Использование нечетких онтологий в задаче анализа отзывов пользователей

Н. С. Кожевникова, Е. Ю. Данилова
ФГБОУ ПГНИУ «Пермский государственный университет»
kozhevnikovans@gmail.com, ket-eref@yandex.ru

***Аннотация*. Одним из ключевых параметров в выборе маркетинговой политики для бизнеса является мнение конечных пользователей. С ростом популярности социальных медиа, растет количество разнообразных отзывов в этих сервисах. Пользователи пишут свои впечатления о сервисах и товарах в режиме реального времени, и контроль за упоминаниями и своевременное реагирование стали нормой современного маркетинга. При этом задача анализа отзывов пользователей в социальных медиа связана с такими проблемами как выделение наиболее влиятельных мнений, оценка позитивности отзыва, выделение из отзывов ключевых аспектов и их группировка в едином отчете. В своей работе мы предлагаем гибридный подход, основанный на лексико-синтаксических шаблонах и нечетких онтологиях, к задаче анализа отзывов пользователей. При этом в рамках исследования особое внимание уделяется анализу метрик социальных медиа, как способу измерения качества и достоверности отзыва, а также степени его влияния.**

Ключевые слова: нечеткая онтология; анализ текста; естественный язык; анализ отзывов

I. Введение

В современном мире социальные медиа являются важным источником информации. Социальные сети, микроблоги, фото и видео блоги, вопросно-ответные сервисы накопили и продолжают накапливать большие объемы данных. Пользователи обмениваются мнениями, оставляют отзывы, передают знания. При этом как бизнес, так и обычные пользователи используют эту информацию для принятия решений. Например, потенциальный покупатель квартиры в новостройке собирает информацию из отзывов об интересующем его жилом комплексе, застройщике, районе и т.д. Строительная компания собирает отзывы о своих строительных объектах для анализа качества работы подрядчиков.

Автоматизация агрегирования и анализа отзывов пользователей усложнена тем, что большая часть информации представлена в виде текста на естественном языке, например, сообщений пользователей или диалогов в комментариях. Необходимо также отметить, что важную роль при оценке этих материалов играет сама социальная составляющая. Так, при взаимодействии с публикациями в социальных медиа пользователи могут выразить поддержку отметкой “Нравится” (лайки), сохранением (репосты) или комментариями, а также используя упоминания и хэштеги. Пользователи могут, вместо написания похожего отзыва, поставить "Нравится" уже существующему. Комментарием к сохраненным материалам, они акцентируют внимание на том, что вызвало наибольший отклик. Тексты связанные хештегами оказываются объединены одной идеей, несмотря на разрозненность их публикации. Поэтому актуальной является задача автоматизации извлечения знаний из текстов в социальных медиа, с учетом их специфики.

II. Обзор

Существует ряд коммерческих продуктов, которые ставят своей целью агрегацию данных из социальных сетей. Например, Klear [10], Brandwatch [4], Traackr [15] – конкурирующие продукты, предоставляющие услуги агрегирования данных для брендов, агентств и предприятий. Однако они в основном концентрируются на улучшении продвижения в социальных сетях. При этом для анализа текста используется кластеризация по ключевым словам и темам, без детализации. Учитываются метрики социальных сетей для выявления наиболее влиятельных отзывов. Однако анализ текста выявленных отзывов, оценка того, какие фразы в тексте вызвали наибольший резонанс, по-прежнему, полностью осуществляется человеком. Еще два похожих продукта Mentionmapp [11] и Tweetreach [16], предоставляют сервис только для Twitter. При этом данные сервисы также концентрируются на взаимосвязи твитов через хештеги, темы и ретвиты, но не позволяют разобрать текст более детально. Важно отметить, что все представленные инструменты не поддерживают русский язык, что не позволяет анализировать русскоязычный контент.

