

Елена Георгиевна БРУНОВА¹
Юлия Владимировна БИДУЛЯ²

УДК 81'322

**КЛИЕНТ ВСЕГДА ПРАВ:
АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТА
В ОТЗЫВАХ О КАЧЕСТВЕ
БАНКОВСКОГО ОБСЛУЖИВАНИЯ**

¹ доктор филологических наук,
заведующая кафедрой иностранных языков
и межкультурной профессиональной коммуникации
естественнонаучных направлений,
Тюменский государственный университет
egbrunova@mail.ru

² кандидат филологических наук,
доцент кафедры информационных систем,
Тюменский государственный университет
bidulya@yandex.ru

Аннотация

Целью исследования является разработка алгоритма на основе правил для анализа тональности текста. Материалом послужили отзывы на русском языке о качестве банковского обслуживания из народного рейтинга банков (сайт www.banki.ru). Анализ тональности текста рассматривается как задача классификации, т. е. отнесение текста к одному из двух классов — с положительной и отрицательной оценкой. В основу алгоритма положено использование определенных лексико-грамматических конструкций наряду с оценочным лексиконом, содержащим классы слов с положительной и отрицательной оценкой, а также три служебных класса.

Эффективность предложенного алгоритма оценивается с помощью показателей точности (Precision), полноты (Recall) и меры F Ван Ризбергена в сравнении с результатами

Цитирование: Брунова Е. Г. Клиент всегда прав: анализ тональности текста в отзывах о качестве банковского обслуживания / Е. Г. Брунова, Ю. В. Бидуля // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2017. Том 3. № 1. С. 72-89. DOI: 10.21684/2411-197X-2017-3-1-72-89

работы другого алгоритма, широко применяемого для анализа тональности — Наивного Байесовского классификатора. Для оценки эффективности использован корпус из 200 отзывов о качестве банковского обслуживания. Значения точности, полноты и F-меры у предложенного алгоритма оказались на 5-8% выше, чем у Наивного Байесовского классификатора.

Ключевые слова

Анализ тональности текста, обработка естественного языка, контент-анализ, классификатор на основе правил, алгоритм, пользовательский контент, лексикон.

DOI: 10.21684/2411-197X-2017-3-1-72-89

Введение

Человеку свойственно интересоваться мнением других — родных, друзей, соседей и т. п. Производителей товаров и услуг волнуют предпочтения клиентов, а клиенты, в свою очередь, прежде чем выбрать товар или услугу, жаждут узнать мнения других клиентов. В современном мире вместо того, чтобы спрашивать у друга или эксперта, человек обычно забывает название товара или компании в поисковую систему и читает отзывы. Поэтому исследования анализа тональности текста являются одной из бурно развивающихся областей прикладной лингвистики. Анализ тональности текста (англ. *sentiment analysis, opinion mining*) — метод извлечения из текста мнений и эмоций, а также их последующей обработки, относится к методам контент-анализа и является одним из средств изучения субъективности в естественном языке.

Мнение каждого клиента, безусловно, является субъективным. Однако большая совокупность таких субъективных мнений в условиях конкуренции является существенным фактором, который может заставить производителя товара изменить некоторые его свойства, а поставщика услуг — предпринять определенные меры организационного или кадрового характера. В этом контексте известный девиз «Клиент всегда прав» (англ. *The customer is always right*), впервые прозвучавший в 1924 г. в сети магазинов Гордона Селфриджа в Великобритании и США [4], приобретает новый смысл — мнение клиента стоит того, чтобы быть извлеченным, проанализированным и услышанным, что при больших массивах данных может быть реализовано только автоматизированными методами.

