УДК 519.876

РАЗРАБОТКА ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ОТЗЫВОВ В БИЗНЕС СФЕРЕ

[1]Казаков Н. И., [2]Курчеева Г. И.
[1]Новосибирский государственный технический университет, Новосибирск, Россия, kazakoff2014@yandex.ru
[2]Новосибирский государственный технический университет, Новосибирск, Россия, kurcheeva@yandex.ru

Введение. Поток текстовой информации, непрерывно проходящий через сеть Интернет, достаточно труден для восприятия людьми. В первую очередь это связанно с большим объемом данных, которые человек вручную обработать не в состоянии из-за ограниченной способности читать, выделять необходимые знания и сортировать информацию. В связи с этим в различных компаниях становится все более востребованным использование каких-либо автоматизированных способов обработки и фильтрации текста, позволяющие упростить и значительно ускорить процесс анализа.

Интеллектуальный анализ текста объединяет методы компиляции и организации большого количества документов разного объема для получения новой информации, и выявления новых связей между текстами одной тематики. Помимо использования статистических методов, анализ текста требует использования знаний компьютерной лингвистики, а также может использовать технологии искусственного интеллекта и обработки естественного языка (NLP) [1].

Возможные задачи для интеллектуального анализа данных могут быть достаточно разнообразны: извлечение неявных данных, визуализация информации и закономерностей, оценка больших текстовых данных, сентимент анализ, структурирование научных публикаций. Помимо этого, перспективный потенциал имеется в области управления персоналом и взаимоотношениями с потребителями, а также для получения информации о действиях конкурентов [3].

В ходе исследования, проводимого в рамках магистерской диссертации, разработана информационная подсистема, функционал которой заключается в сборе текстовых данных из сети Интернет и их анализе. Для анализа собранных данных используется библиотека, разработанная с учетом семантических особенностей русского

языка, и обученная на достаточно большой обучающей выборке, включающей в себя сленг и ненормативную лексику.

Анализ текста и анализ данных

Интеллектуальный анализ текста имеет весомые отличия от анализа данных, хоть и использует некоторые его приемы и методы. Основная разница между ними заключается в типе обрабатываемых данных. Данные для «традиционного» анализа данных находятся в структурированном виде, соответствуя при этом первой нормальной форме (в терминологии реляционных баз данных). Поля БД содержат в себе атомарные значения, которые не поддаются дальнейшему разделению. Если же рассматривать интеллектуальный анализ текста, то данные в нем имеют неявную структуру, вытекающую грамматики текста или структуры самого документа, состоящего из абзацев и заголовков. Несмотря на это, данные для анализа текста принято считать неструктурированными.

Интеллектуальный анализ данных применяется уже к структурированным данным, применяя методы статистики, ИИ и машинного обучения. Помимо этого, используются и следующие функции моделирования данных:

− Кластеризация – организация данных, группировка по схожим признакам, получение новых фактов об этих данных;

− Классификация – показывает шаблоны для прогнозирования класса, к которым относятся данные;

− Ассоциации – позволяет определить вероятность исхода в зависимости от случая в течении времени;

− Регрессия – предсказание значения в зависимости от переменных в наборе данных.

В отличии от анализа данных, интеллектуальный анализ текста требует применения еще одного шага для достижения аналогичного результата. Для применения функций моделирования данных и дальнейшей аналитики, текстовые данные необходимо привести в структурированный вид. Это, в свою очередь, требует применения сложных лингвистических и статистических методов, и иногда применения технологий ИИ и NLP, что позволяет системе «понимать» человеческий язык.

При анализе и структурировании документов, они насыщаются метаданными, такими как дата, краткое содержание, ключевые слова, автор, и т. д. После наполнения данных метаданными, они могут быть переведены в воспринимаемый машинами формат и подвергнуты анализу.