Научное сообщество также активно занимается задачей выделения информации из текстов в социальных сетях. При этом задача извлечения мнений входит в класс задач извлечения информации. Соответственно подходы, использующиеся для этих целей зачастую те же, а именно лингвистический анализ текста, методы машинного обучения [13]. Однако выделение и суммирование отзывов пользователей ставит перед исследователями новые задачи, например, выделение в тексте фактов, отражающих мнение, оценка положительного/отрицательного отношения и его степени, проблема анализа текстов сравнительно малого объёма и т.д. [14]. Обнаруженные нами исследования в основном предлагают четкое разделение отзывов на положительные и отрицательные,

при этом в анализе используются предварительно заданные шаблоны, спецификаторы, что не позволяет корректировать работу программы, в зависимости от специфики задачи. Так одной из тематически близких работ является исследование Hu M. и Liu B. [12], предложившими трехэтапный подход к анализу и оценке отзывов, основанный на анализе частотности фраз. А также представляет интерес метод Opizer-E [5], который выделяет и группирует основные аспекты отзыва, основываясь на его позиции в тексте и близости к заданным спецификаторам. В [8] авторы предложили дополненную версию данного метода. Однако, отсутствие контекста у отзыва и его сравнительно малый объём влияет на качество анализа нестандартных отзывов, таких как сравнение двух продуктов. Также Agichtein и соавторы в [7] отмечают, что социальное ранжирование являются важным фактором в оценке качества источника. Однако мы заметили, что работы, посвященные извлечению и последующему анализу отзывов в социальных медиа не используют эти данные. Поэтому мы предлагаем гибридный подход, основанный на нечетких онтологиях и лингвистическом анализе. Это позволит с одной стороны группировать факты, основываясь на уже существующих знаниях, а с другой учитывать социальный фидбек, благодаря коэффициентам доверия, вычисляемым из значений социальных метрик. При этом в своей работе мы делаем акцент на анализе именно русскоязычного контента.

III. Описание предлагаемого подхода

В основе предлагаемого нами подхода лежит выделение понятий из текстов в социальных сетях с помощью контекстно-свободных грамматик (КСГ) и их представление в виде онтологии. С помощью лексико-синтаксических шаблонов Томита-парсера [1] реализовано извлечение различных структур текста (ключевых слов и словосочетаний): аббревиатур, имен собственных, фраз в кавычках. Парсер включает в себя три стандартных лингвистических процессора: токенизатор (разбиение на слова), сегментатор (разбиение на предложения) и морфологический анализатор (mystem). Основные компоненты парсера: газеттир, набор контекстно-свободных грамматик и множество описаний типов фактов, которые порождаются этими грамматиками в результате процедуры интерпретации. Факты извлекаются с помощью структур фактов Томита-парсера. Возможно пополнение системы новыми паттернами.

Для учёта особенностей предметных областей в системе предусмотрено разделение паттернов и их модульное подключение в зависимости от обрабатываемого текста. Реализация разделения паттернов основана на газеттирах Томати-парсера. Извлечённые ключевые фразы сохраняются в базу данных. Разделение текстов по темам выполняется с помощью классификатора от проекта Zamgi [17]. Данный алгоритм классифицирует текст, относя его к одному из 13 классов.

Извлечённые из текста ключевые слова и словосочетания представляются в виде онтологии отзыва, отображаемой в дальнейшем на внешнюю предметную онтологию – открытый проект Wikidata [18]. Для отображения используется API Wikidata, предоставляющее, в том числе, методы поиска сущностей по запросам, в качестве которых служат извлеченные из текста ключевые слова и словосочетания. Использование онтологии Wikidata позволяет учитывать даже те знания, которые подразумеваются человеком, но не включены в конкретный отзыв. Отображение на внешнюю онтологию оставляет возможность отображения онтологий отзывов между собой, позволяя тем самым находить схожие мнения. Также, результат отображения позволяет пополнять газеттиры Томита-парсера, содержащие паттерны, синонимами, повышая качество извлекаемых в дальнейшем ключевых слов и словосочетаний.

Поскольку тексты отзывов могут быть противоречивы, как мнения разных людей, эти онтологии должны отражать нечёткость знаний. Существует ряд подходов к созданию нечеткой онтологии и в основном они отличаются тем, какой вид нечеткости используется [6]. Для описания нечеткой онтологии будем использовать язык OWL 2, по аналогии с предложенным Bobillo и Straccia [3] подходом. Устанавливать выявленным фактам коэффициенты доверия будем, используя атрибуты отношений. Подробно методика создания нечеткой онтологии описана в [2].

