Е.О. Смирнов, Д.О. Клементьев, Е.В. Решетова, А.Г. Бурова[[1]](#footnote-1)

**Анализ отзывов клиентов**

Рассмотрен алгоритм, который на вход принимает отзывы клиентов по различным продуктам компании, а на выходе предоставляет негативные отзывы или отзывы с предложениями об изменениях в продуктах компании. Таким образом, проведен анализ отзывов клиентов с целью определения содержания информации о дефектах в продукте или сервисе компании. Полученные данные направлены на выявление закономерностей, позволяющих ответить на вопросы что интересно клиентам, какие изменения в продуктах могут понравиться клиенту.

*Ключевые слова:* анализ данных, анализ текста, машинное обучение, предобработка текстовых документов, лемматизация.

*JEL:* C81, C82, C83

**Информация об авторах:**

**Введение**

В настоящее время развитие компьютерных технологий и увеличение роли информации отвело методам анализа текста особую роль. Методы анализа текста применяются при поиске, систематизации, оценке, отборе информации, анализе и прогнозировании событий или поведения субъекта. Возможности применения методик анализа текста очень широки, в частности их можно применять при обработке отзывов на различные продукты компании.

[Отзывы клиентов](https://a1z.ru/otzyvy-kliyentov.html) - это эффективный маркетинговый инструмент, потому что они формируют мнение целевой аудитории о вашей компании и работают на повышение ее имиджа и престижа. Помимо хороших отзывов, могут присутствовать негативные отзывы. Но жалоба – это вовсе не помеха в работе предприятия, а желанная обратная связь, предоставляющая ценнейшую информацию. В связи с этим, анализ и оценка отзывов клиентов является особенно актуальными в наше время.

Целью данной работы является изучение методологических принципов анализа текста, теоретической и математической основы построения методик, а также изучение различных реализаций и готовых библиотек.

Кроме того, необходимо провести анализ с целью вывести потенциальные возможности улучшения существующих реализаций, сравнение их работы и результатов, и реализовать методы оптимизации и улучшении качества работы алгоритма.

**Постановка задачи**

Основная задача проекта – алгоритм, анализирующий отзывы клиентов по продуктам или сервисам компании.

Алгоритм должен определять является ли отзыв негативным и содержит ли он информацию об ошибке. Если таковая информация имеется, то необходимо выделить ключевые слова по данной ошибке. Данный алгоритм отлично бы порекомендовал себя в компаниях, производящих различные товары и услуги и нацеленных на улучшение качества своей продукции.

При разработке данного алгоритма было проведено исследование уже готовых методик анализа и оценки отзывов клиентов. Было выявлено, что полноценной альтернативы данному алгоритму нет, большинство компаний используют традиционные опросные техники, которые наименее эффективны, чем анализ отзывов клиентов.

После исследования предметной области и анализа готовых реализаций были поставлены следующие задачи:

* Нахождение негативных отзывов об ошибках сервиса
* Предобработка текста отзывов клиентов
  + Разделение отзывов на фрагменты
  + Лемматизация текста
  + Удаление стоп-слов
* Определение ключевых слов об ошибках сервиса
* Анализ отзыва и определение ошибки
* Статистический анализ верно распознанных ошибок

**Методика**

**Анализ отзывов клиентов**

**Определение негативных отзывов**

Проведя анализ текста отзывов клиентов явно выделаются три класса данных: дата, название сервиса или продукта, язык, оценка сервиса или продукта (рейтинг) и отзыв. Однако наибольший интерес и пользу представляет рейтинг и отзыв.

В таблице 1 приведены несколько реальных примеров отзывов клиентов.

Таблица 1. Фрагмент текста отзывов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Дата** | **Продукт или сервис** | **Язык** | **Рейтинг** | **Текст отзыва** |
| 2016-12-04 | Сбербанк Онлайн | ru | 4 | Очень удобно |
| 2017-06-14 | Сбербанк Онлайн | ru | 5 | Всё быстро и удобно |
| 2017-08-13 | Сбербанк Онлайн | ru | 1 | Очень долго идёт загрузка, потом вылет |
| 2017-09-21 | Сбербанк Онлайн | ru | 3 | Кнопка выход из приложения неудобно расположена |

Рейтинг имеет шкалу от 1 до 5. После поверхностного анализа отзывов легко заметить, что негативные отзывы, описывающие ошибки сервиса или продукта имеют низкую оценку, как правило, менее 4. Этот фактор и используется для определения того, является ли отзыв негативным.

