УДК 681.3

АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ТОНАЛЬНОСТИ ТЕКСТА И ЕГО ПРИМЕНЕНИЕ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ПЕРЕХОДОВ ПО РЕЛЕВАНТНЫМ ОБЪЯВЛЕНИЯМ

И.А. Минаков

Институт проблем управления сложными системами РАН
443020, г. Самара, ул. Садовая, 61
E-mail: cscmp@iccs.ru

Рассматриваются различные подходы к анализу эмоциональной тональности текста, проводится оценка применимости метода при решении практической задачи оптимизации интернет-рекламы и приводятся практические рекомендации по настройке метода в зависимости от специфики предметной области, задачи и используемого языка документов.

Ключевые слова: *сентимент-анализ, анализ тональности, онлайн-реклама, оптимизация.*

Введение

Анализ тональности текста (sentiment analysis) – область компьютерной лингвистики и интеллектуального анализа текста (text mining), ориентированная на извлечение из него субъективных мнений и эмоций. Технология может использоваться для автоматической оценки новостных событий, новых продуктов и услуг, оценки действий человека, компании или страны. К типовым задачам относятся распознавание и интерпретация мнения, сегментация и классификация текстов по разным типам и категориям эмоциональной окраски мнения; прогнозирование и даже формирование мнений в зависимости от контекста.

Данная технология существует уже много лет под различными именами и синонимами (sentiment metrics, brand monitoring, opinion mining, social media analysis, appraisal extraction, subjectivity analysis, polarity classification и др.), но особое развитие получила в последние годы с распространением Интернета и возрастанием популярности социальных сетей, блогов, твитов.

Оперативность появления новой информации (отзывы доступны для анализа немедленно после исследуемого события) и широкая аудитория сети Интернет (в развитых странах доходящая до 90 % от числа жителей) позволяют применять эту технологию со значительно большей степенью точности и достоверности результатов.

Сейчас подход все более востребован в таких областях, как социология, политология и маркетинг, отвечая на следующие типы вопросов: «Что клиенты думают о продукте?», «Как в динамике воспринимается репутация нашей компании?», «Насколько вновь внедренный сервис/услуга повлиял на мнение клиентов о фирме?», «Чем клиентов привлекают продукты конкурентов?» и т. п.

Подходы к классификации тональности

Обзор существующих методов и подходов к анализу тональности текста приведен в [1] и [2]. Известные подходы можно разделить на следующие категории.

Игорь Александрович Минаков (д.т.н.), старший научный сотрудник.

1. Подход, основанный на тональных словарях. Содержит список слов и словосочетаний со значением тональности, как положительной, так и отрицательной (пример реализации в [3]). При этом используется способ представления документа либо в виде набора слов (bag-of-words), либо в виде набора N-грамм (т. е. комбинаций пар, троек и т. д. синтаксически связанных слов). Общая тональность текста может определяться либо формульным путем (например среднеарифметическое за вычетом стоп-слов), либо более сложными методами (например обучение классификатора с использованием нейронной сети или генетических алгоритмов, чтобы точнее подстроить веса).

2. Подход с использованием эвристических правил и шаблонов. Набор вручную сформированных шаблонов правил ЕСЛИ – ТО, где в части ЕСЛИ описывается условие (как простое унитарное условие, например «содержит слово из положительного набора», так и набор условий, например «не содержит негативных слов» + «нет отрицаний» + «нет нераспознанных слов»), а в части ТО – вес принадлежности к какой-либо группе (см., например, [4]).

3. Машинное обучение. Обучение классификатора на тестовой выборке размеченных текстов, а затем использование сформированной модели для последующего анализа. Включает целый спектр технологий, в том числе латентно-семантический анализ, метод опорных векторов, байесовские классификаторы, метод Rocchio, нейронные сети и другие (см., например, [2]).

Достоинства и недостатки подходов, выявленные на основе нашего опыта их использования, приведены в табл. 1.