Анализ текстовых данных можно разделить на две основные категории: статистический анализ, подразумевающий под собой сбор информации о частоте встречаемых слов и конструкций, и лингвистический анализ (анализ естественного языка), состоящий из следующих методов:

− Морфологический анализ. В ходе данного анализа происходит уменьшение объема исходного текста для проведения дальнейшего анализа;

− Семантический анализ – выявляет контекстные знания, позволяющие разбить текст на отдельные смысловые единицы;

− Синтаксический анализ – выделение отдельных блоков текста или предложений.

Синтаксический анализ заключается в аннотировании выделении определенных фрагментов – маркеров, для отдельных частей текста. Для каждой части речи ставится свой отдельный тег, а также учитывается положение слова в предложении, что позволяет выделить в нем объекты, субъекты и предикаты. Данный метод имеет особое преимущество перед вышеперечисленными, так как позволяет целенаправленно собирать информацию из синтаксических единиц.

Процесс интеллектуального анализа текста

Если рассматривать анализ текста как процесс в целом, в нем можно выделить следующие основные этапы [1]:

1. Определение задачи – определение проблемы и целей, для которых будет проводиться анализ текста;

2. Выбор исследуемых документов. В зависимости от выбранных целей анализа, выбирается необходимый набор документов, который может состоять из текста разных стилей и жанров. Также, при использовании в анализе ИИ или обучаемых машин, документы могут использоваться в качестве обучающей выборки;

3. Обработка документов. Как было сказано ранее, анализ текста требует предварительного структурирования данных. На этапе обработки документов производится извлечение признаков и терминов, которые в дальнейшем будут служить представлением документа. Состоять такой термин может как из одного отдельно взятого слова, так и из нескольких слов, несущих конкретное значение. Для извлечения терминов также могут использоваться и методы обработки естественного языка. Наиболее распространенная моделью представления документов в виде терминов основана на векторах в n-мерном векторном пространстве. Размер данного пространства

соответствует словарю коллекции документов, и может быть графически представлен в виде матрицы.

Таблица 1 – Пример матрицы

	Термин 1	Термин 2	Термин 3
Документ 1	…	…	…
Документ 2	…	…	…
Документ 3	…	…	…

Запись в ячейке может нести необходимую для анализа информацию, такую как частота встречаемого слова, взвешенная частота, показывающая значение термина для данного документа, или же вовсе содержать в себе двоичное значение, показывающее наличие термина в нем;

4. Анализ текста. После того, как текст принял определенную структуру и из него были извлечены термы, к данным можно применять методы классического интеллектуального анализа данных (классификация, сегментация и пр.);

5. Оценка результатов;

6. Применение результатов.

Постановка задачи

Целью данной статьи является разработка информационной подсистемы сбора и обработки текстовых отзывов из социальных сетей о каком-либо бренде для дальнейшего проведения семантического анализа. В качестве интернет-ресурса для сбора отзывов выбрана социальная сеть Вконтакте.

Объектом исследования выступает компания S7 Airlines. Оценка отзывов производится по бинарной шкале, разделяя на их на «положительные» и «отрицательные».

В ходе данной работы будут выполнены следующие этапы:

1) Разработка общего алгоритма сбора текстовых отзывов из социальной сети Вконтакте;

2) Разработка общего алгоритма сбора текстовых отзывов из социальной сети Вконтакте

3) Реализация разработанного алгоритма на языке Python;

4) Тестирование реализованной технологии на выборке из собранных отзывов;

5) Проанализировать собранную информацию;

6) Оценить работу алгоритма.

Анализ инструментов сбора текстовой информации

На данный момент существует ряд популярных и хорошо проработанных готовых библиотек для сбора текстовой информации с web-страниц – так называемых веб-краулеров. Но несмотря на их удобство и широкий функционал, работать в рамках социальных сетей с ними не представляется возможным. Большинству краулеров необходимо задать начальный список URL-адресов для дальнейшего сбора информации с указанных web-страниц.

Социальная сеть Вконтакте имеет достаточно функциональное API, что позволяет искать и собирать отзывы напрямую с сайта, без указания URL-адресов, поэтому для проведения данного исследования будет разработан собственный краулер.

