САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ – ПРОЦЕССОВ УПРАВЛЕНИЯ

Мурадов Аветик Георгиевич

Выпускная квалификационная работа

Система определения тональности текста

Направление 02.06.01

«Математическая кибернетика»

Научный руководитель,
д-р физ.-мат. наук,
профессор Богданов А. В.

Рецензент,
д-р техн. наук,
СПбГЭТУ «ЛЭТИ» Щеголева Н. Л.

Санкт-Петербург
2018

Оглавление

Аннотация 4

Введение 5

Постановка задачи 7

Обзор литературы 8

Глава 1. Предметная область 10

1.1 Практическое применение анализа тональности текста 11

1.3 Методы и подходы для задачи анализа тональности текста 13

Глава 2. Задача автоматической классификации текста, существующие алгоритмы 15

2.1 Задача классификации 15

2.2 Представление и обработка данных 16

2.3 Наивный Байесовский классификатор 18

2.4 Метод опорных векторов 19

Глава 3. Существующие системы 22

3.1 Система I-Teco 22

3.1.1 Стадии работы системы I-Teco 22

3.1.2 Тональные словари 23

3.1.3 Глагольные и неглагольные лексемы и коллокации 24

3.1.4 Коллокаций и лексемы 25

3.1.5 Объектно-ориентированный подход 26

3.1.6 Оценка результата 28

3.2 Система определения тональности SentiScan 29

3.2.1 Главные элементы системы 29

3.2.2 Оценка качества работы системы SentiScan 31

3.3 Система Sentiment Analysis Service 33

3.4 Выбор инструмента тональности 35

Глава 4. Практическая реализация 36

4.1 Архитектура приложения для сбора и анализа данных 36

4.2 Используемые системы и инструменты 39
4.3 Пользовательский интерфейс для сбора и анализа данных 43
4.4 Тестирования для приложения «ВКонтакте» 43
4.5 Определения тональности для приложения «Сезам» 45
4.6 Тестирование системы и выводы 47
Вывод 49
Заключение 51
Список цитируемой литературы 52
Приложение 57
Приложение 1 57
Приложение 2 58
Приложение 3 59
Приложение 4 60
Приложение 5 61
Приложение 6 62

Аннотация

В процессе поиска информации, для дальнейшего принятия решения по отношению к выбранному субъекту или объекту, важную роль играет мнение других людей. С большим темпом развиваются и создаются новые интернет-ресурсы: социальные сети, блоги, мобильные приложения и многие другие источники информации, где люди могут делиться своим мнением.

В рамках данной работы рассматривается проблема автоматического определения тональности текста, исследование существующих систем, подходов и методов для выявления положительной или отрицательной окраски текста. Цель работы — разработка программного обеспечения для сбора данных и автоматического распознавания тональности текста в мобильных приложениях и социальных сетях, в частности для мессенджера (система обмена сообщениями) «Сезам» и социальной сети «ВКонтакте».

Введение

При принятии какого-либо решения человек стремится узнать мнение окружающих людей. До того, как интернет стал всеобще доступным и популярным, люди собирали рекомендации, относящиеся к различным отраслями обыденной жизни, среди друзей и знакомых. Но с развитием всемирной паутины появилась возможность с помощью интернета находить всю необходимую информацию, касательно различных товаров, услуг, политики и многих других сфер деятельности.

С каждым днем увеличивается количество пользователей интернета, глобальный прирост аудитории сильно заметен в социальных сетях (см. приложение 1). Недавнее исследование, проведенное аналитической компанией «We Are Social» совместно с «Hootsuite», опубликовали следующие результаты[1]:

- В 2018 году количество пользователей всемирной паутины достигло более четырех миллиарда человек, что на 7% больше по сравнению с прошлым годом;
- Социальными сетями пользуются 3.196 миллиардов человек, данный показатель превышает оценку 2017 года на 13%;
- Мобильными устройствами в 2018 году пользуются свыше пяти миллиардов людей, что на 4% больше показателей прошлого года.

В России количество активных интернет-пользователей составляет 87 миллионов человек. Приблизительно 47% населения зарегистрировано в различных социальных сетях, большая часть которых отдает свое предпочтение социальной сети «ВКонтакте».

С увеличением пользователей стали активно развиваться и разрабатываться различные интернет-сообщества, социальные сети, интернет-магазины, а также мобильные приложения.

[1] https://www.web-canape.ru/business/internet-2017-2018-v-mire-i-v-rossii-statistika-i-trendy, 03.06.2018

Необходимо выделить мобильные приложения и социальные сети, которые стали неотъемлемой частью нашего современного общества. При помощи этих ресурсов люди обмениваются новостями, сообщениями высказывают свою точку зрения о различных аспектах жизни.

Всемирно известные организации, компании, университеты разрабатывают свои личные веб-страницы, сообщества, блоги. Они размещают свои товары, услуги, предложения, и решения различных задач, затем, при помощи мониторинга определяется мнение общества о предложенной ими информации.

Учитывая вышеперечисленное, есть потребность в разработке программного обеспечения, для выполнения автоматического анализа предложенной обществом информации для выявления отношения людей к данным товарам и услугам. Чтобы распознать мнение, изложенное пользователем в своем тексте, т.е выявить отрицательную или положительную окраску текста, требуется выполнить анализ тональности текста.

Сентимент анализ (англ. sentiment analysis, анализ тональности текста, эмоциональная окраска текста) — обработка естественного языка (англ. NLP, natural language processing,), цель которого является извлечение эмоционального содержания из текста[2].

[2] Прикладная и компьютерная лингвистика. О. В. Митрениной, И.С Николаева, Т. М. Ландо. 2017. С. 245

Постановка задачи

Цель данной выпускной квалификационной работы является изучение и сравнение существующих систем и методов для определения тональности русскоязычных текстов в автоматическом режиме. Также одной их задач текущей работы является создание программного обеспечения для анализа данных социальной сети «ВКонтакте» и мессенджера «Сезам».

Для достижения данной цели были поставлены следующие задачи:

1. Изучение одного из направлений обработки естественного языка — анализ тональности текста;
2. Исследовать существующие подходы, методы и системы для обработки русскоязычных текстов и автоматического определения эмоциональной составляющей текста;
3. Провести сравнительный анализ между существующими системами;
4. Разработка программного обеспечения для сбора и анализа данных из социальной сети «ВКонтакте»;
5. Построение графа с зависимостями между сообществами социальной сети «ВКонтакте»;
6. Разработка программного обеспечения с целью определения эмоциональной окраски текста для мессенджера «Сезам»;
7. Разработка пользовательского интерфейса для использования системы определения тональности текста.

Обзор литературы

Автоматический анализ тональности теста привлек к себе большое количество исследователей в области обработки текстов на естественном языке. В связи с актуальностью данной задачи и проявленному интересу к компьютерной лингвистике и искусственному интеллекту со стороны исследователей данного направления, было написано много статей. Так как речь идет об обработке текста на естественном языке, необходимо отметить, что большая часть литературы написана с целью обработки англоязычных текстов.

В книге Sentiment Analysis and Opinion Mining, автором которой является Bing Liu, хорошо описана задача автоматического определения тональности текса, проблемы с которыми чаще всего сталкиваются исследователи, показана важность применения данного анализа во всех сферах бизнеса и социальной сферы [3].

Большая часть книг, написаны на английском языке. Необходимо отметить работу авторов О. В. Митрениной, И.С Николаева, Т. М. Ландо, которые представили книгу на русском языке — Прикладная и компьютерная лингвистика. Книга дает возможность найти ответы на общие вопросы, связанные с компьютерной лингвистикой [1].

Для решения задачи определения тональности текста используется несколько подходов [8]:

- Машинное обучение;
- Метод, основанный на правилах;
- Гибридный метод.

Наиболее распространенным решением, является машинное обучение, так как при помощи алгоритмов классификации можно определить сентимент анализ без использования лингвистического анализа для естественного языка. В исследовании Клековкина М. В и Котельникова Е. В, рассмотрены следующие методы машинного обучения: наивный байесовский

классификатор (NB), k ближайших соседей, классификатор Rocchio, метод опорных векторов (SVM) на основе ключевых слов и его комбинаций [33]. В работе J.B. Teraiya и S.M. Vohra проводится сравнительный анализ нескольких алгоритмов, в том числе и вышеперечисленных SVM и NB. Результаты исследования показали следующее: точность работы NB достигает да 87%, а SVM до 85.2% [14].

Исследование Соловьева А. Н. и Пазельской А. Г. [28] показало высокий результат в определении эмоциональной окраски текста, используя методы лингвистического анализа с применением тональных словарей. В работе подробно описан процесс синтаксического и морфологического анализа. Также в статье описывается объектно-ориентированный подход для определения тональности. Система в автоматическом режиме определяет объект в предложении, если он не был указан заранее.

Гибридный метод — совокупность нескольких подходов для определения сентимента. В своей работе Разработчик исследования Дмитрий Кан [10] использует гибридный подход лингвистического анализа и метод машинного обучения. Алгоритм, описанный автором, учитывает анафорические ссылки в предложении, противительные союзы и отрицания, что сильно влияет на качество результата.

В текущей работе используются все вышеперечисленные материалы. Для достижения поставленных задач применяется метод машинного обучения и метод, основанный на тональных словарях.

Глава 1. Предметная область

Обработка естественных языков (англ. Natural Language Processing) используется в таких задачах, как распознавание тем, анализ тональности, распознавание языка, классификация документов и извлечение ключевых фраз [5]. Текст — один из основных источников информации. Именно с его помощью миллиарды людей ежедневно получают и передают информацию.

Анализ тональности текста, также используется термин сентимент анализ, (англ. sentiment analysis) — одно из направлений обработки естественных языков, цель которой является определение эмоциональной составляющей текста, выраженное автором[3].

Данное направление начало активно развиваться в 2000-х годах. Компании стали использовать анализ тональности текста для оценки популярности своей продукции среди партнеров и клиентов. Компании оценивали отзывы тысяч покупателей, чтобы определить их эмоции, которые скрыты в тексте.

Развитие интернета и количества информации, растущей с каждым годом, стало толчком для растущего интереса к анализу тональности текста. Компания Radicati прогнозирует, что количество пользователей электронной почты и текстовых сообщений вырастит на 10.8%[4]. Сентимент анализ в центре внимания социальных медиа-исследований. Данная область анализа нашла свое применение в различных отраслях экономики, менеджмента, социальной науки и политологии.

Естественно, что анализ тональности сильно зависит от языка. В интернет-пространстве существует много решений и работ для английского языка. В России данная область стала развиваться в 2010 гг. Сентимент анализ для русского языка вызвал большой интерес среди исследователей.

[3] http://ai-news.ru/2017/04/sovremennye_metody_analiza_tonalnosti_teksta.html, 03.06.2018
[4] http://www.marketwired.com/press-release/the-radicati-group-releases-email-statistics-report-2017-2021-2193516.htm, 03.06.2018

1.1 Практическое применение анализа тональности текста

Коммерческие фирмы, производящие продукт или предоставляющие услуги, интересуются мнением потребителей об их продукте либо услуге. Полученные данные используются для повышения качества их сферы деятельности, определения целевой аудитории и для выявления главных недостатков и достоинств их конкурентов. Людям также интересно мнение общества, например, для того, чтобы проголосовать за того или иного кандидата на политических выборах. Зачастую, учитывая мнения других, люди принимают свое личное решение.

Ряд исследований по сентимент анализу был произведен в области медицины, для выявления отношения людей к вакцинации. Исследование было проведено в социальной сети «Твиттер». Оно показало, что активнее распространяется негативное отношение к вакцинации, нежели позитивное. Также анализ тональности применялся к отзывам о медицинских услугах и товарах. Например, в социальных сетях, часто обсуждают кино [23], книги, политические вопросы, а также и выборы. Зачастую, активность проявляется в предвыборный период. Обсуждаются программы и проекты, предложенные кандидатами. Избиратели строят свои решения исходя из информации, находящейся в сети, и в дальнейшем голосуют за определенного кандидата. Аналитиков, экспертов и представителей предвыборных компаний интересует, какой настрой у общества по отношению к ним. Таким образом, для того, чтобы проанализировать большое количество информации в сети и выявить настрой общества к определенной деятельности, используется анализ тональности текста.

1.2 Проблемы анализа естественных языков

Несмотря на достаточно большое количество работ и систем, проведенных в сфере анализа тональности текста, точность определения сентимент анализа не находится на совершенном уровне. Ниже представлены проблемы, которые часто возникают при разработке и использовании

автоматических алгоритмов определения тональности русскоязычного текста[5]:

1. При проведении анализа, необходимо учитывать предметную область. Для получения высоких результатов требуются более высокоточные алгоритмы. Например, провести анализ тональности продукции или услуги намного легче, нежели анализировать область политики, т.к. там существенно больше различной терминологии и словосочетаний;
2. В процессе выполнения анализа, исследователи сталкиваются с рядом таких проблем как: синтаксические и грамматические, ошибки, сокращения слов, сленги, которые в дальнейшем играют большую роль на качество определения тональности;
3. Выявление сарказма и иронии — одна из больших проблем сентимент анализа, т.к. текста и предложения, где содержатся вышеперечисленные составляющие, могут быть по смыслу отрицательными, но сам текст, по факту, написан в положительном ключе. Например, "Самый быстрый телефон на свете, такого еще никто не видел!";
4. Противительные союзы, могут изменить всю тональность предложения. В русском языке существует небольшое количество союзов такого типа: а, да (в значении но), зато, но, однако. Например, "Мальчик любит ездить на велосипеде, но после того, как он получил серьезную травму, он перестал кататься";
5. Определение анафор в тексте, например, "Кате были очень рады ее друзья. Она готова помочь в любой ситуации". Местоимение *она* ссылается на Катю;

[5] Брунова Е. Г. Автоматизированный контент-анализ мнений трех предметных областей // Тамбов: Грамота, 2014. С. 43-47.

