СЕМАНТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ, НА ПРИМЕРЕ ОТЗЫВОВ О ТОВАРАХ

Ю. С. Кузьменков, И. А. Мурашко
Кафедра "Информационные технологии Гомельский государственный технический университет
Гомель, Республика Беларусь
E-mail: the$_y$uri@mail.ru

Статья посвящена вопросу исследования методов анализа эмоциональной окраски текстов. В данной работе представлена реализация метода анализа тональности текста, на примере отзывов о товарах. Предложенный метод может применяться для систем автоматического понимания текста, где текстовые данные слабо структурированы и принадлежат разговорному стилю.

Введение

Количество генерируемого пользователями контента в интернете выросло экспоненциально за последнее десятилетие. Весь этот контент несет в себе огроное количество информации, которою мы регулярно получаем, анализируем и используем. Для владельцев информационных ресурсов жизненно важно знать мнение пользователей будь это оценка людьми нового продукта в интернет магазине или отношение к свежей новости на новостном сайте [1]. Однако, вся эта информация представляет собой большой объем текстовых данных. Для решения этой проблемы необходимы системы анализа тональности текста. Целью данной статьи является исследование и разработка метода анализа тональности слабо структурированых текстов на примере отзывов о товаре. Для достижения данной цели были поставлены следующие задачи: 1. Провести обзор существующих методов автоматического анализа эмоциональной окраски текстов. 2. Провести исследование текстовых особенностей отзывов о товарах в контексте разработки методов анализа их эмоциональной окраски. 3. Разработать метод автоматического анализа эмоционально окрашенных сообщений.

I. Обзор методов определения тональности

Для автоматического определения тональности текста можно выделить следующие подходы [1]: 1) на основе правил с использованием шаблонов (rule-based with patterns). Подход заключается в генерации правил, на основе которых будет определяться тональность текста. Для этого текст разбивается на слова или последовательности слов (N-grams). Затем полученные данные используются для выделения часто встречающихся шаблонов, которым присваивается положительная или отрицательная оценка. Выделенные шаблоны применяются при создании правил вида «ЕСЛИ условие, ТО заключение»; 2) машинное обучение без учителя (unsupervised learning) [2]. Данный подход основан на идее, что наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции. Выделив данные термины и определив их тональность, можно сделать вывод о тональности всего текста; 3) машинное обучение с учителем (supervised learning) [2]. В этом подходе требуется наличие обучающей коллекции размеченных в рамках эмотивного пространства текстов, на базе которой строится статистический или вероятностный классификатор (например, байесовский); 4) гибридный метод (hybrid method) [2]. Данный подход сочетает все или несколько из рассмотренных выше принципов и заключается в применении классификаторов на их основе в определенной последовательности.

II. Выбор алгоритма определения тональности текста

В ряде исследований по определению полярности текста, высокую эффективность показали методы обучения с учителем. Эти методы использовались как в ранних работах по определению полярности документа, так и в современных работах, где анализируется предложения и короткие текстовые сообщения. Для решения поставленной задачи были выбраны 2 алгоритма: Метод опорных векторов относится к семейству линейных классификаторов. Целью линейной классификации является поиск гиперплоскости в пространстве признаков, разделяющей все объекты на два класса. Основная идея метода опорных векторов состоит в поиске разделяющей гиперплоскости, максимально удаленной от ближайших к ней точек в пространстве признаков. Наивный байесовский классификатор — вероятностный классификатор, основанный на теореме Байеса и (наивном) предположении о статистической независимости случайных величин. Основное достоинство данного классификатора заключается в низкой вычислительной сложности, а также в оптимальности, при условии действительной независимости признаков.

III. Тестирование эффективности алгоритма

Для тестирования алгоритма определения полярности использовался метод кросс-валидации. Процедура кросс-валидации проис-

ходила следующим образом: 1. Было зафиксировано множество разбиений обучающего множества на тренировочное и контрольное подмножества. 2. Для каждого разбиения происходило обучение алгоритма на тренировочном множестве, затем тестирование на контрольном. 3. Результатом кросс-валидации алгоритма стало среднее значение проведенных результатов тестирования на контрольном множестве. В данной работе разбиение на множества производилось случайным образом. Попадание каждого предложения в одно из двух множеств равновероятно. Наиболее эффективной конфигурацией в терминах точности и полноты оказался метод опорных векторов, обученный на юниграммах слов и эмотиконов.