Алгоритм сбора и анализа текстовых отзывов

Обобщенный алгоритм сбора и анализа отзывов можно представить следущим образом:

1) Формирование перечня слов, тематически связанных с объектом исследования;

2) Поиск «постов», содержащих слова из составленного перечня;

3) Добавление текста, содержащегося в «посте» в базу данных;

4) Очистка текста от «мусора», разбиение на коллекции;

5) Помещение коллекций в базу данных;

6) Обработка коллекций классификатором тональности;

Разработка парсер-модуля

После определения перечня слов для поиска интересующей нас информации, разрабатывался парсер-модуль для сбора отзывов из социальной сети Вконтакте.

Данный модуль производит поиск интересующей нас информации среди публикаций сообществ и пользователей, а также комментариев под постами, после чего заносит полученные отзывы в базу данных.

На рисунке 1 представлен пример того, как выглядит отзыв в интерфейсе сайта Вконтакте, на рисунке 2 – отзыв, обработанный парсером.

Помимо самого отзыва, парсер сохраняет дату и URL отзыва. После того, как все отзывы сохранены в базу данных, парсер объединяет их в один текстовый объект и проводит подготовительную обработку. Для дальнейшего анализа полученного текста из него удаляются стоп-слова, знаки препинания. Помимо этого текст подвергается приведению к общему регистру и лемматизации (приведение всех слов к нормальной форме для составления словаря)

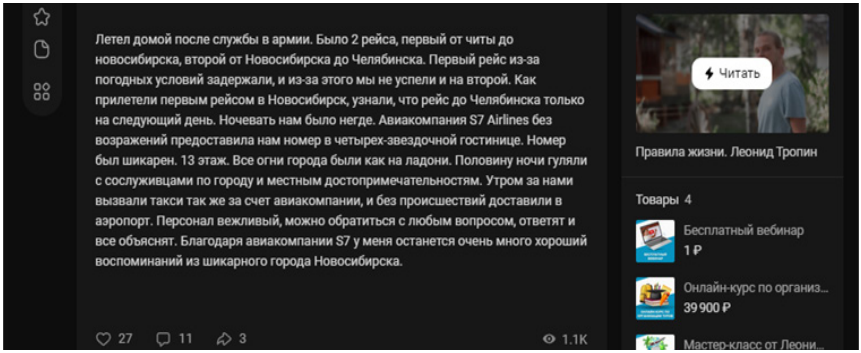

Рисунок 1 – Пример отзыва с сайта Вконтакте

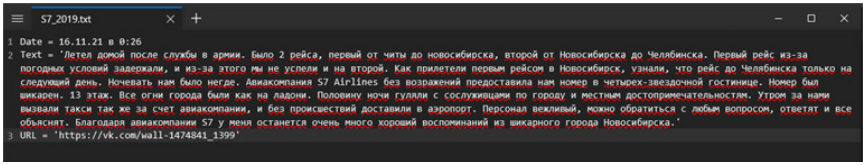

Рисунок 2 – Результат работы парсера

Модели и методы семантического анализа

В качестве семантического тонового анализатора использовалась библиотека языка Python – Dostoevsky, основанная на моделях BoW и n-gram и линейном классификаторе.

Линейный классификатор

Пусть вектор $\vec{x}$ (нормализованный вектор из частот слов в документе) представляет собой входные данные, а на выходе классификатора вычисляется показатель по формуле:

$$y = f(\vec{w} \cdot \overrightarrow{x}) = f\left(\sum_i w_j x_j\right) \quad (1)$$

$\vec{w}$ – действительный вектор весов той же размерности, что и признаковое пространство (рначения весов вектора определяются в ходе обучения на тестовых выборках), f – функция преобразования скалярного произведения.