6. Определение тональности текста по заданному объекту. Например, "Егору нравятся смартфоны от производителя X, но ему не нравится продукция компании Y". Если в качестве объекта взять «X», тональность текста будет позитивная, если же рассматривать объект «Y», то соответственно отрицательная;
7. Использование смайликов, при написании. Зачастую, смайлики не соответствуют тексту, что может привести к неправильной оценке тональности текста;
8. Применение машинного перевода, может привести к искажению смысла исходного текста, что повлияет на качество определения тональности.

1.3 Методы и подходы для задачи анализа тональности текста

Ниже представлены существующие подходы классификации текстов для выявления тональности текста[29]:

- Машинное обучение с учителем;
- Машинное обучение без учителя;
- Подход, построенный на словаре;
- Подход, построенный на правилах (англ.Rule-based approach);
- Объектно-ориентированный подход;
- Гибридный подход.

Обучение с учителем – способ машинного обучения, который часто используется для сентимент анализа. Обучающая выборка должна быть разделена на классы: негативный и позитивный. Выборка используется для получения классификатора. Построенная модель будет определять эмоциональную окраску текста для новых предложений и документов. Наиболее часто применяются наивный Байесовский классификатор и метод опорных векторов. Используя машинное обучение с учителем, можно получить хорошие результаты, точность работы алгоритмов доходит до 90%.

Сложность данного решения, заключается в тестовой выборке, т.к. для того, чтобы получить высокую точность, нужно создать размеченную обучающую выборку.

Обучение без учителя – один из разделов машинного обучения для автоматической кластеризации данных. Но для определения тональности теста используется реже, т.к. точность определения намного ниже по сравнению с машинным обучением с учителем. Отличие данного подхода заключается в том, что нет нужды заранее разделять тестовую выборку на классы. На практике часто применяют алгоритм k-средних (от англ. k-means).

Подход, построенный на словаре. Для данного подхода требуется наличие тонального словаря, где размечены позитивные и негативные слова. Каждое слово должно иметь свой собственный вес, например, значение от -5 до 5. Анализируя документ, к каждому слову необходимо присвоить его значение в соответствии со словарем. Итоговой оценкой является среднее арифметическое из всех слов в документе. Недостаток подхода — не универсальность, для каждой области необходимо создавать свой словарь.

Подход, построенный на правилах. Для реализации данного подхода, необходимо сгенерировать набор правил, которые будут использоваться при анализе тональности. Для получения высокой точности необходимо описать большое количество правил, что является длительным и трудоемким процессом.

Объектно-ориентированный подход. Данная задача считается наиболее тяжелой и требует более сложных алгоритмов, т.к. при определении общей тональности текста важно лишь соотношение положительных и отрицательных терминов в документе, а при определении тональности по отношению к определенному объекту большое значение имеет синтаксическая зависимость объекта с тональной лексикой.

Гибридный подход. Используется во всех коммерческих системах. Данный подход содержит в себе все вышеперечисленные методы и подходы для решения задачи анализа тональности текста.

Глава 2. Задача автоматической классификации текста, существующие алгоритмы

В текущей главе будет рассмотрена задача классификации и наиболее часто используемые алгоритмы машинного обучения с учителем для анализа тональности текста.

Автоматическая классификация документа, то есть обозначение принадлежности текста к какой-либо категории — весьма актуальная и интересная задача с учетом непрерывно растущей информации в интернет пространстве. Принадлежность может определяться к некоторым классам по разным признакам: общая тематика текста, использование определенных понятий, также по эмоциональной окраске текста и многими другими условиями. Классификация применяется для решения различного рода практических задач: определение наличия спама, подбор контекстной рекламы, а также для сентимент анализа.

2.1 Задача классификации

Задачу классификации текста для анализа тональности можно определить следующим образом [30]:

Рассмотрим два класса: отрицательный и положительный, в соответствии c_1 $и$ c_2, $C = \{c_1, c_2\}$ и некое множество документов:

$$D = \{d_1, d_2, \ldots, d_n\}$$

Определим неизвестную целевую функцию:

$$F: C \times D \to \{c_1, c_2\}$$

значение функции известно исключительно для обучающей выборки, т.е. к какому классу относится определенный документ. Необходимо найти функцию (классификатор) F', которая будет распределять документы в соответствующие классы:

$$F': C \times D \to \{c_1, c_2\}$$

Существуют несколько видов классификации:

- $F' : C \times D \to \{0, 1\}$ — точная;
- $F' : C \times D \to [0, 1]$ — ранжированная.

В первом случае, каждому документу сопоставляется нулевое значение, для определения соответствия документа с конкретным классом. В случае ранжирования определяется степень принадлежности — число из диапазона $[0, 1]$. Соответственно, чем больше число, тем больше документ относится к классу.

2.2 Представление и обработка данных

Большинство методов, в процессе обработки естественного языка, в качестве признака используют слова, тем самым, зачастую, игнорируя семантику и синтаксис, которые выявляются из структуры предложений. Модель «мешок слов» (от англ. bag of words)[6] — набор слов, представленный в виде вектора, без учета их порядка. Векторное представление документов можно объединить в кластеры, на которых обучается модель, и используются кластеры для классификации задач. Но если необходимо извлечь более качественное семантическое представление, то потребуются механизмы обработки текстов, работающие с синтаксической структурой предложений или инструментов, которые сохраняют последовательность слов в предложении. Каждое слово из модели «bag of words» имеет свой собственный вес. Для нахождения веса слова используют несколько подходов:

- Первый подход — бинарный подход. Формируется словарь из уникальных слов всего корпуса (всех документов), причем берется основа слова. Затем для каждого слова из документа

[6] http://lab314.brsu.by/kmp-lite/kmp2/job/cmodel/bow-q.htm, 03.06.2018

определяется вес: к слову присваиваем значение 1, если оно присутствует в словаре, в противном случае соответственно 0;

- Второй подход — количество вхождений слова в документ. Предполагает непропорциональность оценки веса для текстов с разной длиной — больший вес будут получать объемные текста, т.к. в них большое количество слов;
- Третий подход — нахождение статистической меры $TF * IDF$ [36] . TF от английского Term frequency — частота слова, формуле для вычисления

$$TF = \frac{n_i}{\sum_k n_k}$$

n_i — число вхождений слова в документе;

$\sum_k n_k$ - длина документа.

IDF от англоязычного термина Inverse Document Frequency — обратная частота документа. Для вычисления используется формула

$$IDF = \log \frac{|D|}{|d_i \supset t_i|}$$

$|D|$ - общее количество документов;

$|d_i \supset t_i|$ - количество документов, где содержится i - ый терм.

Таким образом, каждое слово принимает свой собственный вес

$$w_i = TF * IDF$$

Существует еще один способ представления текстовых данных, входящих в состав обработки естественного языка это $N - граммы$. $N > 0$.

1. **Юниргаммы**. Предложение представляется в виде вектора, например, "Петя хороший друг", текст будет разбит на слова: ["Петя ", "хороший", " друг "];
2. **Биграммы**. При использовании биграмм, вектор будет представлен следующим образом: ["Петя хороший ", " хороший друг "];

3. **Триграммы**. Соответственно ["Петя хороший друг"].

На практике чаще всего используют биграммы.

2.3 Наивный Байесовский классификатор

Наивный Байесовский классификатор, НБК (англ. Naive Bayes, NB) — классификатор, построенный на основе теоремы Байеса. Модели, основанные на НБК, достаточно просты и полезны при работе с большими наборами текстовых данных. Алгоритм часто используется для определения тональности теста [35].

Используя теорему Байеса, можно получить условную вероятность для классов.

$$P\ (c \mid d\ ) = \frac{P(c)P(d \mid c)}{P(d)} \quad (1)$$

Класс с самой большей вероятностью c^*, которому относится документ $d = \{w_1\ , w_2, \dots , w_n\}$:

$$\mathrm{c}^* = arg\ max_c\ P\ (c \mid d).$$

Из формулы (1) следует:

$$\mathrm{c}^* = arg\ max_c\ P\ (d \mid c) * P\ (c)$$

$$\mathrm{c}^* = argmax_c\ P\ (w_1\ , w_2, \dots , w_n \mid c) * P\ (c).$$

При подсчете c^* Можно не учитывать знаменатель $P(d)$, т.к. поиск идет для всех классов $\boldsymbol{c}$ и одного документа d.

В данном классификаторе допускается применять следующие упрощения — предполагается, что признаки (слова) в документе d независимы и последовательность слов не влияет на результат алгоритма.

Условную вероятность выражения $P\ (d \mid c)$ можно подсчитать следующей формулой:

$$P(w_1|c) * (w_2|c)*..*(w_n|c)=\prod_i P(w_i|c_j)$$

Получим:

$$\mathrm{C} = argmax_c \left[P(c_j) * \prod_i P(w_i|c_j) \right],$$

где

$$P(c_j) = \frac{D_c}{D}$$

Обозначим функцию $f(w_i, c)$, которая подсчитает количество вхождений слова w_i в класс $\underline{c}$.

$$P(w_i|c_j) = \frac{f(w_i, c) + 1}{\sum_{w \in V} f(w, c_j) + |V|}$$

V– словарь тестовой выборки.

В случае, если $f(w_i, c) = 0$, добавляется единица для того, чтобы вероятность не обратилась в ноль, а имела хотя бы незначительное значение для определения класса — сглаживание по Лапласу.

Как было указано выше, НБК хорошо подходит для решения задачи определения тональности текста, но для того, чтобы получить высокую точность, необходимо иметь качественную выборку для обучения.

2.4 Метод опорных векторов

Метод опорных векторов (Support Vector Machine — SVM) — мощный и универсальный метод машинного обучения, выполняющий как линейную, так и нелинейную классификацию. Эта модель очень популярна в машинном обучении. SVM хорошо подходит для классификации средних или небольших наборов данных. В данном разделе описана ключевая концепция SVM. Как было указано, метод решает задачу линейной классификации, в случае с задачей тональности текста — распределение на классы: негативный и позитивный [31].

Для использования метода опорных векторов, потребуется тестовая выборка: рассмотрим множество векторов $[x_1, x_2, \ldots, x_n] \in R^n$ и числа $[y_1, y_2, \ldots, y_n] \in \{-1,1\}$. Каждому вектору x_1 задано заранее число y_1.

Метод заключается в нахождении оптимальной разделяющей гиперплоскости, которая отделяет классы с максимальной точностью. Гиперплоскость, разделяющая два класса, выбирается так, чтобы расстояние между ближайшими точками, которые расположены по различные стороны от плоскости, являлись максимальными. Если у алгоритма не получается сразу найти оптимальную плоскость, то он начинает добавлять новое измерение для дальнейшего разделения. Процесс будет продолжаться, пока не произойдет разделение на два отдельных класса, например, позитивный и негативный. Для этого необходимо получить такой вектор w и число b, чтобы некоторого $\varepsilon > 0$ выполнялось:

$$w * x_i \geq b + \varepsilon \Rightarrow y_i = 1$$
$$w * x_i \leq b - \varepsilon \Rightarrow y_i = -1$$

В алгоритме не произойдет никаких изменений, в случае умножения вектора w и число b на одну и ту же константу. Воспользовавшись вышесказанным, можно найти константу, чтобы выполнялось условие:

$$w * x_i - b = y_i$$

Неравенство умножается на $1/\varepsilon$, ε приравняем к единице. Аналогичным образом производится для всех векторов x_i из тестовой выборки:

$$w * x_i - b \geq 1, \text{если } y_i = 1$$
$$w * x_i - b \leq 1, \text{если } y_i = -1$$

Данное условие: $-1 < w * x_i - b < 1$ определяет полосу, которая разделяет два класса. Можно заметить, что объекты из тестовой выборки не

расположены на этой полосе. На (*рис. 1)* представлен пример разбиения гиперплоскости.

