

Аспектный анализ тональности текстов на примере отзывов о банковской деятельности

Алиева А. В.

Федеральное государственное бюджетное образовательное учреждение высшего образования «Уфимский государственный авиационный технический университет»

Аннотация

В данной статье предлагается подход к аспектному анализу тональности текстов на примере анализа текстовых отзывов о банковской деятельности, размещенных в сети Интернет. Для обработки и анализа данных предлагается использовать подход на основе нейронных сетей. На основе предложенного подхода проведен эксперимент по анализу реальных отзывов о деятельности нескольких банков Республики Башкортостан. Предложенный подход реализован в разработанном для решения рассматриваемой задачи программном обеспечении.

Ключевые слова: аспектный анализ, анализ тональности текстов, нейронные сети, методы машинного обучения, анализ банковской деятельности.

Введение

Современный этап развития российской экономики характеризуется постоянным ростом спроса на банковские услуги как со стороны юридических лиц, так и со стороны физических лиц. При этом с каждым годом усиливается конкуренция на рынке банковских услуг. Деятельность в условиях жесткой конкуренции вынуждает банки расширять спектр предоставляемых услуг и операций, заставляет искать наиболее эффективные способы их внедрения, а также проводить работу по улучшению качества банковского обслуживания физических лиц, что позволило бы кредитной организации выйти на новый уровень в работе с клиентами.

При этом в сети Интернет есть большое количество отзывов о банках, в которых клиенты высказывают свое мнение относительно того или иного направления деятельности банка. Собрав данные оценки клиентов можно вычислить средние оценки относительно каждого аспекта деятельности каждого банка. Данную задачу можно решить с использованием аспектного анализа тональности текстов. Чтобы провести данный анализ, необходимо решить ряд задач: выделение сущностей, извлечение аспектов и тональностей, агрегирование аспектов и тональностей, а именно, является ли мнение, высказанное о данном аспекте, положительным, отрицательным или нейтральным.

На основе анализа существующих подходов к аспектному анализу тональности текстов было разработано собственное решение в виде программного обеспечения для аспектного анализа тональности отзывов в области банковской деятельности.

Предлагаемый подход

Предлагаемый подход к решению задачи аспектного анализа тональности текстов включает в себя следующие этапы:

1. Сбор данных;

2. Предобработка собранных данных;
3. Векторизация данных;
4. Извлечение аспектных терминов и тональностей;
5. Агрегирование извлеченных аспектов и тональностей;
6. Оценка тональности категорий;
7. Визуализация результатов.

Задачу сбора данных предлагается решить путем парсинга сайта с отзывами о банковской деятельности «Банки.ру». На выходе данного этапа будет получен корпус отзывов, который будет помещен в базу данных для последующего использования.

Задачу предобработки данных предлагается стандартными методами: удалением стоп-слов, знаков препинания, приведением слов к нормальной форме и токенизацией (разбиением текстов отзывов на предложения, а предложения на слова). На выходе данного этапа будут получены тексты отзывов, удобные для последующего представления в векторной форме.

Задачу векторизации данных предлагается решить с использованием предобученной модели ELMo [1]. На выходе данного этапа каждое слово в отзыве будет представлено в векторной форме для последующей обработки методами машинного обучения.

Задачу извлечения аспектных терминов и тональностей предлагается решить путем применения нейронной сети в качестве модели. Предлагается поэкспериментировать с архитектурой нейронной сети LSTM (как наиболее часто использующуюся для решения подобных задач [2]) и выбрать наиболее подходящую для решения задачи. На выходе данного этапа будет получено множество аспектных терминов и тональностей.

Задачу агрегирования извлеченных аспектов и тональностей предлагается решать с использованием метода прохождения по предложениям контекстным окном. На выходе данного этапа будут получены пары «аспект-тональность».

На этапе оценки тональности категорий будут рассчитаны оценки по каждой категории для каждого анализируемого банка на основе пар «аспект-тональность», полученных на предыдущем этапе.

После проведения всех предыдущих этапов полученные результаты необходимо представить в удобной форме пользователям. Предлагается представить полученные результаты на странице веб-браузера.

