов пользователей с использованием фрагментных правил. РОМИП 2011.
7. Клековкина М.В., Котельников Е.В. Метод классификации текстов по тональности, основанный на словаре эмоциональной лексики
8. Вишневская Н.И. Программа анализа тональности текстов на основе методов машинного обучен-ия // Дипломная работа, М. 2013, 9 – 17.