Допустим, что $y^*: X \to Y$ является целевой зависимостью, известной только на объектах обучающей выборки $X^l = (x_i, y_j)_{i=1}^l, y_i = y^*(x_i)$, и требуется найти вектор весов w, при котором алгоритм $a(x, w)$ аппроксимирует целевую зависимость $y^*(x_i)$.

Данная задача является ни чем иным, как поиском вектора w и доставляющего минимум функционала:

$$Q(w) = \sum_{i=1} L(a(x_i, w), y_i) \to \min_w \tag{2}$$

L(a, y) – функция потерь.

Для минимизации Q(*w*) методом градиентного спуска выбирается некоторое начальное приближение для вектора весов *w*, после чего в цикле, с каждым шагом вектор *w* изменяется в направлении наиболее быстрого убывания функционала Q (противоположно вектору градиента).

$$\begin{gathered} \nabla Q(w) = \left(\frac{\partial Q(w)}{\partial w_j}\right)_{j=1}^{n}; \; w := w - \eta \nabla Q(w) \\ w := w - \eta \sum_{i=1} L'_a(a(x_i, w), y_i) f'(\langle w, x_i \rangle) \, x_i \end{gathered} \tag{3}$$

где $\eta > 0$ – величина шага в направлении антиградиента (темп обучения).

Каждый прецедент (x_i, y_i) вносит аддитивный вклад в изменение вектора *w*, но сам он изменяется только после перебора всех l объектов.

Для повышения скорости сходимости данного процесса прецеденты выбираются случайным образом (x_i, y_i) , для каждого делается градиентный шаг и сразу обновляется вектор весов:

Их инициализация производится подбором небольших случайные значений.

Метод стохастического градиента хорошо приспособлен для динамического обучения, и позволяет настраивать веса больших выборках. Это происходит за счет случайной подвыборки, которой может быть достаточно для обучения. Допускаются различные стратегии обучения. Для большой выборки

допускается вовсе не сохранять объекты обучения, а на малых – повторно использовать их.

BoW (Bag of Word)

Построение модели состоит из: создания словаря очищенных от мусора «слов» и определения вектора документа. Данная компонента является ни чем иным, как количество раз, которое слово используется в документе. Размерность же вектора соответствует мощности составленного «чистого» словаря.

С помощью данной модели можно перейти к такому представлению документа, что: слово $w_i \in V$ словаря V в документе d_i имеет количество вхождений равное n_i, и любой документ может быть представлен в виде:

$$\overline{d_1} = n_1(w_1) + n_t(w_t) + \cdots + n_m(w_m) \tag{4}$$

где m – количество слов в документе d_i.

Данный подход прост в реализации, но имеет ряд существенных недостатков. Одним из них является неверное определение настроения текста. Для примера предложение «питание не очень вкусное» будет отражать негативную оценку. Если же представить данной отзыв в виде словаря, окрас его станет положительным. Решить данную проблему неточного определения семантического окраса предложения позволяет использование метода n -gram.

Метод n-gram

Представляет из себя математическую модель представления текстов в виде набора последовательностей, состоящую из N слов.

Модель n-gram бывает нескольких видов: униграммы $P(w_i)$, биграммы $P(w_i|w_{i-1})$, триграммы $P(w_i|w_{i-2}, w_{i-1})$

Таким образом применение метода n-gram сводится к определению вероятности появления цепочки слов в тексте $P = (w_1, w_2, \dots, w_t)$.

$$P = (w_1, w_2, \dots, w_t) = \prod_{i=1}^{t} P(w_i|w_1, w_2, \dots, w_{i-1})$$
$$P = (w_1, w_2, \dots, w_t) = \frac{C(w_{i-n+1}, \dots, w_i}{C(w_{i-n+1}, \dots, w_{i-1})} \tag{5}$$

где C – количество появлений последовательности слов в обучающем корпусе.

Мера TF-IDF

Для того, чтобы избежать увеличения количества употребления одних и тех же слов в документах большего объема подобного несоответствия используется TF-IDF мера.

Использование данной меры позволяет повысить значимость слова для конкретного документа, если оно не встречается в других документах той же тематики [10].

