САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Магистерская программа
«Прикладная и экспериментальная лингвистика»

АВТОМАТИЧЕСКОЕ ПОПОЛНЕНИЕ ПРЕДМЕТНО-ОРИЕНТИРОВАННЫХ ТОНАЛЬНЫХ СЛОВАРЕЙ (НА МАТЕРИАЛЕ ОТЗЫВОВ О БАНКОВСКИХ ОРГАНИЗАЦИЯХ)

Диссертация
на соискание степени магистра
по направлению 45.04.02 «Лингвистика»

Юшиной Татьяны Андреевны

Научный руководитель –
кандидат филологических наук
доцент О.В. Митренина

Санкт-Петербург
2017

Оглавление

Введение

С развитием интернета и появлением социальных сетей пользователи получили возможность выражать свое мнение. Это мнение может быть относительно услуги или товара, фильма или книги, компании или политического деятеля. Возникла потребность обрабатывать огромные объемы информации для определения отношения пользователей к тому или иному объекту. Однако, количество отзывов достигает десятков тысяч, и обработка отзывов вручную оказывается невозможной. В связи с этим широкое распространение получили автоматизированные подходы к анализу тональности (sentiment analysis).

Цель работы – разработка и реализация алгоритма автоматического пополнения предметно-ориентированных тональных словарей.

В соответствии с поставленной целью решаются следующие *задачи*:

– изучить литературу, посвященную анализу тональности;

– рассмотреть существующие методы анализа тональности;

– рассмотреть методы пополнения тональных словарей;

– разработать алгоритм автоматического пополнения тональных словарей для банковской сферы;

– применить полученный алгоритм и оценить качество работы алгоритма.

Актуальность работы. В последнее время большое внимание направлено на решение задачи анализа тональности в различных предметных областях. Автоматизированные подходы к анализу тональности могут быть полезны как для государственных органов и политиков, так и для компаний и простых пользователей. Одной из важнейших задач для анализа тональности является создание словарей оценочных слов.

Многие исследователи создают словари общеупотребительных оценочных слов. Однако известно, что в разных предметных областях

могут применяться достаточно разные наборы оценочных выражений. Так, например, одно и то же слово может выражать противоположные тональности: «The battery life is *long*» (Батарея работает *долго* – положительная тональность) и «The time taken to focus is *long*» (*Долго* фокусируется – отрицательная тональность). Необходимость разработки алгоритма автоматического пополнения предметно-ориентированных тональных словарей обусловливает *актуальность* данной работы.

Научная новизна работы заключается в обращении к ранее мало изученной предметной области – тональным словарям банковской сферы.

Методы исследования. В качестве основных методов исследования следует назвать описательный метод, статистический метод и метод машинного обучения.

Практическая значимость данной работы состоит в том, что разработанный алгоритм можно применять для извлечения тональных словарей для других предметных областей, а также использовать его для маркетинговых исследований.

Материалом для исследования стали отзывы пользователей о банковских организациях, собранные на сайте banki.ru. Объем корпуса – 2500 отзывов.

Глава I. Автоматический анализ тональности как область прикладной лингвистики

1.1. Автоматический анализ тональности в современном мире

В настоящее время основным средством связи является глобальная сеть Интернет. Ежедневно миллионы терабайтов информации распространяются по всему миру, при этом текст является основным посредником социального воздействия и коммуникации. Изучение и анализ таких текстов дают возможность узнать о тенденциях в мире виртуального общения и об актуальных проблемах современного общества.

Одним из таких анализов является автоматизированный анализ эмоциональной окраски текстов (сентимент-анализ или анализ тональности). Сентимент-анализ – это извлечение положительных или негативных эмоций из текста [Pang, Lee 2008]. Анализ тональности – одна из задач компьютерной лингвистики, однако области его применения могут находиться за пределами самой лингвистики.

Системы автоматического анализа тональности могут использоваться как части других систем, например, рекомендательных (recommendation systems). Сентимент-анализ позволяет выявить объекты с негативными отзывами и в дальнейшем не рекомендовать их.

Диалоговые системы (question answering) – еще одна область применения сентимент-анализа, поскольку для свободного диалога человека и компьютера необходимо, чтобы компьютер мог оценивать эмоциональную окраску.

Классической областью применения автоматического анализа тональности является маркетинг, а именно, анализ оценки продукта на основании отзывов потребителей. Людям всегда требовались советы, если они стояли перед выбором. С появлением блогов и социальных сетей стало гораздо проще делиться рекомендациями и мнениями по товарам и услугам. В связи с этим крупные компании стали понимать, что отзывы

одних покупателей могут оказывать огромное влияние на формирование мнений других покупателей, поэтому компании исследуют интернет-информацию относительно своих брендов и товаров. Сентимент-анализ в свою очередь помогает анализировать данные отзывы, что значительно экономит время на просмотре сотен рецензий. Также сентимент-анализ может использоваться для оценки ситуации на рынке в целом.

Другая область применения – социальные и политические науки. В данной области анализ тональности может применяться для оценки отношения к деятельности политиков на основе текстов СМИ или для оценки политиков во время предвыборной кампании, что помогает определить исход выборов.

1.2. Задачи анализа тональности

Основная цель анализа тональности – нахождение мнений в тексте и выявление их свойств. Какие именно свойства будут рассматриваться зависит от решаемой задачи. Например, целью анализа может быть извлечение объекта, о котором написан отзыв, или автора, которому принадлежит мнение.

Задачи анализа тональности можно условно разделить на следующие группы:

1. Классификация документов на основе мнений.

В самом простом случае данный тип задач сводится к определению, является ли текст положительно или отрицательно окрашенным.

2. Анализ мнений на основе аспектов объекта.

Данная задача ссылается на определение мнений, выраженных различными аспектами объектов. Аспект — это атрибут или компонент сущности, исследуемой на тональность, например, у мобильного телефона (экран, камера, громкость) или банка (расположение, обслуживание, интернет-банк).

Это более сложная задача с увеличенной трудоемкостью, поскольку проблема требует решения таких задач, как идентификация сущностей и извлечение их аспектов. Таким образом, на основе отзыва о банке необходимо понять, какие черты понравились («кредитоустойчивость», «получение средств с минимальным пакетом необходимых документов»), а какие нет («затянутое обслуживание клиентов», «грубый сотрудник») [Popescu 2005].

3. Составление словарей оценочной лексики.

В данном случае, словарь – это список слов и приписанных им тональностей, в том числе с указанием частоты нахождения слова в тексте. Иногда данная группа задач является отдельным этапом при разработке автоматической системы.

4. Поиск сравнений в отзывах.

В текстах отзывов может производиться сравнение одного товара с другими по различным аспектам. В рамках данной задачи осуществляется поиск данных сравнений.

Ниже представлен пример отзыва с сайта otzovik.com о «Почта банк», пунктуация и орфография автора сохранены:

«Сегодня обратилась в отделение Почты банка с целью оформить кредит на 400 тыс. руб. Даже не то, чтобы оформить, а сравнить с другим банком, который мне уже одобрил эту сумму (Сбербанк). При такой активной рекламе и явное расхваливание этого банка захотелось сравнить проценты ... А может и вправду, если у Почта банк настолько выгодные условия, то почему бы не взять кредит у него? С такими размышлениями я пришла на почту. Оформление заявки заняло не много времени, из документов у меня попросили только паспорт и СНИЛС. Оставила хорошие впечатления кредитный специалист: доброжелательная и располагающаяся к себе женщина. Оформили, ждём ответ. Первое, что удивило - сумма ежемесячного платежа - она оказалась больше чем в Сбере (13400 тыс. против 11250тыс.) За счет чего такая сумма???? Ведь процент у Почты ниже... Покаия думала, пришел ответ... Ответ пришёл быстро, минут 5, даже меньше. Он меня шокировал... В кредите мне отказали, якобы потому, что у меня "плохая кредитная история". Я была в шоке. Как так??!! Да у меня ОТЛИЧНАЯ КРЕДИТНАЯ ИСТОРИЯ, ни одной просрочки платежа. Некоторые кредиты закрывала даже досрочно Откуда же тогда она у меня плохая?!

Очень странная система проверки у этого банка. Для Сбера, значит, я - надёжный клиент, и они мне одобрили сумму больше, чем 400 тыс., а у Почта-банк я вдруг странным образом оказалась "неплатильщиком"».

Объектом отзыва является банк «Почта банк». К аспектам относятся: *условия выдачи кредита, специалист банка, оформление заявки, ежемесячный платеж, процент, кредитная история* и др. Оценочная лексика (тональность): «*хороший*», «*доброжелательная и располагающаяся к себе*», «*плохой*», «*больше*», «*отличный*» и др. Также в данном отзыве проводится сравнение с другим банком «Сбербанк».

5. Идентификация субъективности/объективности.

Данная задача обычно определяется как отнесение данного текста в один из двух классов: субъективный или объективный. Однако данная задача может быть трудно решаема, поскольку субъективность зависит от контекста, а объективный документ может содержать в себе субъективные предложения (например, новостная статья, цитирующая мнения людей) [Tong 2001].

1.3. Виды классификации тональности

В современных системах автоматического анализа тональности текста чаще всего используются следующие оценки: положительная/позитивная тональность, негативная/отрицательная тональность и нейтральная тональность. Под нейтральной тональностью имеется в виду, что текст не содержит эмоциональной окраски. Также известны и успешные случаи использования и многомерных пространств [Pang, Lee 2008: 16-17]. Рассмотрим более подробно существующие виды классификации.

Классификация по бинарной шкале.

Полярность документа можно определять по бинарной шкале. В этом случае для определения полярности документа используется два класса оценок: позитивная или негативная. Одним из минусов данного подхода является то, что эмоциональную составляющую документа не

всегда можно однозначно определить, т.е. документ может содержать как признаки позитивной оценки, так и признаки негативной. Одной из систем, которая решает эту проблему, является программа SentiStrength [Thelwall 2010]. Данный алгоритм использует словарь эмотивной лексики и дополнительную лингвистическую информацию для автоматического распознавания эмоциональной окраски в коротких текстах на английском языке (система также адаптирована для русского языка). Для каждого текста результатом является две оценки по пятибалльной шкале от 1 до 5. Например, нейтральный текст получит оценку [1, -1], а такое предложение, как *"I love dogs quite a lot but cats I really hate" (Я очень люблю собак и по-настоящему ненавижу кошек)* получит оценку [3, -5], что будет означать, что в тексте выражается умеренно позитивное и сильно негативное отношение.

Классификация по многозначной шкале.

Можно классифицировать полярность документа по многозначной шкале. Так, авторы статьи [Pang, Lee, 2005] расширили задачу анализа отзывов по бинарной шкале до 4-балльной шкалы. В то же время анализ отзывов может проводиться по 5-балльной шкале [Snyder 2007] и даже по шкале от −10 до 10 (от самого отрицательного к самому положительному).

1.4. Проблемы автоматического определения тональности

Проблемы, возникающие при автоматическом определении тональности, можно разделить на общие и специфичные.

Среди общих проблем можно выделить следующие:

1. Зависимость значения тональности от предметной области.

Так, в предметной области фильмов понятие «непредсказуемый» может иметь положительную окраску, однако в области банковского обслуживания все получается с точностью наоборот. При использовании методов «обучения с учителем», алгоритм классификации сам формирует данные о тональности из обучающей выборки. Следовательно, для

правильной классификации необходимо, чтобы обучающий и тестовый корпус имели общую предметную область.

На практике, например, отзывы пользователей не всегда ограничены какой-либо одной предметной областью. В таком случае применяется разбиение текста на категории в два подхода: сначала осуществляется тематическая классификация документа, затем классификация тональности [Цветков 2013].

2. Наличие отрицания (например, частицы «не») может изменить тональность следуемой за ним части высказывания на обратную.

Рассмотрим следующую часть отзыва: «Раньше мне *очень нравилось* пользоваться интернет-приложением Сбербанка. Качество информационной системы было на высоте. Но использование последней версии прошивки *все портит*». В первых двух предложениях автор высказывает положительное мнение, но из-за использования отрицательно окрашенного словосочетания *«все портит»* в третьем предложении общая тональность относительно аспекта «интернет-приложение» отрицательная.

Для методов машинного обучения, использующих модель типа «набор слов» *(bag-of-words)*, в качестве простой эвристики можно искусственным образом к соседним словам добавлять частицу «не». К примеру, для высказывания «Мне не очень нравится работа банка» получится строка «Мне не не_очень не_нравится работа банка».

Но такая модификация не является очень точным моделированием отрицания, в частности, в случаях, когда отрицание может быть выражено неявно [Wiegand, Balahur 2010].

3. Анализ тональности сарказма и иронии

Высказывания, содержащие сарказм, могут иметь общую тональность, обратную тональности отдельных слов («*Отличное* обслуживание для любителей пропустить половину рабочего дня!»). В некоторых случаях сарказм не воспринимается и самими людьми, что

сводит практически к нулю попытки машинного анализатора правильно определить тональность.

Но стоит отметить, что способы решения такой проблемы существуют – в работе [Oren, Dmitry, Ari 2010] авторам удалось добиться точности на уровне 78% на коллекции отзывов на товары, используя метод частичного обучения *(semi-supervised learning).*

4. Значение тональности зависит и от непосредственного заказчика или участника анализа.

Так, например, для банка «Советский» высказывание *«У банка «Советский» отличное обслуживание по сравнению с другими, в частности, Сбербанком»* несёт положительное значение тональности, а для Сбербанка – отрицательное.

К специфичным проблемам для анализа отзывов относятся:

1. Большой словарь употребляемых слов.

Частое использование сленга, намеренное и ненамеренное искажение написания слов, использование разных регистров при написании одного и того же слова, использование смайликов значительно затрудняют автоматическую обработку данных.