Для вычисления коэффициентов доверия мы предлагаем опереться на социальное ранжирование. Так метрика “коэффициент вовлеченности”

$$\frac{likes + saves + comments}{followers}$$

отражает степень популярности мнения пользователя, и, следовательно, степень доверия фактам в тексте коррелирует со степенью реакции на них. При этом анализ комментариев к посту и сохранениям, позволит скорректировать оценку в меньшую сторону, если в них отрицаются факты из основной публикации. Корректировку значений коэффициентов доверия будем осуществлять по формуле, предложенной в [9]:

$$f = f + \frac{f_{new} - f}{Q + 1}, \qquad (1)$$

где Q-количество предшествующих изменений коэффициента.

В результате обработки будет получена онтология со взвешенными фактами по заданному запросу. Дальнейшая работа с такой онтологией может идти в нескольких направлениях: визуализация выделенной информации, выгрузка выделенных фактов в базу данных и использование традиционных методов DataMining, использование нечеткой онтологии как экспертной системы, для проверки бизнес-гипотез и т.д.

IV. Пример использования

Рассмотрим пример представления знаний извлеченных из текстов социальных медиа с помощью описанного подхода. Проведем анализ отзыва об объекте недвижимости: "Нам этот жк нравится. Красивое здание с качественным интерьером. С представителями застройщика общаться приятно, все вежливо и по делу. Вчера смотрели стройку, делают дороги, проложен новый асфальт." Допустим, у данного поста 135 отметок "Нравится", и 1 репост. Пост написан в группе "ЖК Ива лучшее жилье" с 500 подписчиками.

На первом этапе будут выделены ключевые слова и словосочетания. При этом концепт "ЖК" будет соотнесен с Wikidata и мы сможем найти похожие понятия: "жидко-кристаллический" и "жилой комплекс". При анализе темы поста текст будет соотнесен с тематикой недвижимости. В Wikidata концепт "ЖК" имеет атрибут "Категория" со значением "Жилищно-коммунальное хозяйство". Этот концепт в свою очередь является подклассом категории "Недвижимость". Так, благодаря кластеризации по темам и связи с метаонтологией Wikidata будут соотнесены даже те понятия, значение которых может трактоваться неоднозначно.

В процессе анализа мы получаем также контекст, доступный благодаря ссылкам и упоминаниям в социальных медиа. Так, поскольку пост находится в группе с названием "ЖК Ива лучшее жилье" мы можем конкретизировать о каком ЖК речь, так как "ЖК Ива" — это экземпляр "ЖК". После анализа ключевых слов с помощью знаний из Wikidata, происходит формирование газеттиров и дальнейший анализ текста с помощью КСГ. Тогда в результате анализа будут выделены следующие факты: жк нравится, красивое здание, качественный интерьер, представитель застройщика общается приятно, представитель застройщика общается вежливо, представитель застройщика общается по делу, дороги делают, проложен новый асфальт.

Далее происходит оценка коэффициента доверия из собранных метрик. Получим доверие фактам этого отзыва равным 0,272. Далее, анализируя репост мы отмечаем наличие комментария. Например, "не согласен, красивым здание уж точно не является" с одним лайком. Выявив отрицание, мы понимаем, что необходим пересчёт доверия к факту "красивое здание". Оценив количество лайков и комментариев к этому сохранению и количество друзей автора получим доверие к этому отрицанию. Например, получим 0,002. При этом исходное значение будет пересмотрено на 0,27. Тогда по формуле 1 это значение станет равным 0,271. Допустим, что в другом отзыве также встретился факт "красивое здание" с коэффициентом доверия 0,45. Тогда после отображения этих понятий пересчитанный для этого факта коэффициент составит 0,331. В результате мы получим два противоречащих факта, но с разным уровнем доверия. Это позволит не откидывать при дальнейшем анализе даже минимальные противоречия, которые, однако, могут играть важную роль при выборе маркетинговой политики. При этом, за счет использования нечеткой логики, сама онтология не становится противоречивой.