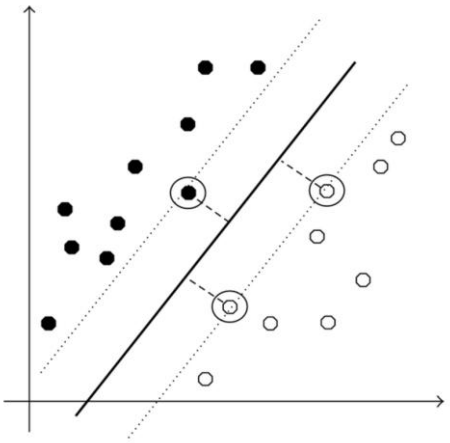


Рис.1 Разделение гиперплоскости

Подводя итоги по вышеперечисленным методам, этапы реализации алгоритма для сентимент анализа можно представить следующим образом (см. *рис. 2*), [**Ошибка! Источник ссылки не найден.**]:

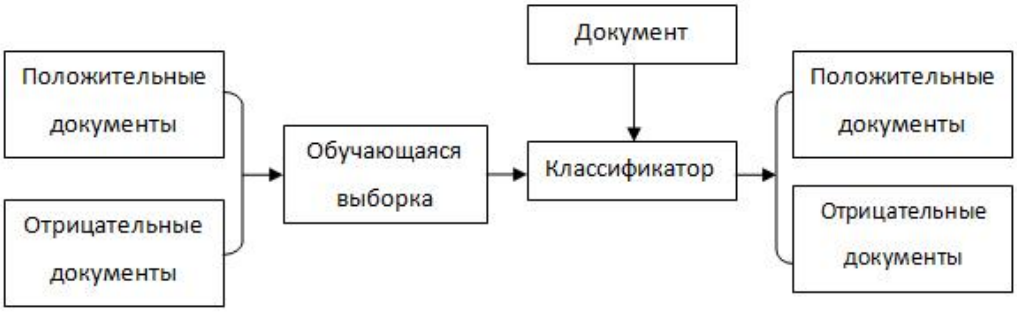


Рис. 2 Этапы реализации

В качестве обучающейся выборки, необходимо иметь два корпуса документов. В одной выборке должны находиться только положительные текста, в другой отрицательные. В зависимости от выбранного алгоритма строится классификатор, который принимает документ для определения окраски текста. Описанные выше методы, метод опорных векторов и байесовский классификатор, хорошо справляются с задачей определения тональности текста. По итогам исследования, точность составляет у NB (81 - 87%), SVM (85 - 87%).

Глава 3. Существующие системы

Сентимент анализ — относительно молодое направление, но, несмотря на этот факт, он стал сильно востребован и популярен во многих сферах деятельности. Существует достаточно большое количество систем для мониторинга и определения тональности текста. Большинство ресурсов работают с англоязычными текстами: Google Natural Language API, Microsoft Text Analytics API, IBM AlchemyAPI, Stanford CoreNL и многие другие системы. Но в данной работе, рассматриваются решения для определения окраски теста русского языка. В текущей главе описаны следующие инструменты: SentiScan, SentimentAnalysisService и I-Teco.

3.1 Система I-Teco

В системе рассматриваются методы и подходы автоматического определения тональности текста для СМИ на русском языке. В основе разработанного алгоритма лежит набор комбинаторных правил, как для отдельных слов, так же словосочетаний. Используются тональные словари. При помощи лингвистического анализа, система определяет эмоциональную окраску текста по отношению к объекту[28].

3.1.1 Стадии работы системы I-Teco

1. На первой стадии отрабатывает лингвистический модуль: производится лемматизация для всего документа, затем морфологический анализ, также учитывается характеристика (падеж, число, лицо);
2. На втором этапе словам и словосочетаниям задается тональность отрицательная либо положительная, по заранее составленным словарям. В случае, если словосочетание либо слово не нашлось ни в одном из словарей, то такому слову задается нейтральное значение;

3. Следующий этап — синтаксический анализ: выделяются субъекты и объекты, затем объединение слов и словосочетаний, определяются анафоры и обороты;
4. На завершающем этапе отмечается объект и производится анализ эмоциональной окраски, в соответствии от месторасположения выделенного объекта в документе.

3.1.2 Тональные словари

В процессе определения тональности теста, зачастую используют тональные словари. В системе I-Teco учитываются четыре словаря. Все они разделены по частям речи: глаголы, существительные, прилагательные, наречия и коллокации (устойчивые выражения, часто встречающиеся словосочетания) глагольные и неглагольные. Части речи разделены на подклассы.

Словари были составлены экспертно и постоянно пополняются. Было собрано большое количество данных из СМИ. После сбора было начислено около 100 млн. словоупотреблений, после обработки всех слов количество составило 15 тысяч слов и словосочетаний. Слово записывается в словарь исключительно в случае, если оно содержит в себе эмоциональную составляющую.

Слово не может встречаться в нескольких словарях. Они распределяются с учетом части речи и в зависимости от тональности. В алгоритме учитываются омонимии. Рассмотрим два предложения: «Болеть гриппом» и «Болеть за любимую команду». В обоих предложениях используется слово «болеть», но предложения различны по эмоциональной окраске. В этом случае применяется глагольное управление и коллокация.

3.1.3 Глагольные и неглагольные лексемы и коллокации

Зачастую слова не содержат в себе какой-либо эмоциональной окраски, но они могут сильно повлиять на слова, к которым они относятся. Например, «финальный бой», «архитектура». Подобного рода словосочетания сами по себе не определяют тональность, но могут повлиять на результат, в зависимости от их расположения в тексте. Рассмотрим имя существительное «архитектура», оно не содержит никакой окраски, но в тексте «Архитектура повысила производительность», она приобретает положительный характер.

Учитываются также отглагольные существительные, которые влияют на тональность слов, расположенных после них. Например, слово *прекращение,* в случае если, после него следует текст с негативной окраской, то тональность становится положительной, и наоборот, если текст негативный, то результат будет положительным. Например, «*Прекращение криминальных войн*» — положительный текст.

Для объектно-ориентированного определения тональности текста система применяет три компонента:

1. Определение эмоциональной окраски самого объекта;
2. Действие и поведение, относящееся к объекту;
3. Определение тональности остальной части текста.

Выделенный объект, в предложении, определяется двумя параметрами:

1) Отношение объекта к остальной части предложения;
2) Значение объекта относительно глагола.

Указанные параметры могут не всегда влиять на тональность текста. Эмоциональная окраска текста зависит непосредственно от глагола. В системе I-Teco выделены некоторые классы глаголов:

1) 1 и 2 класс — негативные и позитивные глаголы, где эмоциональная окраска объекта обуславливается окружающей средой выделенного объекта;
2) 3 и 4 класс — негативные и позитивные глаголы, где эмоциональная окраска объекта не зависит от окружающей среды;

3) 5 и 6 класс — негативные и позитивные глаголы, где эмоциональная окраска объекта зависит от окружающей среды и от его роли в предложении;
4) 7 и 8 класс — негативные и позитивные глаголы, где эмоциональная окраска объекта определяется в независимости от окружающей среды и роли;
5) 9 класс — глаголы объединяющие тональность субъекта и объекта;
6) 10 и 11 класс — негативные и положительные глагольные коллокации (проламывать сквозь, сразить врага).

Для всех отдельных глаголов и для глагольной коллокации была приписана сила тональности. В случае если, тональность текста зависит исключительно от глагола, то данное преобразование поможет в определении тональности контекста.

3.1.4 Коллокаций и лексемы

В данном разделе рассматривается модуль для определения правил сочетаемости лексем и коллокаций. Неглагольные слова и словосочетания из контекста, объединяются в пары, после этого процесса присоединяются к глаголу. Предусмотреть все правила, очень сложно: например, если в предложении встретятся два слова, с положительной и отрицательной окраской, то система не определит тональность. Подобного типа словосочетания, которые часто встречаются в предложениях, добавляются в список коллокаций.

Также система учитывает анафоры, отрицания — *нет, не, без*, различные комбинации членов предложения. При необходимости большие предложения разбиваются на более простые. В дальнейшем они описываются в виде структуры, каждый элемент этой структуры является последовательной цепочкой словоформ с определенным сентиментом.

3.1.5 Объектно-ориентированный подход

Объектно-ориентированный подход, как упоминалось ранее, играет важную роль при определении тональности. Объект может быть задан вручную, либо система может определить его автоматически в процессе анализа. В автоматическом режиме в предложении ищется существительное неодушевленное и одушевленное.

При любых обстоятельствах система предусмотрена для работы с одним объектом, учитывая и тот случай, если в предложении автоматически выделилось два из них.

В системе предусмотрено свыше двадцати правил, определяющих тональность текста по отношению к объекту. В зависимости от его расположения и роли в предложении, определяется тональность, с учетом всех правил.

Было выбрано 40 предложений для тестирования. Все текста выбирались вручную с различными комбинациями слов и коллокаций.

Обозначения:

1. adj — прилагательное;
2. noun — существительное;
3. verb — глагол;
4. invertor — отрицание;
5. prep — предлог;
6. pos — позитивный;
7. neg — негативный;
8. negp — чисто негативный;
9. ppos — потенциально позитивный;
10. posp — чисто позитивный;
11. neut — нейтральный;

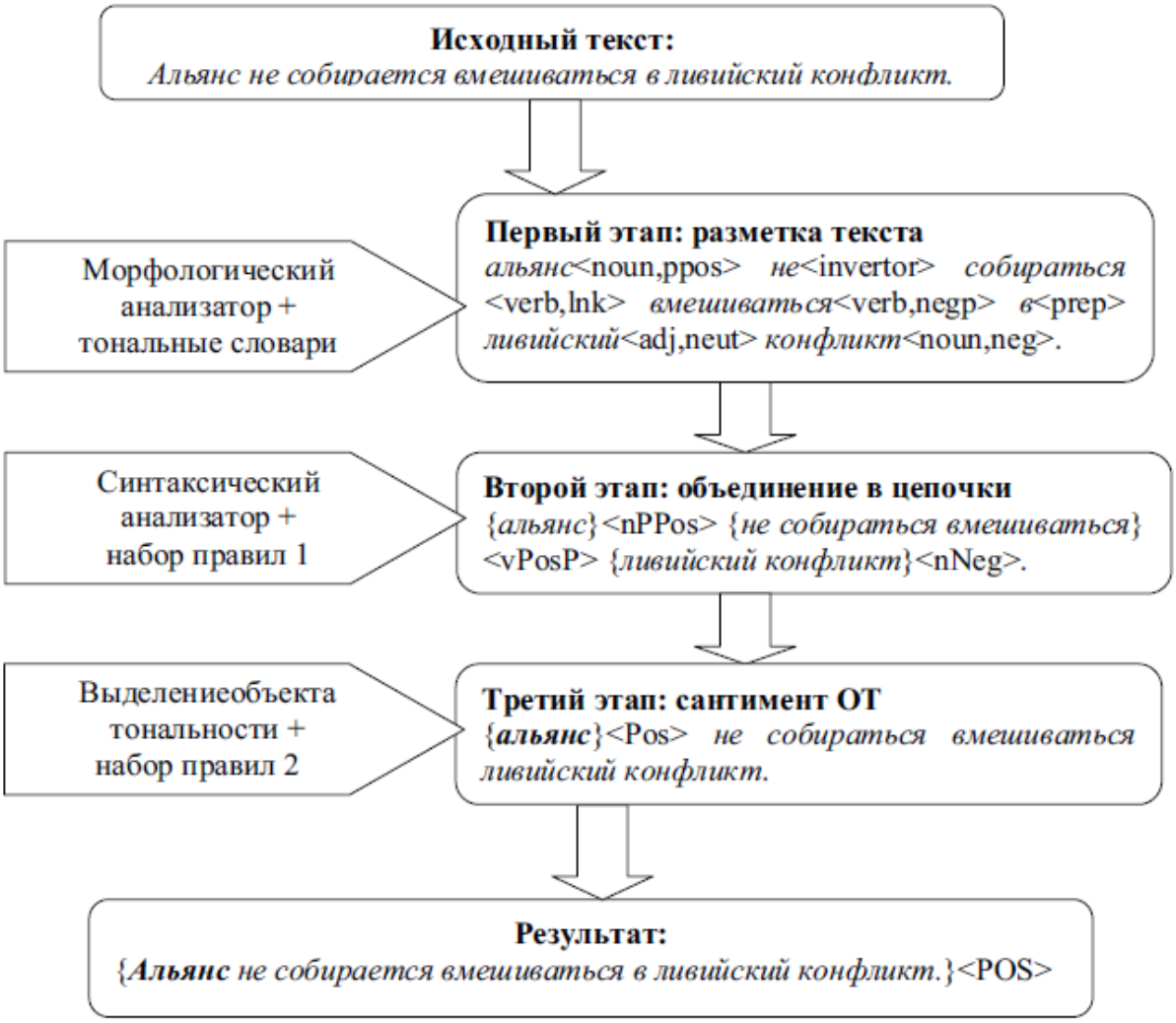


Рис. 4 Последовательные этапы анализа текста для определения тональности.

На (*рис. 4)* Схематичное представление определения тональности предложения.

3.1.6 Оценка результата

Точность работы алгоритма оценивается со стороны эксперта. Эксперту выдается коллекция текстов с определенной тональностью. Его цель состоит в том, чтобы определить насколько точно система справилась с задачей определения эмоциональной окраски текста. В случае, если эксперт не согласен с решением системы, помечает тип ошибки:

1. Вместо того, чтобы пометить предложение как негативное, система определила его положительным и наоборот;
2. Предложение, где отсутствует тональность, система пометила как тональное;
3. Пропуск тональности — система не определила эмоциональную окраску текста.

Для подсчета точности и полноты, вводятся обозначения

1) A — количество документов, где система определила тональность верно;
2) B — пропуск тональных предложений, где находится объект;
3) C — некорректное определение тональности;
4) D — отсутствие объекта или тональности в тексте.

Формулы для точности и полноты:

$$\text{полнота} = \frac{A}{A + B + C}$$

$$\text{точность} = \frac{A}{A + C + D}$$

Точность системы равняется значению от 85% до 90%, полнота варьируется от 75% до 80%. Метод основан на лингвистическом анализе текста для русского языка.

