АЛГОРИТМЫ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Н.С. Иванин
Студент кафедры информатики БГУИР

А.И. Гербик
Студент кафедры информатики БГУИР

Е.А. Макович
Студент кафедры информатики БГУИР

М.В. Аксамит
Студентка кафедры информатики БГУИР

П.Е. Дорошкевич
Студент кафедры информатики БГУИР

А. И. Свито
Студент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: nikivnik@gmail.com, alexander.gerbik@gmail.com, egormakovich@rambler.ru, rikka1128@gmail.com, pavel.darashkevich@gmail.com, alexandervirk@gmail.com

Abstract. Sentiment Analysis (SA) is an ongoing field of research in text mining field. SA is the computational treatment of opinions, sentiments and subjectivity of text. In this paper various algorithms for sentiment analysis are studied and challenges and applications appear in this field are discussed.

Рост информатизации общества и проникновение технологий во все сферы деятельности человека повлек за собой накопление больших объемов данных, в частности текстовых. Текстовая информация играет важную роль в деятельности человека, так как это наиболее распространенный и универсальный способ представления знаний человека об окружающем мире. Поэтому в настоящее время является актуальной задача обработки текстовых данных. Ручная обработка невозможна из-за большого объема накопленных данных, автоматическая обработка осложняется отсутствием структурированности в текстовых данных, неоднозначностью трактовки значений слов, наличием многочисленных исключений из правил естественного языка и т.д.

Задача анализа тональности текста (Sentiment analysis) является одной из задач обработки естественного языка (Natural Language Processing). Целью анализа тональности является нахождение мнений в тексте и определение позиции автора относительно упомянутой темы. Позиция автора может быть различной, и тональная оценка может принимать различные значения. Например: “положительная”, “отрицательная” и “нейтральная” либо “положительная” и “отрицательная”. Данную задачу можно рассматривать как задачу классификации на три и два класса соответственно, далее мы будем рассматривать задачу с двумя возможными вариантами тональной оценки, так как задача классификации на три и более класса является более

сложной в техническом отношении. Для решения задачи классификации эффективными являются методы машинного обучения с учителем.

Для того, чтобы методы решения задач классификации можно было применить для анализа тональности текста, необходимо текст представить в виде математического вектора. С этой целью применяется векторная модель “мешок слов” - модель текста, предложенная в 1975 году Дж. Солтоном, и в настоящее время одной из самых распространенных в различных областях лингвистических исследований. Текст в данной векторной модели рассматривается как неупорядоченное множество слов. Вектор, являющийся модельным представлением текста в векторном пространстве, образуется упорядочением весов всех слов (включая те, которых нет в конкретном тексте). Размерность этого вектора равна количеству различных слов во всей коллекции, и является одинаковой для всех текстов коллекции.

Так как в естественных языках встречаются устойчивые словосочетания, смысл которых отличается от смысла входящих в их слов (например, “give up” в английском языке) целесообразно применять модель не “мешок слов”, а “мешок N-грамм”. N-грамма - последовательность из N элементов (слов). Последовательность из трех элементов называют триграмма, последовательность из двух элементов называется биграмма. N-граммы меньшей длины, как правило, дают лучшие результаты, чем N-граммы большей длины, т.к. обучающая выборка в большинстве случаев недостаточно большая для нахождения статистических закономерностей N-грамм большой длины. Для многих практических приложений можно получить хороший результат, используя в качестве признаков одиночные слова и биграммы.

Модель “мешок слов” некорректно работает со словами, меняющими тональность выражения на противоположное. Например, фразы “мне нравится этот фильм” и “мне не нравится этот фильм” будут иметь положительную тональность, хотя у второй фразы она должна быть отрицательной. Чтобы решить эту проблему, можно объединять слово “не” со следующим словом, в результате в данном примере мы получим слово “не-нравится” и модель будет работать корректно. Также эту проблему можно решать при помощи N-грамм, но, как правило, это вынуждает использовать N-граммы большей длины.

Для ликвидации неоднозначности, вызванной возможностью одного и того же слова быть различными частями речи, применяется тегирование частей речи - определение для каждого слова в предложении его части речи по положению в предложении и/или грамматической форме.

Полученную задачу классификации можно решить различными методами машинного обучения: наивный байесовский классификатор, логистическая регрессия, метод опорных векторов, методы нейронных сетей и т.д. Сравнив их временную сложность, качество полученных моделей, масштабируемость можно выбрать наиболее подходящий для конкретных данных и конкретной задачи.

Алгоритмы анализа тональности текстов предназначены для определения тональности целого текста либо его фрагмента. В таком подходе предполагается, что исходный текст является мнением автора о каком-то одном конкретном объекте, например, ресторане или книге. Однако в некоторых доменах в отзыве о объекте или сущности так же содержится мнение автора о ее составляющих. Например, в отзывах о мобильных телефонах могут быть оценены такие части телефона, как экран, камера, аккумулятор, батарея. Поэтому после решения задачи анализа тональности текстов начали разрабатываться алгоритмы для анализа мнений по конкретным свойствам или частям (такие свойства или части называются аспектами). О таких свойствах или частях автор отзыва может высказывать мнения с различной тональностью и основной задачей алгоритмов анализа тональности является выделение аспектов и определение тональности отзыва пользователя об каждом аспекте.

