4. СОВРЕМЕННЫЕ МОБИЛЬНЫЕ И ИНТЕРНЕТ-ТЕХНОЛОГИИ

А.Н. Зухритдинов, Ю.В. Бидуля

Тюменский государственный университет, г. Тюмень

УДК 004.85

АСПЕКТНО-ОРИЕНТИРОВАННЫЙ АНАЛИЗ МНЕНИЙ С ПРИМЕНЕНИЕМ МЕТОДА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Аннотация. Целью работы является выявление факторов, влияющих на мнение пользователей о товарах и услугах путем автоматического анализа текста отзывов, опубликованных в сети Интернет. Предлагается исследование мнений пользователей с применением тематического моделирования методом LDA для получения так называемых аспектов с последующим ранжированием их в соответствии с позитивностью или негативностью мнений по отношению к этим аспектам.

Ключевые слова: Интернет-ресурсы, сентимент-анализ, аспектно-ориентированный анализ мнений, компьютерная лингвистика, мера взаимной информации (PMI), латентное размещение Дирихле (LDA).

Развитие Интернет-ресурсов стимулирует неуклонно возрастающий поток пользовательского контента, представленного в виде отзывов и мнений, публикуемых как в социальных сетях, так и в специализированных разделах Интернет-магазинов, а также на сайтах, посвященных обзорам о товарах и услугах. Анализ пользовательских оценок позволяет оперативно оценивать работу компании, выявлять

проблемные места в предоставлении товаров и услуг и принимать решения об изменениях в целях повышения спроса и увеличения прибыли.

Аспекты – это характеристики оцениваемых услуг или продуктов. Например, в сфере банковского обслуживания физических лиц для услуги «Дебетовые карты» слова «сотрудники», «обслуживание», «карта» представляют аспекты данной услуги, относительно которых может производиться качественная оценка, выраженная в тональных терминах. Автоматическое выделение аспектных слов является актуальной задачей, в решении которой заинтересованы поставщики товаров и услуг.

Для обнаружения аспектных слов используется метод тематического моделирования [1]. Вероятностная тематическая модель выявляет тематику набора документов, каждый документ представляется дискретным распределением вероятностей тем, а каждая тема — дискретным распределением вероятностей слов.

Цель данного исследования: разработать метод для обнаружения аспектных слов, ранжирования их по тональности контекста и классификации по обобщенным категориям. Материалом для исследования послужили отзывы пользователей банка о качестве обслуживания, собранные на сайте banki.ru (http://banki.ru). На этом же ресурсе ранее был апробирован подход к извлечению аспектов с помощью паттернов с последующим вычислением меры взаимной информации (PMI) по отношению к тональности отзывов, в результате чего определялись классы аспектов [2].

Гипотезой данного исследования является предположение, что каждая тема (топик) отзыва характеризует определенный аспект и состоит из соответствующих аспектных слов. Для выделения топиков здесь применяется метод латентного размещения Дирихле (LDA - Latent Dirichlet allocation) [3]. Подбор оптимальных характеристик тематической модели производился путем вычисления когерентности – меры, задающая

количественную оценку согласованности тем в документах по словам [4]. На первом этапе при помощи тематического моделирования извлекается список аспектных терминов, приписанных к определенному топику. На втором этапе для каждого аспектного термина определяется тональность при помощи вычисления меры взаимной информации [5]:

$$SCORE(W) = PMI(W,POS) - PMI(W,NEG) \quad (1)$$

Оценка PMI для слова в положительных отзывах вычисляется по формуле:

$$PMI(W,POS) = log_2\left(\frac{count(w,pos)*N}{count(w)*count(pos)}\right) \quad (2)$$

где N – общее количество слов во всех отзывах,

count(w,pos) – частота слова в окрестности положительного сентиментного слова,

count(w) – частота исследуемого слова во всех отзывах,

count(pos) – частота положительных слов во всех отзывах.

Величина *PMI(W,NEG)* вычисляется аналогичным образом.

Для реализации предложенного метода было разработано программное приложение, состоящее из следующих модулей.