IV. Описание процесса определения полярности

Метод определения полярности, реализуемый в данной статье, не предполагает использования априорных предположений о том, какие слова или символы могут содержать сообщения, принадлежащие к какому-либо классу. Это значит, что все признаки априорно равнозначны. Часто используемыми признаками при решении задачи определения полярности являются н-граммы слов. В данной работе под словом подразумевается любая последовательность букв алфавита, а под н-граммой порядка n - разделенная пробелами последовательность из n слов. Например, сообщение «Сегодня шел дождь. ;)» содержит только следующие н-граммы слов первого и второго порядка: «Сегодня», «шел», «дождь», «Сегодня шел» и «шел дождь». В ряде работ в качестве признаков используются части речи. Это объясняется тем, что мнение содержит субъективную лексику. Например, в работе [3] составляется словарь прилагательных и наречий, как терминов, выражающих эмоцию. По этой же причине было так же решено выбрать в качестве признаков н-граммы из частей речи. В связи с тем, что в сети интернет часто используется сленг и эмотиконы, в качестве признаков был выбран набор часто употребляемых эмотиконов. Для использования методов обучения с учителем требуется обучающая выборка. Обычно обучающее множество составляется из примеров той области, в которой будет применяться классификатор. В качестве обучающей и проверочной выборки был составлен корпус, состоящий из 8000 предложений, для которых определена полярность. Часть этих предложений была извлечена из размеченного корпуса, предоставленного авторами [3] для свободного доступа. Другая часть была получена с помощью онлайновой системы Sentiment140 анализа эмоциональной окраски. Все примеры полученного обучающего множества получены из мнений об электронной технике, а именно о мобильных телефонах, планшетах, плеерах.

V. Оценка полученных результатов

Эффективность алгоритмов извлечения аспектов формулируется в терминах точности и полноты. В контексте решаемой задачи эти метрики имеют следующий смысл. Алгоритм извлечения аспектов проверяет каждый термин документа на принадлежность множеству аспектов. Тогда точностью этого алгоритма называется отношение числа правильно определенных аспектов к числу всех терминов, отнесенных к классу аспектов, а полнотой – отношение числа правильно определенных аспектов к числу аспектов в документе. Для тестирования используются предложения, взятые из размеченного корпуса, предоставленного авторами работы [3]. Из корпуса взяты 1700 предложений из отзывов о смартфонах и 1100 отзывов о ноутбуках. В каждом из этих предложений выделены аспекты. Результаты тестирования реализованного алгоритма показаны в таблице :

Таблица 1 – Результаты работы алгоритма

Домен	Точность	Полнота
Смартфоны	0.67	0.79
Ноутбуки	0.61	0.71

VI. Заключение

Проанализировав полученные результаты, можно сделать следующие выводы: для решения задачи определения полярности предложений и коротких сообщений эффективны алгоритмы обучения с учителем. Проблемой обучения с учителем является составление тренировочного корпуса с примерами из предметной области, в которой будет использоваться классификатор. Однако схожей проблемой обладают и словарные методы: веса терминов словаря, составленного для одной предметной области, могут оказаться неадекватными для другой. Задача определения полярности текста успешно решается с помощью методов обучения с учителем. Для увеличения эффективности этих методов используются лингвистические и частотные фильтры, позволяющие отсеивать слова, не имеющие отношения к аспектам.

Список литературы

1. Конторович С.Д. Методика мониторинга и моделирования структуры политически активного сегмента социальных сетей / С. Д. Конторович, С. В. Литвинов // Инженерный вестник Дона. – 2011. – № 4.
2. Ahuja, Y. M. Corporate blog as e-CRM tools Building Consumer engagement through content management. Journal of Database Marketing Customer Strategy Management / Y. M. Ahuja, J. P. Medury // IEEE 24th International Conference on Data Engineering Workshop. – 2012. – P. 11–14.
3. Thelwall, A. V. Sentiment strength detection in short informal text / A. V. Thelwall, J. K. Buckley // Journal of the American Society for Information Science and Technology. – 2010