Таблица 1

Сравнение методов анализа эмоциональной тональности текста

	Достоинства	**Недостатки**
1. Словари	Простота реализации Легкость масштабирования на новые области и языки Объяснимость результатов Легкое подключение разных языков	Низкая точность
2. Правила	Высокая точность при корректной подстройке Прозрачность принятия решения Объяснимость результатов Хорошая поддержка стемминга и лемматизации	Сложность настройки на новую предметную область Новые правила для каждого языка Сложность и противоречивость для сложной системы
3. Машинное обучение	Легкость настройки Хорошие результаты в случае, если велико количество классов, на которые делится тональность	Необходимость обучающей выборки Усложнение подстройки под задачу/предметную область

На наш взгляд, наибольшей перспективой обладают гибридные методы, в идеале совмещающие подходы машинного обучения и эвристических правил и шаблонов.

Хороший обзор платных и бесплатных систем анализа тональности текста приведен в [5].

Использование анализа эмоциональной окраски текста в задаче оптимизациия интернет-рекламы

В качестве практической проблемы исследовалась задача оптимизации интернет-рекламы в ее частном случае – повышение вероятности перехода пользователя по рекламной ссылке (постановка проблемы и общие методы оптимизации в области интернет-рекламы рассматривались в работах [6, 7]).

В качестве тестовой выборки анализировались рекламные кампании, которые идут на сайтах, принадлежащих разным клиентам и даже находящихся в разных странах. Всего в выборке участвовала 1031 кампания, каждая в среднем идущая на 40 страницах.

Для каждой кампании измерялся CTR (Click-through rate, число нажатий на рекламную кампанию за тысячу показов).

В силу того, что рекламные кампании некорректно сравнивать друг с другом, все сравнения проводились на каждой кампании независимо, а затем результаты нормировались. При этом при нормировании учитывались внешние факторы (то, что кампании с разными типами рекламных баннеров могут иметь CTR, различающийся в несколько раз; что CTR зависит от региона, времени и т. п.). Т. е. нормировка осуществлялась среди кампаний со схожими параметрами, чтобы максимально исключить внешние факторы, воздействующие на CTR.

Оценка эмоциональной тональности осуществлялась с помощью бесплатной системы оценки тональности текста AlchemyAPI, а также ряда собственных систем классификации, построенных по правилам, приведенным в п. 3 данной статьи.

Результаты анализа приведены в табл. 2.

Таблица 2

Средний CTR рекламных кампаний в зависимости от тональности текста

Тон	А Анализ тональности	Б Тональность + ориентация на контент	В Тональность + поведенческий таргетинг
Нейтральный	0.3 (базис)	0.54	0.67
Позитивный	0.28	0.66	0.79
Негативный	0.38	0.41	0.55

Анализ выявил несколько очень интересных закономерностей.

А – можно видеть, что если реклама показывается «случайным» образом, без привязки к смыслу страницы, то число переходов на рекламу больше в случае, когда общая эмоциональная окраска негативна. Вероятно, это объясняется тем, что большинство людей предпочитают избегать большого количества негативных новостей путем переключения на другую тематику.

Б – когда реклама связана с содержанием страницы, то ситуация меняется и пользователи чаще переходят по рекламному объявлению в том случае, если тональность текста страницы позитивна. Данный результат легко объясним – если, к примеру, рекламируются туристические путешествия, то с куда большей вероятностью они привлекут внимание в статье «*Как правильно отдохнуть*», нежели в статье «*Автобус с туристами упал в пропасть*».

В – использование поведенческого таргетинга. Повышается отклик пользователя, поскольку релевантность еще более возрастает. Но тональность в данном случае

влияет так же, как и в предыдущем варианте.

Прагматические выводы просты – следует стремиться показывать позитивную рекламу (без привязки к контексту) в случае, если смысл страницы не распознан. А когда смысл ясен и существует рекламное объявление, релевантное контенту страницы, следует стремиться показывать его только тогда, когда эмоциональная окраска текста положительна.