2. Недостаточное количество слов в отзыве.

Зачастую на текстовые поля ставятся ограничения по размеру для более экономного расходования памяти базой данных. Из-за такого ограничения при наборе текста пользователем, с одной стороны, имеют место всевозможные сокращения слов, с другой – обрывистые данные с незаконченной мыслью, так как система изначально принимает длинное сообщение, а затем при внесении в базу данных обрезает для корректного пополнения.

1.5. Методы определения тональности текстов

Подходы к автоматической оценке тональности текста можно разделить на две группы: основанные на правилах (*rule-based*) и

использующие методы машинного обучения *(machine learning)*. Рассмотрим каждую из групп подробнее.

1. Подходы с использованием правил и словарей

1.1. Метод, основанный на правилах (rule-based).

Данный подход заключается в том, что существует набор заранее разработанных правил, которые описывают некую предметную область. Необходимо определить тональность текста, применив данные правила. Для этого текст разбивается на слова или последовательности слов (n-grams).

Затем полученные данные используются для выделения часто встречающихся шаблонов, которым присваивается тональная оценка (положительная или отрицательная). Выделенные шаблоны применяются при создании правил вида «ЕСЛИ условие, ТО заключение». [Liu 2004, Клековкина, Котельников 2012]. Например, следующее правило *«если цепочка содержит глагол из списка («любить», «нравиться» и др.) и не содержит глагола из другого списка («ненавидеть», «отвращать» и др.) или отрицания»* приписывает предложению положительную тональность.

При использовании отрицания перед найденной цепочкой тональность может меняться на противоположную – указанную правилом или определенную в словаре.

Для получения итоговой оценки тональности текста необходимо сделать расчет общей суммы весов. При этом, сумма тональностей фрагментов может быть не равна общей окраске всего текста, в частности, в случае наличия сарказма. Например, фраза: *«Отличное обслуживание для любителей пропустить половину рабочего дня!»* имеет в основе сарказм.

Также сложности возникают в случаях, когда сработали сразу несколько правил. Для разрешения предусмотрены механизмы комбинации правил на основе частоты и позиции в документе (как часто

правило используется в документе, на какой позиции встречается и др.) [Хохлова 2016, Пазельская 2011].

Данный алгоритм показывает высокие результаты при большом наборе правил.

К недостаткам можно отнести то, что создание большого набора правил затратно по выделяемой памяти и ресурсам, поэтому словари зачастую описывают лишь определенную узкую тематику, например, рестораны, фотоаппараты и т.д., поскольку тональная оценка может зависеть от предметной области. Так, прилагательному «замысловатый» в обзоре фильмов или ресторанов скорее будет соответствовать положительная тональность («замысловатый сюжет» или «замысловатый интерьер»), в то время как в текстах, посвященных техническим устройствам, отрицательная («замысловатая настройка») [Хохлова 2016, Kan 2011].

1.2. Метод, основанный на использовании словарей оценочной лексики.

Метод основан на поиске эмоционально окрашенной лексики в тексте по заранее составленным словарям. Также данный подход можно рассматривать в рамках вышеупомянутого подхода, основанного на правилах, так как эти методы могут использоваться вместе.

Создаются специальные словари (лексиконы), в которых приводится лексика (слова и их сочетания) с присвоенными ей весами в зависимости от степени позитивной или негативной окраски.

Одним из способов пополнения таких словарей является разработка правил, которые используются для извлечения новых оценочных слов из текстов (т.е. тех, которые не попали в словарь). Так, если прилагательные объединены сочинительным союзом «и» и первое из них содержится в словаре, то и второму можно приписать такой же вес. На выходе список этих слов пополнит лексикон. Например, в предложении *«This car is beautiful and spacious»* («Эта машина красивая и просторная»)

прилагательное «beautiful» («красивая») имеет положительную тональность, значит, можно сделать вывод, что «spacious» («просторная») тоже обладает положительной тональностью. Это объясняется тем, что одно и тоже мнение выражается по обе стороны соединительного союза. Следующее предложение выглядит неестественно: «*The car is beautiful and difficult to drive*» («Эта машина красивая и ее сложно водить»). Но если мы заменим союз *and* на *but*, данное предложение будет допустимо [Hatzivassiloglou 2000; Liu 2010].

Также необходимо посчитать частоту слова или словосочетания с положительно или отрицательно окрашенной лексикой. При этом лучше использовать максимально простые варианты, например «хороший» /«хорошо» и «плохой» /«плохо».

В качестве инструмента для измерения совместной встречаемости лексических единиц используются статистические меры — хи-квадрат, мера поточечной взаимной информации (PMI) и др.

Для каждого слова, встречаемого в документе, из словаря получают значение тональности. Чтобы получить итоговую тональность необходимо взять среднее арифметическое или вычислить сумму значений тональности всех слов из документа [Цветков 2013, Ding 2008].

Для данного алгоритма желательно иметь достаточно большой корпус. К достоинствам алгоритма можно отнести его простоту использования.

Среди недостатков нужно отметить отсутствие универсальности: для каждой предметной области требуются свои словари [Hatzivassiloglou 1997].

2. Подходы с использованием машинного обучения (machine learning)

Существует 2 основных подхода с использованием машинного обучения – обучение с учителем и без него. Рассмотрим каждый из вариантов.

2.1. Машинное обучение с учителем (supervised learning).

При данном подходе требуется наличие заранее размеченного набора текстов. Каждый размеченный текст представляет собой пару – вектор признаков текста и приписанную ему тональность. Под вектором признаков понимается представление текста как набора терминов (слов, сочетаний, фраз) с соответствующими им весами.

На основе вышеуказанной выборки строится вероятностный или статистический классификатор, который впоследствии используется для определения тональности новой коллекции документов [Joachims 1999].

К достоинствам данного алгоритма следует отнести:

– высокую точность определения тональности;

– на основе обучающей выборки классификатор самостоятельно выделяет признаки, влияющие на тональность. Таким образом, проблема зависимости от предметной области решается с помощью использования обучающей выборки из той же области.

Недостатки подхода:

– Требуется размеченная обучающая выборка;

– Результаты могут сильно зависеть от выбранного алгоритма, его параметров, обучающей выборки [Цветков 2013].

2.2. Машинное обучение без учителя (unsupervised learning).

Главная идея подхода заключается в том, что наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции.

При выделении терминов с наибольшим весом и определении их тональности можно сделать вывод о тональности всего текста [Turney 2002, Клековкина, Котельников 2012].

Обучение без учителя является ещё одним из разделов машинного обучения. Отличие состоит в том, что в этом случае для тренировки алгоритма используется обучающая выборка, состоящая из документов,

классы которых заранее неизвестны (или известны, но эта информация не используется алгоритмом).

Недостатки подхода: точность обычно значительно ниже, чем у алгоритмов, основанных на обучении с учителем.

Достоинство подхода: для обучения не требуется размеченная выборка.

Точность и качество системы анализа тональности текста оценивается в соответствии с тем, насколько хорошо она согласуется с мнением человека относительно эмоциональной оценки исследуемого текста. Для этого могут использоваться такие метрики, как точность и полнота.

Полнота определяется как отношение верно приписанных тональностей (т.е. тех, что совпали с оценкой эксперта) к общему количеству тональностей (приписанных и не приписанных).

Точность – это отношение верно определенных тональностей ко всем определенным системой тональностям. [Хохлова 2016].

1.6. Обзор работ по автоматическому составлению тональных словарей

Многие системы сентиментного анализа используют словари, составленные вручную или полуавтоматическими методами, также известны и варианты комбинирования подобных словарей.

Существуют специальные тезаурусы, в которых размечена эмоциональная составляющая лексики. Для английского языка это SenticNet, SentiWordNet и WordNet-Affect.

SenticNet [Cambria, Havasi, Hussain 2012] представляет собой семантический *тезаурус*, в котором отражена не только тональность лексики, но и когнитивная информация. Последнее стало возможным благодаря специальным вычислениям, использующим алгоритмы семантических сетей и искусственного интеллекта. Так, для понятия «празднование дня рождения» (“birthday party”) системой будет выданы

его принадлежность к домену верхнего уровня «события», а также набор семантически связанных понятий (например, «клоун» или «сладкое»).

В тезаурусе SentiWordNet [Esuli, Sebastiani, 2006] представлены результаты работы по автоматической разметке синсетов WordNet. Синсетам из WordNet приписывается тональная оценка: «положительно/нейтрально/отрицательно окрашенный

Для русского языка подобная работа ведется в рамках проекта RussNet [Дегтева, Азарова 2013].

В тезаурусе WordNet-Affect [Strapparava, Valitutti 2004] наряду с метками, описывающими эмоции (такими как «физическое состояние», «настроение», «поведение», «отношение», «чувство» и др.), синсетам вручную были приписаны валентности (позитивная, негативная, неоднозначная и нейтральная). Также синсетам были сопоставлены так называемые эмоциональные категории: «радость», «страх», «гнев», «печаль», «отвращение» и «удивление». Синсеты тезауруса переведены с английского языка на русский и румынский [Sokolova, Bobicev 2009].

Для русского языка существующие ресурсы неоднократно обсуждались участниками соревнования SentiRuEval (URL: http://www.dialog-21.ru/evaluation/2016/sentiment/); все свободно распространяемые словари оценочных слов [Kotelnikov, Bushmeleva 2016] собирались полуавтоматически различными методами.

Полуавтоматические методы предполагают отбор определённого количества слов и словосочетаний для словаря и последующую экспертную разметку. Отбор кандидатов в оценочные выражения производится с помощью различных методов машинного обучения и некоторыми другими способами (в частности, перевод с английского языка с последующей коррекцией [Ulanov, Sapozhnikov 2013]), используются начальные словари (seed dictionary), содержащие небольшое количество тонально окрашенных лексических единиц [Протопопова, Букия, Митрофанова 2016].

Так, в работе [Chetviorkin, Loukachevitch 2012] описываются методы построения словаря на основе текстов отзывов с использованием классификатора для извлечения схожих выражений. Авторы предлагают использовать совокупность статистических и лингвистических признаков, позволяющих выявлять оценочные слова, и комбинировать эти признаки с помощью алгоритмов машинного обучения. Метод базируется на обучении алгоритма извлечения русской оценочной лексики в одной предметной области (фильмы), и затем переносе обученной модели на другие предметные области. Авторы применяют модель к нескольким предметным областям и затем из оценочных словарей отдельных предметных областей собирают единый словарь оценочной лексики, рассматривая его как словарь оценочной лексики в широкой области товаров.

Извлечение оценочных слов в заданной предметной области основано на нескольких текстовых коллекциях: коллекции отзывов о продуктах с оценками пользователей, коллекции описаний продуктов и контрастной коллекции (например, новостная коллекция). Для каждого слова в коллекции отзывов вычисляется набор статистических и лингвистических признаков. Для обучения алгоритмов необходимо размеченное множество слов. Для этого было вручную размечено множество всех слов с частотой выше трех из предметной области о фильмах (18362 слова). Слово относили к категории оценочных в случае если могли представить его в каком-либо оценочном контексте. Авторы решали задачу классификации на два класса: разделение всех слов на оценочные и неоценочные. Для этих целей использовались следующие алгоритмы: Logistic Regression, LogitBoost и Random Forest.

Используя данные алгоритмы, были получены списки слов, упорядоченные по вероятности оценочности слов. Для оценки качества этих списков использовалась мера точности. В результате точность автоматического извлечения оценочных слов в области фильмов составила

81.5%. Для использования системы в новой предметной области необходимо собрать аналогичный набор коллекций, как и предметной области о фильмах. Авторы также применили модель извлечения оценочных слов в таких областях, как книги, игры, цифровые камеры, мобильные телефоны.

Таким образом, авторы создали русскоязычный список оценочных слов для широкой области товаров и показали его полезность в задачах, связанных с настройкой систем анализа тональности на новую предметную область. В дальнейшем исследователи планируют опубликовать полученный список оценочных слов, и это будет первый общественно доступный список оценочной лексики для русского языка [Chetviorkin, Loukachevitch 2012: 593-596].

В работе [Blinov, Kotelnikov 2014] с целью пополнения тонального словаря используются векторные представления слов. Идея алгоритма – автоматически расширить первоначально заданный тональный словарь. Так, для каждого слова в векторном пространстве авторы выделяли ближайших соседей с помощью косинусного коэффициента (cosine similarity). Схожесть между двумя векторами $\vec{a} = (a_1, \dots, a_n)$ и $\vec{b} = (b_1, \dots, b_n)$ определяется по формуле:

$$similarity\ (\vec{a}, \overrightarrow{b}) = \cos(\theta) = \frac{\Sigma_{i=1}^{n} a_i b_i}{\sqrt{\Sigma_{i=1}^{n} a_i^2}\sqrt{\Sigma_{i=1}^{n} b_i^2}},$$

где θ – угол между веторами, n – размерность пространства.

Поскольку в тональные словари могут входить не только слова, но и словосочетания, потребовалась дополнительная обработка данных. С помощью простых правил $<$ очень|не $>$ $+<$ очень|не $>$ $+<$ прилагательное|глагол|наречие $>$ были сформированы сложные лексические единицы, как например, *не_готовый, очень_сытный, очень_не_приятный, не_очень опрятный* и т.д. Конечно, данный подход не рассматривает все тонально окрашенные слова в полном объеме, но он

позволяет рассмотреть необходимое количество таких слов для дальнейшего анализа тональности с применением полученных словарей [Blinov, Kotelnikov 2014].