Однако очевидная простота выражения своего мнения и ознакомления с мнениями других в сети Интернет порождает ряд социальных и технологических проблем:

- избыточность информации. Часто мнения скрываются в пространственных комментариях на форумах и блогах, и требуется потратить много времени на поиски нужной информации. Это определяет необходимость разработки систем автоматического извлечения и обработки мнений, одной из задач которых является устранение избыточности;

- полезность информации. Проблема устранения избыточности не означает, что лучший отзыв — это краткий отзыв. Читателю отзыва важно не только мнение (нравится, не нравится), но и аргументация (почему нравится, почему не нравится);
- возможная недостоверность информации. Было бы наивным ожидать, что человек не попытается обмануть такие системы. Многочисленные фальшивые отзывы в сети могут являться скрытой рекламой «нужного» товара или средством компрометации «ненужного» товара, производимого конкурентами.

Актуальность проводимого исследования определяется широким спектром возможных областей применения анализа тональности текста, в том числе:

- лингвистика (изучение языковых средств выражения субъективности);
- информационные технологии (составление рейтинга мнений для поисковых систем; автоматическое определение флейма (перебранки) на сайтах и форумах и его последующая фильтрация);
- финансовый анализ (оценка финансовых ожиданий);
- реклама (оценка эффективности рекламных кампаний);
- политика (отслеживание общественного мнения);
- СМИ (оценка отклика общественности на новости).

Для извлечения мнений о качестве банковского обслуживания мы предлагаем алгоритм на основе правил для анализа тональности текста, который предполагает использование лексико-грамматических конструкций наряду с оценочным лексиконом. Эти конструкции, извлекаемые на уровне предложения, считаются факторами, влияющими на полярность как отдельного предложения, так и текста в целом.

Целью данного исследования является разработка алгоритма на основе правил для анализа тональности текста.

Предшествующие публикации

Информационные технологии традиционно относят информацию, содержащуюся в текстах на естественном языке, к одному из двух классов: факты и мнения. Мнения (субъективная информация), в отличие от фактов (объективной информации), подразумевают описание не самих сущностей и событий, а их оценку (одобрение или неодобрение) и эмоции человека по отношению к ним. В первые десятилетия существования Интернета задачи информационных технологий в области поиска, извлечения и обработки информации из текстов на естественном языке были сосредоточены на фактах, и только с начала 2000-х гг. стали появляться исследования, посвященные поиску, извлечению и обработке мнений [7, 9, 10, 11, 12, 14, 16]. Подавляющее большинство таких исследований использовали английский язык как материал для анализа. Однако в последнее десятилетие, в связи с бурным развитием неанглоязычного сегмента Интернета и появлением соответствующего пользовательского контента, стали публиковаться

исследования, материалом для которых послужили другие языки, в том числе — русский. Так, на ежегодной Международной конференции по компьютерной лингвистике и информационным технологиям «Диалог» (Москва) выделяется целая секция, посвященная анализу тональности [2, с. 3-76].

Задачу извлечения и обработки субъективной информации можно свести к задаче классификации корпуса текстов на два класса: с положительной (одобряет) и отрицательной (не одобряет) оценкой. Некоторые исследователи добавляют еще третий класс — с нейтральной оценкой [3] и даже четвертый — со смешанной оценкой [13], однако в основе любой гибридной классификации мы обнаруживаем бинарный принцип.

Традиционно выделяются два подхода к методологии анализа тональности текста: на основе машинного обучения и на основе лексикона. Машинное обучение предусматривает формирование тренировочного набора данных, который впоследствии используется для обучения классификатора. В рамках подхода на основе лексикона полярность определяется в процессе сопоставления текста отзыва с положительным и отрицательным классами оценочного лексикона [4].

Эффективность анализа тональности текста, осуществляемого на основе лексикона, может быть существенно повышена с помощью классификатора на основе правил, что, в частности, наблюдается для отдельных предметных областей. Так, алгоритм на основе правил был успешно применен для классификации финансовых новостей в [15], где определялось соотношение положительной и отрицательной полярности для каждого предложения, а также рассчитывалась тональность всего текста.