Все этапы, приведенные выше, вместе составляют концепцию программного обеспечения, которая представлена на рисунке 1. Подробно каждый этап решения описан в следующих разделах.

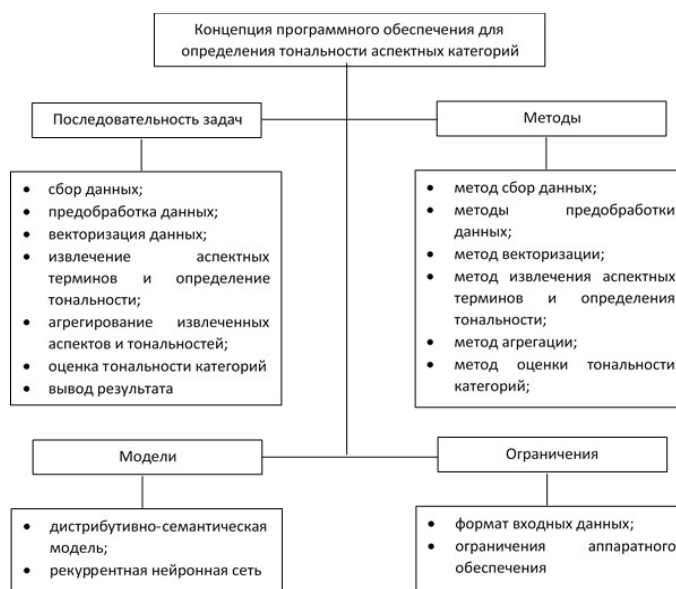


Рис.1 Концепция программного обеспечения для определения тональности аспектных категорий

Сбор данных

Сбор данных был осуществлен путем парсинга отзывов с крупнейшего финансового портала российской сети Интернет «Банки.ру». Для последующего анализа были собраны отзывы о следующих банках: Сбербанк, Альфа-банк, Банк УралСиб, РСХБ, Газпромбанк, ВТБ. Данные банки были выбраны для анализа как наиболее широко представленные банки в Республике Башкортостан. Для сбора данных использовались следующие библиотеки: Requests для отправки HTTP-запросов на сайт и получения данных и BeautifulSoup для извлечения данных из файлов HTML. Полученные тексты отзывов были загружены в базу данных SQLite.

Предобработка данных

Предобработка данных осуществлялась стандартными методами: удалением стоп-слов, знаков препинания, приведением слов к нормальной форме и токенизацией (разбиением текстов отзывов на предложения, а предложения на слова). Для реализации данных методов использовался пакет библиотек NLTK, который содержит множество методов для предобработки текстов.

Векторизация слов

Для векторизации слов, выделенных из текстов на предыдущем этапе, использовалась модель ELMo (Embeddings from Language Models). ELMo – это нейронная сеть, которая требует минимальной предобработки текстовых данных – лемматизации и токенизации – и позволяет векторизовать слова, учитывая контекст до и после анализируемого слова за счет того, что в ее основе лежит двунаправленная модель долгой краткосрочной памяти (BiLSTM). Для обучения данной модели использовались предобученные векторы из контекстуализированной модели «Araneum», размерностью 1024. Для извлечения векторов из модели используется библиотека simple_elmo, разработанная RusVectores.

Извлечение аспектных терминов и тональностей

Для извлечения аспектных терминов использовалась модель нейронной сети, состоящая из слоя двунаправленной LSTM и линейного слоя Dense – данная архитектура показала лучшие результаты в экспериментах. На вход данной нейронной сети подаются вектора размерностью 1024. Выходные значения BiLSTM конкатенируются. Далее эти значения подаются на вход линейного слоя, где происходит предсказание меток. На выходе полученные значения проходят через функцию softmax для категориального распределения по меткам классов. Принцип работы модели отображен на рисунке 2.

