Исследование подходов к аспектному анализу тональности текстов и существующих программных решений

Алиева А. В.

Федеральное государственное бюджетное образовательное учреждение высшего образования «Уфимский государственный авиационный технический университет»

Аннотация

В данной статье рассмотрены основные подходы к аспектному анализу тональности текстов, выявлены их особенности. Рассмотрены существующие программные средства, применяющиеся при решении задачи выделения сущностей из текстов и определения их тональности.

Ключевые слова: извлечение аспектных терминов, машинное обучение, нейронные сети, обработка текстов на естественном языке, определение тональности текста.

Введение

В настоящее время задача анализа тональности текстов на естественном языке чрезвычайно востребована. Анализ тональности сейчас – это мощный инструмент для масштабной обработки мнений людей, полученных из различных источников, таких как блоги, веб-сайты, социальные сети и др. Уметь пользоваться этим инструментом особенно важно коммерческим компаниям: от того, насколько быстро и качественно компания может анализировать потребительские отзывы, зависит уровень продаж, успех компании и её превосходство над конкурентами. При этом важно уметь анализировать как позитивную, так и негативную информацию о продуктах компании, чтобы вовремя на нее реагировать и корректировать направление деятельности компании в нужную сторону.

Основной формой передачи информации в сети Интернет является текст. Однако представленная в нем информация является слабо структурированной, содержит орфографические, грамматические и пунктуационные ошибки. Кроме того, количество платформ, на которых размещается подобная информация, а также количество текстов, которые необходимо проанализировать, настолько огромно, что ручной анализ данной информации – это практически невозможная для решения задача. Именно поэтому задача автоматизации обработки и анализа тональности текстов является сейчас актуальной задачей.

Первые походы к анализу тональности текстов заключались в том, чтобы определить тональность всего текста или его фрагмента. Такой подход и сейчас удобно применять, когда в тексте речь идет об одном продукте и необходимо выделить мнение пользователя о продукте целиком, не обращая внимания на детали. Например, такой подход представлен в работе [1], в которой рассматривалась задача о разделении отзывов о продуктах Amazon на положительные и отрицательные. При этом для представления слов в векторном виде использовался один из простейших методов «Bag of words», а классификация отзывов на позитивные и негативные осуществлялась при помощи наивного байесовского классификатора. В другой работе [2] подобная задача разделения отзывов на положительные и отрицательные была решена при помощи нейронной сети. В данном случае рассматривалась бинарная классификация отзывов о еде в ресторанах, для

представления слов в векторном виде использовалась модель Glove, для классификации использовалась сеть RNN.

Однако в настоящее время знание только тональности отзывов не дает практически никакой прикладной информации для последующего анализа, так как чаще всего в отзывах рассматривается сразу несколько сторон объекта, мнение о каждой из которых может быть противоположным и должно рассматриваться отдельно друг от друга. Например, в одном отзыве о ресторане посетитель может выразить положительное мнение о качестве еды, но при этом высказаться негативно об обслуживании официантов. Для решения этой проблемы появился другой подход к анализу тональности текстов, а именно анализ тональности по отношению к сущностям (аспектам), упомянутым в тексте. При этом каждая сущность в каждом предложении характеризуется различными словами и выражениями, которые называются аспектными терминами.

В данной статье будут рассмотрены различные подходы к анализу тональности текстов по аспектам, а также существующие программные средства, использующиеся для решения данной задачи.

Подходы к анализу тональности текстов по аспектам

Задача анализа тональности текстов по аспектам включает в себя следующие подзадачи: выделение аспектных терминов, выделение оценочных суждений (тональностей), сопоставление тональностей аспектным терминам. Рассмотрим несколько работ, в которых использовались различные подходы для решения данных задач.

В работе [3] для выделения аспектных терминов использовался классический частотный подход, который заключается в том, чтобы определить наиболее часто встречающиеся в предложениях слова. Для определения тональности выявленных аспектных терминов был использован подход на основе словаря оценочной лексики и правил. Стоит отметить, что такой подход к выделению аспектов хорошо работает в том случае, если аспект постоянно описывается одними и теми же терминами. Однако он плохо показывает себя в работе с низкочастотными терминами (например, аспект «еда» включает в себя названия различных блюд).

В работе [4] анализировалось влияние отзывов других покупателей на покупку или отказа от покупки клиентом на сайте товаров. Авторами использовался тот же частотный подход для выделения аспектных терминов, но с отличием: аспектный термин включался в рассмотрение только в том случае, если мнение пользователей по данному аспектному термину влияло на оценку товара в магазине (например, пользователи часто критиковали плохой «сигнал связи» на телефоне iPhone, однако это не повлияло на высокую оценку товара, а значит данный термин исключался из рассмотрения). Для определения тональности выделенных аспектов использовался метод, основанный на машинном обучении, а именно метод опорных векторов (SVM). Авторами отмечается, что выделение только важных аспектных терминов существенно повлияло на итоговое качество определения тональности.

Такой же подход, только уже на русскоязычных текстах, был опробован в работе [5]. Кроме классификатора SVM были также рассмотрены другие популярные классификаторы, такие как наивный байесовский классификатор (бернуллевский (NB) и мультиномиальный (MNB)), логистическая регрессия (LogReg), дерево решений и случайный лес. Авторами было отмечено, что лучшими классификаторами являлись SVM и NB. К недостатку данного подхода стоит отнести необходимость в использовании

предобученного корпуса текстов, подготовка которого для качественной классификации занимает большое количество времени.