Таблица 1

Возможные классы текстов при анализе тональности

Table 1

Possible text classes when analyzing the tonality

Класс	Обязательные классы		Необязательные классы	
	Положительная полярность (+)	Отрицательная полярность (-)	Смешанная полярность (+ и -)	Нейтральная полярность (ни +, ни -)
Интерпретация	Одобряет	Не одобряет	Что-то одобряет, что-то не одобряет	Отсутствие оценки

Материал для исследования и оценочный лексикон

Материалом для исследования послужили отзывы на русском языке о качестве банковского обслуживания, взятые из народного рейтинга банков, публикуемого на сайте www.banki.ru. Клиенты банков публикуют свои отзывы и выставля-

ют оценки от 5 (отлично) до 1 (очень плохо). Впоследствии эти оценки могут быть подтверждены или отвергнуты модератором. Требования к отзывам для народного рейтинга банков включают информативность, поэтому авторы стараются предоставить как можно больше подробностей и аргументов при описании событий, вызвавших у них положительные или отрицательные эмоции. Самый короткий текст из нашего корпуса содержит три предложения. В этом заключается существенное отличие материала для данного исследования от таких коротких сообщений, как твиты, где, как правило, текст отзыва состоит из одного предложения.

Кроме того, особенности изучаемой предметной области могут существенно влиять на применение оценочного лексикона. Некоторые слова могут относиться к противоположным классам в зависимости от предметной области [7], например, слово «долго» входит в положительный лексикон при оценке срока службы заряженного аккумулятора в предметной области «Смартфон», но то же самое слово будет отнесено к отрицательному лексикону при оценке времени ожидания обслуживания в предметной области «Качество банковского обслуживания». Речь идет о параметрической лексике [6], т. е. словах, характеризующих объем некоего параметра, специфичного для определенной предметной области (срок службы аккумулятора и срок ожидания обслуживания соответственно).

Специфические черты выявляются не только на лексическом, но и на синтаксическом, и даже текстовом уровнях. Так, отзыв о товаре (автомобиле, смартфоне и т. д.) обычно представляет собой дескриптив, где свойства объекта вполне эксплицитны. Отзыв о качестве обслуживания чаще всего является нарративом, где автор повествует о фактах и событиях, вызвавших у него положительные или отрицательные эмоции, в этом случае свойства объекта имплицитны и трудно извлекаемы. Таким образом, типичный текст отзыва о качестве банковского обслуживания представляет собой нарратив, состоящий из нескольких предложений, а иногда — из нескольких абзацев.

Для эксперимента было случайным образом отобрано 200 отзывов о качестве банковского обслуживания в соответствии с критерием длины отзыва: минимальная составила 50 токенов (под токеном подразумевается слово или знак препинания), максимальная — 500. Мнение автора отзыва считается положительным, если отзыв оценивается на 4 или 5, или отрицательным, если отзыв оценивается на 2 или 1. Отзывы с оценкой 3, отзывы без оценки, а также отзывы, оценки которых не были засчитаны модератором, не рассматривались.

Для оценочного лексикона был вручную сформирован базовый (seed) лексикон из 100 слов на основе миникорпуса из 50 отзывов (24 положительных и 26 отрицательных), отобранных случайным образом. Затем базовый лексикон был расширен с помощью синонимов, антонимов и методики согласованности тональности (sentiment consistency technique) [8]. В этой методике, впервые предложенной в [7] используется список базовых оценочных прилагательных

с известной полярностью и набор лингвистических ограничителей (и, но, или-или, ни-ни) для автоматического выявления других оценочных слов и определения их полярности. Например, если имеется предложение «Этот айфон красивый и удобный» и известно, что слово «красивый» относится к положительному классу оценочного лексикона, то подразумевается, что «удобный» также относится к положительному классу. С другой стороны, если имеется предложение «Этот айфон красивый, но дорогой» и известно, что слово «красивый» относится к положительному классу оценочного лексикона, то подразумевается, что «дорогой» относится к отрицательному классу. В результате расширения оценочный лексикон составил 286 положительных и 385 отрицательных лексических единиц.

Алгоритм на основе правил для анализа тональности текста

Один из самых известных методов анализа тональности текста на основе лексикона — Наивный Байесовский классификатор основан на упрощенном понимании текста как набора слов (англ. bag of words), при этом даже место слова в предложении и тексте не имеет значения [17]. Простота его применения одновременно является и наиболее существенным недостатком, поскольку не учитываемые данным классификатором синтаксические отношения могут оказывать существенное влияние на смысл предложения и текста в целом, а следовательно — и на его оценку.