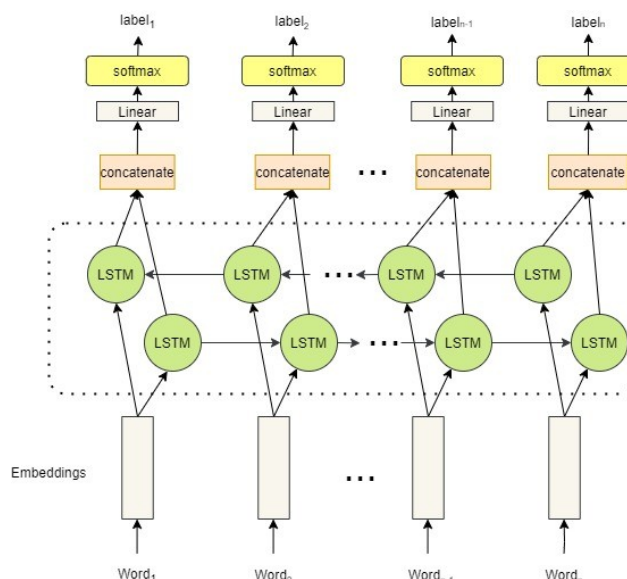


Рис.2 Принцип работы модели для извлечения аспектов и тональностей

Для обучения нейронной сети необходим набор размеченных данных. Такой набор данных был подготовлен путем разметки 400 отзывов с использованием инструмента для создания аннотаций doccano. В качестве аспектов деятельности банков (классов для нейронной сети) были выделены следующие категории: обслуживание в офисах, уровень поддержки клиентов, работа ДБО (дистанционного банковского обслуживания, включая те-лебанк, мобильный и интернет-банкинг), работа банкоматов и терминалов, кредитование, кредитные карты, карточные продукты (дебетовые, моментальные карты), депозиты (деPOSITные, расчетные, накопительные счета), РКО (расчетно-кассовое обслуживание, включая переводы и платежи, обмен валюты и монет). Помимо аспектов также были выделены метки тональности: положительные и отрицательные. Пример разметки отзыва представлен на рисунке 3.

Здравствуйте! 18 Августа при попытке внести наличные в указанном отделении банка, банкомат не зачислил 40 000 руб. на счет моей карты. Обращение по горячей линии было зарегистрировано под номером №002222-2015-164013. В данном отделении (Демский район г. Уфы ул. Ухтомского 10) часть банкоматов постоянно выдают ошибки и периодически отключены, что создает большие очереди в помещении с банкоматами. При обращении в отделение с вопросом о возврате денежных средств мне было предложено написать заявление и ждать после неограниченный срок возврата средств на карту. Денежные средства предназначались для покрытия долгов по кредитам в Сбербанке. В отделение банка по этому вопросу мне ответили следующее: "Что-нибудь придумаем!" и даже не извинились за сложившуюся ситуацию. Вопрос следующий: Что же я должен придумать... может быть взять еще один кредит, чтобы успеть закрыть свои ежемесячные платежи... и кто мне компенсирует потом эти проценты по кредиту. И когда уже вернут мои деньги (хочется услышать четкий срок). Видимо, придется обращаться в суд впоследствии с целью компенсации Сбербанком России моих затрат.

Рис.3 Пример размеченного отзыва

После разметки тексты отзывов были предобработаны стандартными методами и также была добавлена BIO-разметка, где В – начало аспектного термина или тональности, I – продолжение аспектного термина или тональности, О – слово не относится ни к аспектному термину, ни к тональности. Данные размеченные отзывы использовались для обучения описанной выше нейронной сети.

Агрегирование извлеченных аспектов и тональностей

Так как в одном предложении может находиться несколько аспектных категорий, то был использован метод прохождения по предложению контекстным окном, который учитывал тональность не только на уровне предложения, но и на уровне фрагментов предложений. В ходе анализа полученных результатов было выявлено, что наиболее подходящий размер контекстного окна равен трем словам, так как тональность от аспекта может быть удалена на несколько слов, но при этом случаи, когда расстояние между аспектом и тональностью больше трех слов очень редки.