Общие практические рекомендации по применению подхода к анализу тональности текста

По итогам практической работы с технологией в задачах разной тематики сформировался некоторый набор рекомендаций по использованию подхода в зависимости от задачи. Данные выводы сугубо эмпирические. Кроме того, они применимы только к европейским языкам. Восточные языки (арабский, китайский, японский и др.) требуют дополнительного исследования.

Классификатор

– Не существует наилучшего классификатора, подходящего под любые задачи.

– Выбор напрямую определяется типом данных. В частности, для блогов лучше подходит метод опорных векторов, а для «Твиттера» хорошие результаты показывают байесмовские методы (предположительно потому, что выполняется основное допущение – слова в твите практически независимы, а смысл понимается из общего набора).

– Рекомендуется использовать фильтр FCBF [8] – он хорошо отбирает атрибуты с минимальной взаимной информацией. Также подходит Mutual Information [9].

– При слишком долгой тренировке классификатора есть риск «перетренировки» – признаки становятся слишком специфичными для обучаемой коллекции данных. При этом перетренировка в основном зависит от классификатора, а не от признаков. Такие методы, как опорные векторы или деревья решений, менее подвержены проблеме перетренировки.

Классификации по группам

– Наиболее корректно алгоритмы отрабатывают в случае, когда деление осуществляется на три группы (позитив, негатив, нейтральное).

– Использование двух групп приводит к большому шуму (из-за необходимости принять бинарное решение для пограничных случаев).

– Использование большого дерева решений (с учетом степени уверенности, эмоциональности, объективности-субъективности) в основном не приносит особой выгоды из-за общей высокой погрешности методов, обусловленной неопределенностью, нерелевантностью, спамом и т. д.

Выделение признаков

– Наилучшее практическое использование дают биграммы и триграммы, при этом для французского и русского языка лучше работают триграммы, а для английского – биграммы. Для немецкого в силу специфики словоформирования возможны и униграммы.

– Использование униграмм дает плохое качество результатов, использование комбинаций более чем трех слов существенно повышает нагрузку на производительность без ощутимого прироста результата.

– Отдельно следует обрабатывать сокращения, аббревиатуры и слова из 2-3 букв.

– При определении веса признака рекомендуется использовать delta tf-idf [10]. Обычный TD-IDF не очень применим, поскольку частотность не так важна, как при поиске.

– Отрицательные конструкции желательно прикреплять к соседним словам. При этом для русского языка достаточно прикреплять «не» к глаголу и «нет» к существительному. Для английского учитывать модальные глаголы. Для французского и немецкого необходим более сложный синтаксический разбор, т. к. отрицание может отстоять от объекта.

Использование на практике

– Поскольку каждый из методов обладает рядом недостатков, повышающих недостоверность результата, для реальных задач рекомендуется использовать наборы классификаторов, где каждый участвует в итоговом решении с некоторым весом.

– В случае если анализируется тренд (т. е. изменение отношения во времени), рекомендуется принцип «не уверен – исключи из рассмотрения». Понятно, что он применим только при большой выборке.

– В практических задачах рекомендуются методы, позволяющие расширения за счет эвристик. Например, во французском и немецком крайне желателен синтаксический разбор до формирования N-грам, в английском полезен анализ времени глаголов и условных времен, а в русском – использование морфологии (первое или третье лицо, местоимения и пр.).

Ограничения метода

Необходимо помнить, что все используемые методы все равно обладают определенной степенью погрешности. Наше практическое исследование показало, что даже лучшие подходы позволяют добиться точности не более 70-75 % для случая трех классов разделения. Основные проблемы, препятствующие корректному анализу, следующие:

– использование сарказма (негатив трактуется как позитив);

– смешанность формулировок (в одной фразе содержится как позитив, так и негатив, – например, когда часть функций нравится, а часть нет);

– смешанность адресата или сравнение (когда упоминаются два объекта, но не распознано, по отношению к кому/чему проявлен негатив);

– целенаправленный спам (все чаще используется в политтехнологиях, где применяются однотипные записи в поддержку того или иного кандидата);

– некорректность контекста (когда позитив и негатив имеются, но не относятся к объекту исследования);

– невыполненное условие (отзыв был бы позитивным / негативным, если бы выполнилось определенное условие; иногда условие уже невыполнимо, иногда оно зависит от будущих факторов);

– использование смайликов, меняющее контекст фразы.