Мы разработали набор правил, которые используются для корректировки окончательной тональности текста после того, как он подвергается предварительному анализу по принципу «набора оценочных слов» [1]. Каждое правило представляет собой шаблон для извлечения определенных лексико-грамматических структур из предложений.

Определение тональности текста отзыва состоит из предварительной обработки, расчета черновой полярности или счета (англ. count) для каждого предложения, корректировки полярности каждого предложения с помощью правил, применяемых в определенной последовательности, суммирования скорректированных величин и нормализации окончательной величины тональности текста. Этапы определения тональности текста представлены на рис. 1.

Предварительная обработка включает следующие шаги:

Шаг 1. Разбить текст на предложения, пометить начало и конец каждого предложения.

Шаг 2. Разбить каждое предложение на фрагменты, заменить знаки препинания внутри предложения, кавычки и союзы специальными символами.

Шаг 3. Разбить текст на слова, используя пробелы.

Шаг 4. Подсчитать количество слов в тексте.

Шаг 5. Пометить слова, набранные прописными буквами.

Шаг 6. Заменить все прописные буквы строчными.

Шаг 7. Подвергнуть все слова стеммингу.

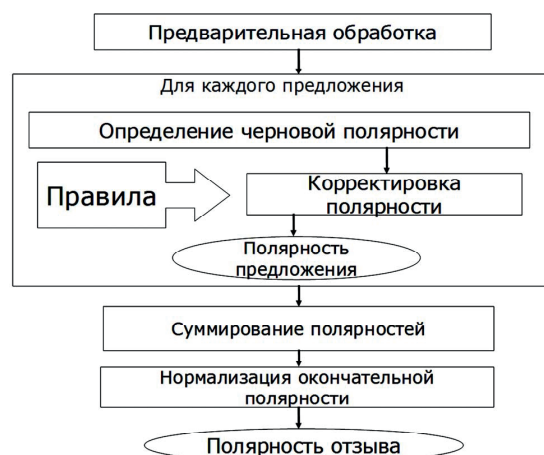


Рис. 1. Этапы определения тональности текста

Fig. 1. The steps of determining the text tonality

Под фрагментом понимается отрезок предложения от начала предложения / знака препинания / союза до конца предложения / знака препинания / союза. Под стеммингом понимается усечение слова до основы, например, «красивый / красивая / красивые → красив».

Черновая полярность каждого предложения рассчитывается после того, как все слова, входящие в положительный или отрицательный класс оценочного лексикона, будут заменены служебными символами (POS или NEG соответственно). Для корректировки черновой полярности и применения правил был вручную составлен корректирующий лексикон. Он включает три класса: модификаторы полярности, усилители полярности и антимодификаторы полярности. Модификаторы полярности — это слова, изменяющие полярность предложения на противоположную. Усилители полярности — это слова, увеличивающие счет полярности предложения. Антимодификаторы полярности — это слова, отменяющие изменение полярности предложения на противоположную.

Служебные символы для корректирующего лексикона представлены в таблице 2. Правила для корректировки черновой полярности представлены ниже. Каждое правило сопровождается примером.

Правило 1.1. $\langle \text{ALT} \rangle \langle \text{POS} \rangle \{n\} \rightarrow \langle \text{NEG} \rangle \{n\}$

Если в промежутке от начала предложения или знака препинания или союза «и, или» до следующего знака препинания или союза «и, или» имеется модификатор полярности, то положительная полярность всех слов, входящих в оценочный лексикон, в данном промежутке изменяется на отрицательную. Порядок указанных элементов (модификатор, слово с полярностью, любое другое слово) значения не имеет [1]: «Специалисты *не могут* дать *внятного* ответа на *простой* вопрос». Count = -3.