3.2 Система определения тональности SentiScan

SentiScan — система для автоматического анализа тональности текста для русских и англоязычных текстов. Авторы продукта используют гибридный метод, в своем решении они применяют машинное обучение (англ. machine learning) и метод, основанный на правилах (англ. rule-based approach). Плюс данной технологии заключается в объектной ориентированности, т.е. данная система может определять окраску текста по заданному объекту. Задача SentiScan найти заданный объект в документе или в предложении, выделить контекст и распознать сентимент по отношению к объекту[10].

3.2.1 Главные элементы системы

SentiScan использует тональные словари для определения тональности отдельных слов в документе. Два отдельных словаря для положительных и отрицательных слов. Каждый из них представлен в виде леммы. Лемматизация (англ. lemmatization) — это один из методов морфологического анализа, который приводит слово к ее изначальной словарной форме (лемме). Например, «важнейшие» преобразуется в «важн». Для формирования словаря для русского языка, с первоначальной формой слова, используется грамматический словарь, составленный Зализняком А. А., в котором содержится приблизительно 100 тыс. словоформ.

Тональные словари заполняются в полуавтоматическом режиме. В процессе заполнения, разработчики поставили перед собой задачу добавлять исключительно те слова, которые содержат в себе однозначную тональность. Например, такие как: «превосходный», «унылый» и т.д.

Алгоритм для анализа текста и выявления общей эмоциональной окраски представлен в виде псевдокода (*приложение 1)*.

Основные блоки системы для определения общей тональности теста:

1. Наличие словарей: тональные словари для негативных и позитивных слов;
2. Принимать во внимание отрицания, которые влияют на определение тональности текста. В случае, если в предложении встречается отрицание, слово, расположенное за ним, автоматически меняет свою эмоциональную окраску. Например, «Мне не нравятся телефоны с кнопками». Отрицание «не» меняет всю тональность предложения;
3. Противительные союзы также играют важную роль при определении окраски текста. В системе учитывается данный кейс, что повышает точность определения. Например, «Моему другу нравится телефон с кнопками, а мне нет, так как он выглядит несовременно!".

Объектно-ориентированный подход для определения тональность текста.

1. На первом этапе производится поиск на существование указанного объекта в предложении;
2. Второй этап — в случае, если заданный объект не был найден в предложении, то запускается процесс на проверку существования противительных союзов, разделителей, например, пунктуационные знаки;
3. Если, указанные действия из второго пункта не были определенны системой, то срабатывает алгоритм из (*приложения 1)* для определения общей тональности предложения;

4. В случае, когда системе удалось успешно обозначить предложение с заданным объектом, применяется алгоритм (*приложении 1)* для выделенного предложения.

Также система использует модуль для определения анафор в предложении, что играет большую роль в случае определения сентимента по заданному объекту. Например, «Коллеги были рады встретить Ольгу из отпуска, но, к сожалению, она этого не поняла». В примере местоимение «она» относится к Ольге.

Рассмотрим еще один пример: «Андрей написал письмо Ольге, чтобы она переслала его Алексею». Для разбора данного предложения применяется структурный подход и система строит гипотезу: местоимение «она» женского рода, именительный падеж единственного числа, относится к Ольге, учитывая то, что Ольга отсутствует во второй части предложения. Использование анафор помогает придерживаться логической цепи, тем самым повышает качество определения сентимента.

3.2.2 Оценка качества работы системы SentiScan

Для расчета оценки качества применяются метрики: точность (англ. precision), полнота (англ. recall) и F-мера.

Точность системы — доля документов, которые действительно относятся к определенному классу, относительно всех объектов отнесенных методом к этому классу. Полнота — доля найденных системой объектов, которые принадлежат классу относительно всех объектов этого класса. *F-мера* — объединение точности и полноты, для определения баланса.

Качество алгоритма для двух классов оценивается следующим образом:

$$P = \frac{P_N + P_P}{2}, \qquad R = \frac{R_N + R_P}{2}, \qquad F = \frac{F_N + F_P}{2}$$

Вычисление для позитивного класса:

$$P_P = \frac{tp}{tn + fn}, \qquad R_P = \frac{tn}{tn + fp}, \qquad F_P = 2 * \frac{P_p + R_p}{P_p * R_p}$$

1. tp — истинно-положительное (верно определено в позитивный класс);
2. tn — истинно-отрицательное (неправильно определено);
3. fp — ложно-положительное (верно определено в негативный класс);
4. fn — истинно-отрицательное (неправильно определено).

Вычисление для негативного класса:

$$P_N = \frac{tn}{tn + fn}, R_N = \frac{tn}{tn + fp}, F_N = 2 * \frac{P_N + R_N}{P_N * R_N}$$

Общая точность, вычисляется по формуле:

$$A = \frac{tp + tn}{tp + tn + fp + fn}$$

P	R	F	A	P_p	R_p	F_p	P_N	R_N	F_N
0.5845	0.6581	0.5574	0.6378	0.9171	0.6288	0.7461	0.2519	0.6875	0.3687

Рис. 3 Результаты системы для двух классов

Подводя итоги системы SentiScan, следует отметить, что система использует подход, основанный на правилах, которые позволяют проводить глубокий анализ предложения и строить гипотезы. Также системы отлично справляются с задачей определения тональности текста с заданным объектом, т.е. учитывается объектно-ориентированность. Разработанный

алгоритм, демонстрирует высокие результаты в случае определения положительной тональности, 90% (*рис. 3*). Общая полнота (R) — 65%, а точность (A) — 63%.

3.3 Система Sentiment Analysis Service

Автоматическое определение эмоциональной окраски текста подразумевает обозначение тех фрагментов текста, которые содержат в себе негативную или позитивную составляющую по отношению к объекту. В качестве объекта могут выступать имена собственные, название продукта, услуги, организации, профессии, по отношению к которым производится анализ текст. Объект может быть обозначен как один в целом документе, с учетом анафорических и синонимических употреблений, или определяться в предложениях как имя нарицательное или собственное. Тональность текста можно определить тремя факторами: объект тональности, субъект тональности, тональная оценка. Под субъектом тональности предполагается автор текста, под объектом — тот, о ком высказывается автор и тональная оценка — мнение автора к указанному объекту[7].

Сентимент анализ, разработанный данной системой, можно разделить на следующие этапы.

1. На первом этапе отрабатывает лингвистический модуль, который производит морфологический анализ, лемматизацию и определяет части речи для каждого слова, его морфологические характеристики (число, лицо, падеж). Также определяется роль слова в предложении: для существительных — подлежащие, дополнение, обстоятельство, для глаголов — деепричастие, причастие и тип слова. Например, для существительных: юридическое лицо, физическое лицо, географическое название и др.;

[7] https://github.com/zamgi/SentimentAnalysisService, 03.06.2018

2. Затем все слова и словосочетания (коллокации) размечаются по заранее составленным словарным спискам тональной лексики. Каждое слово имеет два атрибута: тональность и/или сила тональности. Если слово не удалось найти в списках тональной лексики, то оно помечается как нейтральное;
3. На третьем этапе отрабатывает модуль — синтаксический анализ: строится цепочка из слов и словосочетаний и выделяются субъект, объект и предикат, определяются причастные и деепричастные обороты, анафорические связи, подчинительные предложения. Учитываются предложения обобщенно-личные и неопределенно-личные, также с нулевой формой глагола, сказуемые, выраженные неглагольной формой;
4. Последний этап — выделение объекта тональности и определение его сентимента в зависимости от роли и местоположения этого объекта в предложении.

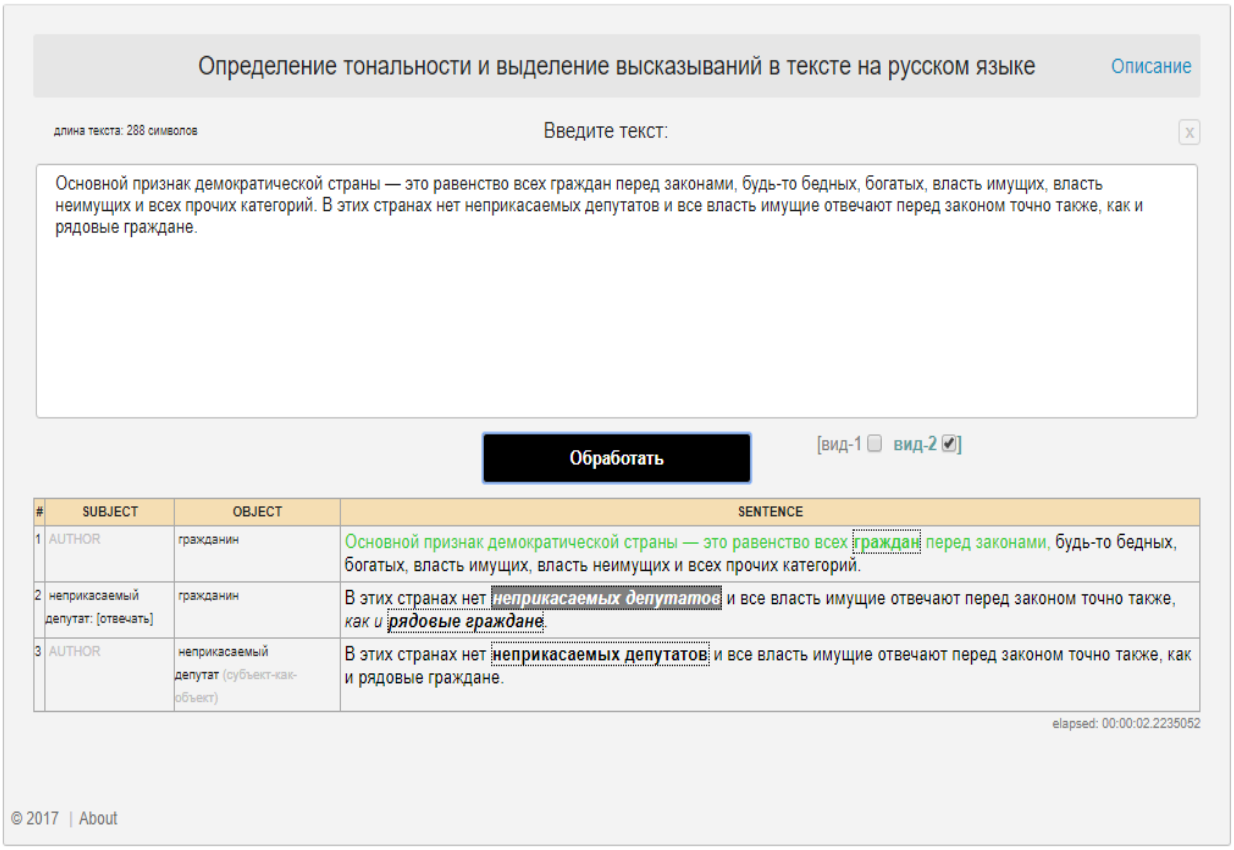


Рис. 5 Интерфейс приложения Sentiment Analysis Service

3.4 Выбор инструмента тональности

Инструмент *I-Teco* основан на наборе правил (англ. Rule-based approach). В системе производится глубокий лингвистический анализ. Используются различные словари для разных частей речи: как для позитивных слов, также и для отрицательных. Система ориентирована для определения тональности в *средствах массой информации*. Один из плюсов системы — определение тональности по отношению к объекту. Также система автоматически определят объект в предложении, если он заранее не был указан.

Система SentiScan использует гибридный подход для определения тональности текста: лингвистический анализ и машинное обучение. Данная система хорошо справляется с определением тональности текста для различных сфер деятельности.

Система Sentiment Analysis Service использует подход, основанный на правилах. Технологии SentiScan и I-Teco являются коммерческими, а Sentiment Analysis Service — продукт с открытым исходным кодом. У коммерческих решений существуют тестовые режимы, где можно провести тестирование с небольшим количеством данных.

В данной работе протестированы все указанные системы, но за основу было взято решение Sentiment Analysis Service, так как программное обеспечение позволяет делать неограниченное количество запросов.

Глава 4. Практическая реализация

В четвертой главе будут представлены инструменты, которые использовались для разработки программного обеспечения. Можно будет ознакомиться с архитектурой приложения и всеми этапами ее реализации. Также описан пользовательский интерфейс для использования системы.

4.1 Архитектура приложения для сбора и анализа данных

В данном разделе отображен каждый модуль архитектуры см. (*рис. 6*). Инструменты, выбранные для решения поставленных задач, будут изложены в подглаве **4.2**.