Кратко, алгоритм прохождения по предложению контекстным окном заключается в следующем: если в контекстном окне есть слово с меткой аспекта, то оно назначается текущим аспектным термином, если же есть слово с меткой тональности, то оно назначается текущей тональностью. Текущему аспекту назначается текущая тональность, если у аспекта уже есть тональность, то тональности складываются. Если у аспектного термина в пределах контекстного окна слева и справа нет тональности, то ему назначается нейтральная тональность. При этом положительной тональности соответствует число «1», нейтральной – «0», отрицательной – «-1». После прохождения контекстным окном по всему предложению оценки найденных аспектов приводятся к 1, 0 и -1. После прохождения контекстным окном по всем предложениям отзыва значения тональности полученных аспектов также складываются и приводятся к значениям 1, 0 и -1. Такой способ позволяет избежать завышенных и заниженных оценок аспектов.

Оценка тональности категорий

Оценка тональности по каждой категории по каждому банку рассчитывается следующим образом:

$$aspect_{bank} = \frac{N_{pos} - N_{neg}}{N}$$

где N_{pos} – количество положительных оценок, N_{neg} – количество отрицательных оценок, N – общее количество оценок. Таким образом, оценка банка по аспектной категории имеет значение в интервале от -1 до 1. Для удобства интерпретации результатов полученный интервал преобразуется в интервал от 0 до 100 (изображение результата в процентах).

Чтобы иметь представление о количестве нейтральных, положительных и отрицательных отзывов по каждой категории, также была рассчитана доля этих оценок по формуле:

$$aspect_{pos} = \frac{N_{pos}}{N} * 100 \%$$

где $aspect_{pos}$ – доля положительных оценок по аспектной категории для банка.

Архитектура программного обеспечения

Для удобного представления результатов анализа было разработано программное обеспечения в виде веб-приложения с клиент-серверной архитектурой. Пользователю предоставлен доступ к клиентской части приложения, в которой реализовано представление данных и взаимодействие с пользователем. Данные передаются клиенту от сервера приложений, которые получает данные через запросы к веб-серверу. Веб-сервер получает запросы от сервера приложения и передает данные в ответе на запрос. Внутри сервера выполняются необходимые вычисления, анализ тональности по аспектам. Данные хранятся в базе данных в виде таблиц. На диаграмме размещения отображено аппаратное размещение компонентов программного обеспечения (рисунок 4).

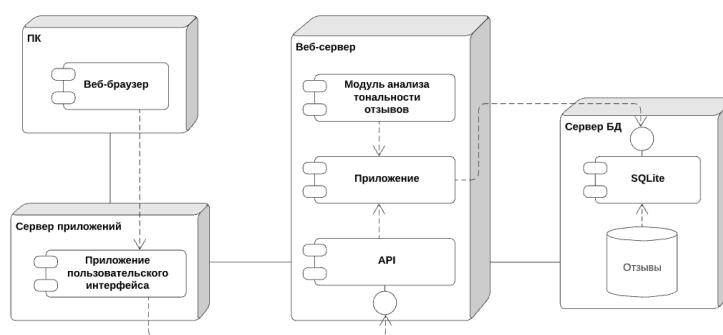


Рис. 4 Диаграмма размещения компонентов ПО

В качестве языка для разработки серверной части приложения был выбран язык Python, в качестве фреймворка для веб-разработки выбран Django. В качестве языка для разработки клиентской части приложения был выбран язык Javascript, в качестве фреймворка для разработки выбран ReactJS.

Эксперимент

Размеченный и предобработанный набор текстов отзывов был предварительно разделен на предложения (в итоге получено 3929 предложений), полученные предложения были поделены на обучающую и тестовую выборки в отношении 70/30. Таким образом, размер обучающей выборки составил 2750 предложений, размер тестовой выборки составил 1179 предложений.

Модель была создана при помощи высокоуровневой библиотеки Keras, построенной на основе TensorFlow, используемой для построения нейронных сетей.