Таблица 2

Служебные символы
для корректирующего лексикона

Table 2

System symbols
for the correcting lexicon

Корректирующий лексикон	Примеры	Служебные символы
Модификаторы полярности	не, нет	ALT
	без	WT
	? ?! (вопросительный знак, вопросительный знак с восклицательным знаком)	QM
Усилители полярности	“ “ (кавычки)	Q
	очень, совершенно, никогда, нигде	INC
	Слово, набранное прописными буквами	CAP
	! (восклицательный знак)	EM
	Антимодификаторы полярности	так, такой
Прочие символы	Предложение	S
	Фрагмент	F
	союз и / или	CJ
	, ; — (запятая, двоеточие, точка с запятой, тире)	Z
	Слово, не входящее в лексикон	w

Правило 1.2. <ALT>(<NEG>){n} → <POS>{n}

Если в промежутке от начала предложения или знака препинания или союза «и, или» до следующего знака препинания или союза «и, или» имеется модификатор полярности, то отрицательная полярность всех слов, входящих в оценочный лексикон, в данном промежутке изменяется на положительную. Порядок указанных элементов (модификатор, слово с полярностью, любое другое слово) значения не имеет [1]: «Никто не заставил меня *ждать*». Count = +1.

Правило 2.1. (<INC><NEG>) → <NEG><NEG>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеются слова, входящие в отрицательный

лексикон, а также имеется усилитель, то каждый усилитель засчитывается за одно слово отрицательного лексикона. Порядок указанных элементов (усилитель, слово с полярностью, любое другое слово) значения не имеет: «Банк *совершенно не в состоянии* провести поиск счетов в течение месяца». Count = -2.

Правило 2.2. (<INC><POS>) → <POS><POS>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеются слова, входящие в положительный лексикон, а также имеется усилитель, то каждый усилитель засчитывается за одно слово положительного лексикона. Порядок указанных элементов (усилитель, слово с полярностью, любое другое слово) значения не имеет: «Тут могу отметить *очень удобную* возможность погашения кредита». Count = +2.

Правило 3.1. (<TH><ALT><NEG>) → <NEG>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеется модификатор полярности (ALT) и антимодификатор полярности (TH), то полярность слов в данном промежутке не изменяется, т. е. Правило 1 не применяется. Порядок указанных элементов (модификатор, антимодификатор, слово с полярностью, любое другое слово) значения не имеет: «*Таких проблем* с ипотекой я *не* ожидал». Count = -1.

Правило 3.2. (<TH><ALT><POS>) → <POS>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеется модификатор полярности (ALT) и антимодификатор полярности (TH), то полярность слов в данном промежутке не изменяется, т. е. Правило 1 не применяется. Порядок указанных элементов (модификатор, антимодификатор, слово с полярностью, любое другое слово) значения не имеет: «*Такого чуткого* отношения к клиентам я нигде *не* встречала». Count = +1.

Правило 4. <POS>{1}<QM> → <NEG>

Если предложение состоит из одного слова с положительной полярностью и заканчивается ? или ?!, то полярность данного слова изменяется на отрицательную: «Оперативно?». Count = -1.

Правило 5. <QM> → <NEG>

Если предложение заканчивается вопросительным знаком или вопросительным знаком с восклицательным знаком, при этом в предложении нет слов, входящих в оценочный лексикон, то полярность данного предложения равна -1: «Но на каком основании мне была подключена данная услуга?». Count = -1.

Правило 6. <Q><POS><Q> → <NEG>

Если слово с положительной полярностью заключено в кавычки, то полярность данного слова изменяется на отрицательную. Кавычки являются модификатором полярности: «“Забота” о клиенте». Count = -1.

Правило 7.1. $\langle WT \rangle (w | INC) \{0, \} \langle POS \rangle \rightarrow \langle NEG \rangle$

Если перед словом с положительной полярностью имеется слово «без», то полярность данного слова изменяется на отрицательную. Слово «без» является модификатором полярности. Между словом «без» и словом с полярностью может быть усилитель (особый, всякий, единый и т. п.): «Банк начислил пени без всяких разъяснений». Count = -1.