Частота термина-обратная частота документа *tf-idf* – это статистическая мера, используемая для оценки важности слова в контексте документа (Term Frequency-Inverse Document Frequency)

$$tf - idf = tf_{j,i} \ln\left(\frac{N}{df_i}\right) \quad (6)$$

$tf_{j,i}$ – отношение количества вхождений слова к общему числу терминов документа, df_i – число документов из коллекции, в которых встречается слово, N – число документов в коллекции.

Итоги работы классификатора

С использованием тонового классификатора (библиотеки Dostoevsky) был проведен анализ отзывов пользователей сайта Вконтакте, пользовавшихся услугами компании S7 Airlines.

Для анализа выбирались отзывы пользователей, оставленные с 2018 по 2020 года. В результате работы парсера была сформирована база данных, содержащая в себе 2720 отзывов реальных пользователей.

В результате анализа отзывы были поделены на две выборки: 1606 положительных отзывов, и 1114 отрицательных.

Таблица 2 – Фрагменты классифицированных отзывов на положительные и отрицательные

Направление	Класс обслуживания	Положительный отзыв	Отрицательный отзыв
Россия	Бизнес-класс	Возвращалась после двухнедельного отдыха бизнесс-классом из Ларнака в Москву. Осталась очень довольна качеством рабоы персонала и обслуживанием. Так держать S7!!! Вы молодцы))	Задержки рейсов стали частыми на от 2 до 3 часов. Летели как из Симферополя так и из Москвы. Суть уже изложена, добавить ничего не хочу.

Продолжение таблицы 2

Россия	Эконом-класс	Вчера летели рейсом 3504 в 16.35 из Бургаса. Хочу выразить огромную благодарность всей команде стюардов и обслуживающему персоналу и в Бургасе и в Москве. Все очень отзывчивы и доброжелательны! Летела после травмы позвоночника и я благодарна за их отношение и обслуживание. Низкий поклон.	Летала этой компанией трижды и два раза из них рейсы переносились, из-за чего один раз пришлось переносить стыковочный рейс, а второй раз просидеть в аэропорту 5 часов. Летели из Кишинева в сторону Москвы.
Между-народный	Эконом-класс	Летали туда и обратно на отдых в Хургаду вашей компанией. Туда летели на новом Аэробусе 320 нео. Прямо классный., удобный, на каждом кресле зарядное устройство , на спинах держатели для гаджетов. Кормили вкусно, персонал отличный!!! Взлет и посадка хорошая. Особенно понравился взлет и приземление!! Очень хочется объявить благодарность за это командиру лайнера!! Это было 17 11 рейс 6623	Купили билеты 3шт на 23 июня, прямой рейс в Баку, отменили рейс, пока возврат не оформили попросили перенаправить пересадками другими рейсами, ведь подготовились, справки ковид примерно 9т обошлись, терять деньги не хотели, короче нам ответили что самолеты этой авиакомпании летать не будут в Баку, а в другие авиакомпании мы не перенаправляем . А потом выяснилось что они летят, почему нормальную информацию не дают когда звонят клиенты??
Между-народный	Бизнес-класс	В Бангкок и обратно летаем только на регулярном рейсе авиакомпании S7. Возможно это был 10 год с s7 , а может просто повезло и нам при посадке в самолет авиакомпания сделала апгрэйд до бизнес класса. Было очень приятно. Обслуживание отличное, меню конечно не Turkish airlines…	Супруга летела на срочную операцию в Питер 11.01.19. Взяли бизнес-класс. Багаж не стали оформлять транзитным т. к. хотели сами быстро успеть на ближайший рейс до Питера из Домодедово по прилету. Прилетела Москву (Домодедово) в 8-40. До ближайшего рейса на Питер было 1 час и 10…

По результатам анализа отзывов можно сделать вывод о том, что компания S7 Airlines учитывает отзывы своих потребителей при построении плана своего развития. Несмотря на тяжелую эпидемиологическую ситуацию в стране и снижение количества перевозимых

пассажиров, с 2018 года качество предоставляемых услуг улучшилось.