Также хороший обзор проблем анализа тональности приведен в [11].

Заключение

В связи с экспоненциальным ростом информационных материалов, популяризацией социальных сетей и переходом бизнес-активности в Интернет интерес к области анализа эмоциональной тональности текстов возрос многократно.

Наш опыт работы с данной технологией подтверждает ее пользу и применимость при решении практических задач.

В то же время технология находится еще на ранних стадиях развития и ни один из методов неприменим «из коробки» и нуждается в существенной доработке под предметную область и специфику задачи – будь то трудновыполнимое требование к разметке значительного корпуса текстов или дорогостоящая подстройка правил под

предметную область.

Существующие известные системы для анализа эмоциональной окраски, как платные, так и бесплатные, тоже не гарантируют приемлемого результата.

Отдельным вопросом является поддержка разных языков одновременно, зачастую востребованная при анализе.

На наш взгляд, перспективы заключаются в открытом и дополняемом инструментарии, позволяющем конструировать инструмент анализа тональности и выбора из множества известных методов компьютерной лингвистики, включая морфологический и синтаксический анализ, работу со словарями, методы машинного обучения, деревья принятия решений и ряд других.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. http://habrahabr.ru/post/149605/
2. *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» (Бекасово, 30 мая – 3 июня 2012 г.). – Вып. 11 (18). – М.: Изд-во РГГУ, 2012.
3. *Strapparava, C. and Vlitutti, A.* (2004). Wordnet-affect: and affective extension of wordnet. In Proceedings of the 4th International Conference on Language Resources and Evaluation.
4. Yang, Shih A rule-based approach for effective sentiment analysis – http://pacis2012.org/files/papers/pacis2012_T25_Yang_288.pdf
5. *Прохоров А., Керимов А.* Сентимент-анализ и продвижение в социальных медиа / // Компьютер-Пресс. – 2012. – № 7. – С. 98-105.
6. *Якушин А.В., Вольман С.И. , Минаков И.А.* Разработка системы поддержки принятия решений при оптимизации хода рекламных кампаний в сети Интернет // Проблемы управления и моделирования в сложных системах: Тр. XI Междунар. конф. – С. 68-72.
7. *Минаков И.А., Якушин А.В., Кочуров А.В., Хайрутдинов А.Р., Вольман С.И.* Разработка системы моделирования динамики поведения пользователей для оптимизации рекламных кампаний в сети Интернет // Проблемы управления и моделирования в сложных системах: Тр. XI Междунар. конф., Самара, 22 июня – 24 июня 2009. – Самара: СНЦ РАН, 2009. – С. 644-651.
8. http://web.itu.edu.tr/~cataltepe/pdf/2008_ISCIS_BarisFeatSelect.pdf
9. http://nlp.stanford.edu/IR-book/html/htmledition/mutual-information-1.html
10. Justin Martineau, and Tim Finin. Delta TFIDF: An Improved Feature Space for Sentiment Analysishttp://ebiquity.umbc.edu/_file_directory_/papers/446.pdf
11. http://www.semanticforce.net/ru/blog/article/10-problem-analiza-tonalnosti/

Статья поступила в редакцию 12 января 2013 г.

SENTIMENT ANALYSIS: PRACTICAL RECOMMENDATIONS AND ITS USAGE FOR IMPROVING USER-REACTION ON RELEVANT ONLINE ADVERTISEMENTS

I.A. Minakov

Institution of the Russian Academy of Sciences Institute for the Control of Complex Systems of RAS
61, Sadovaya st., Samara, 443020

The paper discusses and classifies different approaches for sentiment analysis, shows its implementation for solving a problem of optimizing online advertisements and gives practical recommendations for SA setup depending on problem domain specifics, task requirements and languages of documents.

Keywords: *sentiment analysis, online advertisements, optimization.*

Igor A. Minakov (Dr. Sci. (Techn.)), Senior Research.