Правило 7.2. $\langle WT \rangle (w | INC) \{0, \} \langle NEG \rangle \rightarrow \langle POS \rangle$

Если перед словом с отрицательной полярностью имеется слово «без», то полярность данного слова изменяется на положительную. Слово *без* является модификатором полярности. Между словом «без» и словом с полярностью может быть усилитель (особый, всякий, единый и т. п.):

«Оплата с моей стороны происходила *без задержек*». Count = +1.

«Я смог *без особых проблем* снять деньги со счета». Count = +2.

Правило 8.1. $\langle POS \rangle \langle EM \rangle \rightarrow \langle POS \rangle \langle POS \rangle$

Если предложение заканчивается восклицательным знаком, а счет предложения положительный, то восклицательный знак приравнивается к одному слову с положительной полярностью: «Научились, делают, *молодцы!*». Count = 2. «Великолепное отношение к клиентам!». Count = 2.

Правило 8.2. $\langle NEG \rangle \langle EM \rangle \rightarrow \langle NEG \rangle \langle NEG \rangle$

Если предложение заканчивается восклицательным знаком, или вопросительным знаком, или вопросительным знаком с восклицательным знаком, а счет предложения отрицательный, то восклицательный знак, или вопросительный знак, или вопросительный знак с восклицательным знаком приравнивается к одному слову с отрицательной полярностью: «Умопомрачительная халатность сотрудников!». Count = -2. «Почему сотрудники данного филиала не знают об этом?!». Count = -2. «И зачем полгода мурыжили?» Count = -2.

Правило 9.1. $\langle POS \rangle \langle CAP \rangle \rightarrow \langle POS \rangle \langle POS \rangle$

Если в предложении имеются слова из прописных букв, а счет предложения положительный, то каждое слово из прописных букв приравнивается к одному слову с положительной полярностью: «Уровень сервиса — *BAU!*». Count = +3.

Правило 9.2. $\langle NEG \rangle \langle CAP \rangle \rightarrow \langle NEG \rangle \langle NEG \rangle$

Если в предложении имеются слова из прописных букв, а счет предложения отрицательный, то каждое слово из прописных букв приравнивается к одному слову с отрицательной полярностью: «Сотрудники *НЕ ЗНАЮТ* свою работу». Count = -2.

Правила применяются в определенной последовательности. Результатом применения алгоритма является суммирование скорректированных полярностей всех предложений, нормализация данной величины относительно количества слов в тексте отзыва и вывод системы (Отзыв положительный или Отзыв отрицательный). Алгоритм представлен на рисунке 2.