Несмотря на то что данный метод очень прост, он дает высокие результаты.

Следующим шагом производится предобработка текста отзывов.

**Метод предобработки текста отзывов клиентов**

При нахождении негативного отзыва необходимо произвести предобработку текста данного отзыва для наилучших результатов последующей обработки.

Предобработка состоит из трех последовательных этапов: разделение отзыва на фрагменты, лемматизация и удаление стоп-слов.

**Разделение отзыва на фрагменты**

На данном этапе происходит разделение конкретного негативного отзыва на фрагменты, которые разделяются знаками препинания.

Таким образом проще выделить фрагмент, который содержит ключевые слова. Необходимость данного этапа отражена ниже.

Был проведен анализ фрагментов отзывов:

* *«Очень долго идёт загрузка, потом вылет»*

В данном отзыве присутствует ошибка - *«очень долго идет загрузка».* Это мы и должны получить в результате. Этот фрагмент отзыва отделен от другого запятой.

* *«Спасибо Вам! После обновления не могу зайти в приложение, выдает ошибку. И как теперь быть?»*

В данном отзыве также присутствует ошибка - *«после обновления не могу зайти в приложение»*. Этот фрагмент отзыва отделен от остальных знаками препинания.

**Лемматизация текста**

Лемматизация реализовывалась с использованием словарей, в которых находятся леммы и список форм для леммы.

За основу для будущего словаря выбран частотный словарь М. Хагена [9].

Метод заключается в поиске каждого слова отзыва в словаре среди форм лемм, затем преобразование данного слова в исходный формат.

Данный метод необходим, т.к. алгоритм может посчитать формы одного слова за разные слова, таким образом подсчет ключевых слов может быть некорректным. Также метод упрощает работу с текстом, потому что все слова принимают начальную форму.

**Удаление шумовых слов**

Качество работы алгоритма зависит от наличия общеупоребляемых слов, которые должны отбрасываться.

Для удаления стоп-слов использовался список со стоп-словами на русском языке, который можно расширять, в зависимости от специфики текстов.

Было найдено несколько списков частотных слов [10][11] русского языка, они содержат большинство шумовых слов русского языка.

Метод удаления заключается в поиске каждого слова отзыва в словаре среди стоп слов, при нахождении данное слово удаляется из текста отзыва.

Метод упрощает обработку текста, т.к. удаляются слова, которые не несут в себе нужной информации и сокращается размер текста.

**Определение ключевых слов об ошибках**

При написании отзыва клиенты употребляют ограниченное количество слов и фраз, которые похожи по смыслу. Также круг ошибок, возникающих у пользователей ограничен.

Таким образом, проведя анализ отзывов был составлен список ключевых слов и фраз, которые сигнализируют об ошибке в продукте или сервисе. Также каждому ключевому слову соответствует количество баллов (рейтинг), что подразумевает его «значимость».

Некоторые из них представлены в таблице 2.

Таблица 2. Фрагмент ключевых слов

|  |  |
| --- | --- |
| **Слово** | **Рейтинг** |
| вылет | 50 |
| неудобный | 20 |
| приложение | 10 |
| кнопка | 30 |
| бесполезный | 40 |
| вылетать | 50 |
| разрыть | 40 |

Второй вариант добывания ключевых слов – это прогон библиотеки BigARTM [1] по коллекции отзывов. Данная библиотека выдает список слов, которые наиболее часто употребляются в отзывах. Эти слова также будут являться ключевыми словами.

**Анализ отзыва и определение ошибки**

Для каждого слова в отзыве создается счетчик, который определяет значимость слова. Если слово не удалилось на предыдущих шагах – лемматизации и удаления стоп-слов, значит необходимо добавить ему 10 баллов. Затем проверяется, присутствует ли слово в списках ключевых слов. Если это условие выполняется, то к рейтингу слова добавляется рейтинг, указанный для ключевого слова.