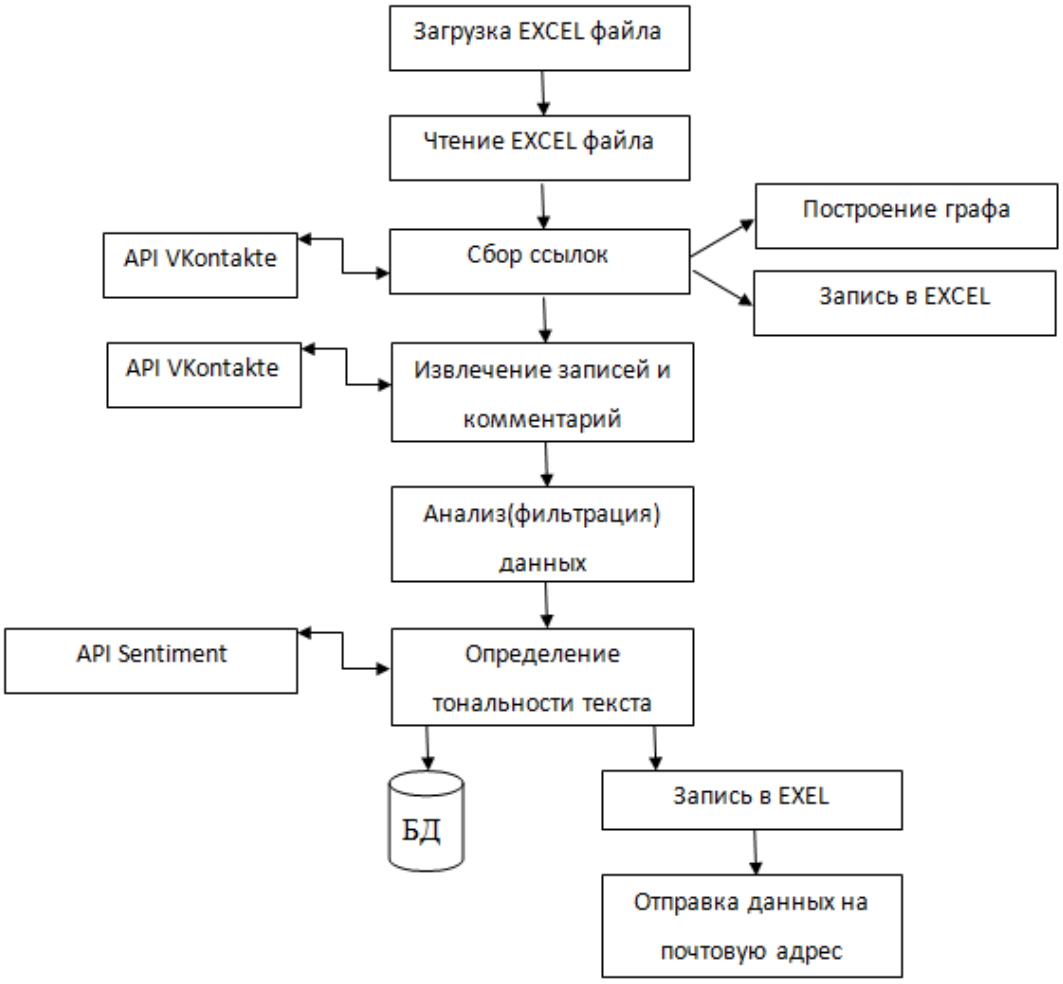


Рис. 6 Архитектура приложения

1. ***Загрузка файла.*** Для использования программы, необходимо загрузить файл в систему, который содержит в себе различные группы социальной сети «ВКонтакте»;
2. ***Чтение файла***. После успешной загрузки файла на сервер, система начинает его считывать, затем выделяются ссылки на группы и помещаются в очередь;
3. ***Сборщик ссылок***. В группах сети «ВКонтакте» существует отдельный раздел, который называется «*ссылки».* В разделе указаны ссылки на другие сообщества в сети с релевантной тематикой. Модуль учитывает только те из них, которые находятся в загруженном файле, все остальные игнорируются. Система их исключает с целью выявить взаимосвязь между сообществами, которые были указаны для анализа данных, и построить граф со связями. Дополнительные ссылки можно получить при помощи «API VKontakte»;
4. ***Взаимосвязи сообществ в виде графа***. После завершения работы модуля, отвечающего за сбор ссылок, начинается процесс визуализации и построения графа. Полученный объект передается вместе со ссылками библиотеке Graph Library Dracula для визуализации графа. Пример графа *(см. приложение 2)*;
5. ***Взаимосвязи сообществ в табличном виде.*** Помимо визуализации в виде графа, также взаимосвязь представляется в виде квадратной матрицы $A_{n\times n}$ и записывается в Excel. В случае, если матрицы пересекаются, ячейка заполняется цифрой «1», в ином случае поле остается пустым;
6. ***Записи и комментарии.*** Используя «API VKontakte» извлекаются посты и комментарии к ним. Количество постов можно указать в интерфейсе. Сначала собираются посты, затем комментарии к каждому их них. Для каждой записи извлекается $\underline{n}$-е количество высказываний. По окончанию работы модуля получаем объекты с записями и

комментариями. Каждый объект представляет из себя сообщество социальной сети;

7. ***Анализ и фильтрация данных.*** В пользовательском интерфейсе можно задать объект, по которому будет происходить анализ тональности текста. Все предложения приводятся к нижнему регистру и для каждого слова выделяется его базовая форма. Затем, если пользователь указал конкретный объект, то идет поиск предложения, текстов, где он присутствует;
8. ***Сентимент анализ.*** Полученные данные из вышеопределенного пункта необходимо проверить на эмоциональную окраску и поместить в соответствующий класс: позитивный, негативный или нейтральный. Для определения тональности текста, будет использована система, построенная на лингвистическом анализе. В результате работы модуля, будут сформированы три документа для *позитивных, негативных и нейтральных предложений*;
9. ***Запись результатов в базу данных.*** После успешного анализа информации, результаты необходимо записать в базу данных. В текущей работе используется ООБД (Объектно-ориентированная база данных). Все модели представлены в виде объектов;
10. ***Запись результатов в Excel файлы.*** Соответственно, полученные результаты записываются в три документа: файлы с нейтральными, негативными и позитивными предложениями;
11. ***Отправление файлов на электронный адрес.*** На последнем этапе, полученные результаты отправляются на электронную почту. Электронный адрес можно заранее указать в интерфейсе.

4.2 Используемые системы и инструменты

Для реализации программного обеспечения, сбора и анализа данных были использованы следующие инструменты:

1. Программный код был написан в среде разработки *IntelliJ IDEA* [38]. Среда поддерживается на всех современных операционных системах. Для данного инструмента написано множество расширений, которые облегчают работу при разработке. Поддерживает большое количество языков программирования: Java, , Python, JavaScript и т.д.;
2. Большая часть кода была реализована на языке программирования *Java* [12]. Выбор был сделан в пользу данного языка по ряду причин: объектно-ориентированного подхода, который позволяет сохранять модульность приложения с целью разделения логики, доступно множество разработанных пакетов и библиотек, которые упрощают процесс разработки. Код, написанный на языке Java, транслируется в байт-код, который можно запустить на любой современной компьютерной архитектуре. Для этого достаточно установить *Java Virtual Machine*. Java предназначен для высоконагруженных систем и является одним из востребованных в сфере информационных технологий;
3. *Сервлет (англ. Servlet)* — простой Java интерфейс, реализация которого расширяет функциональные возможности сервера. Сервлет работает с клиентом посредством принципа запрос-ответ (*см рис. 7*). Процесс работы можно представить следующим образом: как только сервлет получает запрос со стороны клиента, веб-сервер использует файл с определенной конфигурацией, тем самым определяя, какой сервлет должен обработать запрос клиента [39]. Затем, запускается JVM, который выполняет код сервлета. Задача сервлета — сгенерировать страницу, размеченную HTML тегами и передать ее веб-серверу. Сервер должен отправить клиенту HTML страницу;

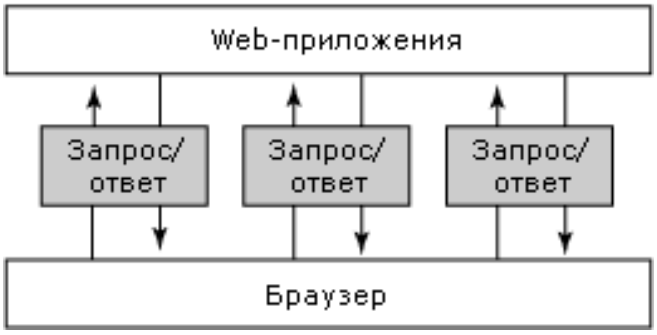


Рис. 7 Сервлет взаимодействие с клиентом

4. Сервлеты получают запросы со стороны клиента, затем выполняют некоторую работу и выводят результаты на экран. В сервлете можно обработать сложную логику и выполнять запросы к базе данных, что является необходимым для приложения. Но выводить данные на экран из самого сервлета неудобно. Для отображения вывода данных на пользовательском интерфейсе используется технология JSP (от англ. Java Server Pages). Технология является расширением сервлетов для упрощенной работы с веб-содержимым. Используя JSP, легко можно разделить статическую и динамическую часть приложения. Статическую часть приложения можно описать HTML тегами;
5. Для получения данных из объекта Java на страницах JSP, используют *скриплеты* — прямые вставки Java кода на страницах. Но данный подход является анти-шаблоном и на практике применяется крайне редко. Более распространенное решение — библиотека JSTL *(англ. Java Server Pages Standard Tag Library),* которая помогает избежать написания чистого Java кода внутри JSP файлов, заменяя их специальными тегами для получения информации из объектов;
6. В качестве веб-сервера используется контейнер сервлетов *Apache Tomcat.* Инструмент обеспечивает работу статической и динамической части приложения. Tomcat разработан на языке Java[47];

7. *Dracula.js* — набор инструментов для компоновки и отображения интерактивных графиков, а также разных алгоритмов из области теории графов. В реализации используется простой JavaScript и векторная графика [37];
8. *Vis.js* — библиотека визуализации для браузеров. Инструмент предназначен для обработки больших объемов динамических данных и их манипулирования [40];
9. Для придания приложению интерактивности, используется объектно-ориентированный язык программирования *JavaScript*. В текущей работе, *JavaScript* применяется для подключения сторонних библиотек с целью реализации визуального графа;
10. В данном пункте рассматривается текстовый формат JSON. Формат используется для хранения данных. Любой JavaScript объект можно представить в виде текста. Но, не смотря на то, что формат основан на JavaScript, его активно используют практически все языки программирования, так как зачастую данные, которые возвращает веб-сервер клиенту, представлены в виде формата JSON [50];
11. *API VKontakte* — это программный интерфейс, с помощью которого, можно получить информацию из базы данных социальной сети «ВКонтакте», используя http-запросов к специальному серверу. Тип и синтаксис запросов, возвращаемый с API, строго определен на стороне сервиса. При помощи инструмента можно разработать приложение, которое дает возможность получить доступ к публичной информации из базы данных vk.com. Например, данные из сообществ социальной сети, в случае если они не закрыты со стороны администратора группы. В данной работе использовались следующие методы: *group.getById* — информация о сообществе (можно указать сразу несколько групп), *wall.get* — список постов со стены сообщества либо пользователя,

wall.getComments — список комментариев к определенным постам. Результат запроса API возвращается в двух форматах: XML(EXtensible Markup Language) и JSON.[42];

12. Russian Sentiment API SentiScan — система для автоматического определения эмоциональной окраски текста с поддержкой русского языка. Разработка является коммерческой, но систему можно использовать в бесплатном режиме с определенными ограничениями [43]. Система предоставляет удобный программный интерфейс для осуществления запросов;
13. С*теммер Портера для русского языка* — это алгоритм стемминга. Оригинальная версия была разработана для английского языка. *Стемминг* — процесс нахождения базовой формы (основа) слова. Базовая форма может не совпадать с морфологическим корнем слова. Если применить стемминг к слову «валялись», то получим основу «валя». [45];
14. *API JavaMail* — программный интерфейс для отправки и получения электронной почты. Инструмент использует протоколы POP3, SMTP, и IMAP [46]. JavaMain — это часть Java Enterprise Edition;
15. *DB4O (Database for Objects)* — система управления баз данных для объектов с публичным доступом [49]. Инструмент используется в языках программирования Java и C#. Система позволяет разработчикам использовать вышеуказанные языки программирования для написания запросов к базе данных, тем самым, исключая возможность SQL-инъекций.

4.3 Пользовательский интерфейс для сбора и анализа данных

Одной из главных целей данной работы является создание пользовательского интерфейса для использования программным обеспечением. Веб-интерфейс был разработан с применением технологий HTML, CSS, JavaScript и ReactJS. Приложение состоит из четырех веб-страниц:

1. На первой странице сайта представлена вводная информация о ресурсе и функциональная возможность;
2. Вторая страница отвечает за загрузку файла со списком ссылок на сообщества социальной сети «ВКонтакте». Также предоставляется опциональная возможность с указанием параметров для сбора информации: количество записей со стены сообщества социальной сети, количество сообщений к каждой из записи, электронная почта;
3. На третей странице веб-приложения размещена документация по использованию данного ресурса, т.е. в каком виде должны быть записаны ссылки в *Excel* файл, для дальнейшего анализа, за что отвечает каждый из параметров;
4. Последняя страница — форма обратной связи, где пользователь системы имеет возможность обратиться к разработчику приложения.

Пользовательский интерфейс интуитивно понятный и простой в использовании (*приложение 3)*.

4.4 Тестирования для приложения «ВКонтакте»

В ходе тестирования приложения проводилось несколько замеров на время работы системы с разными входными параметрами. Качество работы сервиса Sentiment Analysis Service, для определения эмоциональной окраски текста, протестированы на двух коллекциях с отрицательными и

положительными предложениями. В каждом корпусе содержатся по 50 предложений.

Тестовая выборка была подобрана из социальной сферы: совокупность отраслей, организаций, предприятий, задачей которых является улучшение уровня жизни населения. К социальной сфере относится социальное обеспечение, здравоохранение, коммунальное обслуживание и т.д. Результат тестирования показал, что точность определения сентимента составляет 81%.