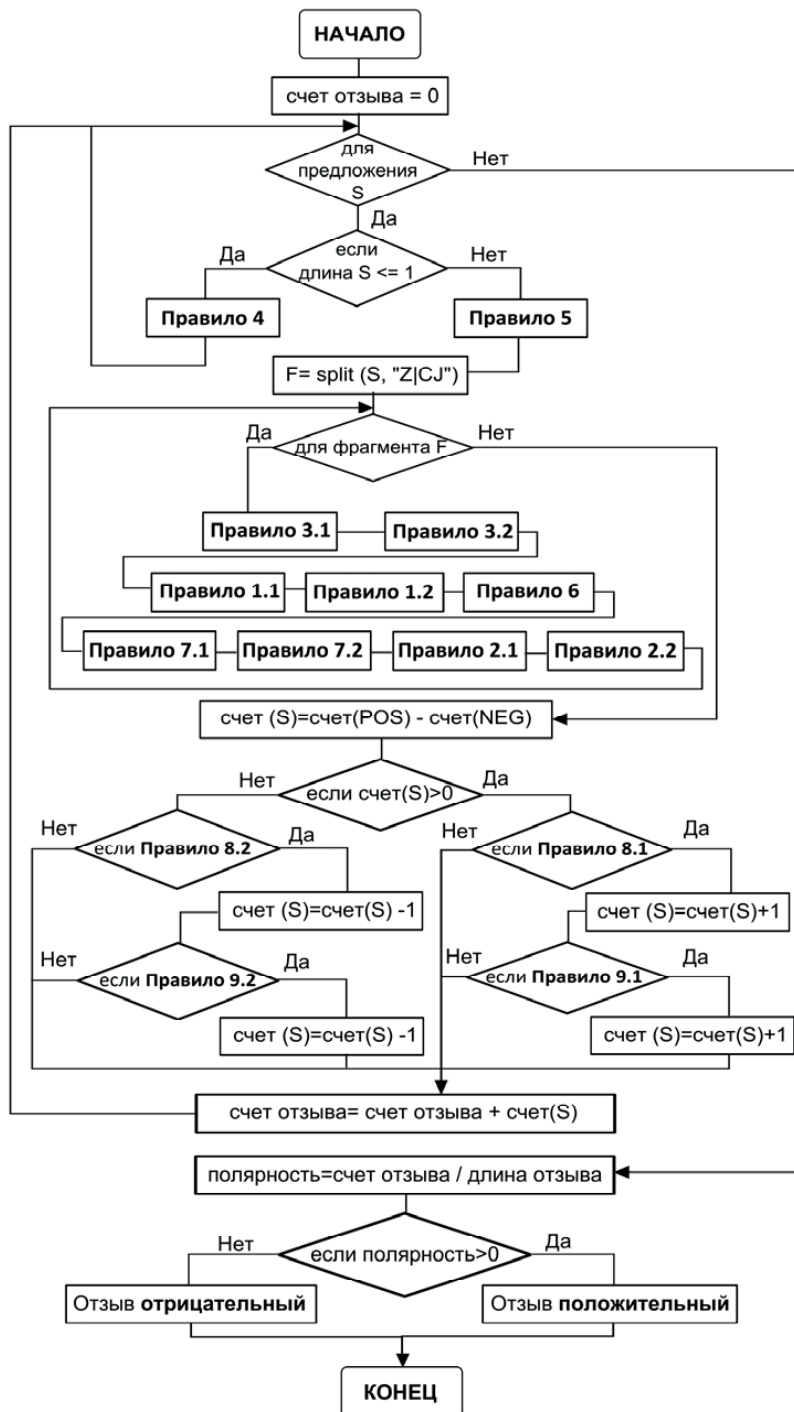


Рис. 2. Алгоритм для определения тональности текста отзыва

Fig. 2. The algorithm for determining the tonality of review texts

Результаты и обсуждение

Для оценки эффективности результатов эксперимента было использовано сопоставление выводов системы с выводами человека (авторская оценка) с помощью матрицы неточностей [10], представленной в таблице 3.

Таблица 3

Table 3

Матрица неточностей

The errata matrix

	Система говорит «да»	Система говорит «нет»
Человек говорит «да»	tp	fn
Человек говорит «нет»	fp	tn

Здесь tp — количество найденных релевантных документов, «истинные да» (true positive);

tn — количество найденных нерелевантных документов, «истинные нет» (true negative);

fp — количество ненайденных нерелевантных документов, «ложные да» (false positive);

fn — количество ненайденных релевантных документов, «ложные нет» (false negative).

Левый столбец матрицы (tp + fp) — это общее количество документов, найденных системой. Первая строка (tp + fn) — это общее количество релевантных документов. Вторая строка (fp + tn) — это общее количество нерелевантных документов.

Применительно к оценке классификаций релевантным считается результат правильного отнесения документа к тому или иному классу, нерелевантным — результат ошибочного отнесения документа к тому или иному классу.

Результаты эксперимента были верифицированы с помощью точности (Precision), полноты (Recall) и меры F. Точность представляет собой отношение количества найденных релевантных документов к общему количеству документов, найденных системой. Полнота представляет собой отношение количества найденных релевантных документов к общему количеству релевантных документов.

Точность и полнота рассчитываются с помощью формул (1), (2), (3) и (4):

$$\text{Precision(pos)} = a/(a+b), \quad (1)$$

$$\text{Precision(neg)} = d/(d+c), \quad (2)$$

$$\text{Recall(pos)} = a/(a+c), \quad (3)$$

$$\text{Recall(neg)} = d/(d+b), \quad (4)$$

где: а — это «истинные да» для отзывов с предполагаемой положительной полярностью;

b — это «истинные нет» для отзывов с предполагаемой положительной полярностью;

c — это «истинные да» для отзывов с предполагаемой отрицательной полярностью;

d — это «истинные нет» для отзывов с предполагаемой отрицательной полярностью.