Следующим шагом вычисляется суммарное количество баллов для каждого фрагмента отзыва и выбирается тот фрагмент, в котором рейтинг максимален. Данный фрагмент и будет ключевым, то есть содержать ключевую информацию об ошибке.

**Результаты алгоритма для анализа отзывов клиентов**

**Положительные отзывы**

1. **Отзыв:** *«У меня нет ни одного замечания к работе этого приложения. С появлением копилок и автоплатежей, все стало только лучше»*

**Оценка:** 5

**Результат:** false

Данный результат значит, что отзыв не является негативным и не рассматривается. Все верно, т.к. оценка больше 4.

1. **Отзыв:** *«Очень удобное приложение )никаких ошибок))»*

**Оценка:** 5

**Результат:** false

Аналогичный результат, отзыв не рассматривается, т.к. его оценка выше 4.

**Негативные отзывы**

1. **Отзыв:** *«Что вы сделали с приложением .. оно теперь просто бесполезно ..куда Дели переводы клиентам ?»*

**Оценка:** 1

**Результат:** true

**Ключевые слова:** куда дели переводы клиентам

В данном случае описана ошибка продукта или сервиса, клиент недоволен отсутствием функции. Алгоритм выделил ключевую часть среди остальных, не несущих информацию, частей.

1. **Отзыв:** *«Тормозит при запуске. Не могу решить между рутом и сбером. Дали бы хоть возможность под нашу ответственность полный запуск функций с рутом»*

**Оценка:** 1

**Результат**: true

**Ключевые слова:** тормозит при запуске

Аналогичный случай, клиент недоволен загрузкой сервиса. Алгоритм выделил основную часть проблемы.

**Другие отзывы**

1. **Отзыв:** *«При загрузке вылетает и заново входишь.»*

**Оценка:** 4

**Результат:** false

В данном случае отзыв считается положительным, хоть в нем присутствует жалоба клиента по причине того, что клиент поставил оценку 4. Данная оценка считается положительной и отзыв не рассматривается алгоритмом.

1. **Отзыв:** *«за телефон не заплатишь. жене денег не переведешь. и жкх . ВСЕ ПРОПАЛО!!!!!(((((( »*

**Оценка:** 5

**Результат:** false

Аналогичный случай. Клиент недоволен сервисом, однако ставит отличную оценку, в связи с чем отзыв не рассматривается как негативный.

В таблице 3 приведен фрагмент результатов работы алгоритма.

Таблица 3. Фрагмент результатов

|  |  |  |
| --- | --- | --- |
| **Review** | **IsBag** | **Keywords** |
| Не дождешься смс никак :-( ужасно ужасно и ещё раз ужасно | true | Не дождешься смс никак |
| После обновления приложение перестало нравится. Через wi-fi невозможно зайти? | true | Через wi-fi невозможно зайти |
| Пишет ошибка 24. Обьясните что это такое? Хотя до этого пользовалась прогой | true | Пишет ошибка 24 |
| Мне с Вами удобно | false | - |

Алгоритм дает хороший результат, в большинстве случаев, если оценка указана верно – находит ошибку.

**Статистические результаты определения ошибок**

В результате статистического анализа, после распознавания ошибок, мы получаем 3 числа:

* 0.77218492231 ~ 77% верно распознанных отзывов о присутствии или отсутствии ошибки
* 0.694349501854 ~ 69% распознанных ошибок
* 0.800072956565 ~ 80% распознанных отзывов об отсутствии ошибки

**Выводы**

Таким образом, были изучены методики анализа текста и разработан алгоритм анализа и оценки отзывов клиентов по продуктам компании, а так же алгоритм был успешно применен на отзывах клиентов на сервис «Сбербанк-Онлайн».

Было значительно ускорено время работы, а так же улучшено качество работы алгоритма за счет предобработки текста. Количество слов в отзывах после предобработки уменьшается в 2-3 раза, как следствие время работы алгоритма уменьшается из-за уменьшения объема данных, с которыми ему придется работать. После предобработки по топу слов в каждом отзыве понятно, содержит данных отзыв ошибку, без дополнительных обработок в виде регуляризаций.