В статье [Dubatovka, Kurochkin, Mikhailova 2016] предложен алгоритм создания предметно-ориентированных тональных словарей с использованием графовой модели с применением синтаксических шаблонов. Важно заметить, что данный метод не требует предварительно размеченных данных, нужен только достаточно большой корпус текстов предметной области на русском языке. Алгоритм предполагает разбиение узлов построенного графа на положительные и отрицательные слова. Сначала собирается небольшой набор прилагательных, тональность которых не зависит от предметной области, например, *хороший, плохой, замечательный* и т.д. Затем выполняется следующая итерация: если вершина графа имеет наиболее «сильную» связь с уже существующим прилагательным из начального набора, то добавляется эта вершин графа. Данный алгоритм также может быть адаптирован и для других языков.

Простой способ извлечения слов для последующей тональной разметки предложен в статье [Ivanov, Tutubalina 2015]. Авторы собрали отзывы о ресторанах и машинах, а затем сформировали из полученных корпусов два раздела «Достоинства» и «Недостатки»: в первый подкорпус попали отзывы с пользовательской оценкой 5 (по 5-ти балльной шкале), во второй – отзывы с пользовательскими оценками 1 и 2. Из каждого раздела выделяются наиболее частотные прилагательные, наречия и глаголы, таким образом, в словарь после ручной оценки попадают только однословные единицы. Таким образом, получились словари для предметной области «рестораны» (741 – положительные, 362 – отрицательные) и «машины» (1576 – положительные, 741 – отрицательные) [Ivanov, Tutubalina 2015].

Похожий метод составления тональных словарей использовали авторы статьи [Протопопова, Букия, Митрофанова 2017]. Авторы собрали

корпус отзывов на фотоаппараты сервиса «Яндекс.Словари», который организован на основе бинарной структуры. Таким образом было сформировано два подкорпуса одинакового объёма (20758 отзывов), состоящих из текстов – описаний достоинств и текстов – описаний недостатков. Критерий, на основе которого выделяются тонально окрашенные слова – оценка корреляции двух случайных событий: «отзыв содержит слово *w*» и «отзыв описывает достоинства». Для оценки корреляции была построена таблица сопряженности и использовалась статистик Хи-квадрат. Были получены тональные словари, включающие около 1,5 тысяч положительно окрашенных единиц и около 2 тысяч отрицательно окрашенных.

1.7. Выводы к главе I

В первой главе мы рассмотрели проблему анализа тональности и области его использования, обозначили основные задачи, которые могут решаться в рамках анализа тональности, а также основные проблемы, возникающие при анализе отзывов. Мы выделили основные подходы сентимент-анализа: подход, основанный на словарях и на правилах, и подход, основанный на машинном обучении, рассмотрели их достоинства и недостатки.

В главе был проведен обзор последних основных исследований, посвященных автоматическому составлению тональных словарей:

– Ulanov, Sapozhnikov (2013): перевод тональных словарей с английского языка с последующей коррекцией;

– Chetviorkin, Loukachevitch (2012): метод построения словаря на основе текстов отзывов с использованием классификатора для извлечения схожих выражений;

– Blinov, Kotelnikov (2014): использование векторных представлений слов для пополнения тонального словаря;

– Dubatovka, Kurochkin, Mikhailova (2016): извлечение оценочных слов при помощи графов, построенных с применением синтаксических шаблонов;

– Ivanov, Tutubalina (2015): использование корпусов отзывов «Достоинства» и «Недостатки» с последующим выделением наиболее частотных слов;

– Протопопова, Букия, Митрофанова (2017): также использование корпусов отзывов «Достоинства» и «Недостатки» с последующим обучением на основе критерия согласия Пирсона (Хи-квадрат) для таблиц сопряженности для извлечения тональных словарей.

Глава II. Разработка системы автоматического пополнения тональных словарей для банковской сферы

2.1. Постановка задачи и описание алгоритма

Материалом для практической части работы стал корпус отзывов о банках с веб-ресурса banki.ru. Объем корпуса – 2500 отзывов.

Отзывы представлены в виде следующего набора данных: дата и время написания отзыва, информация о пользователе, заголовок, текст, наименование банка, о котором написан отзыв, ссылка на веб-ресурс и оценка пользователя по 5-тибалльной шкале. Пример такого отзыва приводится ниже:

Дата: *2016-11-16 11:36:45*

Пользователь: *sahea*

Заголовок отзыва: *Быстрая реакция службы безопасности и операторов в разных ситуациях*

Текст отзыва: *Произошел со мной недавно неприятный случай. Подруга попросила через соц.сеть перевести ей деньги. И вроде болтали мы с ней по сети, и сумма небольшая - 5000. В общем, попалась я как крайне доверчивое существо. Перевела ей сумму на счет, что она прислала. Через минуту позвонила служба безопасности банка, которая стала уточнять про перевод, уверена ли я, что эта информация верна. В процессе общения выяснили, что общение идет по сети. В общем, заблокировали карту по риску мошеннических действий. Позвонила подруге и узнала, что ее аккаунт взломали. В результате действий службы безопасности этого банка удалось сохранить деньги и не попасть в крайне нехорошую ситуацию. А вообще мне пока банк нравится. Когда была за границей и появилась необходимость переприязать к местной сим-карте, ребята тоже оперативно среагировали. Сами позвонили - переприязали.*

Наименование банка: *Тинькофф Банк*

Оценка пользователя: 5

Ресурс: *http://www.banki.ru/*

Для анализа тональности мы используем трехзначную шкалу оценивания (положительная тональность, отрицательная и нейтральная).

Нам необходимо получить словари различных частей речи (прилагательные, наречия, глаголы и существительные), а также словари конструкций, имеющих положительную или отрицательную тональность, наиболее применимых к предметной области «банковская система», для дальнейшего автоматического пополнения исходных словарей и их использования в анализе отзывов.

В нашей работе за основу мы будем использовать методы, представленные в работах [Ivanov, Tutubalina 2015] и [Протопопова Букия, Митрофанова 2017], в силу удобства в использовании при решении задач анализа тональности и простоты реализации.

Для успешного решения поставленной задачи необходима реализация следующего алгоритма:

1. Формирование двух подкорпусов: текстов с описаниями достоинств и текстов с описаниями недостатков.

Формирование каждого из подкорпусов будет основано на пользовательской оценке (по 5-тибалльной шкале): если оценка равна 1, тогда отзыв описывает недостатки; если оценка равна 5, тогда отзыв описывает достоинства. В отличие от отзывов на товары народного потребления банковские отзывы редко бывают смешанными.

Объем подкорпуса достоинств составил 833 отзыва, объем подкорпуса недостатков – 1345.

2. Отбор слов-кандидатов на включение в тональные словари на основе полученных подкорпусов.

Элементы в тональные словари дополняются путем анализа каждого слова и его частоты в подкорпусе достоинств/недостатков. Такой же анализ применяется и для анализа тональных конструкций: «прилагательное + существительное», «глагол + наречие» и т.д.

3. Заполнение тональных словарей на основе проведенного анализа слов и тональных конструкций.

В основу анализа распределения слов по тональным словарям входит оценка корреляции двух случайных событий: «слово содержится в отзыве» и «отзыв описывает недостатки/достоинства». И если одно событие с большой вероятностью влечет другое, то, скорее всего, лексема имеет положительную тональность. В результате необходимо построить таблицу сопряженности, выбрав наиболее подходящую статистику.

После распределения слов по словарям для выделения тональных лексем и конструкций необходимо эмпирически подобрать порог значения статистического критерия, который позволил бы отделить нейтральные единицы словаря от окрашенных.

Полученные списки слов из словарей наших корпусов мы и добавляем в исходные словари.

4. Оценка работы алгоритма на основе метрик полноты и точности с использованием кросс-валидации.

Для оценки качества работы нашего алгоритма был выбран метод кросс-валидации по трем блокам. Корпус был разбит на три равные части по 726 отзывов, где 2/3 корпуса использовались в качестве обучающей выборки и 1/3 – в качестве тестовой выборки. На каждой итерации были посчитаны полнота и точность, затем вычислено их среднее значение.

Тональность отзывов считается следующим образом. Каждое слово отзыва принимает значение либо +1 (оно находится в словаре достоинств), -1 (находится в словаре недостатков), 0 (нейтральное – не находящееся ни в

одном тональном словаре). Далее подсчитывается сумма по всему отзыву за каждое слово. В зависимости от знака суммы оно попадает либо в корпус достоинств, либо в корпус недостатков, либо считается не рассматриваемым (в случае 0).

2.2. Инструменты и технологии

Для успешной реализации работы алгоритма использованы следующие инструменты и технологии.

Язык Python

Python – это высокоуровневый расширяемый и встраиваемый язык программирования с поддержкой открытого кода. Python является простым и, в то же время, мощным интерпретируемым объектно-ориентированным языком программирования. Он предоставляет структуры данных высокого уровня, имеет изящный синтаксис и использует динамический контроль типов, что делает его идеальным языком для быстрого написания различных приложений, работающих на большинстве распространенных платформ [Россум 2001].

В программировании на Python реализована поддержка структурной, функциональной, императивной, объектно- и аспектно-ориентированной парадигм.

Основные архитектурные черты – динамическая типизация, автоматическое управление памятью, полная интроспекция, механизм обработки исключений, поддержка многопоточных вычислений и удобные высокоуровневые структуры данных [Россум 2001].

Python – язык универсальный, он широко используется для различных целей – обработка текста и базы данных, создание Web-

приложений (клиентских и серверных), встраивание интерпретатора, программирование пользовательского интерфейса.

Python выбран в качестве языка разработки в связи с наличием текстовых библиотек, наиболее подходящих для реализации алгоритма по функционалу, затрачиваемой памяти и скорости выполнения.

Среда разработки IDLE

IDLE (Integrated DeveLopment Environment) – это кроссплатформенная (Windows, MAC, Linux) интегрированная среда разработки на языке Python, которая позволяет решать задачи просмотра, редактирования, запуска и отладки программного кода Python [URL: www.python.org].

Имеет ряд дополнительных функций, связанных с успешной интеграцией и настройкой программных библиотек, интерактивное перемещение по буферу обмена, автоматическая поддержка отступов, подсветка кода и т.д.

Отличительные черты IDLE Python:

– среда запрограммирована на Python с использованием GUI-инструментария;

– кроссплатформенность: работает на Windows и Unix;

– многооконный текстовый редактор с функцией многократной отмены, подсветкой синтаксиса Python и многими другими свойствами;

– отладчик (IDLE имеет встроенную систему отладки, позволяющую запускать программу построчно, что облегчает процесс поиска ошибок).

Данная среда разработки была использована как наиболее популярная и имеющая весь необходимый инструментарий для программной реализации алгоритма [URL: www.python.org].

Библиотеки Python

Pymorphy2. Класс MorphAnalyzer.

Pymorphy2 – библиотека для морфологического анализа. Выполняет лемматизацию и анализ слов, способен осуществлять склонение по заданным грамматическим характеристикам слов.

Анализатор Pymorphy2 позволяет:

– приводить слово к начальной форме («изменил» -> «изменить»);

– ставить слово в нужную форму (ставить слово во множественное число, менять падеж слова и т.д.);

– возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т.д.) [URL: https://pymorphy2.readthedocs.io].

MorphAnalyzer – специальный класс для морфологического анализа русских слов.

Данная библиотека была выбрана в качестве наиболее подходящей для приведения слов в отзывах к нормальной форме и добавления их в тональные словари.

Библиотека Codecs

Codecs – Python-библиотека (модуль), осуществляющая связь между интерпретатором и файловой системой операционной системы.

Этот модуль определяет базовые классы для кодеров и предоставляет доступ к внутреннему реестру кодеров, позволяет выполнить различного рода функции, связанные с чтением файла, изменением его структуры, удалением, сжатием и т.д.

Библиотека используется в связи с наличием повышенной скорости и больших возможностей обработки информации.

Библиотека Math

Math – Python-библиотека для работы с математическими функциями, представляющая обширный функционал для работы с числами.

Существует три встроенных числовых типа данных:

– целые числа (int);

– вещественные числа (float);

– комплексные числа (complex).

Представленные в библиотеке функции представляют результаты расчета в оптимальное, установленное время.

Библиотека используется в данной работе при статистическом анализе данных.

Таблицы сопряженности

Таблица сопряженности в статистике — метод представления совместного распределения двух переменных, предназначенный для исследования связи между ними. Таблица сопряженности является наиболее универсальным средством изучения статистических связей, так как в ней могут быть представлены переменные с любым уровнем измерения. Таблицы сопряженности часто используются для проверки гипотезы о наличии связи между двумя признаками с использованием точного теста Фишера или критерия согласия Пирсона [Аптон 1982: 7-15, Шитиков, Розенберг 2003].

Критерий согласия Пирсона

Критерий согласия Пирсона или критерий согласия $\chi 2$ (Хи-квадрат) – один из наиболее популярных критериев для проверки гипотезы о соответствии эмпирического распределения предполагаемому

теоретическому распределению F(x) при большом объеме выборки (n ≥ 100) [Лемешко 2011].

Использование критерия χ2 предусматривает разбиение размаха варьирования выборки на интервалы и определения числа наблюдений (частоты) n_j для каждого из e интервалов. Для удобства оценок параметров распределения интервалы выбирают одинаковой длины.

Критерий вычисляется по следующей формуле:

$\chi^2 = \sum_{j=1}^{e} \frac{(n_j - np_j)^2}{np_j}$, где p_j – вероятность попадания изучаемой случайной величины в j-и интервал, вычисляемая в соответствии с гипотетическим законом распределением F(x) [Лемешко 2011, Степнов 1985].

Критерий выбран как наиболее подходящий и универсальный вариант для проверки гипотезы соответствия теоретическому распределению F(x), применимый для любых видов функций, даже при неизвестных значениях их параметров.