Замеры времени:

Построение визуального графа занимает *8 секунд*.

Количество постов с каждого сообщества	Количество сообщений для каждого поста	Число сообщений со всех сообществ	Число комментарий со всех сообществ	Время, затраченное на обработку в минутах
200	200	27513	31273	18,2
400	200	294907	53767	163,85
800	200	100210	87132	321

Таблица 1. Замеры работы системы по времени

В таблице указаны временные замеры с целью сбора информации со 100 сообществ социальной сети «ВКонтакте». Программный интерфейс «API VKontakte» позволяет отправлять 3 запроса в течение одной секунды. Соответственно, время работы выполнения по сбору необходимых данных напрямую зависит от сервера.

4.5 Определения тональности для приложения «Сезам»

Одна из поставленных задач для данной работы — создание программного обеспечения для системы обмена сообщениями «Сезам».

Сезам — первое приложение на русском языке, с открытым исходным кодом, которое дает возможность, обмениваться специализированными пиктограммами [41]. Из схематических картинок с подписями можно составлять полноценное предложение. В приложении доступно свыше 500 черно-белых пиктограмм международного формата. Сезам имеет простой и удобный интерфейс. Данное приложение прошло тестирование на группе детей с аутизмом и получила положительные отзывы.

Как было указано выше, данный проект разрабатывался для людей с ограниченными возможностями, и некоторые из них находятся под наблюдением врачей. Анализ тональности текста поможет выявить, насколько положительно влияет данный вид общения, и как правильно пациент использует инструмент для общения с друзьями и родственниками.

Авторами проекта была предоставлена коллекция из пиктограмм и, соответствующие слова и словосочетания к ним, а также обучающая выборка из 850 предложений. Экспертным путем каждому слову было присвоено значение (вес слова) от -5 до 5 в зависимости от его эмоциональной окраски. Значение задавалось словам с однозначной окраской, например, «отлично с весом 5» и «ужасно с весом -5 ». Словам, не содержащим в себе сентимента, присваивалось значение 1. Все предложения были проанализированы и разделены на два класса. Корпус предложений с положительной окраской и отрицательной.

С целью решения задачи, было выбрано два подхода:

- Метод, основанный на словаре;
- Машинное обучение (Байесовский классификатор).

Разработан простой пользовательский интерфейс, который позволяет проверять эмоциональную окраску предложения, а также добавлять новые предложения в тестовую выборку. Для ввода в интерфейсе предназначено два поля: первое для положительных предложений, второе для отрицательных. После внесения алгоритм снова начинает обучаться на новой выборке.

В интерфейс добавлены сторонние ресурсы для автоматического анализа данных и выявления эмоциональной окраски контекста. Это сделано с целью сравнения полученных результатов разработанной системы и существующих инструментов.

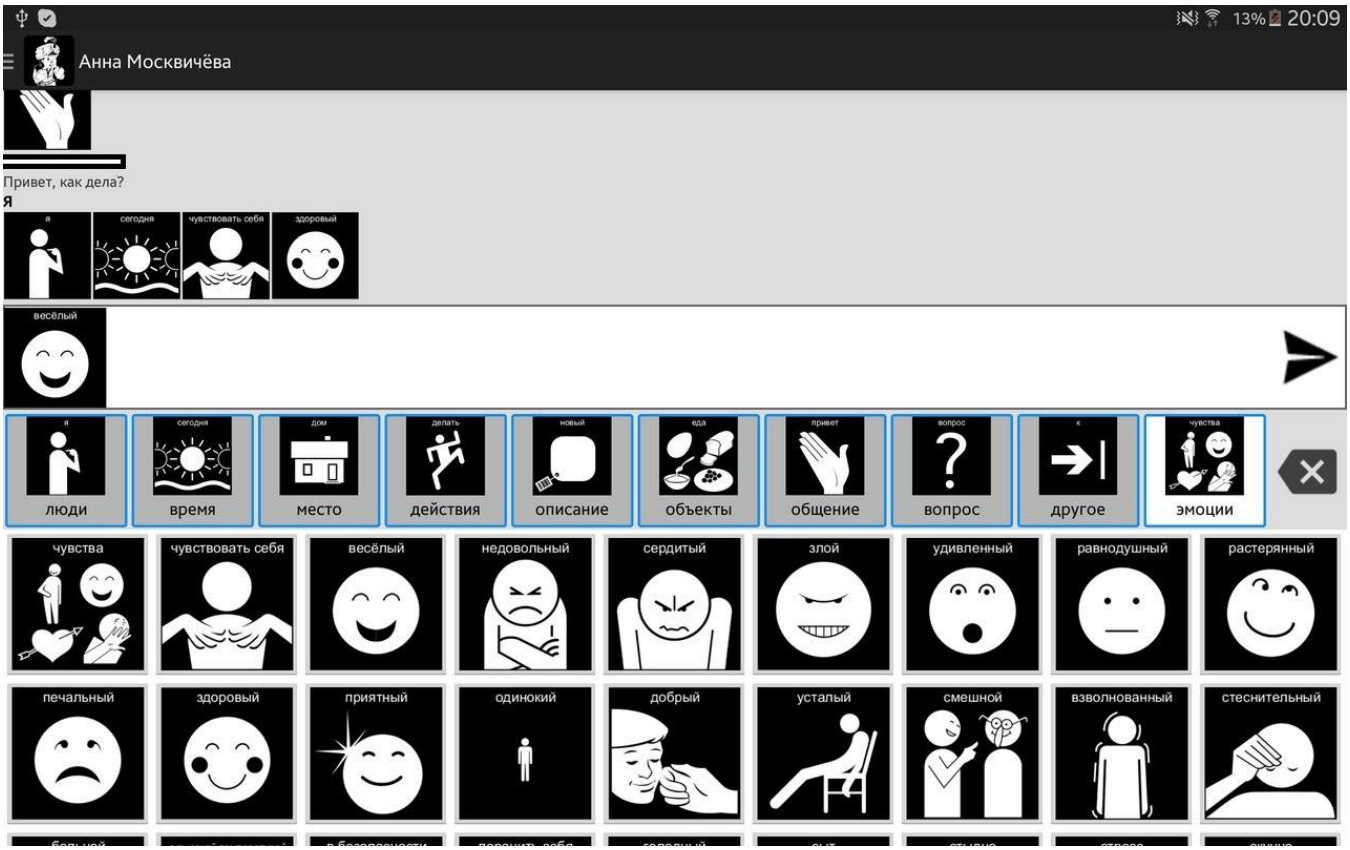


Рис. 7 Сервлет взаимодействие с клиентом

4.6 Тестирование системы и выводы

Для тестирования приложения, разработанного на основе выборки сообщений из программы «Сезам», были выбраны 50 положительных предложений и 50 негативных. Также, в качестве сравнения результатов, выбранные предложения, были протестированы в системе SentiScan, которая использует гибридный подход (*лингвистический анализ, тональные словари и машинное обучение*) для определения тональности текста.

Точность работы составляет:

	Naive Bayes	Dictionary	SentiScan
positive	0.88%	0.84%	0.92%
negative	0.76%	0.79%	0.69%

Для определения точности, вводятся обозначения:

1) A — число документов, где система определила тональность верно;
2) B — количество документов, где система неверно определила сентимент;
3) C — пропуск тональных предложений

$$\text{точность} = \frac{A}{A + B + \text{C}}$$

Тестовая выборка была использована в рамках данного исследования. Разработанная система, на данный момент, не интегрирована в приложение «Сезам».

В рамках данной работы было реализовано два приложения:

- Первое приложение является комплексной системой для сбора и анализа информации социальной сети «ВКонтакте». Данная разработка является совокупностью из существующих инструментов и отдельно разработанного модуля с целью сбора информации из сообществ социальной сети. Программное обеспечение может быть использовано в разных сферах деятельности. Например, в отрасли предоставления услуг и товаров, для выявления мнения потребителей по отношению к определенной продукции;
- Второе приложение разработано с целью определения эмоциональной окраски текста для приложения «Сезам». Приложение предназначено для людей с ограниченными возможностями и различными отклонениями. Общение между пользователями происходит при помощи специализированных пиктограммам. Пользователи приложения, могут находиться под присмотром врачей и коммуницировать с ними посредством данного мессенджера. Учитывая этот факт, на основе составленных предложений, можно проводить сентимент анализ с целью выявления настроения пациента, его предрасположенности к общению в зависимости от различных факторов. Данное приложение является исследовательским и на данных момент не применяется на практике.

Вывод

Проведенное исследование задачи автоматического определения эмоциональной окраски текста позволило прийти к следующим выводам:

1. В течение последнего десятилетия сентимент анализ вызвал большой интерес среди исследователей и имеет свое практическое применение в разных сферах деятельности: маркетинг, экономика, психология, политология и т.д.;
2. Существует несколько подходов для решения задачи сентимент анализа:
    - Машинное обучение с учителем;
    - Машинное обучение без учителя;
    - Подход, построенный на словаре;
    - Подход, построенный на правилах;
    - Объектно-ориентированный подход;
    - Гибридный подход.
3. Обзор существующих инструментов, показал, что, в основном, готовые решения предназначены для работы с английским языком. Также существуют разработанные системы, определяющие сентимент для русскоязычных текстов. Практически все программные интерфейсы и программные обеспечения для русского языка являются платными;
4. Коммерческие системы имеют бесплатные (тестовые) режимы, с определенными ограничениями, которые не позволяют работать с большим количеством данных. Например, система SentiScan позволяет делать 3000 бесплатных запросов в месяц, а Indico 10000 запросов.

В большинстве коммерческих систем для определения тональности текста используют гибридный подход. Данный подход содержит в себе методы машинного обучения, лингвистический анализ и применение специальных словарей. Лингвистический анализ текста, позволяет системе, автоматически обозначать объект, если он заранее не был задан пользователем и определяет сентимент по отношению к нему. Лингвистический анализ учитывает анафорические ссылки, противительные союзы, отрицания, что повышает качество определения сентимента.

На практике наиболее часто встречаются решения с машинным обучением, которые показывают высокие результаты, достигающие точности свыше 90% для определения тональности текста [14]. В текущей работе, для реализации практической части, используется метод машинного обучения без учителя, и метод, основанный на тональном словаре. Для автоматического определения эмоциональной окраски в тексте используется наивный байесовский классификатор.

Заключение

В рамках текущей работы была рассмотрена задача автоматического определения тональности текста, исследована предметная область, существующие инструменты и методы для решения поставленной задачи. Реализованы программные обеспечения для автоматизированного сбора данных и проведения анализа тональности текста

В ходе данной работы были решены следующие задачи:

1. Исследование предметной области: существующие методы и подходы для реализации сентимент анализа;
2. Рассмотрены часто используемые алгоритмы и инструменты для классификации данных;
3. Протестировано несколько инструментов для автоматического определения эмоциональной окраски русскоязычного текста;
4. Реализован инструмент для автоматического сбора данных с публичных сообществ социальной сети «ВКонтакте»;
5. Для выявления взаимосвязей между сообществами строится визуальный граф;
6. Полученные данные, по средствам существующих инструментов классифицируются на классы: негативные, позитивные и нейтральные;
7. Разработан пользовательский интерфейс для работы с программным обеспечением;
8. На тестовой выборке из системы обмена сообщениями «Сезам» — приложение для людей с нарушениями письма или речи, был реализован наивный байесовский классификатор. Полученное программное обеспечение определяет сентимент анализ. Для использования программы создан простой пользовательский интерфейс.

Учитывая полученные результаты, можно прийти к выводу, что поставленные задачи в текущей работе полностью выполнены.

Список цитируемой литературы

1. Прикладная и компьютерная лингвистика / Под ред. О. В. Митрениной, И.С Николаева, Т. М. Ландо. Изд. 2-е — М.:ЛЕНАНД, 2017. — 320с.
2. Автоматическая обработка текстов на естественном языке и анализ данных / Под ред. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. 2017. — 269с.
3. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012 P. 10 - 108.
4. Bing Liu. Sentiment Analysis: A Multi-Faceted Problem // IEEE Intelligent Systems, 2010.
5. Bing Liu. Sentiment Analysis and Subjectivity // Handbook of Natural Language Processing, 2010.
6. Jiang L., Yu M., Zhou M., Liu X., Zhao T. Target-dependent Twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, US. 2011. P. 151-160.
7. Popescu A., Etzioni O. and Opinions from Reviews // EMNLP, 2005.
8. Pang B., Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008 - P.1-135.
9. Singh P. K., Husain M. S. Methodological study of opinion mining and sentiment analysis techniques // International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014 P. 11 - 22.
10. Dmitry K. // Rule-based approach to sentiment analysis at ROMIP 2011
11. Walaa M., Ahmed H., Hoda K. // Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 2014 P. 1094 - 1115.
12. Bruse Eckel. Thinking in Java 4th ed. 2006.
13. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts // Stanford University

14. Mr. S.M. Vohra, Prof. J.B. Teraiya // A comparative study of sentiment analysis techniques, journal of information, knowledge and research in computer engineering, issn: 0975 – 6760| nov 12 to oct 13 p. 313-317.
15. Abrahamyan S., Balyan S., Muradov A., Kulabukhova N. and Korkho V.: A Concept of Unified E-Health Platform for Patient Communication and Monitoring // Lecture Notes in Computer Science, 2017. — Vol. 10408, — P. 448–462
16. Abrahamyan S., Balyan S., Muradov A., Vladimir Korkhov, Moskvicheva A., and Jakushkin O. Development of M-Health software for people with disabilities // Lecture Notes in Computer Science, 2016. — Vol. 9787, – P. 468–479
17. Фролов А.В.,Поляков П.Ю., Плешко В.В. Использование семантических категорий в задаче классификации отзывов о книгах// ООО «ЭР СИ О», Москва, Россия РОПИМ 2012.
18. Лукашевич Н.В. Четвёркин И.И. Открытое тестирование систем анализа тональности на материале русского языка // МГУ им. М.В.Ломоносова С. 1 - 9.
19. Е.В. Тутубалина, В.В. Иванов, М.А. Загулова, Н.Р. Мингазов, И.С. Алимова, В.А. Малых. Тестирование методов анализа тональности текста, основанных на словарях // Электронные библиотеки. 2015. Т. 18. № 3-4 139.
20. Абраамян С. А., Балян С. Г., Мурадов А. Г. Программное обеспечение совместного принятия решений на основе мобильных инфраструктур // Процессы управления и устойчивость, 2015. — Т. 2, — № 18. — С. 333–339
21. Ю. Лифциц. Классификация текстов. 2005.
22. Ю. Лифшиц. Метод опорных векторов 2006.