В итоге в процентном соотношении результаты работы разработанного алгоритма получились очень качественными, что делает его удобным для использования.

**Список литературы**

1. Автоматическая обработка информации и анализ данных на основе тематического моделирования «BigARTM» / Бурова А.Г., Смирнов Е.О., Андреева О.В., Багиров М.Б. // Экономика и менеджмент систем управления. - 2017. - №4.1. С.159-164.
2. Багиров М.Б., Андреева О.В. // Использование Stanford Core Natural Language Processing для определения частей речи в русском языке //13-я Международная молодежная научно-техническая конференция «Современные проблемы радиоэлектроники и телекоммуникаций, РТ – 2017» Сборник тезисов докладов 20-24 ноября 2017г. Севастополь. 2017.
3. Янина А.О., Воронцов К.В., Бунина Е.И., Разведочный поиск [Электронный ресурс] // Мультимодальные тематические модели статей коллективных блогов для разведочного поиска. – Москва, 2016-2017. URL:

<http://www.machinelearning.ru/wiki/images/5/52/Yanina2016Bsc.pdf>

(Дата обращения: 22.11.2017)

1. Воронцов К. В. [Аддитивная регуляризация тематических моделей коллекций текстовых документов](http://www.machinelearning.ru/wiki/images/d/d7/Voron14dan-rus.pdf) // Доклады РАН. 2014. — Т. 455.
2. Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. — 2013 (в печати).
3. Документация к библиотеке bigARTM: [Электронный ресурс] // 2015, URL: https://bigartm.readthedocs.io/en/stable/
4. Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. — 2012. — Т. 4, № 4.
5. П.А. Ромов А.О.Янина М.А.Суворова М.А.Апишев К.В.Воронцов, А.И.Фрей. Bigartm: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций. Труды XVII Международной конференции DAMDID/RCDL’2015 «Аналитика и управление данными в областях с интенсивным использованием данных», 2015.
6. Хаген М. Словарь форм слов и их лемм: [Электронный ресурс] // Хаген М. Полная парадигма. Морфология. Частотный словарь. Совмещенный словарь. URL: http://www.speakrus.ru/dict/#morph-paradigm. (Дата обращения: 22.11.2017).
7. Список стоп-слов: [Электронный ресурс] // Букварикс – программа для быстрого подбора ключевых слов, 2014-2017. URL: <http://www.bukvarix.com/bukvarix-stop-words.html>. (Дата обращения: 22.11.2017).
8. Рейтинги частотности слов: [Электронный ресурс] // Русский Викисловарь, 2014-2017. URL: <https://ru.wiktionary.org/wiki/Приложение:Рейтинги_частотности_слов>. (Дата обращения: 22.11.2017).

**ANALYSIS OF CUSTOMER REVIEWS**

E.O. Smirnov, D.O. Klemetyev, E.V. Reshetova, A.G. Burova

Nizhny Novgorod State Technical University n.a. R.E. Alekseev

An algorithm is considered that accepts customer feedback on various products of the company, and at the output provides negative feedback or feedback with suggestions about changes in the company's products. Thus, the analysis of customer feedback was conducted to determine the content of information about defects in the product or service of the company. The obtained data are aimed at revealing regularities that allow answering questions that are interesting to customers, what changes in products the client may like.

*Key words:* data analysis, machine learning, preprocessing of text documents, lemmatization.

1. Смирнов Егор Олегович ([email@domen.com](mailto:email@domen.com)), студент, студент Нижегородского Государственного Технического Университета им.Р.Е.Алексеева (Нижний Новгород); Клементьев Дмитрий Олегович, студент, студент Нижегородского Государственного Технического Университета им. Р.Е.Алексеева; Решетова Елена Владимировна, студент, студент Нижегородского Государственного Технического Университета им.Р.Е.Алексеева; Бурова Анжела Григорьевна, студент, студент Нижегородского Государственного Технического Университета им.Р.Е.Алексеева. [↑](#footnote-ref-1)