На вход нейронной сети подается вектор $W(w_1, w_2, \dots, w_n)$, где W – набор векторных представлений слов, полученных из контекстуализированной модели Araneum, обученной при помощи алгоритма ELMo на веб-корпусе русскоязычных текстов, размерность вектора для каждого слова равна 1024, n – количество слов в предложении. Число n было определено экспериментально. На рисунке 5 представлена диаграмма, показывающая распределение длин предложений отзывов. Как видно из диаграммы, большинство отзывов имеют длину не более 30 слов, поэтому было выбрано $n=30$ слов. Предложения, которые имели длину более 30 слов обрезаются до 30 слов.

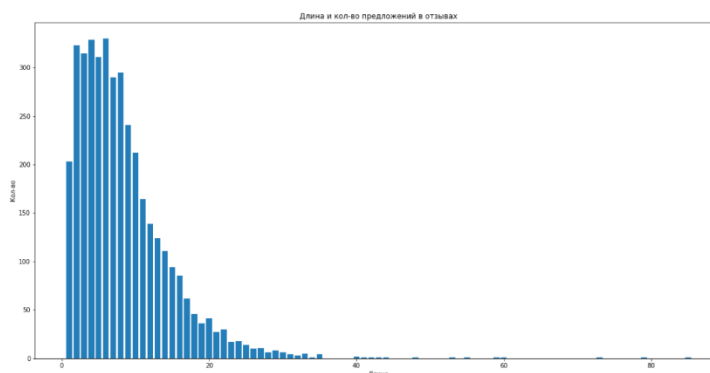


Рис. 5 Количество предложений с определенной длиной в размеченном наборе отзывов

Отметим, что при проведении экспериментов не было разделения на теги, соответствующие аспектам, и теги, соответствующие тональностям, так как для модели это всего лишь набор меток. Таким образом, метки аспектов и тональностей были предсказаны одновременно. Были проведены эксперименты обучения модели с одним слоем LSTM и одним слоем Dense. В одном эксперименте были удалены все предложения, содержащие только тег «О», таким образом, осталось 3137 предложений. В другом эксперименте был добавлен слой Dropout, использующийся для уменьшения переобучения сети. В еще одном эксперименте слой LSTM был заменен на двунаправленную сеть LSTM (BiLSTM), также с добавлением слоя Dropout и удалением предложений, содержащих только тег «О». Результаты всех экспериментов приведены в таблице 1.

Таблица 1 Результаты обучения на тестовой выборке

Аспекты / тональность	LSTM	LSTM без предложений с тегом «О»	LSTM без предложений с тегом «О» с dropout=0.1	BiLSTM без предложений с тегом «О» с dropout=0.1
Other	0,98	0,98	0,98	0,98
В-работа банкоматов и терминалов	0,71	0,78	0,77	0,83
В-обслуживание в офисах	0,8	0,85	0,8	0,8
В-карточные продукты	0,84	0,83	0,82	0,82
В-кредитование	0,78	0,84	0,79	0,82
В-кредитные карты	0,73	0,65	0,66	0,62
В-работа ДБО	0,75	0,76	0,82	0,76
В-депозиты	0,56	0,57	0,7	0,76
В-РКО	0,69	0,67	0,77	0,72
В-уровень поддержки клиентов	0,72	0,7	0,63	0,74
И- карточные продукты	0,55	0,66	0,61	0,74
И- кредитование	0,77	0,82	0,8	0,79
И- кредитные карты	0,88	0,85	0,81	0,79
И- работа ДБО	0,45	0,46	0,4	0,47
И- депозиты	0,76	0,74	0,5	0,61
И- РКО	0,55	0,63	0,59	0,57
И- уровень поддержки клиентов	0,85	0,9	0,95	0,88

В-негативная тональность	0,47	0,58	0,57	0,64
В-позитивная тональность	0,72	0,73	0,68	0,75
F мера	0,96	0,95	0,95	0,96

Как видно из таблицы, в целом F-мера примерно одинакова во всех случаях, что связано с хорошим результатом предсказания класса «Other» во всех случаях, который, однако, не является показательным для нашей классификации. Также можно отметить, что результаты без использования предложений с тегом «О» лучше, так как при обучении нейросети используется меньше неинформативных предложений. Стоит отметить низкие показатели для тегов «I-депозиты», «I- работа ДБО». В связи с этими результатами были приняты следующие решения:

- исключить из рассмотрения тег «I-работа ДБО», оставить только метки «В- работа ДБО»;
- оставить метки «I-депозиты», так как модель правильно определяет саму категорию, но ошибается с «I» и «В» тегами, что не существенно в данном случае, так как в дальнейшем эти теги будут объединены в одну категорию.