1. Парсер отзывов о качестве банковского обслуживания, опубликованных на сайте banki.ru. Для извлечения текстов отзывов использовалась библиотека Python BeautifulSoup [5]. Всего было собрано 84345 отзывов по 438 банкам, снабженных оценкой пользователя от 1 до 5.
2. Модуль предобработки, включающий функции токенизации (библиотека Gensim [6]), удаления стоп-слов (NLTK.corpus [7]), лемматизации и POS-разметки (PyMorpy2 [8]).
3. Модуль тематического моделирования с использованием класса LdaMulticore библиотеки Gensim. Результатом работы модуля является список топиков, к каждому из которых приписан ряд характеризующих

слов. Оптимальное количество топиков было определено при помощи оценки когерентности модели и выбрано равным 10. Для дальнейшего рассмотрения использовались слова первого топика. На рисунке 1 продемонстрировано распределение слов по топикам. Данная визуализация показывает, насколько различны топики по своему составу, а также «мощность» каждого топика, выраженная в относительной частоте его слов.

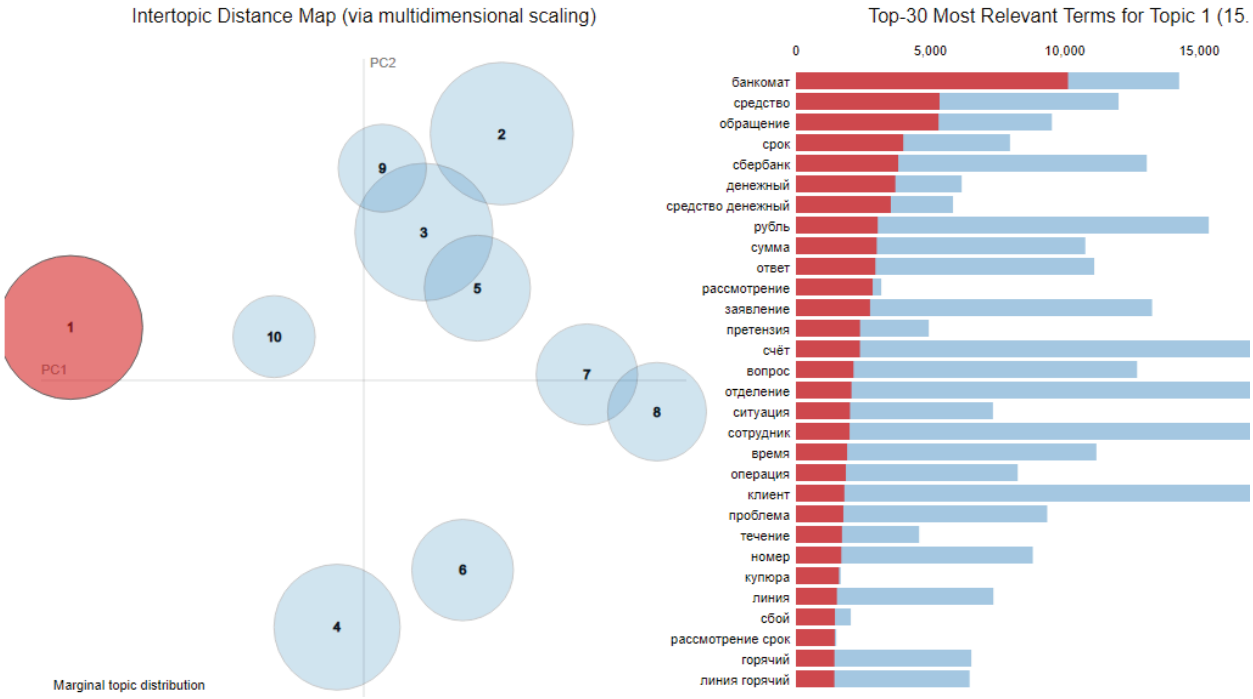

Рис. 1. Визуальное представление тематической модели.

4. Модуль оценки PMI, вычисляющий для слов из топиков меру взаимной информации и величину *score* по приведенным выше формулам (1) и (2). Если величина имеет отрицательное значение, то можно сделать вывод, что данное слово чаще встречалось в негативных отзывах. На рисунке 2 приведен пример вывода на диаграмме списка аспектных слов с отрицательной тональностью.

