ИСПОЛЬЗОВАНИЕ НАИВНОГО МЕТОДА БАЙЕСА ДЛЯ АВТОМАТИЧЕСКОГО АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ

Лой А. А.
Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків, вул.. Пушкінська, 79/2, тел. 707-63-63,
e-mail:loy.alyna@gmail.com

Анализ тональности текстов или Sentiment Analysis – это область компьютерной лингвистики, которая занимается изучением мнений и эмоций в текстовых документах. Целью анализа тональности является нахождение мнений в тексте и определение их атрибутов.

В системах сентимент анализа применяются различные методы определения тональности текстов, среди которых можно выделить векторный анализ, поиск эмотивной лексики по заранее составленным тональным словарям с применением лингвистического анализа, метод на основе правил с использованием шаблонов, различные статистические методы и д.р.

В данном исследовании для классификации мнений, выражаемых по отношению к определенному объекту высказывания, мы используем наивный метод Байеса. Наивная Байесовская модель является вероятностным методом классификации на основе обучающей коллекции. Метод обычно применяется для определения принадлежности документа к определенному классу на основе частотного распределения термов. При этом, вероятность того, что документ *d* попадёт в класс *c*, пропорциональна произведению весов термов документа для данного класса. “Наивность” метода заключается в предположении, что вероятности попадания термов в определенный класс независимы друг от друга.

Использование этого метода для задачи Sentiment Analysis имеет ряд особенностей. Как и любой статистический метод, наивный метод Байеса сопровождается трудностями в обработке контекста, в частности сарказма. Частично решить эту проблему помогает особый подход к обработке смайликов и кавычек, тогда как при обычной классификации текстов подобные символы были бы отсеяны. Осуществляемая нами классификация ограничивается тремя классами: *positive*, *negative* и *neutral*. Причем класс neutral обычно определяется по равенству вероятностей попадания текста в классы *positive* и *negative*.

Предлагается следующий алгоритм реализации метода Байеса для анализа тональности текстов:

Формирование обучающей коллекции, представляющей набор текстов с заранее известной тональностью.

Лингвистическая обработка текстов обучающей коллекции.

Определение частотного распределения термов по классам *positive* и *negative* для обучающей коллекции.

Лингвистическая обработка анализируемого теста.

Классификация анализируемого текста.

Определение тональности на основе результатов классификации.

Для лингвистической обработки используется специально разработанный лингвистический процессор (ЛП), который включает следующие этапы: предлингвистический, проблемно-ориентированное удаление стоп-слов, морфологическая обработка, дополнительное исключение стоп слов.

На этапе предлингвистической обработки из текста исключаются числительные (кроме круглых чисел, так как круглые большие числа усиливают позитивную или негативную окрасу выражения), знаки препинания и другие разделители (особой подвергаются символы, которые могут обозначать началом смайлика: они отсеиваются только после считывания следующего символа), все буквы приводятся к нижнему регистру.

На первом этапе лингвистической обработки исключаются стоп слова. Проблемно-ориентированная особенность удаления стоп слов заключается в отличие данного этапа от подобного этапа традиционного лингвистического анализа, так как некоторые местоимения могут иметь эмотивный оттенок (часто в английском языке употребление "*that*" вместо "*this*" может отражать негативное отношение автора к объекту высказывания). При последующей лингвистической обработке учитывается, что эмотивное (то есть нужное нам) значение часто несет не корень слова. Например, однокоренные слова *useful* и *useless*, *necessary* и *unnecessary*, обладают противоположными тональностями. Поэтому значимые морфологические элементы не должны быть отсеяны.

На этапе морфологической обработки неправильные глаголы приводятся к форме инфинитива, и осуществляется замена изменяемых окончаний (например *-lies* на *–ly*) и отсечение незначимых окончаний *(-s, -es, -ies, -ly* и т. д.). Под окончанием здесь понимается не морфологическая категория, а набор букв в конце слова.

Существительные, меняющие корень во множественном числе, приводятся к форме единственного, и существительные, меняющие корень в женском роде, приводятся к форме мужского рода. Затем осуществляется повторное исключение стоп слов, которые могли видоизмениться в процессе лингвистической обработки (например, *ones* после отсечения окончания становится *one* и исключается как стоп слово).

В ситуации, когда объект и субъект высказываний известны или не важны, при использовании разработанного специализированного лингвистического процессора, наивный метод Байеса не только прост в реализации, но и предоставляет достаточно точные результаты.

Список літератури

1. Андреева А.Н., Никитина М.С. Сентимент-анализ брендов в российской блогосфере как инструмент маркетинговых исследований. // Бренд-менеджмент– М., 2012. – № 4 (64).

2. Керимов А., Прохоров А. Сентимент-анализ и продвижение в социальных медиа // Журнал «Компьютер Пресс» – М. 2012. – № 7.

3. Субботин С.В., Большаков Д.Ю. Применение байесовского классификатора для распознавания классов целей. // "Журнал Радиоэлектроники" – М., 2006. – № 4.

4. Хомив Б. Проблемы анализа тональности // SemanticForce . Електронний ресурс: http://www.semanticforce.net/ru/blog/article/10-problem-analiza-tonalnosti/