Согласно [1] мнение (или регулярное мнение) это набор из пяти элементов $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$, где e_i – это имя сущности; a_{ij} – это один из аспектов сущности; oo_{ijkl} – это тональность мнения автора о аспекте a_{ij}, относящемуся к сущности e_i; h_k – автор мнения;

t_l – время, когда автор h_k высказал свое мнение. Регулярное мнение – это позитивное или негативное настроение, отношение, эмоция об объекте или аспекте объекта, которое высказал автор. Тональность мнения oo_{ijkl} может быть положительной, отрицательной или нейтральной, либо же измеряться в некотором интервале, например, от 0 до 1.

Довольно часто в отзывах пользователей можно встретить мнение об объекте в целом, например, «отличный смартфон». В [2] делается предположение о том, что такую категорию можно рассматривать как аспектную. Также существует возможность объединить аспекты в аспектные категории. Для случая смартфонов такие аспекты как разрешение, цветопередача, диагональ могут быть объединены в аспектную категорию дисплей.

Когда человек высказывает свое мнение о чем-либо, то его высказывание имеет некоторую цель. Такой целью служит аспект или тема, которые в дальнейшем будут извлекаться из высказывания. Таким образом, основой задачи извлечения аспектов является определение оборотов, характеризующих отношение автора, и аспекта, к которому автор высказывает свое отношение. Обороты, выражающие настроение, могут выполнять две функции: показывать положительное или отрицательное отношение и быть неявным аспектом, например «этот телефон большой», «большой» это прилагательное, характеризующее отношение автора, но также это неявный аспект размер. Как правило, в качестве аспектов выступают существительные и именные группы[3]. Длина извлекаемых именных групп при этом обычно не превосходит трех.

В [4] выделяют 4 основных подхода к извлечению явных аспектов:

– извлечение на основе часто встречающихся существительных и именных групп;

– извлечение на основе отношений между оценочными оборотами и аспектами;

– извлечение на основе машинного обучения с учителем;

– извлечение на основе статистических тематических моделях.

В подходе извлечения аспектов на основе часто встречающихся существительных и именных групп осуществляется поиск явных оценочных оборотов, как было отмечено выше, это существительные и именные группы. Они извлекаются из большого числа отзывов из определенной области.

В работе [3] для извлечения аспектов используется алгоритм, основанный на CBA[5]. Перед началом работы к каждому отзыву необходимо осуществить предобработку. Это необходимо для исключения слов, которые обычно не являются аспектами. Предобработка включает удаление стоп-слов, стемминг, лематизацию и исправление написания слов. На следующем шаге алгоритм CBA извлекает часто встречающиеся множества элементов. Каждый элемент в этом множестве это возможный аспект. Для извлечения полезных и подлинных аспектов используется фильтрация. Авторы предлагают использовать два типа фильтрации: фильтрация на основе компактности (среди кандидатов длины 2 и более удаляются те, составляющие которых отстоят друг от друга на большом расстоянии) и фильтрация лишних кандидатов (среди кандидатов длины 1 удаляются те, которые определенное число раз входят в кандидаты большей длины). Затем осуществляется поиск оценочных оборотов. Для каждого отзыва, который содержит аспект, извлекается ближайшее прилагательное. Если такое прилагательное найдено, то оно рассматривается как оценочный оборот. Так же данный подход позволяет извлекать аспекты, упомянутые только несколькими пользователями. Для этого из каждого отзыва, который не содержит аспектов, но содержит оценочный оборот, извлекается наиболее близкое к оценочному обороту существительное или именная группа.

В [6] предлагается система, построенная на основе домен независимой системы извлечения информации KnowItAll[7]. На первом этапе своей работы предложенная система извлекает существительные и именные группы из отзывов, оставляя при этом только те, частота встречаемости которых больше определенного уровня. После этого каждой именной группе присваивается оценка с помощью оценивающего модуля. Оценка выставляется на основе вычисления PMI[8]. Система использует явные аспекты для извлечения оценочных оборотов.

Если в предложении содержится аспект, то она использует определенные шаблоны извлечения оценочных оборотов. В системе, описанной выше, использовалась похожая идея, однако в данной системе для извлечения применяется парсер, генерирующий синтаксические зависимости.

Как было описано выше, оценочные обороты имеют цель. Довольно часто найти оценочные обороты не является сложной задачей, поэтому для извлечения аспектов достаточно найти цели. На этой идее основан метод извлечения аспектов на основе отношений между оценочными оборотами и аспектами.