В результате было принято решение использовать модель с двунаправленной LSTM сетью с добавлением слоя Dropout и с удалением предложений, содержащих только тег «О». На рисунке 6 отображена нормализованная матрица ошибок модели извлечения аспектных терминов и тональностей. Как видно из данной матрицы, чаще всего модель ошибается в различении классов с категорией «Other» («О»), а также с различением «I» и «В» тегов (например, «I-a_deposit» и «B-a_deposit»).

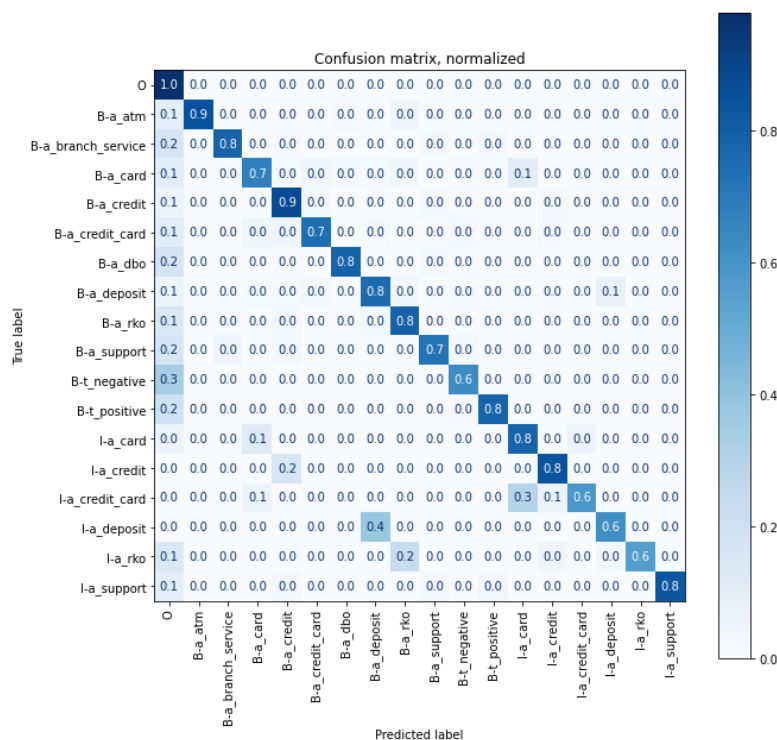


Рис. 6 Нормализованная матрица ошибок модели извлечения аспектных терминов и тональностей

Пример результата классификации представлен на рисунке 7.

	Слово	Предсказанный тег	Истинный тег		Слово	Предсказанный тег	Истинный тег
0	хотеть	О	О	0	начитаться	О	О
1	выразить	О	О	1	отзыв	О	О
2	большой	О	О	2	интернет	B-a_dbo	О
3	благодарность	B-t_positive	B-t_positive	3	весь	О	О
4	кредитный	B-a_credit	B-a_credit	4	волокига	B-t_negative	B-t_negative
5	инспектор	I-a_credit	I-a_credit	5	документ	B-a_credit	B-a_credit
6	далее	О	О	6	объект	О	О
7	грамотный	B-t_positive	B-t_positive	7	многий	О	О
8	ведение	B-a_branch_service	О	8	отказывать	О	B-t_negative
9	наш	О	О	9	кредит	B-a_credit	B-a_credit
10	ипотечный	B-a_credit	B-a_credit	10	мучить	B-t_negative	B-t_negative
11	сделка	I-a_credit	I-a_credit	11	документ	B-a_credit	B-a_credit
				12	продавец	О	О
				13	риелтор	О	О

Рис.7 Верно и неверно классифицируемые теги аспектов и тональностей в предложении

Выделенным аспектам были сопоставлены соответствующие тональности с использованием метода агрегирования. При этом В и I-категории, соответствующие одному аспекту, были объединены в одну. Далее была вычислена тональность аспекта на основе предложения, после этого была вычислена тональность аспекта на основе отзыва. Результат анализа одного из отзывов представлен на рисунке 8.