Кросс-валидация (cross-validation)

Кросс-валидация, или перекрестная проверка, – техника валидации модели для проверки того, насколько успешно применяемый в модели статистический анализ способен работать на независимом наборе данных. Кросс-валидация дает значимые результаты только когда тренировочный набор данных и тестовый набор данных берутся из одного источника.

Кросс-валидация используется в случае, когда получение дополнительных данных затруднительно или невозможно.

Обычно кросс-валидация используется в ситуациях, где целью является предсказание, и хотелось бы оценить, насколько предсказывающая модель способна работать на практике. Один цикл кросс-валидации включает разбиение набора данных на части, затем построение модели на

одной части (называемой тренировочным набором), и валидация модели на другой части (называемой тестовым набором). Чтобы уменьшить разброс результатов, разные циклы кросс-валидации проводятся на разных разбиениях, а результаты валидации усредняются по всем циклам [Arlot, Sylvain 2010].

Распространенные типы кросс-валидации

Кросс-валидация по K блокам (K-fold cross-validation)

В этом случае исходный набор данных разбивается на K одинаковых по размеру блока. Из K блоков один оставляется для тестирования модели, а остающиеся K-1 блока используются как тренировочный набор. Процесс повторяется K раз, и каждый из блоков используется один раз как тестовый набор. Получаются K результатов, по одному на каждый блок, они усредняются или комбинируются каким-либо другим способом, и дают одну оценку.

Преимущество такого способа перед случайным сэмплированием (random subsampling) в том, что все наблюдения используются и для тренировки, и для тестирования модели, и каждое наблюдение используется для тестирования в точности один раз.

Валидация последовательным случайным сэмплированием (random subsampling)

Этот метод случайным образом разбивает набор данных на тренировочный и тестовый наборы. Для каждого такого разбиения, модель подгоняется под тренировочные данные, а точность предсказания оценивается на тестовом наборе. Результаты затем усредняются по всем разбиениям.

Преимущество такого метода перед кросс-валидацией на K блоках в том, что пропорции тренировочного и тестового наборов не зависят от числа повторений (блоков).

Недостаток метода в том, что некоторые наблюдения могут ни разу не попасть в тестовый набор, тогда как другие могут попасть в него более,

чем один раз. Другими словами, тестовые наборы могут перекрываться. Кроме того, поскольку разбиения проводятся случайно, результаты будут отличаться в случае повторного анализа.

Поэлементная кросс-валидация (Leave-one-out, LOO)

Здесь отдельное наблюдение используется в качестве тестового набора данных, а остальные наблюдения из исходного набора – в качестве тренировочного. Цикл повторяется, пока каждое наблюдение не будет использовано один раз в качестве тестового. Это то же самое, что и K-блочная кросс-валидация, где K равно числу наблюдений в исходном наборе данных [Arlot, Sylvain 2010].

В связи с небольшим объемом корпуса для оценки качества работы нашего алгоритма мы выбрали метод кросс-валидации по K блокам. В нашем случае k = 3.

2.3. Реализация алгоритма

1. Формирование двух подкорпусов.

Были созданы следующие два подкорпуса:

corpusNeg – подкорпус отрицательных отзывов,

corpusPos – подкорпус положительных отзывов.

Для формирования корпуса использовалась процедура load_corpus, параметром которой является наш исходный файл с отзывами о банках.

```
#Заполнение корпуса
def load_corpus(file):
#открываем файл на чтение
   with codecs.open(file, 'r', 'utf8') as file_dict:
#для каждого отзыва в файле
      for line in file_dict:
#массив columns получается после преобразования строки с данными отзыва к нижнему регистру и разбивкой на элементы табуляцией
         columns = line.lower().rstrip('\r\n').split('\t')
#если строка отзыва не пустая
         if len(columns)>6:
#comment – текст отзыва, score – пользовательская оценка
```

```
            comment = columns[4].lower()
            score = int(columns[7])
        #Если оценка равна 1, то добавляем отзыв в корпус недостатков,
            if score=1:
               corpusNeg.append(comment)
            else:
        #Если оценка равна 5, то добавляем отзыв в корпус достоинств
               if score=5:
                  corpusPos.append(comment)
        #читаемый файл по окончании использования закрывается
          file_dict.close()
```

2. Отбор слов-кандидатов на включение в тональные словари на основе полученных подкорпусов.

Определим тональные словари следующим образом:

dictPosAdj – тональный словарь прилагательных положительной окраски,

dictNegAdj – тональный словарь прилагательных отрицательной окраски,

dictPosAdv – тональный словарь наречий положительной окраски,

dictNegAdv – тональный словарь наречий отрицательной окраски,

dictPosNoun – тональный словарь существительных положительной окраски,

dictNegNoun – тональный словарь существительных отрицательной окраски,

dictPosVerb – тональный словарь глаголов положительной окраски,

dictNegVerb – тональный словарь глаголов отрицательной окраски,

Для их заполнения воспользуемся процедурой word_list с 2-мя параметрами: корпусом и его тональностью.

```
        #Заполнение списков тонально окрашенных слов (по частям речи)
        def word_list(corpus,tone):
        #Определяем переменную – морфологический анализатор
          m = MorphAnalyzer()
          ne = False
        #для каждого отзыва в корпусе:
```

```
    for comment in corpus:
#для каждого слова в отзыве
        for w in simple_word_tokenize(comment):
#если не, то пропускаем слово
            if w == 'не':
                ne = True
                continue
#передаем переменной w все ее морфологические параметры
            w = m.parse(w)[0]
#n – начальная форма слова
            n = w.normal_form
#задаем переменную текущего словаря в зависимости от части речи слова w и тональности корпуса corpus
            tekDict={}
#если w – прилагательное, то заполняем тональные словари прилагательных
            if 'ADJF' in w.tag and not ('Anum' in w.tag or 'Apro' in w.tag):
                if tone=='-':
                    tekDict=dictNegAdj
                else:
                    tekDict=dictPosAdj
#если w – существительное, то заполняем тональные словари существительных
            if 'NOUN' in w.tag:
                if tone=='-':
                    tekDict=dictNegNoun
                else:
                    tekDict=dictPosNoun
#если w – глагол, то заполняем тональные словари глаголов
            if 'VERB' in w.tag:
                #pair.append(n)
                if tone=='-':
                    tekDict=dictNegVerb
                else:
                    tekDict=dictPosVerb
#если w – наречие то заполняем тональные словари наречий
            if 'ADVB' in w.tag:
                if tone=='-':
                    tekDict=dictNegAdv
                else:
                    tekDict=dictPosAdv
#если начальной формы слова w нет в рассматриваемом на текущей итерации словаре, тогда мы ее добавляем, иначе увеличивает значение частоты на 1
            if n not in tekDict:
```

```
            if not ne:
                tekDict[n]=1
        else:
            if not ne:
                tekDict[n]+=1
```

Аналогично составляются словари тональных конструкций:

dictPosAdj_Noun – тональный словарь конструкций «прилагательное + существительное» положительной окраски,

dictNegAdj_Noun – тональный словарь конструкций «прилагательное + существительное» отрицательной окраски,

dictPosAdv_Verb – тональный словарь конструкций «глагол + наречие» положительной окраски,

dictNegAdv_Verb – тональный словарь конструкций «глагол + наречие» отрицательной окраски,

dictPosVerb_Noun – тональный словарь конструкций «глагол + существительное» положительной окраски,

dictNegVerb _Noun – тональный словарь конструкций «глагол + существительное» отрицательной окраски,

Пополнение происходит с одним принципиальным отличием – в конструкциях проверяется соответствие морфологическими характеристиками слов, получаемых с помощью морфологического анализатора *m (MorphAnalyze*r). Например, при выделении словосочетания «прилагательное + существительное» выполняется следующая проверка:

```
#Осуществляется проверка двух подряд размещенных слов на соответствие необходимой части речи (в нашем случае «прилагательное + существительное» и «существительное + прилагательное», а также проверяем, что прилагательное не является численным или местоименным)
        if ('ADJF' in w1.tag and 'NOUN' in w2.tag and not ('Anum' in w1.tag or 'Apro' in w1.tag))
        or
        ('ADJF' in w2.tag and 'NOUN' in w1.tag and not ('Anum' in w2.tag or 'Apro' in w2.tag)):
```

#Далее проводится проверка на соответствие морфологических признаков (в нашем случае это род, число, падеж)

if (w1.tag.gender == w2.tag.gender) and (w1.tag.number == w2.tag.number) and (w1.tag.case == w2.tag.case):

3. Заполнение тональных словарей на основе проведенного анализа слов и тональных конструкций.

Нами выбрана наиболее подходящий для анализа статистический критерий – критерий согласия Пирсона (χ^2).

Таблица сопряженности представляет собой набор статистик (a, b, c, d), где a – частота данной лексемы в корпусе достоинств, b – в корпусе недостатков, c – частота прочих лексем в корпусе достоинств и d – частота прочих лексем в корпусе недостатков. Статистика для таблиц сопряженности 2×2 определяется следующим образом:

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)},$$

где $n = a + b + c + d$.

Для выполнения пункта 3 алгоритма воспользуемся функцией *SortKhiDict* с 2-мя параметрами в виде словаря *dict1* и *dict2*, при этом *dict1* – словарь, в который будет внесена коррекция с учетом подсчета χ^2.

```
def SortKhiDict (dict1, dict2):
# считаем сумму частот всех слов каждого из этих словарей
   total1 = sum(dict1.values())
   total2 = sum(dict2.values())
# Определяем переменную, которая будет хранить всю информацию об отсортированных значениях слов
   sortDict={}
#Присваиваем значения переменных вышеуказанной формулы критерия Пирсона и рассчитываем значение χ²
   for word in dict1:
      a = dict1[word]
      b = 0
      if  word in dict2:
         b = dict2[word]
      c = total1 - a
```

```
        d = total2 – b
    #если частота данной лексемы в dict1 больше, чем в dict2, тогда добавляем ее в sortDict, иначе она не является значимой для тонального словаря
        if a>b:
            sortDict[word] = khi(a, b, c, d)
    #производим итоговую сортировку по значениям слов и возвращаем словарь sortDict
        l = lambda x: x[1]
        sortDict = sorted(sortDict.items(), key=l, reverse=True)
        return sortDict
```

Затем было эмпирически подобраны пороговые значения критерия «Хи-квадрат» и отобраны слова для пополнения словарей (прилагательных, существительных, наречий, глаголов) и тональных конструкций («существительное + прилагательное», «существительное + глагол», «глагол + наречие»).

Пороговое значение критерия «Хи-квадрат» для положительно окрашенных прилагательных – 4,9352; для наречий – 4,8354; для существительных – 8,9252; для глаголов – 6,0403.

Пороговое значение критерия «Хи-квадрат» для отрицательно окрашенных прилагательных – 0,6079; для наречий – 1,2413; для существительных – 10,6405; для глаголов – 3,0637.

Таким образом, было выделено 253 положительно окрашенных слова, из них 93 – прилагательные, 63 – наречия, 49 – существительные, 48 – глаголы и 449 отрицательно окрашенных слов, из них 163 – прилагательные, 71 – наречия, 107 – существительные, 108 – глаголы.

Среди положительных слов были выделены: *профессиональный, приятный, быстрый, вежливый, грамотный, качественный; оперативно, удобно, терпеливо; благодарность, отзывчивость; понравиться, помочь, радовать* и другие. Были выделены слова, которые не имеют

положительной тональности, а скорее являются нейтральными: *втбшный, тинькофф, скайп, анна.*

Среди отрицательных слов были выделены: *ипотечный, лишний, хамский, судебный; неожиданно, некорректно, неправомерно; ипотека, задолженность; нарушать, блокировать.* Однако, с наиболее высокими показателями «Хи-квадрат» оказались слова, имеющие на первый взгляд положительную тональность, как например, *хороший, выгодный, выгодно.*

Если мы обратимся к исходному корпусу, то увидим, что *выгодный, выгодно* и правда имеет неоднозначную тональность. Так, то, что выгодно для банка может быть совсем невыгодно для его клиентов:

В условиях без бутылки и оператора не разберешься. Некоторые вещи здравому смыслу не поддаются - например автоматическая отмена автотраты через 30 дней. Хотя если начинаешь думать - почему это ***выгодно*** *банку….*

Почему тогда Сбербанк распорядился капиталом, так как ***выгодно*** *ему!!??*

Банк блокирует ваши деньги и требует, под прикрытием закона №115 (отмывание денег), предоставить обоснование их происхождения, не верит этим обоснованиям, и заставляет вас перечислить деньги обратно отправителю за от суммы ВАШИХ денег. Представляете себе, какой ***выгодный*** *бизнес для банка?!*

Нет же, им лишь бы подороже и ***повыгоднее*** *для себяпродать и навязать услуги.*

Неоднозначная тональность и у прилагательного *хороший.* Так, в корпусе много примеров использования данного прилагательного с сарказмом, чем можно объяснить его попадание на первое место в списке отрицательных прилагательных.

Какой уж тут ***хороший*** *день если не можешь воспользоваться своими деньгами и совершить запланированные платежи!*

Возникла необходимость забрать вклад. Позвонил (идентифицировали по телефону, сразу по имени-отчеству обратились), попросил заказать на завтра деньги. Потребовали серию и номер

паспорта, без этого никак. Ссылаются на некий внутренний регламент, согласно которому меня надо "идентифицировать как клиента", хотя, насколько я знаю, идентификация необходима в случае проведения операции, а в данном случае, никакой операции не производится. А при этом я им должен по телефону выложить все свои паспортные данные. ***Хорошая*** *безопасность…*

Как этот банк принимает сотрудников? Она походу никогда не слышала о качестве обслуживания? Мы с семьей держим вклады с хорошими суммами в этом банке и из-за таких сотрудников как Ц-нко П.С. хочется убежать в другой банк. Кредитную карту нужно было продать, а не втюхать с доп.услугами! Я бы взяла ее и пользовалась, если бы знала как правильно с ней работать. Как говорится ***хорошее*** *впечатление о банке. Клиент приведет как минимум 5-х клиентов, а из-за плохого, уведет 10-х! Управляющая должна была уладить конфликт, я понимаю что у всех планы, но нужно их грамотно предлагать, а не втюхивать. Примите меры, я очень обижена на этот банк из-за такого отношения.*

Полученные списки представлены в Приложениях 1 и 2.