23. Аксенов А.В. Анализ тональности текстовых сообщений социальной сети Twitter // Научно-технический журнал «Теория. Практика. Инновации» 2016.

24. Фролов А.В.,Поляков П.Ю., Плешко В.В. Использование семантических категорий в задаче классификации отзывов о книгах// ООО «ЭР СИ О», Москва, Россия РОПИМ 2012.

25. И.И. Четверкин, Н.В. Лукашевич. Тестирование систем анализа тональности на семинаре // Семинар РОПИН 12.

26. И.И. Четверкин, Н.В. Лукашевич. Автоматическое извлечение оценочных слов для конкретной предметной области.

27. Ю. В. Адаскина, П. В. Паничева, А. М. Попов. Использование синтаксиса для анализа тональности твитов на русском языке.

28. Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке // "Диалог 2011". — С. 576 - 586.

29. Посевкин Р. В. Практическое применение методов компьютерной лингвистики // Университет ИТМО, г. Санкт-Петербург.

30. Юсупова Н. И., Богданова Д. Р., Бойко М. В. Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения // Математическое моделирование, численные методы и комплексы программ, УГАТУ, 2012 Т. 16, № 6 (51). С. 91–99.

31. Учителев Н.В. Классификация текстовой информации с помощью SVM // Кафедра «Предсказательного моделирования и оптимизации» МФТИ, ИППИ РАН.

32. Брунова Е. Г. Автоматизированный контент-анализ мнений трех предметных областей // Тамбов: Грамота, 2014. № 12 (42): в 3-х ч. Ч. II. ISSN 1997-2911 С. 43-47.

33. Котельников Е. В. Клековкина М. В. Автоматический анализ тональности текстов на основе методов машинного обучения.

34. Котельников Е. В. Клековкина М. В. Определение весов оценочных слов на основе генетического алгоритма в задаче анализа тональности текстов // Программные продукты и системы. №4 2013.

35. Осокин В.В., Шегай М.В. Анализ тональности русскоязычного текста

36. How to compute Tf-Idf. Режим доступа: http://www.tfidf.com/. Дата обращения 03.06.2018.
37. JavaScript Graph Library. Режим доступа: http://www.graphdracula.net/. Дата обращения 03.06.2018.
38. Среда разработки Intellij Idea. Режим доступа: https://www.jetbrains.com/idea/. Дата обращения 03.06.2018.
39. Сервлеты. Режим доступа: http://www.java2ee.ru/servlets/. Дата обращения 03.06.2018.
40. Vis.js A dynamic, browser based visualization library. Режим доступа: http://www.visjs.org/. Дата обращения 03.06.2018.
41. Описание приложения «Сезам». Режим доступа: https://hi-tech.mail.ru/news/Sezam_for_Android/. Дата обращения 03.06.2018.
42. API VKontakte Режим доступа: https://vk.com/dev. Дата обращения 03.06.2018
43. API Russian Sentiment Analyzer Режим доступа: https://www.mashape.com/dmitrykey/russiansentimentanalyzer. Дата обращения 03.06.2018.
44. API Indico for sentiment analysis. Режим доступа: https://market.mashape.com/indico/indico. Дата обращения 03.06.2018.
45. Russian stemming algorithm Режим доступа: http://snowball.tartarus.org/algorithms/russian/stemmer.html. Дата обращения 03.06.2018.
46. JavaMail API documentation Режим доступа: https://javamail.java.net/nonav/docs/api/. Дата обращения 03.06.2018

47. Apache Tomcat. Режим доступа: http://tomcat.apache.org/. Дата обращения 03.06.2018
48. Определение тональности и выделение высказываний в тексте на русском языке. Режим доступа: http://ston.apphb.com/index.html/. Дата обращения 03.06.2018
49. Jim Paterson, Stefan Edlich. The Definitive Guideto db4o 2006 P. 71 – 125.
50. JavaScript Object Notation. Режим доступа: https://www.json.org/json-ru.html. Дата обращения 03.06.2018.
51. Библиотека JavaScript, ReactJs. Режим доступа: https://reactjs.org/docs/hello-world.html. Дата обращения 03.06.2018.

Приложение

Приложение 1

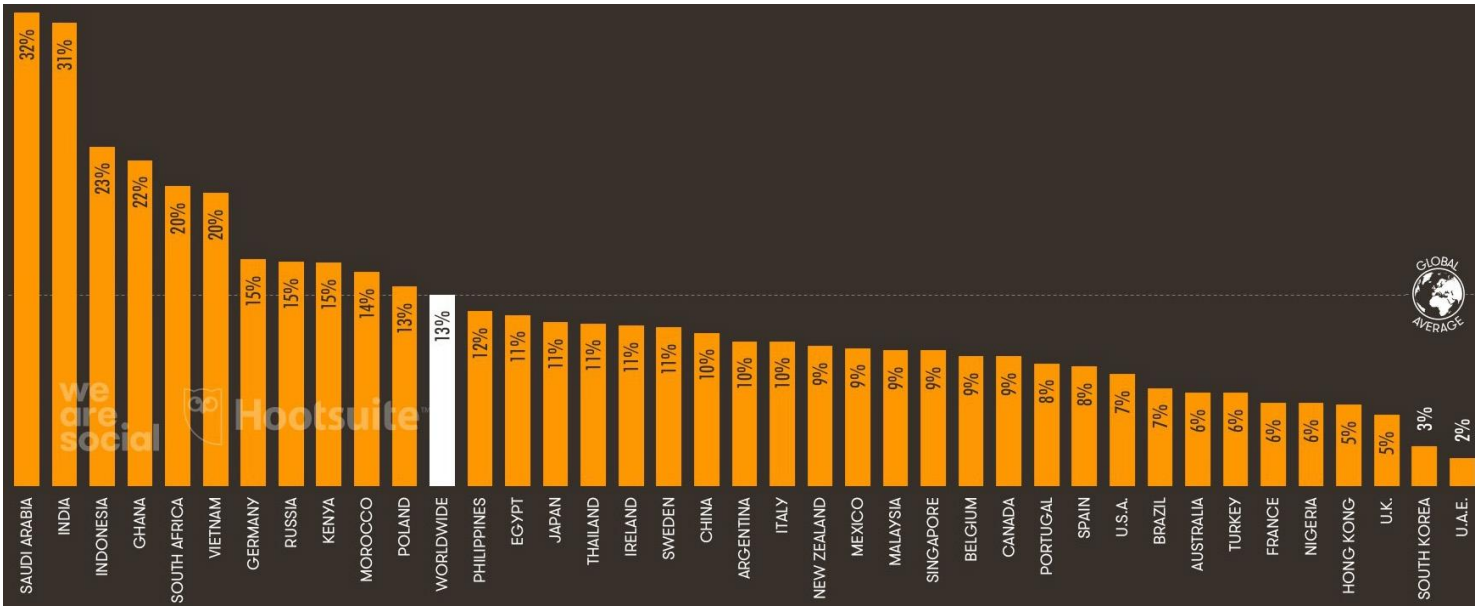


Рис. 1 Прирост аудитории социальных сетей за 2017 год

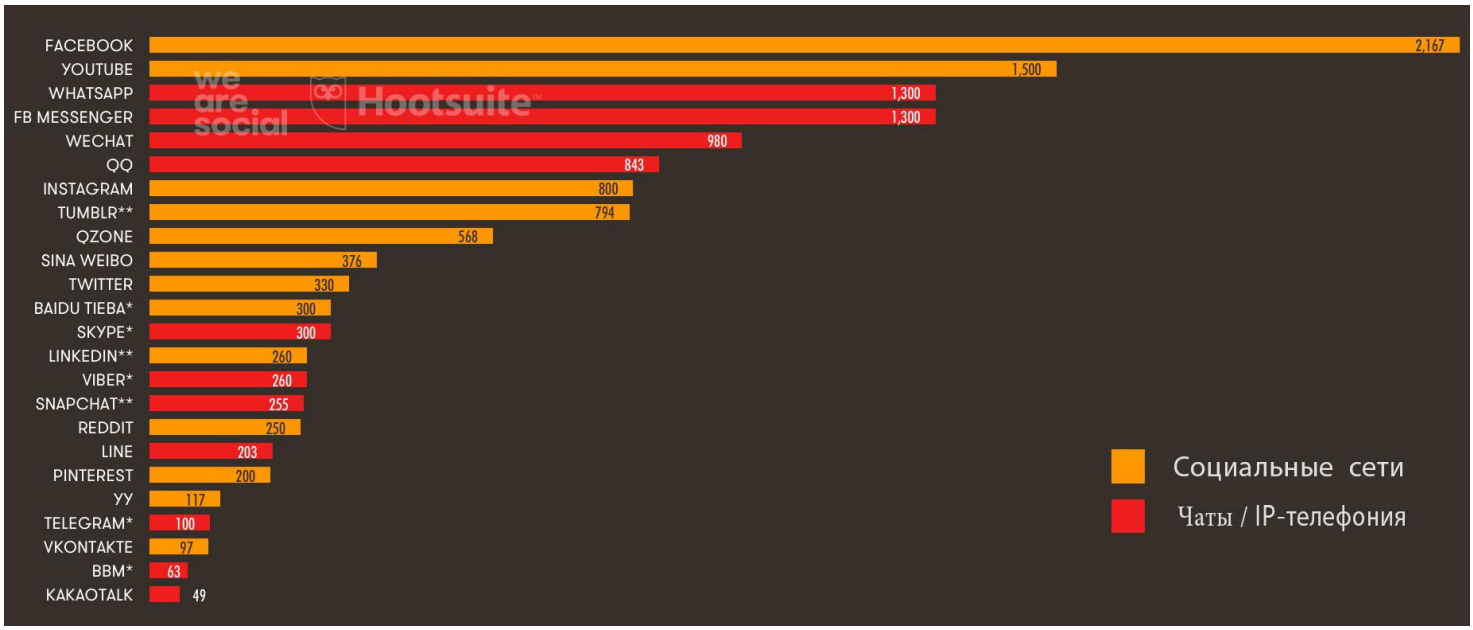


Рис. 2 Количество активных пользователей в социальной сети (в млн.)