Стоит отметить, что решения, приведенные выше, рассматривали задачу поиска аспектов и определения их тональности как две разные задачи, решаемые последовательно. Однако в реальности эти задачи взаимосвязаны и их можно рассматривать вместе. В работе [6] как раз использовался данный подход: авторы использовали SVM для определения пар «аспект-оценка».

Кроме классических методов машинного обучения популярно использование нейронных сетей. В работе [7], например, предложена модель, комбинирующая свёрточную и рекуррентную нейронные сети для извлечения аспектных терминов и выражений, описывающих тональность этих терминов. В работе [8] также использовался нейросетевой подход, но использовалась модель на основе систем переходов (transition-based). Таким образом, авторы извлекают выражения, в которых авторы отзывов высказывают свои мнения об аспектных терминах, а также связи между ними, тем самым предсказывая составные объекты.

Авторы статьи [9] представили гибридное решение для аспектно-ориентированного анализа настроений на уровне предложений с использованием онтологии лексикализованной области и модели регуляризованного нейронного внимания (ALDONAr). В данной модели сочетаются онтология настроений для захвата информации о настроениях, BERT для получения эмбеддингов слов и два слоя CNN для расширенной классификации тональности.

Таким образом, существует множество различных подходов к аспектному анализу тональности текстов. Выбор подхода зависит от задачи, которую необходимо решить, и от данных, которые необходимо проанализировать. Кроме того, можно воспользоваться существующими программными средствами, некоторые из которых будут рассмотрены далее.

Существующие программные решения

В связи с тем, что обработка текстов сейчас является актуальной задачей, на рынке программных продуктов представлено множество программных средств для обработки естественного языка. Рассмотрим некоторые из них.

В первую очередь нужно отметить наличие библиотек, позволяющих проектировать собственные решения. Например, одной из таких библиотек является Intel NLP Architect, в которой реализован распознаватель именованных сущностей, который позволяет выделять такие сущности, как имена, числа, места, валюты, даты и организации. Данная модель основана на двунаправленной LSTM сети и CRF классификаторе. После выделения аспектов также можно определить их тональность.

Компания IBM предлагает собственный сервис для анализа неструктурированного текста Watson Natural Language Understanding и позволяет извлекать сущности, ключевые слова, категории, эмоции, также позволяет определять тональность как конкретных сущностей, так и текста в целом.

Компания Amazon предлагает собственное решение Amazon Comprehend для обработки естественного языка, в котором для обнаружения информации в неструктурированных текстах применяются технологии машинного обучения. Данный сервис позволяет не только извлекать сущности и определять тональности, но и

систематизировать документы по темам, помечая их тегами на основе правил. Стоит отметить отдельный сервис Amazon Comprehend Medical, который идентифицирует и классифицирует медицинские данные (например, названия препаратов и диагнозов).

Компания Microsoft также предлагает решение для анализа текстов - Microsoft Language Understanding Intelligent Service (LUIS), которое включает в себя множество сервисов для решения задачи обработки и анализа текстов. Данный программный продукт позволяет распознавать именованные сущности, такие как имена людей, названия мест и организаций, дата и время, числа, а также более 100 типов личных сведений. Кроме того он позволяет определять язык текста, извлекать ключевые фразы и определять тональность. К удобству можно отнести то, что все решения разворачиваются в облачном сервисе и доступны через API.

Кроме представленных выше сервисов также можно отметить такие веб-сервисы, как MonkeyLearn и AYLIEN, которые предлагают инструменты для аспектного анализа и поддерживают несколько языков. Однако стоит отметить, что данные веб-сервисы ограничены в выборе предметных областей для анализа (отели, рестораны, авиакомпании и автомобили).

Заключение

В данной статье были исследованы различные подходы к аспектному анализу тональности текстов, а также рассмотрены существующие программные продукты для решения данной задачи. Проведенный анализ показал, что на современном этапе развития технологий обработки и анализа текстов существует множество решений, позволяющих решать задачу выделения аспектов и определения их тональности. В заключение стоит отметить, что выбор используемого метода или программного продукта для решения данной задачи зависит от целей задачи и данных, используемых для анализа.

Список литературы

1. Rain C. “Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning”, Swarthmore College, 2013.
2. Wu J., Ji T. “Deep Learning for Amazon Food Review Sentiment Analysis,” 2015.
3. Hu, M. Mining and Summarizing Customer Reviews / M. Hu, B. Liu // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — N.Y. : ACM, 2004. — P. 168—177.
4. Yu J. Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews / J. Yu, M. Zha Z.-J. and Wang, T.-S. Chua // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — Stroudsburg : ACL, 2011. — P. 1496—1505.
5. Проноза, Е.В., Ягунова, Е.В. Аспектный анализ отзывов о ресторанах для рекомендательных систем е-туризма // Компьютерная лингвистика и вычислительные онтологии: сборник научных статей. Труды XVIII объединенной конференции «Интернет и современное общество» (IMS-2015). НИУ ИТМО, 2015. p. 130-141.
6. Kobayashi N. Opinion Mining on the Web by Extracting Subject-Aspect-Evaluation Relations / N. Kobayashi [et al.] // Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. — Menlo Park : AAAI Press, 2006. — P. 86—91.

7. He R., Lee W.S., Ng H.T., Dahlmeier D. An Interactive Multi-Task Learning Network for End-to-End Aspect Based Sentiment Analysis // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, ACL, 2019, pp. 504–515.
8. Грибков Е.И., Ехлаков Ю. П. Нейросетевая модель на основе системы переходов для извлечения составных объектов и их атрибутов из текстов на естественном языке // Доклады ТУСУР. 2020. №1., стр.47-52.
9. Meskele D., Frasincar F. ALDONA: a hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalised domain ontology and a neural attention model // Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, 2019.