Количество выделенных отрицательно окрашенных слов практически в два раза больше выделенных положительных слов. Это объясняется объемами сформированных подкорпусов (подкорпус достоинств – 833 отзыва и подкорпус недостатков – 1345 отзывов).

Аналогичным образом были подобраны эмпирически пороговые значения критерия «Хи-квадрат» для положительно окрашенных биграмм (10,6330081) и для отрицательно окрашенных биграмм (0,4636189). Таким образом, было выделено 248 положительных и 152 отрицательных словосочетаний. Полученные списки представлены в Приложениях 3 и 4.

Среди отрицательно окрашенных словосочетаний удалось выделить следующие: *коллекторское агенство, банкомат съел, банкомат зажевал, технический сбой, длительное ожидание, хамское поведение, крайне возмущать* и другие. Интересно заметить, что словосочетания, связанные с ипотекой, как например, *ипотечный кредит, ипотечный договор* алгоритм оценил, как отрицательные. Также были выделены некоторые устойчивые словосочетания: *шарашкина контора, мелкий шрифт*.

Среди положительных словосочетаний были выделены как общеупотребимые словосочетания (*качественная услуга, высокий уровень, индивидуальный подход, комфортный офис, достойный сервис)*, так и относящиеся к банковской сфере (*быстро одобрить (кредит), приличный кэшбэк, минимальный процент, выгодная конвертация)*.

Однако в словари словосочетаний попали и нейтральные биграммы: *мобильное приложение, ленинский проспект, волгоградский филиал, банковский счет* и др. Это может быть связано либо с высокой частностью данных словосочетаний, либо с тем, что использование оценок пользователей не может служить достаточным основанием для разбиения корпуса на 2 части (подкорпус достоинств и подкорпус недостатков).

2.4. Оценка работы алгоритма

Оценка работы алгоритма на основе метрик полноты и точности была проведена с использованием кросс-валидации.

Для реализации метода кросс-валидации мы последовательно разбили корпус отзывов на 3 части: corpusA, corpusB, corpusC, содержащие одинаковое количество неповторяющихся отзывов, равное 726.

Далее мы рассмотрели 3 варианта компоновки вышеуказанных частей (подкорпусов) для получения значений полноты и точности:

1. CorpusA U corpusB – обучающая выборка, corpusC – тестовая;
2. corpusB U corpusC – обучающая выборка, corpusA – тестовая;
3. corpusA U corpusC – обучающая выборка, corpusB – тестовая.

Каждый вариант имеет одинаковую последовательность действий оценки работы алгоритма, представленную ниже.

Для анализа отзывов введем значения, используемые при оценивании слов:

-1 – слово из словаря отрицательной тональности;

1 – слово из словаря положительной тональности;

Если перед словом стоит отрицательная частица «не», то тональность рассматриваемого слова меняется на противоположную.

Рассмотрим пример оценивания отзывов; орфография и пунктуация авторов сохранены.

Не в первый раз оказываюсь в крайне неприятной (-1) ситуации в связи с тем, что не могу дистанционно (-1) осуществить переводы. Иногда это нужно ну очень срочно! Как так можно? Сегодня с 16.00 система_висит (-1) постоянно. О какой оперативности (1) может идти речь? Эти технические_сбои (-1), внутренние ошибки (-1) происходят в последнее время постоянно. Сегодня я вообще оказалась в ситуации, когда не могу провести важный и обязательный платеж по причине какой-то ошибки (-1). Очень расстроена (-1) тем, что ВТБ24 так "подставляет" (-1) своих клиентов.

Сумма оценок = -7, отзыв – отрицательный.

Пригласили на собеседование в назначенное время, заставили (-1) больше часа ждать в коридоре, после чего препроводили к мерзкой (-1) хамоватой (-1) бабе (-1), которая наглым (-1) тоном в презрительной (-1) форме пыталась оскорбить (-1) и унизить (-1). В конечном итоге не предоставили никакой информации и выпроводили (-1) вон (-1). Испортили (-1) настроение на весь день.

Сумма оценок = -11, отзыв – отрицательный.

Хочу рассказать об опыте знакомства с Почта банком. Мне срочно были нужны деньги, а времени на сбор документов для кредита не было. Подруга посоветовала (-1) обратиться в офис Почта банка на ул. Тухачевского. Я была приятно (1) удивлена. Располагающая_обстановка (1), все сотрудники доброжелательны (1), даже чай мне предложили. Я если честно (1) насторожилась (-1), думала обманут (-1), раз такие добрые (1) значит в чем то есть подвох (-1). Сотрудница Людмила рассказала_подробно (1) условия кредита (-1), и буквально через 15 минут пришло положительное (1) решение. Я была очень довольна (1), не мурыжили (-1 =>1), не_впаривали (-1 =>1) ничего, все быстро_оформили (1), все объяснили (1). Понятно (1) как погашать кредит (-1), куда обращаться, если вдруг вопросы возникнут. Плюс мне еще банк одобрил (1) карту БЕЗ_процентов (1)!!! Я очень довольна (1) и работой сотрудников банка и кредит (-1) мне отличный (1) предоставили под 16.9% и карта, вообще какая то чудесная (1)! В общем никаких нервов, только положительные (1) эмоции. Спасибо сотрудникам офиса на Тухачевского 11 а! Теперь всем буду советовать (1) к Вам обращаться!!!

Сумма оценок = 13, отзыв – положительный.

При автоматизации алгоритма мы получаем следующий набор действий:

```
#для всех отзывов в корпусе рассчитываем сумму оценок каждого слова, принимая изначально значение переменной  sum_otzyv, равное нулю.
for comments in corpus:
    sum_otzyv = 0
#если текущее слово «не», тогда знак слов приобретает отрицательную окраску
if w == 'не':
            sign = -1
            continue
score = 0
#далее анализ по частям речи
sum_otzyv = 0
#Если текущее слово – прилагательное, то присваиваем переменная score принимает значение -1, если  слово существует в словаре прилагательных с отрицательной окраске.
Если же слово в словаре положительной тональности, тогда значение переменной score равно 1, иначе значение score принимает нейтральное значение.
            if 'ADJF' in w.tag and not ('Anum' in w.tag or 'Apro' in w.tag):
               if n in dictNegAdj:
                  score = -1
               else:
                  if n in dictPosAdj:
                     score = 1
                  else:
                     score = 0
         #Аналогично с остальными частями речи
            if 'NOUN' in w.tag:
               if n in dictNegNoun:
                  score = -1
               else:
                  if n in dictPosNoun:
                     score = 1
                  else:
                     score = 0
            if 'VERB' in w.tag:
               if n in dictNegVerb:
                  score = -1
               else:
                  if n in dictPosVerb:
```

```
                score = 1
            else:
                score = 0
        if 'ADVB' in w.tag:
            if n in dictNegAdv:
                score = -1
            else:
                if n in dictPosAdv:
                    score = 1
                else:
                    score = 0
    #Оценка слова зависит от двух параметров: тональности слова и наличия перед словом частицы «не»
        sum_otzyv += score * sign
    #если пользовательская оценка равна 5, и если отзыв положителен согласно сумме полученных оценок, то переменная суммы совпадений goodNumTest  увеличивается на единицу, иначе увеличивается переменная badNumTest
    if sum_score==5:
        if sum_otzyv>0:
            goodNumTest += 1
        else:
            badNumTest += 1
    else:
        if sum_score==1:
            if sum_otzyv<0:
                goodNumTest += 1
            else:
                badNumTest += 1
```

Таким образом проверяются все отзывы.

В завершении алгоритма нам необходимо оценить полноту и точность на каждом из 3-х подкорпусов: corpusA, corpusB, corpusC.

Обозначим R– полнота, а P – точность нашего алгоритма на корпусе отзывов.

Пусть R_i – полнота, а P_i – точность алгоритма анализа тональности отзывов в i-том тестовом подкорпусе, N_i – количество отзывов в тестовом подкорпусе, K_i – количество определенных в подкорпусе тональностей, T_i –

количество верно определенных алгоритмом тональностей, совпадающих с пользовательской оценкой.

Тогда:

$R_i = \frac{T_i}{N_i}$ — полнота алгоритма анализа тональности в i-том тестовом подкорпусе,

$P_i = \frac{T_i}{K_i}$ — точность алгоритма анализа тональности в i-том тестовом подкорпусе, где $i \in \{A, B, C\}$.

Таким образом, мы получили значения полноты и точности при рассмотрении 3-х вариантов, указанных ранее:

CorpusA U corpusB – обучающая выборка, corpusC – тестовая

$N_C = 726, \quad K_C = 622, \quad T_C = 483, \quad R_C = 66{,}5\%, \quad P_C = 77{,}7\%$

corpusB U corpusC – обучающая выборка, corpusA – тестовая;

$N_A = 726, \quad K_A = 617, \quad T_A = 475, \quad R_A = 65{,}4\%, \quad P_A = 76{,}9\%$

corpusA U corpusC – обучающая выборка, corpusB – тестовая.

$N_B = 726, \quad K_B = 634, \quad T_B = 511, \quad R_B = 70{,}4\%, \quad P_B = 80{,}6\%$

Полнота и точность нашего алгоритма при кросс-валидации есть среднее арифметическое из значений, полученных на каждом подкорпусе. Следовательно, полнота R = 67,4%, точность P = 78,4%.

2.5. Выводы к главе II

Во второй главе была описана практическая часть работы. Мы рассмотрели основные шаги разработанного алгоритма, а также инструменты и методы, которые использовались в данной работе.

В качестве языка разработки был выбран Python; использованы библиотеки Pymorphy2, Codecs, Math.

Для оценки корреляции между событиями «слово содержится в отзыве» и «отзыв описывает недостатки/достоинства» мы построили

таблицы сопряженности и использовали критерия согласия Пирсона («Хи-квадрат»).

Подобрав эмпирически пороги значений для полученных списков слов, мы получили словари тонально окрашенных слов и словосочетаний.

Качество работы алгоритма было оценено с помощью полноты и точности на основе метода кросс-валидации. Полученные результаты: полнота – 67,4%, точность – 78,4%.

Заключение

В первой главе мы рассмотрели предметную область анализа тональности, обозначили основные задачи, которые могут решаться в рамках анализа тональности, и основные проблемы, возникающие при анализе отзывов. Были выделены основные подходы сентимент-анализа: подход, основанный на словарях и на правилах, и подход, основанный на машинном обучении (с учителем и без учителя), рассмотрели достоинства и недостатки каждого подхода. Затем мы выполнили обзор работ, посвященных автоматическому составлению тональных словарей.

Во второй главе мы описали разработку и реализацию алгоритма по автоматическому пополнению тональных словарей для банковской сферы.

Основные шаги реализации алгоритма включали в себя формирование двух подкорпусов: текстов с описаниями достоинств и текстов с описаниями недостатков; отбор слов-кандидатов на включение в тональные словари на основе полученных подкорпусов; заполнение тональных словарей на основе проведенного анализа слов и тональных конструкций; оценка работы алгоритма на основе метрик полноты и точности с использованием кросс-валидации.

В качестве языка разработки был выбран Python в связи с наличием библиотек, наиболее подходящих для реализации алгоритма по функционалу, затрачиваемой памяти и скорости выполнения. Для реализации алгоритма использовались библиотеки Pymorphy2 (библиотека для морфологического анализа), Codecs (библиотека, осуществляющая связь между интерпретатором и файловой системой операционной системы), Math (работа с математическими функциями).

В качестве основного критерия анализа распределения слов по тональным словарям была выбрана оценка корреляции двух случайных событий: «слово содержится в отзыве» и «отзыв описывает

недостатки/достоинства». Для оценки корреляции мы построили таблицы сопряженности и использовали критерия согласия Пирсона («Хи-квадрат»).

Затем были эмпирически выбраны пороги значений для полученных списков слов и получены словари тонально окрашенных прилагательных, наречий, существительных и глаголов (702 слова), а также словосочетаний (прилагательное + существительное, наречие + глагол, существительное + глагол и др.) (400 словосочетаний). Удалось выделить как общеупотребительные тонально окрашенные слова и словосочетания (*грубый, отвратительный, доброжелательный, потрясающе, неприятная ситуация, страшный сон*), так и относящиеся к банковской сфере (*ипотечный кредит, приличный кэшбэк, минимальный процент, банкомат зажевал (карту), огромная комиссия).*

Для оценки качества работы алгоритма был использован метод кросс-валидации, показавший достаточно высокие результаты: полнота (67,4%) и точность (78,4%). Также к достоинствам разработанного алгоритма можно отнести простоту реализации и его легкую адаптацию к другим предметным областям.

Были замечены также и некоторые недостатки:

– недостаточно большой исследуемый корпус для полного охвата лексики банковской сферы;

– использование оценок пользователей в отзывах для разбиения корпуса на 2 части (подкорпус достоинств и подкорпус недостатков) влечет большое количество шума.

В заключении стоит отметить, что во время проведения исследования были выявлены следующие факты:

1. На основе полученных словарей можно сделать вывод, что тонально окрашенная лексика банковской сферы, в основном, носит общеупотребительный характер (*хороший, положительный, удобный, грубый*).

2. В собранном корпусе оказалось, что количество отрицательных отзывов более, чем в 1,5 раза превышает количество положительных отзывов. Так, объем подкорпуса достоинств составил 833 отзыва, объем подкорпуса недостатков – 1345 отзывов. Это отражает негативный характер обслуживания в банковской сфере в целом.