Увеличилось количество бортов и штат бортпроводников, открылись новые направления, а на более востребованных из них увеличилось количество рейсов.

При дальнейшей доработке технологии интеллектуального анализа отзывов планируется добавить более гибкий функционал по уточнению классификации, что позволит сделать инструмент более гибким в настройке сбора и анализа информации. Это позволит применять данный инструмент независимо от сферы бизнеса исследуемой компании или продукта.

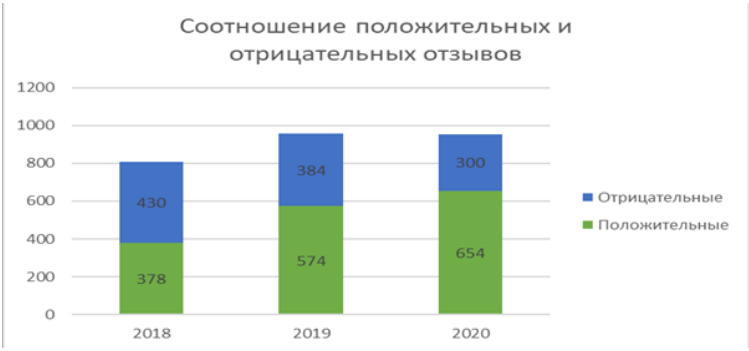

Рисунок 3 – Количество отзывов

Вывод. На сегодняшний день очевиден тот факт, что рост количества информации в Интернете не прекратится, и объемы текстовых данных ежедневно увеличиваются. При этом, данная информация может иметь большой интерес для различного рода компаний, позволяя отслеживать последствия тех или иных решений, в том числе и со стороны конкурентов.

В ходе проводимого исследования была разработана технология сбора и обработки отзывов, работающая на базе API Вконтакте и использующая для сентимент анализа заранее обученную библиотеку Dostoevsky.

Подобные системы, основанные на анализе отзывов из социальных сетей, позволяют не только проводить самоанализ и влиять на принимаемые компаниями решения, но и производить конкурентную разведку.

СПИСОК ЛИТЕРАТУРЫ

1. Text Mining. – Режим доступа: https://gi.de/informatiklexikon/text-mining/ (дата обращения: 25.04.2021)

2. Marti Hearst: What Is Text Mining? – Режим доступа: http://people.ischool.berkeley.edu/~hearst/text-mining.html (дата обращения: 25.04.2021)

3. Text Mining. Michaela Geierhos. – Режим доступа: https://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/technologien-methoden/text-mining (Дата обращения: 23.04.2021)

4. Text Mining. Ian H. Witten – Режим доступа: https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf (Дата обращения: 23.04.2021)

5. Seliverstov Y. A. et al. Development of management principles of urban traffic under conditions of information uncertainty // Conference on Creativity in Intelligent Technologies and Data Science. 2017. pp. 399–418.

6. Шелманов А. О. и др. Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. 2016. № 4. С. 47–61.

7. Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1. С. 72–78.

8. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными // Альфакнига. 2017. 393 с.

9. Нугуманова А. Б., Бессмертный И. А., Пецина П., Байбурин Е. М. Обогащение модели Bag of Words семантическими связями для повышения качества классификации текстов предметной области // Программные продукты и системы. 2016. № 2. С. 89–99.

10. Кипяткова И. С. Программно-алгоритмическое обеспечение создания синтаксическо-статистической модели русского языка по текстовому корпусу // Труды СПИИРАН. 2013. № 1(24). С. 332–348.

11. Шаграев А. Г., Фальк В. Н. Линейные классификаторы в задаче классификации текстов // Вестник Московского энергетического института. 2013. № 4. С. 204–208.

12. Воронцов К. В. Лекции по линейным алгоритмам классификации. URL: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf. (Дата обращения: 16.11.2021).