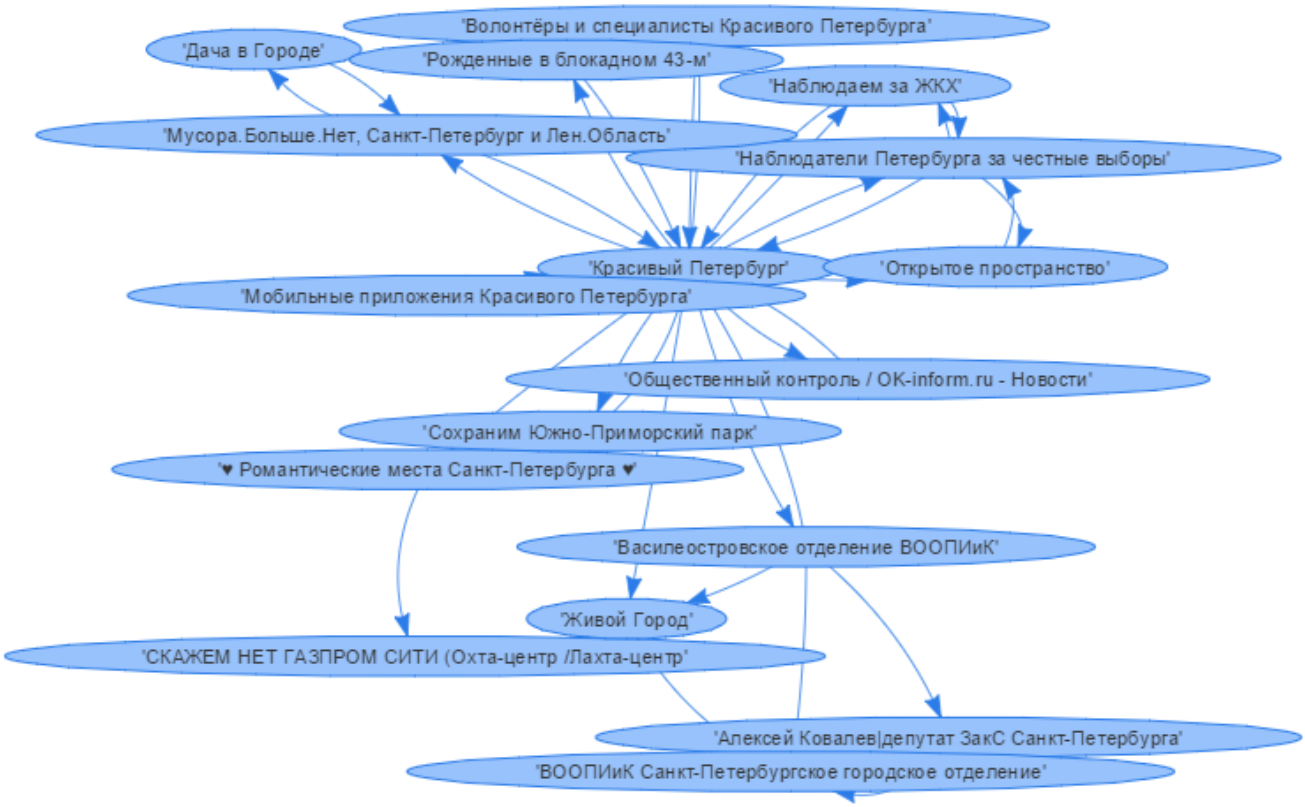
Приложение 2

Псевдокод алгоритма для определения тональности системы SentiScan.

```
func classifySentence(sentence)
{
  positiveScore = 0
  negativeScore = 0
  totalScore = 0
  NEGATION_WEIGHT = 2
  OPPOSITE_CONJUNCTION_WEIGHT = -1
  Let NEGATIONS be set of negation words
  Let P be set of positive words
  Let N be set of negative words
  Let O be set of opposite conjunctions
  words[] = getWords(sentence) // returns list of words in base form
  // 1. analyze negation word window first
  // negation word examples: не, ни, еле, нет
  i = 0
  // iterate over words
  do
    if word_i ∈ NEGATIONS then
      for each word∈[word_{i+1}, word_{i+3}] do
        if word ∈ P then
          positiveScore = positiveScore - NEGATION_WEIGHT
        end if
        if word ∈ N then
          negativeScore = negativeScore - NEGATION_WEIGHT
        end if
      end for
    end if
    i = i + 1
  while i<|words|-3
  // 2. Calculate sentiment rank of separate words - second pass
  i  = 0
  do
    if word ∈ P then
      positiveScore = positiveScore + 1
    end if
    if word ∈ N then
      negativeScore = negativeScore + 1
    end if
    i = i + 1
  while i<|words|
  // 3. Third pass is: global view of a sentence for handling opposite
conjunctions,
  // example of which are: а, но, хотя
  i = 0
  sentimentCount = 0
  oppositeSentimentScore = 0
  do
    if word ∈ P then
      sentimentCount = sentimentCount + 1
    end if
    if word ∈ N then
      sentimentCount = sentimentCount + 1
    end if
    if word ∈ O then
      oppositeSentimentScore = OPPOSITE_CONJUNCTION_WEIGHT * sentimentCount / 2
    end if
    i = i + 1
  while i<|words|
  totalScore = positiveScore - negativeScore + oppositeSentimentScore
  if totalScore > 0 then
    sentiment is positive
  else if totalScore < 0
    sentiment is negative
  else
    sentiment is neutral
}
```

Приложение 3

Визуальный граф со связями между сообществами «ВКонтакте».



Приложение 4

На (см. рис. 3) изображена главная страница пользовательского интерфейса

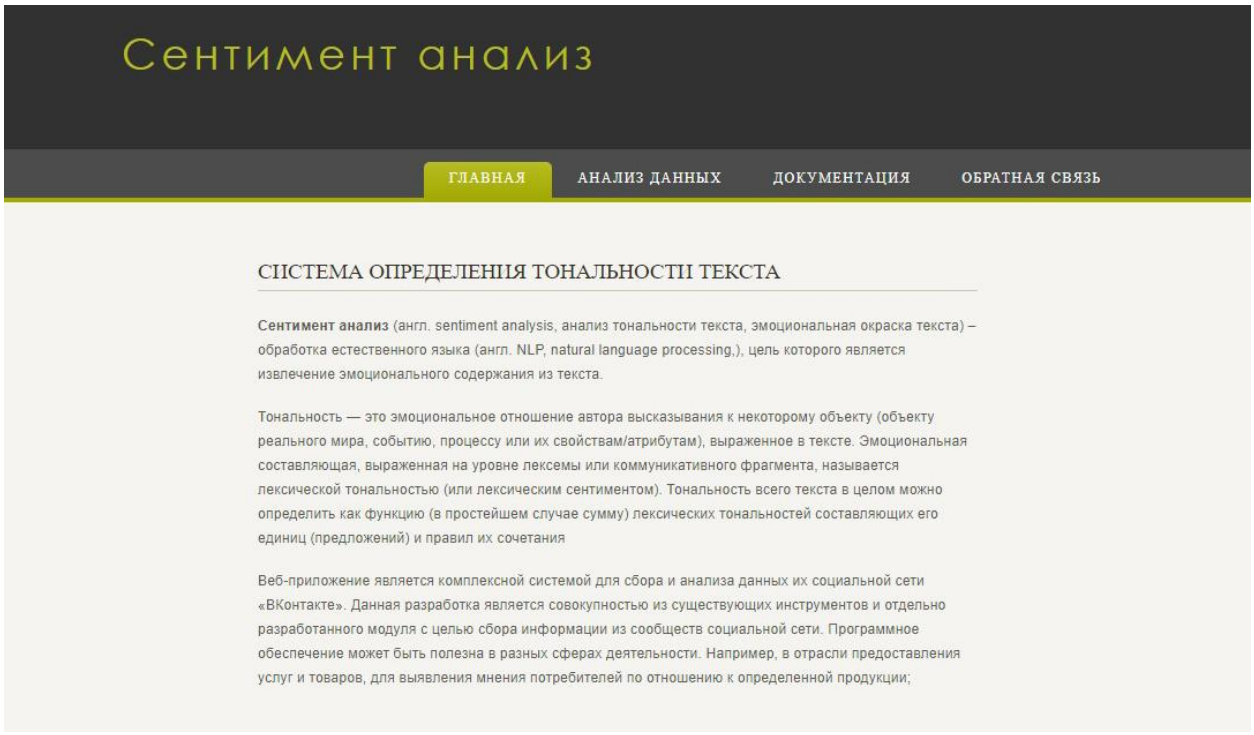


Рис. 3 Главная страница

Приложение 5

На (см. рис. 4) страница для загрузки файла и обработки данных

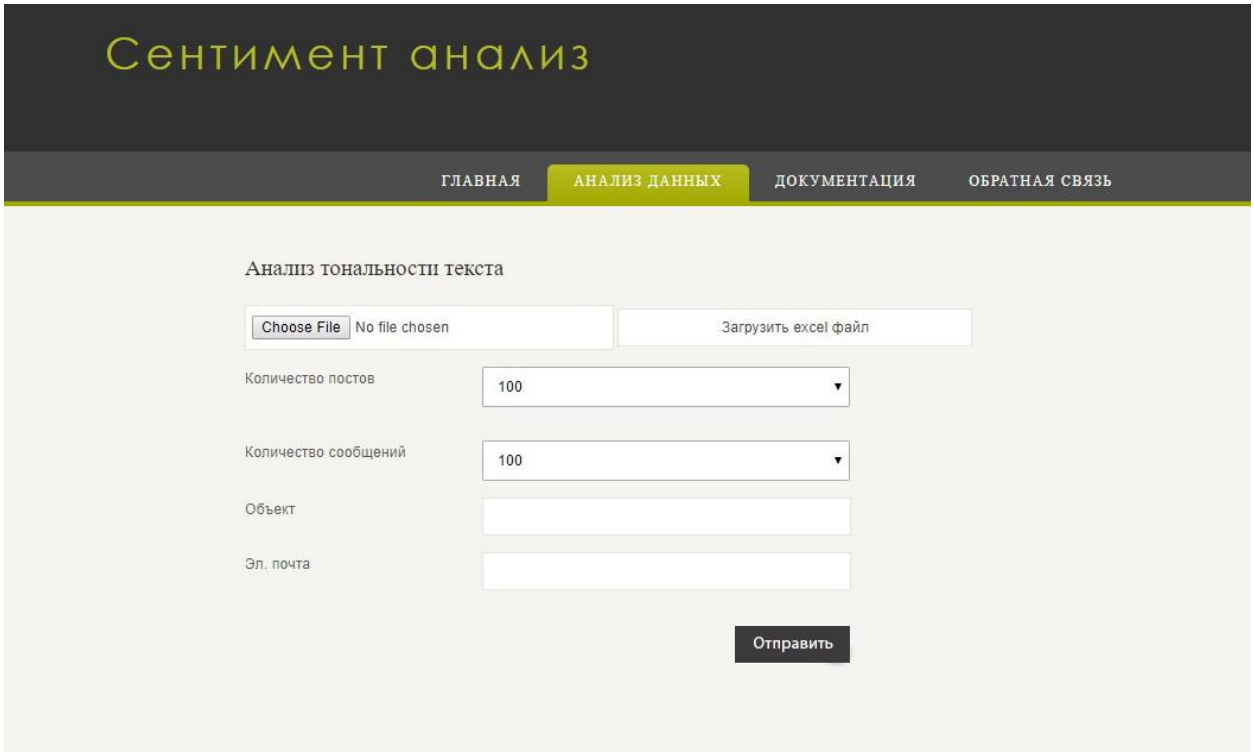


Рис. 4 Анализ данных

Приложение 6

На (см. рис. 5) изображен веб-интерфейс для работы с данными приложения «Сезам»

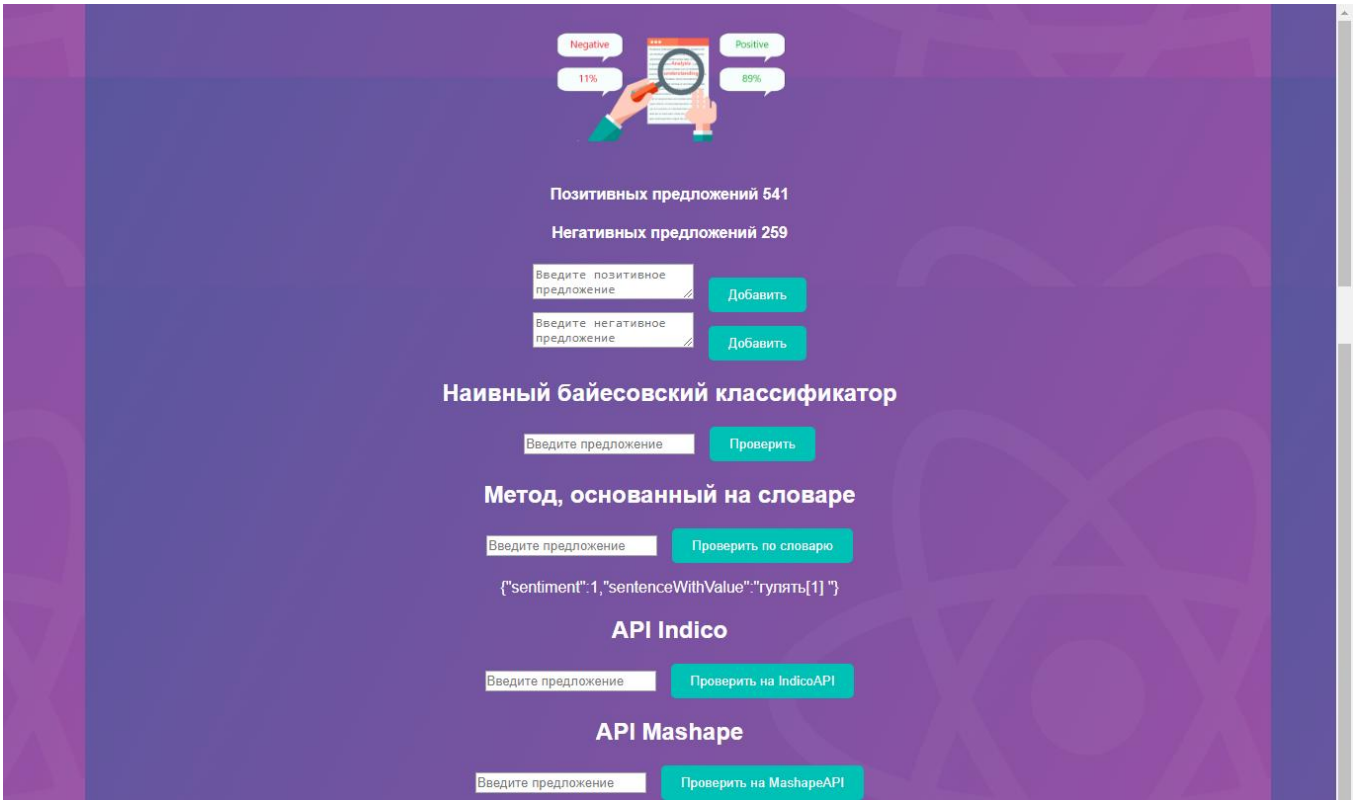


Рис. 5 Интерфейс для обработки данных приложения