Данный алгоритм в дальнейшем можно применить для автоматического составления предметно-ориентированных тональных словарей других предметных областей, а также его можно использовать для маркетинговых исследований банковских услуг.

Список использованной литературы

1. Аптон Г. Анализ таблиц сопряженности. – М., 1982. – 143 с.
2. Дегтева А.В., Азарова И.В. Структура эмоционально-экспрессивного компонента в тезаурусе русского языка RussNet // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). – М., 2013. – С. 200–211.
3. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные м методы и технологии, электронные коллекции» (RCDL-2012). – Переславль-Залесский, 2012. – С. 81-86.
4. Лемешко Б.Ю., Лемешко С.Б., Постовалов С.Н., Чимитова Е.В. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. — Новосибирск, 2011. – 888 с.
5. Морфологический анализатор pymorphy2: [Электронный ресурс] / URL: https://pymorphy2.readthedocs.io.
6. Пазельская А., Соловьев А. Метод определения эмоций в текстах на русском языке // The international conference on computational linguistics and intellectual technologies “Dialogue 2011”. – М., 2011. – С. 510 - 522.
7. Протопопова Е.В., Букия Г.Т., Митрофанова О.А. Автоматическое составление тонального словаря для процедур сентиментного анализа // Материалы XLV Международной филологической конференции 14-21 марта 2016 г. СПб., 2017 (в печати).
8. Россум Г., Дрейк Ф.Л.Дж., Откидач Д.С. Язык программирования Python. 2001 г. – 454 с.
9. Степнов М.Н. Статистические методы обработки результатов механических испытаний: Справочник. – М., 1985. – С. 81–83.

10. Хохлова М.В. Анализ тональности. – СПб, 2016. – 11 с.

11. Цветков А.Д. Анализ тональности сообщений социальной сети Twitter. – Томск, 2013. – 31 с.

12. Шитиков В.К., Розенберг Г.С., Зинченко Т.Д. Таблицы сопряженности и “интервальная” математика. – Тольятти, 2003. – С. 259-266.

13. Applications for Python: [Электронный ресурс]. URL: http://www.python.org.

14. Arlot, Sylvain, Alain Celisse. A survey of cross-validation procedures for model selection // Statistics surveys 4. 2010. Pp. 40-79.

15. Bing Liu. Sentiment Analysis and Subjectivity // Handbook of Natural Language Processing. 2010. 49 p.

16. Blinov P., Kotelnikov E. Using Distributed Representations for Aspect-Based Sentiment Analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2014». № 13 (20). Vol. 2. 2014. Pp. 68–79.

17. Bo Pang, Lillian Le. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. №2. 2008. Pp. 1-135.

18. Bo Pang, Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales // In Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL). – University of Michigan, USA, 2005. Pp. 115–124.

19. Cambria E., Havasi C., Hussain A. SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis // AAAI FLAIRS. Marco Island, 2012. Pp. 202-207.

20. Chetviorkin I.I., Loukachevitch N.V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // Proceedings of COLING 2012: Technical Papers. 2012. Pp. 593–610.

21. Ding X., Liu B., Yu P. A holistic lexicon-based approach to opinion mining // Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008). 2008. 9 p.

22. Dubatovka A., Kurochkin Yu., Mikhailova E. Automatic Generation of the Domain-Specific Sentiment Russian Dictionaries // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016». №15 (22). 2014. Pp. 146–154.

23. Esuli A. and Sebastiani F. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining // Proceedings of the International Conference on Language Resources and Evaluation. Genoa, Italy, 2006. Pp. 417−422.

24. Hatzivassiloglou V., McKeown K. Predicting the semantic orientation of adjectives // Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997). 1997. Pp 174-181.

25. Hatzivassiloglou V., Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity // Proceedings of International Conference on Computational Linguistics (COLING-2000). 2000. Pp. 299-205.

26. Kan D. Rule-based approach to sentiment analysis // Sentiment Analysis Track at ROMIP, 2011. 7 p.

27. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2015». № 14 (21). Vol. 2. 2015. Pp. 22–33.

28. Joachims T. Making large-scale SVM learning practical // Schölkopf B. Advances in kernel methods: support vector learning. 1999. Pp. 41-54.

29. Kotelnikov E.V., Bushmeleva N.A., Razova E.V., Peskisheva T.A., Pletneva M.V. Manually Created Sentiment Lexicons: Research and Development // Computational Linguistics and Intellectual Technologies:

Proceedings of the International Conference «Dialogue 2016» №. 15 (22). 2016. Pp. 300–316.

30. Liu H. MontyLingua: An end-to-end natural language processor with common sense. 2004.

31. Oren T., Dmitry D., Ari R. ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews // AAAI Conference on Artificial Intelligence. 2010. Pp. 162-169.

32. Popescu A., Etzioni O. Extracting product features and opinions from reviews // Proceedings of Conference on Empirical Meth- ods in Natural Language Processing (EMNLP-2005). 2005. Pp. 339-346.

33. Snyder B., Barzilay R. Multiple Aspect Ranking using the Good Grief Algorithm // Proceedings of the Joint Human Language Technology (North American Chapter of the ACL Conference (HLT-NAACL)). 2007. Pp. 300–307.

34. Sokolova M., Bobicev V. Classification of Emotion Words in Russian and Romanian Languages // Proceedings of RANLP-2009 conference. Borovets, Bulgaria, 2009. Pp. 415-419.

35. Strapparava C., Valitutti A. Wordnet-affect: an affective extension of Wordnet // Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, 2004. Pp. 1083-1086.

36. Thelwall M., Buckley K., Paltoglou G. Sentiment strength detection in short informal text // Journal of the American Society for Information Science and Technology, 61(12). 2010. Pp. 2544–2558.

37. Tong R. An operational system for detecting and tracking opinions in on-line discussions // Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification. New York, 2001. Pp. 1-6.

38. Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. 2002. Pp. 417–424.

39. Ulanov A., Sapozhnikov G. Context-Dependent Opinion Lexicon Translation with the Use of a Parallel Corpus // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2013». № 12 (19). 2013. Pp. 165–174.

40. Wiegand M., Balahur A., Roth B., Klakow D., Montoyo. A. A Survey on the Role of Negation in Sentiment Analysis // Workshop on Negation and Speculation in Natural Language Processing. 2010. Pp. 334-337.

Приложение 1. Словарь положительно окрашенных слов, выделенных с помощью алгоритма, со значениями «Хи-квадрат»

Прилагательные

профессиональный	128,3876618
приятный	122,9193977
огромный	96,90717053
быстрый	88,27634295
отличный	85,23887988
удобный	82,5190913
доброжелательный	80,56609718
качественный	72,19217038
грамотный	72,19217038
вежливый	66,32993515
карьерный	59,2565157
высокий	57,90439787
отзывчивый	49,37539838
оперативный	49,10035222
приветливый	42,73046346
внимательный	41,42412258
положительный	34,3259986
уютный	29,61920322
чуткий	29,61920322
сложный	28,05108837
оптимальный	24,68141221
привлекательный	24,52876996
искренний	24,52876996
беспроцентный	24,36209743
позитивный	23,85835488
знающий	19,31680094
дружелюбный	18,93549763
сжатый	14,80733899
классный	14,23623332
высококвалифицированный	14,23623332
комфортный	10,63173878
нестандартный	10,63173878
исчерпывающий	10,63173878
функциональный	9,871056625
дружественный	9,871056625
информативный	9,871056625

вкусный	9,871056625
душевный	9,871056625
уважительный	9,871056625
высокопрофессиональный	9,871056625
внушительный	9,871056625
несомненный	9,871056625
доходный	9,871056625
опытный	8,1149081
бодрый	5,315057227
благоприятный	5,315057227
гибкий	5,315057227
улыбчивый	5,315057227
чистенький	4,935276987
плодотворный	4,935276987
взаимовежливый	4,935276987
благодарственный	4,935276987
интеллигентный	4,935276987
партнерский	4,935276987
паевой	4,935276987
терпеливый	4,935276987
тактичный	4,935276987
неравнодушный	4,935276987
клиенто-ориентированный	4,935276987
разговорчивый	4,935276987
благодарный	4,935276987
молчаливый	4,935276987
учтивый	4,935276987
клиентоориентированный	4,935276987
втбшный	4,935276987
доходчивый	4,935276987
наглядный	4,935276987
трудолюбивый	4,935276987
изящный	4,935276987
вызывающий	4,935276987
извечный	4,935276987
деликатный	4,935276987
опрятный	4,935276987
новомодный	4,935276987
практичный	4,935276987
стильный	4,935276987

стремительный	4,935276987
доверительный	4,935276987
общедоступный	4,935276987
низкопроцентный	4,935276987
обаятельный	4,935276987
добропорядочный	4,935276987
внятный	4,935276987
бонусный	4,935276987
неплохой	4,935276987
законопослушный	4,935276987
идеальный	4,935276987
безукоризненный	4,935276987
культурный	4,935276987
надежный	4,935276987
восхитительный	4,935276987
креативный	4,935276987
усердный	4,935276987
добрый	4,935276987
довольный	4,935276987
чудесный	4,935276987

Наречия

быстро	326,7325621
приятно	237,4520014
оперативно	148,9537115
профессионально	90,31003873
удобно	67,38326815
подробно	50,33827327
особенно	45,37027674
грамотно	41,51596488
терпеливо	32,4271478
редко	30,04125986
доходчиво	27,84364628
доступно	21,93304349
компетентно	18,52075224
качественно	15,36095431
отдельно	15,36095431
доброжелательно	14,50809565
комфортно	13,91795274
прилично	9,671526071
зимой	9,671526071

благожелательно	9,671526071
солидарно	9,671526071
высоко	9,671526071
честно	9,671526071
гладко	9,671526071
незадолго	9,671526071
торжественно	9,671526071
безболезненно	9,671526071
надёжно	9,671526071
неплохо	7,892349122
идеально	5,183436823
бойко	4,835494218
безукоризненно	4,835494218
безгранично	4,835494218
неподдельно	4,835494218
потрясающе	4,835494218
полноценно	4,835494218
культурно	4,835494218
тепло	4,835494218
люто	4,835494218
скептически	4,835494218
катастрофически	4,835494218
посменно	4,835494218
дёшево	4,835494218
светло	4,835494218
вежливо	4,835494218
терпимо	4,835494218
позитивно	4,835494218
квалифицированно	4,835494218
повсеместно	4,835494218
материально	4,835494218
персонально	4,835494218
утомительно	4,835494218
бесцельно	4,835494218
молниеносно	4,835494218
эффективно	4,835494218
тактично	4,835494218
по-русски	4,835494218
опрятно	4,835494218
дотошно	4,835494218

предметно	4,835494218
неуклонно	4,835494218
аккуратно	4,835494218
чудесно	4,835494218
понятно	4,835494218

Существительные

благодарность	561,2721336
тинькофф	288,009898
профессионализм	200,661873
отзывчивость	85,61711388
профессионал	78,84417685
мкб	72,39991935
оперативность	70,15538703
кэшбэк	68,57146987
атмосфера	68,42769923
скайп	66,6647089
плюс	55,59741234
анна	52,85553631
успех	51,55039837
доброжелательность	48,55952868
процветание	42,47528494
молодец	40,42134563
клиентоориентированность	32,99252699
похвала	31,10190947
эмоция	30,89570379
рост	30,54089315
автокредитование	30,299661
удача	29,84989537
компетентность	28,38765792
признательность	24,88131556
нюанс	24,78018313
потребность	22,85287159
скорость	22,85287159
кофе	21,04284991
анатолий	19,90244701
вежливость	19,45058824
восторг	19,45058824
респект	18,66082766
пунктуальность	18,66082766

пятёрка	18,66082766
достоинство	18,66082766
парковка	18,66082766
коммуникабельность	18,66082766
грамотность	18,337309
простота	18,337309
гарантия	15,8698118
любовь	13,12047516
уважуха	12,44044577
честность	12,44044577
чуткость	12,44044577
чёткость	12,44044577
доброта	12,44044577
спецпредложение	10,99619705
конкурент	10,99619705
тонкость	8,925205583

Глаголы

хотеться	157,5472843
понравиться	135,2747002
выражать	118,1390376
помочь	111,2218953
рассказать	88,09797335
желать	74,23309504
нравиться	71,08570059
радовать	47,45527442
консультировать	47,07205787
порадовать	41,97701765
чаять	30,20492144
разъяснить	21,18889767
проконсультировать	19,18189129
реагировать	12,13306651
уменьшать	12,13306651
посмеиваться	12,08096512
отработать	12,08096512
отслеживать	12,08096512
отобрать	12,08096512
познакомиться	12,08096512
накопить	12,08096512

расценить	12,08096512
рефинансировать	12,08096512
продаваться	12,08096512
содействовать	12,08096512
скорректировать	12,08096512
прояснить	12,08096512
остановить	10,59239542
зарегистрироваться	8,605904892
восстановить	6,775946517
сопроводить	6,775946517
учитывать	6,775946517
сохранить	6,775946517
развиваться	6,775946517
обновляться	6,775946517
подняться	6,775946517
выручать	6,775946517
расспросить	6,775946517
обожать	6,040315334
объяснить	6,040315334
выводить	6,040315334
одобрить	6,040315334
способствовать	6,040315334
подталкивать	6,040315334
восхищаться	6,040315334
любоваться	6,040315334
укрепить	6,040315334
посоветовать	6,040315334

Приложение 2. Словарь отрицательно окрашенных слов, выделенных с помощью алгоритма, со значениями «Хи-квадрат»