V. Заключение

В статье мы рассмотрели существующие подходы к автоматическому извлечению отзывов пользователей из социальных сетей. Были отмечены недостатки существующих подходов, такие как сложность анализа нестандартных отзывов, опирающихся на знания, не представленные в тексте явно. Также минусами являются бинарная оценка положительности отзыва и недостаточный учет социальных метрик, связанных с отзывом. Для нивелирования отмеченных недостатков мы предложили подход на основе нечетких онтологий, учитывающий значения метрик социальных сетей для оценки коэффициентов доверия.

Важно отметить, что предложенный подход может быть расширен на другие задачи извлечения данных, помимо анализа отзывов. Кроме того, онтологический подход позволяет предоставить пользователю возможность настраивать систему, для получения более точных результатов, путем расширения онтологии Wikidata, дополнением Wikipedia. Так данный подход позволяет с одной стороны воспользоваться возможностями краудсорсинга семантики, а с другой избежать "холодного старта", характерного для подобных систем.

Дальнейшие исследования предполагается посвятить оценке качества получаемых коэффициентов доверия, а также развитию инструментов анализа полученной нечеткой онтологии.

Список литературы

[1] Яндекс. Томита-парсер — Технологии Яндекса [Электронный ресурс]// URL: https://tech.yandex.ru/tomita (дата обращения 04.04.2018)

[2] Alexopoulos, P., Wallace, M., Kafentzis, K., Askounis, D. IKARUS-Onto: a methodology to develop fuzzy ontologies from crisp ones. // Knowledge and Information Systems. 2012 Т.2, вып. 3. С. 667-695

[3] Bobillo, F., Straccia, U. Fuzzy ontology representation using OWL 2. // International Journal of Approximate Reasoning. 2011. Т. 52, вып. 7. С. 1073-1094.

[4] Brandwatch: Home [Электронный ресурс]// URL: https://www.brandwatch.com/ (дата обращения 04.04.2018)

[5] Condori, R. E. L., Pardo, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. // Expert Systems with Applications.2017. Т. 78. С. 124-134.

[6] El-Sappagh, S., Elmog., M. A fuzzy ontology modeling for case base knowledge in diabetes mellitus domain. // Engineering Science and Technology, an International Journal. 2017. Т. 20. С. 1025–1040

[7] Finding high-quality content in social media. /Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G. // Proceedings of the 2008 international conference on web search and data mining. 2008, February. ACM. С. 183-194.

[8] Improving Opinion Summarization by Assessing Sentence Importance in On-line Reviews. /Anchiêta, R., Sousa, R. F., Moura, R., Pardo, T. // Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology. 2017. С.32-36.

[9] Integrating Fuzzy Logic in Ontologies. / Calegari, S., Ciucci, D. //. ICEIS 2006 - 8th International Conference on Enterprise Information Systems, Proceedings. 2006. С. 66-73.

[10] Klear: Influencer Marketing Software [Электронный ресурс]// URL:https://klear.com/ (дата обращения 04.04.2018)

[11] Mentionmapp [Электронный ресурс]// URL: http://mentionmapp.com/ (дата обращения 04.04.2018)

[12] Mining and summarizing customer reviews / Hu, M., Liu, B. // Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining. 2004. С. 168-177.

[13] Paltoglou, G., Giachanou, A. Opinion retrieval: Searching for opinions in social media. // Professional Search in the Modern World. Lecture Notes in Computer Science. 2014. Т. 8830. С. 193-214.

[14] Pang, B., Lee, L. Opinion mining and sentiment analysis. // Foundations and Trends in Information Retrieval. 2008. Т.2, вып. 1-2. С. 1-135.

[15] Traackr: Influencer Marketing that matters [Электронный ресурс]// URL: http://www.traackr.com/ (дата обращения 04.04.2018)

[16] TweetReach: How Far Did Your Tweets Travel? [Электронный ресурс]// URL: https://tweetreach.com/ (дата обращения 04.04.2018)

[17] Zamgi: Автоклассификация текста на русском языке [Электронный ресурс]// URL: https://github.com/zamgi/lingvo--classify (дата обращения 04.04.2018)

[18] Wikidata [Электронный ресурс]// URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (дата обращения 04.04.2018)Klear: Influencer Marketing Software [Электронный ресурс]// URL:https://klear.com/ (дата обращения 04.04.2018)