В работе [9] для извлечения аспектов используется синтаксический анализатор, который генерирует граф грамматической зависимости. Этот графи используется для получения зависимостей между аспектами и направленными на них оценочными оборотами. В этой системе применяется Stanford Parser (http://www-nlp.stanford.edu/software/lex-parser.shtml). Этот парсер используется для определения наиболее короткого расстояния от аспекта до оценочного оборота. Затем производится стемминг и частеречная разметка. После этого извлекается размеченная часть между аспектом и оценочным оборотом, например, в предложении «This smartphone is great» smartphone является аспектом, а great оценочным оборотом. Это предложение будет размечено следующим образом «smartphone(NN) – nsubj – is(VBZ) – dobj – great (JJ)». После удаления встречаемых редко шаблонов оставшиеся шаблоны используются как шаблоны отношений между аспектами и оценочными оборотами для извлечения аспектов.

В последнее время довольно активное распространение получили алгоритмы машинного обучения. Они находят активное применение в задаче извлечения информации. Извлечение аспектов из отзывов так же относится к этой задаче, что дает возможность применения алгоритмов машинного обучения для извлечения аспектов. Существует два подхода при использовании алгоритмов машинного обучения в нашей задаче: методы, использующие для обучения заранее подготовленный список аспектов и методы, основанные на разметке последовательности слов. Наибольшее распространение получили методы на основе скрытых марковских моделей и условных случайных полей.

В работе [10] были для извлечения аспектов были применены условные случайные поля. Аспекты извлекались из предложений, содержащих оценочные обороты. На вход модели условного случайного поля были переданы следующие параметры:

– токен - этот параметр приставляет собой текущий токен;

– часть речи – этот параметр представляет собой часть речи текущего токена. Так же этот параметр может служить для разрешения лексической неоднозначности;

– путь зависимости – путь, получаемый в синтаксическом дереве между аспектом и оценочным оборотом. Для получения пути зависимости используется Stanford Parser (http://www-nlp.stanford.edu/software/lex-parser.shtml);

– расстояние между словами.

В этой системе возможные метки использовались в соответствии со схемой Inside-Outside-Begin: метка B-target означала начала аспекта, I-target означала продолжение аспекта, а метка O использовалась для обозначения других токенов.

Статистические тематические модели используются для определения тем на основе большой коллекции документов. Тематическое моделирование относится к обучению без учителя, который считает, что текст состоит из некоторого числа тем, а темы являются вероятностным распределением слов. Тематические модели могут быть применены для извлечения аспектов, если каждый аспект будет рассматриваться как униграммная языковая модель [11].

В работе [12] предложена модель, которая является смесью моделей для аспектного анализа и анализа тональности. Это модель состоит из аспектной модели, модели анализа положительной тональности и модели анализа отрицательной тональности. Эти модели были обучены на некоторых тренировочных тестовых данных. Предложенная модель базируется на pLSA [13].

Подходы, основанные на машинном обучении, показывают неплохие результаты. Однако они требуют для обучения размеченные данные, а этот процесс довольно трудоемкий, к тому же полученные модели являются доменно-зависимыми. Подходы, основанные на часто встречающихся существительных и именных группах и на отношениях между оценочными оборотами и аспектами позволяют избежать этих проблем, однако они показывают меньшую точность при работе.

В дальнейшем предполагается реализовать рассмотренные методы, а также рассмотреть возможность применения методов машинного обучения — регрессионного и структурированного вариантов SVM, Gradient boosting.

Литература

[1]. Liu B., Zhang L. A survey of opinion mining and sentiment analysis // Mining Text Data. Springer: US, 2012. P. 415-463.

[2]. Лукашевич Н. В. Автоматический анализ тональности текстов по отношению к заданному объекту и его характеристикам // Russian Digital Libraries Journal. — 2015. — Т. 18, № 3-4 . — С. 88–119.

[3]. Hu M., Liu B., Mining opinion features in customer reviews, Proceedings of the 19th national conference on Artifical intelligence, p.755-760, July 25-29, 2004, San Jose, California.

[4]. B. Liu. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, pages 1--167, 2012.

[5]. Liu, B., Hsu, W., Ma, Y. 1998. Integrating Classification and Association Rule Mining. KDD-98, 1998.

[6]. Popescu A., Extracting product features and opinions from reviews // Natural language processing and text mining. A. Popescu et al. Springer: London. 2007. P. 9-28.

[7]. Etzioni O., Unsupervised named-entity extraction from the Web: An experimental study, Artificial Intelligence, O.Etzioni, [et al], v.165 n.1, p.91-134, June 2005 .

[8]. Turney P.D., Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, Proceedings of the 12th European Conference on Machine Learning, p.491-502, September 05-07, 2001.

[9]. Zhuang L., Jing F., Zhu X. Movie review mining and summarization // Pro-ceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006), 2006. P. 43-50.

[10]. Jakob N, Gurevych I., Extracting opinion targets in a single- and cross-domain setting with conditional random fields, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, p.1035-1045, October 09-11, 2010, Cambridge, Massachusetts.

[11]. Chu W.W. Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities (Studies in Big Data, Springer. 2013.

[12]. Mei Q., Ling X., Wondra M., Su H, Zhai C, Topic sentiment mixture: modeling facets and opinions in weblogs, Proceedings of the 16th international conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada.

[13]. Hofmann T., Probabilistic latent semantic analysis, Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, p.289-296, July 30-August 01, 1999, Stockholm, Sweden.