Прилагательные

хороший	76,88497585
денежный	53,28255464
большой	45,4096593
отдельный	23,51877973
клиентский	23,51877973
выгодный	23,06759106
дебетовый	22,38374583
различный	19,06709743
уважаемый	17,23193246
полный	15,46407439
технический	14,86222039
контактный	14,84741212
небольшой	14,61892882
особый	14,2066762
компетентный	13,53779401
бесплатный	10,2916702
ипотечный	9,60193992
платёжный	8,801875578
очередной	8,649428464
выходной	8,491882665
повышенный	8,422488962
добровольный	8,29087471
доступный	8,15969466
лишний	7,698270539
ежемесячный	7,660318313
последний	6,55299977
низкий	5,843382996
повторный	5,715419776
хамский	5,681830114
расходный	5,681830114
некомпетентный	5,478628245
странный	5,478628245
недавний	5,27861443
карточный	5,275447097
вышеуказанный	5,275447097
грубый	5,275447097

судебный	5,229734134
страшный	5,072286667
бывший	4,869146951
календарный	4,869146951
ближний	4,848531548
противный	4,666027948
понятный	4,660517955
частичный	3,853758987
коллекторский	3,853758987
проблемный	3,827469498
маленький	3,708942936
премиальный	3,650743494
подобный	3,567734902
отвратительный	3,528409916
кредитный	3,452413818
претензионный	3,244774583
физический	3,102915302
ужасный	3,102891221
прежний	3,033568184
головной	2,997920659
наглый	2,838888417
неверный	2,75870394
старый	2,701665824
ошибочный	2,635976356
платный	2,619903472
ненадлежащий	2,23021426
системный	2,23021426
халатный	2,23021426
штрафной	2,23021426
отрицательный	2,148087275
тупой	2,02736422
обманный	2,02736422
неудобный	2,02736422
неправильный	2,02736422
неизвестный	2,006323575
немалый	1,913247397
неправомерный	1,824534847
необоснованный	1,824534847
несчастный	1,824534847
недобросовестный	1,824534847

незаконный	1,64017094
недостоверный	1,621726138
некорректный	1,460624496
непрофессиональный	1,41893809
безвозмездный	1,41893809
больной	1,41893809
бесполезный	1,41893809
несвоевременный	1,41893809
медленный	1,41893809
нервный	1,216170701
нерабочий	1,216170701
недействительный	1,216170701
хамоватый	1,216170701
жалобный	1,216170701
неадекватный	1,216170701
замкнутый	1,216170701
издевательский	1,216170701
секретный	1,216170701
ограниченный	1,216170701
неоплаченный	1,216170701
бедный	1,216170701
несанкционированный	1,216170701
неуважительный	1,013423965
нерадивый	1,013423965
наплевательский	1,013423965
внутрибанковский	1,013423965
глупый	1,013423965
безобразный	1,013423965
левый	1,013423965
глухой	1,013423965
неблагонадёжный	1,013423965
бешеный	1,013423965
виновный	1,013423965
автоматизированный	1,013423965
принудительный	1,013423965
дикий	1,013423965
неграмотный	1,013423965
уголовный	1,013423965
недобровольный	1,013423965
истекший	1,013423965

невнятный	1,013423965
безграмотный	1,013423965
безответственный	1,013423965
непосредственный	1,013423965
транзитный	1,013423965
заказной	1,013423965
устаревший	1,013423965
недопустимый	1,013423965
спорный	0,853546751
затруднительный	0,810697882
неподходящий	0,810697882
недополученный	0,810697882
идентификационный	0,810697882
чудовищный	0,810697882
бестолковый	0,810697882
назойливый	0,810697882
невозможный	0,810697882
мифический	0,810697882
грязный	0,810697882
недостаточный	0,810697882
жалкий	0,810697882
грабительский	0,810697882
покойный	0,810697882
жуткий	0,810697882
свинский	0,810697882
военный	0,810697882
неопределённый	0,810697882
навязчивый	0,810697882
бездарный	0,607992447
противоречивый	0,607992447
жёсткий	0,607992447
наследственный	0,607992447
презрительный	0,607992447
оскорбительный	0,607992447
неподъемный	0,607992447
неполученный	0,607992447
противозаконный	0,607992447
некрасивый	0,607992447
неприемлемый	0,607992447
дерзкий	0,607992447

некачественный	0,607992447
глючный	0,607992447
злостный	0,607992447
невнимательный	0,607992447
сырой	0,607992447
мерзкий	0,607992447
неприятный	0,607992447

Наречия

очень	372.38100773609733
всегда	90.61081781330103
буквально	32.28533142613084
онлайн	17.11398601338235
неожиданно	12.021069203485387
случайно	10.974274619983852
впервые	9.798711946183756
незаконно	8.949404378855343
никогда	8.192256210970456
бесплатно	6.815278656697748
неизвестно	6.662856145706663
невозможно	6.007335975410629
сложно	6.003269520300492
дополнительно	5.55088257933299
частично	5.3849879794322515
долго	5.215859915558338
одновременно	5.0879438191820325
дистанционно	3.933650704742038
лично	3.883037350252151
непонятно	3.8330287984369873
некорректно	3.7264091003213915
выгодно	3.7003579622033085
нагло	3.682944028969584
недостаточно	3.682944028969584
неправомерно	3.674107765276916
ужасно	3.4759287878850516
упорно	3.4427433266266267
настойчиво	3.407366402790054
молча	3.404985403638322
заведомо	3.31199507194224
медленно	2.9409028859392095
неправильно	2.7664202249353576
стыдно	2.7605469476445315
самовольно	2.48344362928676

необоснованно	2.48344362928676
слабо	2.48344362928676
резко	2.459412192586282
возмутительно	2.3369665563733255
наотрез	2.320012980063282
ошибочно	2.2763633741239366
неудобно	2.2763633741239366
дорого	2.2460802589120368
бесконечно	1.8622719634885727
умышленно	1.8622719634885727
халатно	1.8622719634885727
впустую	1.8622719634885727
хамски	1.6552608003306035
печально	1.6552608003306035
нервно	1.6552608003306035
негативно	1.6167687610756645
бессмысленно	1.5934443761376036
раздражённо	1.5308065373526039
дико	1.5179059316134689
жёстко	1.471899394049768
проблематично	1.470810481172733
пошло	1.4482726601569147
пренебрежительно	1.4482726601569147
демонстративно	1.4482726601569147
неохотно	1.4482726601569147
безумно	1.4482726601569147
сухо	1.4482726601569147
нелегко	1.4482726601569147
некрасиво	1.4482726601569147
скрытно	1.361181951710605
мерзко	1.3601988520283823
нежданно-негаданно	1.3399715883266539
отрицательно	1.281131751833459
по-хамски	1.2413075391269286
болезненно	1.2413075391269286
безуспешно	1.2413075391269286
небезопасно	1.2413075391269286
вон	1.2413075391269286

Существительные

сделка	106.76461758264102
заявление	101.39513206415496
помощь	69.20338161079702

работа	65.45273275625031
обслуживание	62.47155703568859
почта	61.77878099602801
ответ	50.5136812812324
ипотека	44.98044220765685
счёт	44.264674556172324
приложение	41.08233768282826
выбор	37.8945794104785
консультация	36.97914898863858
номер	36.342849501678295
сумма	34.79372563191442
договор	33.88239970027029
задолженность	33.83735759920529
обращение	33.29890282737326
специалист	32.60055663851968
менеджер	29.94138255437389
квартира	29.90281997919761
впечатление	28.314460858052467
уровень	28.05188555723785
претензия	27.698051268566946
вопрос	27.22667014525686
суд	26.56652236346923
продукт	26.014368762156273
кредитование	24.939099119746786
сбербанк	20.79019056715991
кредит	20.703278542253262
ставка	20.69858710913087
право	20.57937713593288
деньга	20.073339351637035
клиент	20.073339351637035
звонок	19.984925328622055
средство	19.507304578063916
день	19.05489077635971
телефон	19.00705529473533
сервис	18.423827676791365
срок	18.38166583600195
оформление	18.235460418881235
елена	18.207586932520865
решение	18.013767089649697
уважение	17.892703831735904
погашение	17.786837914963197
опыт	17.76732712696676
подвох	17.387754342866742
баба	17.367508295071183
долг	17.29722124581256

закон	17.187057667711134
сотрудничество	16.924255662627992
оператор	16.924255662627992
скидка	16.42091742034127
настроение	16.258255027059036
персонал	15.927176064438319
месяц	15.879702811839989
отношение	15.738154025158384
рынок	15.714614145858697
отказ	15.688621132311278
рассмотрение	15.435863980683422
филиал	15.303989620169888
офис	15.19681999385822
размер	15.185633382781454
причина	15.113228800850422
лимит	14.964517081608005
бонус	14.839088431902177
процесс	14.811229165557982
друг	14.754151321356472
автосалон	14.74614663111455
случай	14.74517780545298
россия	14.504373781162267
форма	14.091453254321456
качество	13.540058747868386
пора	13.50250268300937
сравнение	13.455591012873112
пара	13.381861330827858
отзыв	13.246705947474071
просрочка	13.206386007897274
заёмщик	13.201712007482508
прокуратура	13.092401239477832
нарушение	12.981827312409735
ошибка	12.859478114505146
сомнение	12.79796016849143
проблема	12.68906129318267
жалоба	12.656857459829064
взыскание	12.646086758996661
штраф	12.59302117260472
пеня	12.506246615029552
заявка	12.434745577111395
оценка	12.360498664997442
дата	12.35673415405732
слово	12.321550768346608
орган	12.321550768346608
распоряжение	12.168497136573519

коллектор	12.090211055660552
заблуждение	11.910383512026472
заинтересованность	11.663978446741304
сбой	11.6214496429849
комиссия	11.463862129768092
сеть	11.45628058299468
недовольство	11.370634373278584
расторжение	11.370634373278584
блокировка	11.260524037606269
разбирательство	11.141189222070446
нужда	11.099564777404726
некомпетентность	10.938608261723584
отписка	10.728290334190822
негатив	10.640517497820126

Глаголы

хотеть	101.65515312854407
сказать	54.18895911383218
рекомендовать	48.29583530527551
пользоваться	43.59532498403919
одобрить	31.975604288578918
сообщить	24.89352594671044
работать	23.460204452088075
звонить	21.244012920980825
удивить	20.2756020074696
оформлять	20.023915147950017
выбрать	19.517820921424217
говорить	19.383545347068136
помогать	19.251545793872857
просить	18.89694657143783
объяснить	16.8547840351227
обманывать	16.0047479862713
требовать	15.672073409892253
обслуживаться	15.33515753039724
оскорбить	12.431146304515392
понять	12.224503952404284
показать	12.143339843968029
потерять	11.611621045529239
написать	11.466695626125244
унизить	10.892975267751677
устраивать	10.638813621362582
подсказать	10.153198430521426
решаться	10.153198430521426

встретить	9.435163803901109
встречать	9.435163803901109
повезти	9.435163803901109
делать	9.175130971005116
состояться	8.817577969727637
бывать	8.79086879990561
внести	8.530249957806898
выбирать	8.193775745503958
заставлять	8.193775745503958
позвонить	8.187385577948103
пригласить	7.99727552771066
получать	7.789726414118925
попадать	7.428808406262155
спрашивать	7.202290337281869
составить	7.202290337281869
сделать	7.153764823922966
понадобиться	7.024460971168842
объяснять	6.912566519793516
ждать	6.399321547621344
ответить	6.370646889682164
подставлять	6.287341864424683
выполнять	6.279142026484534
списывать	6.131958425315409
отправить	5.873709249067022
менять	5.597787224336539
возникнуть	5.537255688542249
иметься	5.514943869206626
позволить	5.461760854699728
взять	5.350372865630545
поднять	5.317344397007427
войти	5.317344397007427
успеть	5.317344397007427
поступать	5.211014487426162
нарушать	5.136732706966496
считать	5.1136383008210275
принять	4.995688251866208
пояснить	4.984677051052638
списать	4.979708293938981
оформить	4.764663402890848
продолжать	4.738635049687639
отнимать	4.716667540633862
расплачиваться	4.716667540633862
скинуть	4.716667540633862
заработать	4.709044161111483
блокировать	4.639243923584604

выяснить	4.628892263972678
выпроводить	4.533694805699878
уверять	4.473432705630451
посещать	4.43360043926284
подготовить	4.43360043926284
обращаться	4.3415996459753305
предложить	4.337330973408582
оплачивать	4.32293390569873
названивать	4.307630674656821
выясняться	3.8102796959844523
сомневаться	3.6043754139539637
снизить	3.5633272068284425
убедить	3.5633272068284425
почувствовать	3.5633272068284425
отдать	3.5425620080782205
отказать	3.5313406343803426
попытаться	3.5216963162330543
знать	3.4574129371302296
принести	3.453793763130896
использовать	3.453793763130896
искать	3.453793763130896
видеть	3.444979956507091
отказаться	3.4052868398434692
утверждать	3.3840737628709947
впарить	3.3724353474463977
изменить	3.3130113726557404
тратить	3.3130113726557404
приходить	3.292028543387804
потребовать	3.2260517971650216
насторожиться	3.1988868162936757
угрожать	3.14727362917241
извиниться	3.14727362917241
подтвердить	3.120071647193565
относиться	3.1126514402742385
испортить	3.063736995290739
мурыжить	3.063736995290739

Приложение 3. Словарь отрицательно окрашенных словосочетаний, выделенных с помощью алгоритма, со значениями «Хи-квадрат»