Впервые обратился в Уралсиб на Революционной 41, для получения кредита, ранее вашим клиентом не являлся. Обслуживание было на высшем уровне, очень приятный сотрудник - К-ва Александра, всё подробно рассказала, ответила на мои вопросы. Кредит мне одобрили и предложили получить кредитную карту на выгодных условиях, от чего я не отказался. Повторная коммуникация была с Валерией, тоже очень приятная и милая девушка, научила пользоваться мобильным банком, предложила открыть пакет карт, но мне ещё необходимо обдумать. Выражаю благодарность сотрудникам за профессионализм и отзывчивость! <pre> a_branch_service = 1.0 a_card = 0.0 a_credit = 1.0 a_credit_card = 1.0 a_dbo = 0.0 a_support = 1.0 </pre>

Рис.8 Результат анализа тональности аспектов на примере отзыва

Как видно из результата анализа, в отзыве было выделено 6 аспектов. При этом таким аспектам как «Обслуживание в офисах», «Кредитование», «Кредитные карты» и «Уровень поддержки клиентов» была поставлена оценка 1, что соответствует положительной тональности данных аспектов. Это подтверждается текстом отзыва («обслуживание на высшем уровне», «кредит одобрили», «кредитную карту на выгодных условиях», «ответила на мои вопросы»). Также можно отметить, что были выделены такие аспекты, как «Работа ДБО» и «Карточные продукты», но они имеют нейтральную тональность (оценка 0), так как в отзыве нет четкой оценки данным аспектам деятельности банка.

На основе полученных результатов по каждому отзыву были также получены результаты по каждому аспекту по каждому из анализируемых банков (были получены средние оценки тональности по каждому аспекту). Для анализа были выбраны наиболее популярные банки, представленные в г. Уфа. Для удобства шкала оценок от -1 до 1 была переведена в проценты (от 0 до 100%). Результаты анализа представлены на рисунке 9.