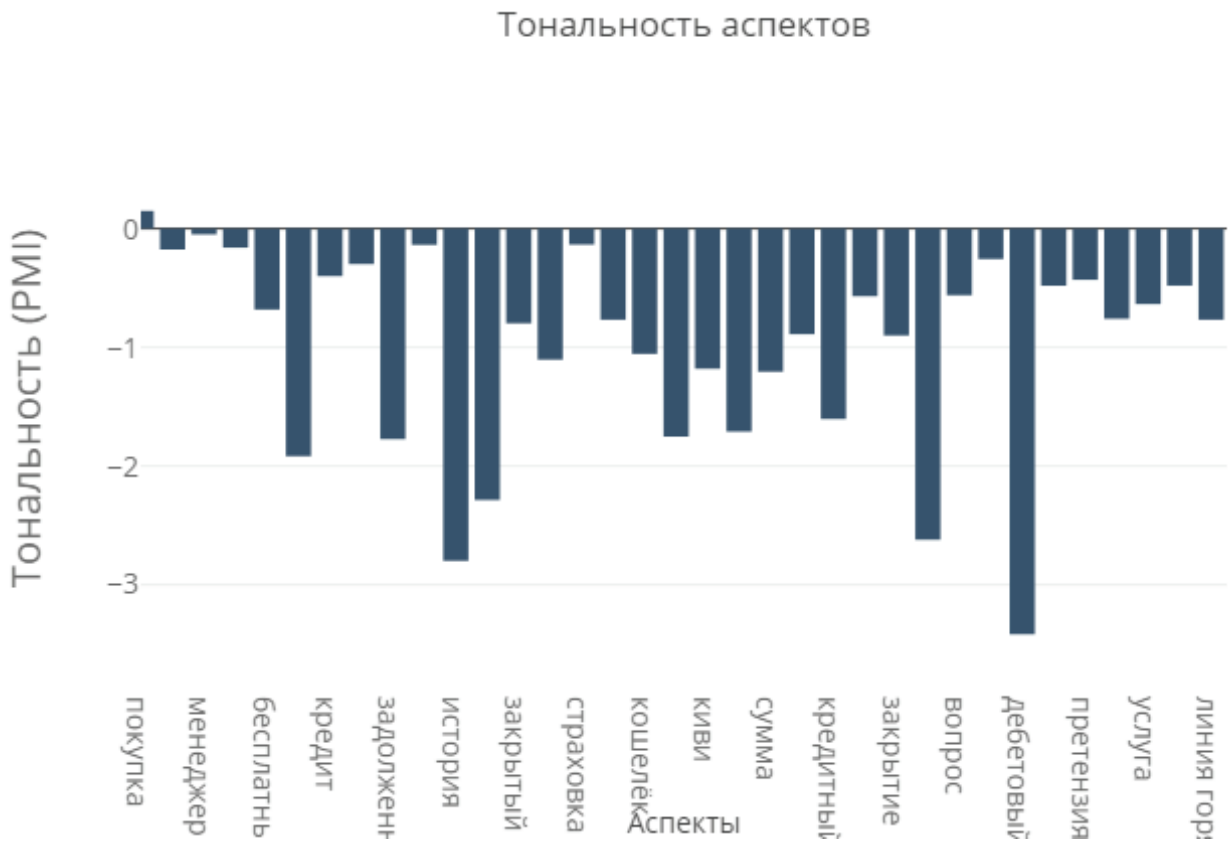

Рис. 2. Негативно окрашенные аспектные слова.

Разработанная методика позволяет в дальнейшем построить временные зависимости изменения тональности мнений по аспектам с целью определения наиболее проблемных факторов, влияющих на качество услуг.

СПИСОК ЛИТЕРАТУРЫ

1. Bagheri A., Saraee M., de Jong F. An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews, in Natural Language Processing and Information Systems. Springer, Berlin, Heidelberg, pp. 140–151, 2013.
2. Brunova, E., Bidulya, Yu.. Aspect Extraction and Sentiment Analysis in User Reviews in Russian about Bank Service Quality (2019) 11th IEEE International Conference on Application of Information and Communication

Technologies, AICT 2017 – Proceedings 10 April 2019, art. no. 8687070. ISBN: 978-153860501-1. DOI: 10.1109/ICAICT.2017.8687070.

3. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — Vol. 3. — Pp. 993–1022.
4. Keith Stevens, Philip Kegelmeyer. Exploring Topic Coherence over many models and many topics. – 2012. URL: https://www.aclweb.org/anthology/D12-1087.pdf (дата обращения 21.05.2020).
5. BeautifulSoup. URL: https://pypi.org/project/beautifulsoup4/ (дата обращения 21.05.2020).
6. Gensim: Topic modelling for humans. URL: https://radimrehurek.com/gensim/ (дата обращения 21.05.2020).
7. Natural Language Toolkit NLTK. URL: https://www.nltk.org/ (дата обращения 21.05.2020).
8. Морфологический анализатор pymorphy2. URL: https://pymorphy2.readthedocs.io/en/latest/ (дата обращения 21.05.2020).