технический сбой	11,1004837
ипотечный кредит	9,6792224
противный случай	8,6802996
валютная карта	4,3854381
страховая премия	4,2194027
кредитная организация	4,1447418
банковский счёт	4,1447418
годовое обслуживание	3,7985371
ужасный банк	3,76744
длительное ожидание	3,3902397
российский капитал	3,0131409
негативный отзыв	3,0131409
лицевой счёт	3,0131409
платная услуга	3,0131409
расходная операция	3,0131409
главный бухгалтер	3,0131409
неверный информация	2,6361436
повторный обращение	2,6361436
страховой взнос	2,6361436
платёжное поручение	2,532163
некорректная информация	2,2592477
неприятная ситуация	2,2592477
страшный сон	2,2592477
жалобная книга	2,2592477
недостоверная информация	2,2592477
замкнутый круг	2,2592477
коллекторское агенство	1,8824532
отвратительный банк	1,8824532
ложная информация	1,8824532
некомпетентный сотрудник	1,8824532
мелкий шрифт	1,8824532
очень сожалеть	1,8560258
длительное поручение	1,50576
сплошной обман	1,50576
ипотечный отдел	1,50576
ипотечный менеджер	1,50576
ипотечный договор	1,50576
очень разочароваться	1,3916316
халатно относиться	1,3916316
уже потерять	1,3916316
грубо ответить	1,3916316
долго рассматривать	1,3916316

негативный опыт	1,3793994
неприятное впечатление	1,3793994
бесплатное снятие	1,3793994
шарашкина контора	1,3793994
отвратительное обслуживание	1,1291682
очередная жалоба	1,1291682
полное отсутствие	1,1291682
банковская ошибка	1,1291682
хамское обращение	1,1291682
ненадлежащее исполнение	1,1291682
банк нарушил	1,0301297
связь оборваться	1,0301297
система зависнуть	1,0301297
незаконно удерживать	0,9274961
нагло врать	0,9274961
постоянно зависать	0,9274961
постоянно перебивать	0,9274961
полностью отсутствовать	0,9274961
устная претензия	0,7526776
крайнее возмущение	0,7526776
истекший срок	0,7526776
медленное обслуживание	0,7526776
плохой сайт	0,7526776
наглый образ	0,7526776
странный банк	0,7526776
тихий ужас	0,7526776
хамское поведение	0,7526776
мотивированная жалоба	0,7526776
отрицательный отзыв	0,7526776
злой умысел	0,7526776
системная ошибка	0,7526776
хамский тон	0,7526776
рутинная операция	0,7526776
плохая репутация	0,7526776
незаконное списание	0,7526776
минусовый баланс	0,7526776
банкомат съесть	0,6889512
претензия быть	0,6864464
отказ прийти	0,6864464
система повиснуть	0,6864464
банкомат зажевать	0,6864464
техническая ошибка	0,51137
очень жалеть	0,4636189
совсем обнаглеть	0,4636189

крайне возмущать	0,4636189
честно пожалеть	0,4636189
неправомерно быть	0,4636189
весьма ухудшиться	0,4636189
неохотно отвечать	0,4636189
незаконно получать	0,4636189
всячески хамить	0,4636189
дорого обходиться	0,4636189
бесследно исчезнуть	0,4636189
всего возмущать	0,4636189
всячески препятствовать	0,4636189
опять пропасть	0,4636189
открыто хамить	0,4636189
постоянно бесить	0,4636189
последовательно ухудшать	0,4636189
всего наобещать	0,4636189
вчера заблокировать	0,4636189
грубо отказать	0,4636189
незаконно обогатиться	0,4636189
незаконно произвести	0,4636189
прямо нарушать	0,4636189
сильно сомневаться	0,4636189
цинично выглядеть	0,4636189
реально издеваться	0,4636189
необоснованно начислить	0,4636189
умышленно умалчивать	0,4636189
давно кредитоваться	0,4636189
ошибочно положить	0,4636189
дооооолго аутентифицироваться	0,4636189
сильно ошибаться	0,4636189
нагло нарушаться	0,4636189
специально затянуть	0,4636189
небрежно относиться	0,4636189
всегда косячать	0,4636189
самовольно подключить	0,4636189
долго проверять	0,4636189
просто-напросто ущемлять	0,4636189
сильно отличаться	0,4636189
далее начинаться	0,4636189
самовольно заблокировать	0,4636189
насильно подключить	0,4636189
оперативно отреагировать	0,4636189
круто развести	0,4636189
нагло лгать	0,4636189
незаконно обуславливать	0,4636189

постоянно навязывать	0,4636189
неоднократно обещать	0,4636189
медленно работать	0,4636189
опять обмануть	0,4636189
сильно огорчить	0,4636189
хамски отвечать	0,4636189
хамски общаться	0,4636189
неприятно удивить	0,4636189
ранее оплачивать	0,4636189
странно работать	0,4636189
трижды позвонить	0,4636189
исправно погашать	0,4636189
ежемесячно списывать	0,4636189
периодически названивать	0,4636189
буквально выталкивать	0,4636189
необоснованно присвоить	0,4636189
неправомерно списывать	0,4636189
безбожно врать	0,4636189
неоднократно повторить	0,4636189
намеренно затягивать	0,4636189
немного разочароваться	0,4636189

Приложение 4. Словарь положительно окрашенных словосочетаний, выделенных с помощью алгоритма, со значениями «Хи-квадрат»

карьерный рост	106,4588794
мобильное приложение	102,9491291
клиентский центр	95,6756706
высокий профессионализм	74,4911384
высокий уровень	71,9848697
индивидуальный подход	61,8757667
удобный интернет-банк	53,1936432
очень порадовать	51,8676796
приятно удивить	43,2110198
отличная работа	42,5491917
профессиональный подход	42,5491917
подробно рассказать	41,9924484
положительный опыт	41,060732
ленинский проспект	37,780146
хороший банк	34,9448254
качественное обслуживание	32,5012111
оперативная работа	31,9076028
приятное впечатление	31,9076028
вежливое обслуживание	31,9076028
качественная работа	31,9076028
удобный банк	31,9076028
вежливый персонал	31,9076028
ведущий менеджер	31,9076028
хорошее настроение	31,9076028
быстрое обслуживание	30,9170778
беспроцентный период	29,3150664
оперативно отвечать	25,9121681
быстро оформить	25,9121681
профессионально объяснить	25,9121681
терминал зажевать	23,3501073
удобное обслуживание	21,2688753
добросовестный сотрудник	21,2688753
отличный специалист	21,2688753
комфортная обстановка	21,2688753
качественная услуга	21,2688753
выгодный процент	21,2688753
маленький процент	21,2688753
слаженная работа	21,2688753
исчерпывающая информация	21,2688753
располагающая обстановка	21,2688753
оценочный компания	21,2688753
двойное списание	21,2688753

отдельная благодарность	21,2688753
профессиональный сотрудник	21,2688753
волгоградский филиал	21,2688753
отличный интернет-банк	21,2688753
вежливый сотрудник	21,2688753
положительное впечатление	21,0264166
быстро решать	17,2699681
понятно объяснять	17,2699681
быстро одобрить	17,2699681
всегда предлагать	17,2699681
всегда помочь	17,2699681
доходчиво объяснять	17,2699681
очень радовать	17,2699681
приветливо встретить	17,2699681
профессионально помочь	17,2699681
грамотно объяснить	17,2699681
неправильно указать	17,2699681
впервые столкнуться	17,2699681
доходчиво ответить	17,2699681
подробно ответить	17,2699681
очень нравиться	16,552199
особенно радовать	16,552199
народ потянуться	11,6698416
подруга порекомендовать	11,6698416
оформлять ипотеку	11,6698416
процедура затянуться	11,6698416
оператор разблокировать	11,6698416
банк зарекомендовать	11,6698416
получить деньги	11,6698416
поисковик выдать	11,6698416
эпопея длиться	11,6698416
проблема решить	11,6698416
услуга отключить	11,6698416
карта заблокироваться	11,6698416
сутки рассматривать	11,6698416
подруга посоветовать	11,6698416
оформление затянуться	11,6698416
ошибка возникнуть	11,6698416
выгодное предложение	11,6388776
грамотная консультация	11,6388776
высокий класс	11,6388776
повышенный кэшбэк	11,6388776
высокий процент	11,6388776
грамотный специалист	11,6388776
отличный сервис	11,6388776

ипотечная сделка	11,6388776
расчетно-кассовое обслуживание	10,6330081
непростая сделка	10,6330081
бесплатная парковка	10,6330081
сложная ситуация	10,6330081
полная самоотдача	10,6330081
шаговая доступность	10,6330081
резкое понижение	10,6330081
отличная консультация	10,6330081
достойный сервис	10,6330081
большая текучка	10,6330081
зарплатный рост	10,6330081
гибкая система	10,6330081
оптимальный вариант	10,6330081
добросовестный труд	10,6330081
бумажная волокита	10,6330081
хороший подбор	10,6330081
огромная благодарность	10,6330081
прекрасный интернет	10,6330081
бесплатный перевод	10,6330081
профессиональное образование	10,6330081
приятное обслуживание	10,6330081
автоматический перевыпуск	10,6330081
разговорчивый человек	10,6330081
симпатичный график-прогноз	10,6330081
платёжная карта	10,6330081
хорошее обслуживание	10,6330081
функциональный интернет-банк	10,6330081
ближний банк	10,6330081
приличный кэшбэк	10,6330081
логичный способ	10,6330081
наёмный сотрудник	10,6330081
персональный подход	10,6330081
непонятная комиссия	10,6330081
великолепный специалист	10,6330081
крупное мошенничество	10,6330081
профессиональная консультация	10,6330081
бесплатный межбанк	10,6330081
твёрдая пятёрка	10,6330081
дружелюбная атмосфера	10,6330081
сомнительный характер	10,6330081
низкопроцентный ставка	10,6330081
быстрое реагирование	10,6330081
ставка хорошая	10,6330081
кэшбэк классный	10,6330081

ставка низкий	10,6330081
удобный банкомат	10,6330081
уникальный предложение	10,6330081
позитивное впечатление	10,6330081
сотрудник отзывчивый	10,6330081
оптимальный предложение	10,6330081
высокопрофессиональное обслуживание	10,6330081
специалист грамотный	10,6330081
удобный сайт	10,6330081
вежливый консультант	10,6330081
хорошее впечатление	10,6330081
светлый офис	10,6330081
приятная атмосфера	10,6330081
высококвалифицированная работа	10,6330081
приятный персонал	10,6330081
прекрасное обслуживание	10,6330081
процесс быстрый	10,6330081
некорректная работа	10,6330081
опрятный парень	10,6330081
соответствующий специалист	10,6330081
финансовый сектор	10,6330081
полноценная консультация	10,6330081
персонал доброжелательный	10,6330081
выгодная конвертация	10,6330081
юридическая поддержка	10,6330081
обслуживание быстрое	10,6330081
виртуальный кошелёк	10,6330081
подробная консультация	10,6330081
атмосфера приятная	10,6330081
вежливый разговор	10,6330081
искренняя забота	10,6330081
отвратительная ситуация	10,6330081
приветливый человек	10,6330081
екатеринбургский филиал	10,6330081
профессиональная работа	10,6330081
валютный перевод	10,6330081
трудолюбивый коллектив	10,6330081
высокомерный тон	10,6330081
приятный бонус	10,6330081
превосходная клиентоориентированность	10,6330081
незаменимый сотрудник	10,6330081
приятная музыка	10,6330081
удобная парковка	10,6330081
излишняя бюрократия	10,6330081

отличный офис	10,6330081
минимальный процент	10,6330081
понятный банк	10,6330081
индивидуальный подход	10,6330081
досрочное погашение	10,6330081
высокопрофессиональный работник	10,6330081
стандартное приветствие	10,6330081
отличный сотрудник	10,6330081
приветливый персонал	10,6330081
стильный дизайн	10,6330081
удобный вариант	10,6330081
благодарственный отзыв	10,6330081
опытный сотрудник	10,6330081
бесплатный перевыпуск	10,6330081
приветливый руководитель	10,6330081
опытный менеджер	10,6330081
высокая оценка	10,6330081
моральная поддержка	10,6330081
хорошая атмосфера	10,6330081
высокая ставка	10,6330081
дежурная улыбка	10,6330081
известный магазин	10,6330081
доброжелательное обслуживание	10,6330081
своевременная помощь	10,6330081
адекватный процент	10,6330081
огромная комиссия	10,6330081
внимательный работник	10,6330081
хорошее качество	10,6330081
быстрая реакция	10,6330081
оперативная помощь	10,6330081
компетентное обслуживание	10,6330081
привлекательная услуга	10,6330081
финансовая грамотность	10,6330081
достойный уровень	10,6330081
банковская гарантия	10,6330081
клиентоориентированный сотрудник	10,6330081
удобная программа	10,6330081
бесплатная альтернатива	10,6330081
дополнительное отделение	10,6330081
уютный кафетерий	10,6330081
выгодный вклад	10,6330081
отдалённое отделение	10,6330081
сложная настройка	10,6330081
приятный сотрудник	10,6330081
грамотная работа	10,6330081

нестабильная ситуация	10,6330081
лояльная работа	10,6330081
быстрая помощь	10,6330081
простенький договор	10,6330081
дружественный персонал	10,6330081
банк отличный	10,6330081
профессиональная помощь	10,6330081
крутой сервис	10,6330081
хороший процент	10,6330081
комфортный офис	10,6330081
классное приложение	10,6330081
отзывчивый подход	10,6330081
особое отношение	10,6330081
широкий выбор	10,6330081
оперативно решить	10,6330081
быстро объяснить	10,6330081
любезно предложить	10,6330081
оперативно попытаться	10,6330081
внимательно отнестись	10,6330081
доступно рассказать	10,6330081
грамотно отвечать	10,6330081
благополучно получить	10,6330081
оперативно обслужить	10,6330081
профессионально относиться	10,6330081
успешно получить	10,6330081
оперативно получать	10,6330081
всегда дозваниваться	10,6330081