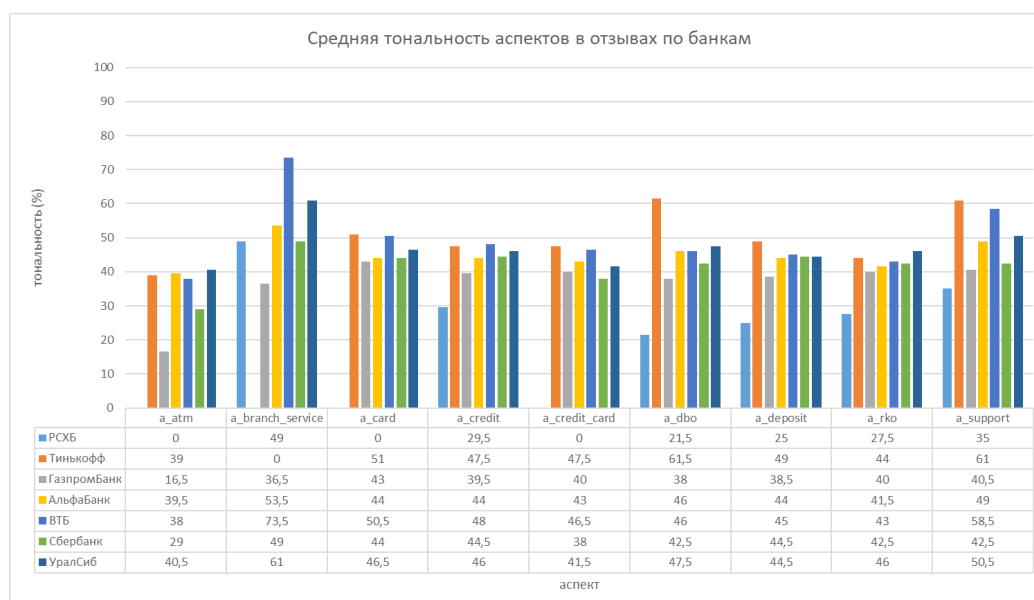


Рис. 9 Средняя тональность аспектов в отзывах по банкам

Из результатов анализа можно заметить, что в большинстве случаев средняя тональность отзывов не превышает 50%, что объяснимо, так как пользователи сайта чаще всего оставляют негативные отзывы, чем положительные. Однако можно заметить, что больше всего положительных отзывов приходится на такие банки как ВТБ и Тинькофф, а больше всего отрицательных отзывов на РСХБ. Также можно сделать следующие выводы:

1. больше всего негативных отзывов касается работы банкоматов и терминалов: по всем банкам наблюдается низкая оценка по данному аспекту;
2. по большинству аспектов тональность отзывов негативная или нейтральная;
3. по уровню обслуживания в офисах лучший показатель имеют такие банки, как ВТБ, УралСиб и АльфаБанк;
4. по уровню поддержки клиентов лучший показатель имеют такие банки, как ВТБ и Тинькофф;
5. по уровню дистанционного банковского обслуживания лидером среди банков является Тинькофф, что закономерно, так как именно этот банк, в отличие от других, специализируется на удаленном оказании банковских услуг.

Визуализация результатов

Для визуализации результатов было разработано программное обеспечение. Интерфейс представляет собой веб-приложение со следующими страницами:

1. главная страница с формой для отправки отзыва пользователем;
2. страницы банков с информацией о банке, оценками по категориям, отзывами о банке;
3. страницы отзывов с текстом отзыва, краткой информацией и оценками по категориям.

Рассмотрим пример визуализации результатов анализа на примере банка РСХБ (Россельхозбанк). Результаты анализа отображены на странице в виде диаграмм (рисунок 10).



Рис. 10 Визуализация результатов анализа

Пользователь также может перейти на страницу определенного отзыва и на данной странице он может видеть выделенные положительные, отрицательные и нейтральные аспекты (рисунок 11).

Сельская ипотека приостановлена на неопределенный срок
Тональность: 2

Оставил заявку на сайте россельхозбанка на сельскую ипотеку, перезвонили через пару часов и сказали, что данный вид ипотеки приостановлен до неизвестного времени. Предложили другую ипотеку под 5,9%, но она не попадает под строительство загородного дома, тогда предложили другую ипотеку под 12%, ну я и послал их вежливо восояси. Зачем вся эта реклама на сайте если такой услуги нет?!!!!

Автор: Игорь
Дата публикации: 2020-08-31

Выделенные аспекты:

Положительные аспекты: Положительных аспектов не найдено	Нейтральные аспекты: ДБО Служба поддержки	Отрицательные аспекты: Кредитование
--	--	---

Рис.11 Страница отзыва

Таким образом, результаты анализа были представлены в удобном формате в виде приложения для веб-браузера.

Заключение

В данной статье был предложен подход к аспектному анализу тональности текстов на примере отзывов о банковской деятельности. Данный подход позволяет проводить исследование удовлетворенности клиентов банков Республики Башкортостан в автоматизированном режиме. На основе предложенного подхода было разработано программное обеспечение для проведения аспектного анализа тональности отзывов о

банках, которое позволяет не только проводить анализ в автоматизированном режиме, но и визуализирует результаты в удобном формате для пользователей. Проведенный эксперимент показал эффективность подхода для решения поставленной задачи на основе реальных отзывов о банковской деятельности, удовлетворительную точность методов машинного обучения и логичность полученных результатов.

Список литературы

1. Peters, M., et al. Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2018).
2. Куликов, Г.С., Грачев Н.С., Кудинов В.А. Анализ тональности текста с помощью рекуррентных нейронных сетей // Научный альманах. – 2016 - №7-1 (21). – С.401-401.