Мера F вычисляется по формуле (5):

$$F=2*Precision*Recall/(Precision+Recall). \quad (5)$$

Эффективность предлагаемого классификатора на основе правил была сопоставлена с эффективностью Наивного Байесовского классификатора [17]. Результаты сопоставления двух классификаторов представлена в таблице 4.

Как мы видим, среднее значение F-меры для нашего классификатора на основе правил выше, чем среднее значение F-меры для Наивного Байесовского классификатора — 0.86 и 0.80 соответственно.

Значения F-меры показывают, что у обоих классификаторов эффективность для положительных отзывов выше, чем для отрицательных. Это означает, что

Таблица 4

**Результаты сопоставления
двух классификаторов**

Table 4

**The comparison results
of the two classifiers**

Полярность	Классификаторы	
	Классификатор на основе правил	Наивный Байесовский классификатор
Положительная		
Точность	0.82	0.75
Полнота	0.93	0.85
F-мера (для положительной полярности)	0.88	0.80
Отрицательная		
Точность	0.93	0.88
Полнота	0.77	0.71
F-мера (для отрицательной полярности)	0.84	0.79
Средняя F-мера	0.86	0.80

система обычно лучше распознает положительные отзывы, чем отрицательные. При выражении своего мнения по поводу качества обслуживания людям свойственно ассоциировать положительные мнения с некоей единой нормой, стандартом. Поэтому автор положительного отзыва, как правило, использует шаблонные, клишированные слова и выражения. И наоборот, отрицательные эмоции воспринимаются как аномалии, которые многочисленны и разнообразны. Следовательно, для отрицательных отзывов характерно значительно большее разнообразие лексических средств, а вероятность того, что используемая оценочная единица входит в используемый лексикон, снижается. Данная особенность просматривается также в извлеченном нами оценочном лексиконе, где число единиц отрицательного лексикона больше, чем положительного — 385 и 286 единиц соответственно.

Заключение и перспективы исследования

Мы разработали алгоритм на основе правил для определения тональности текста на русском языке, который включает девять правил, применяемых в определенной последовательности. Данные правила и соответствующие лексикограмматические конструкции упрощают каждое предложение и представляют текст отзыва как формальную модель.

Для проверки алгоритма был использован корпус из 200 отзывов о качестве банковского обслуживания. Эксперимент показал, что для одного и того же набора отзывов значения точности, полноты и F-меры у нашего алгоритма на 5-8% выше, чем у Наивного Байесовского классификатора. Это означает, что использование лексико-синтаксических моделей в дополнение к положительному и отрицательному лексиконам повышает эффективность анализа тональности текста, что особенно важно для текстов, которые относятся скорее к нарративу, чем к дескриптиву.

Мы применили алгоритм на основе правил к одной предметной области и оставляем экспансию на другие предметные области для будущих исследований. Перспективным также представляется сравнительный анализ применения разработанного алгоритма для отзывов о товарах (дескриптива) и об услугах (нарратива), а также эксперименты по определению тональности текстов других типов дискурса — научного, художественного и т. д.

СПИСОК ЛИТЕРАТУРЫ

1. Брунова Е. Г. Алгоритм с элементами формальной грамматики для контент-анализа мнений / Е. Г. Брунова, Ю. В. Бидуля // Вестник Тюменского государственного университета. Физико-математическое моделирование. Нефть, газ, энергетика. 2014. № 7. С. 242-250.
2. Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Москва, 27-30 мая 2015 г.). М.: Изд-во РГГУ, 2015. Том 2. Вып. 14(21).

3. Лукашевич Н. В. Извлечение и использование оценочных слов в задаче классификации отзывов на три класса / Н. В. Лукашевич, И. И. Четверкин // Вычислительные методы и программирование. 2011. Т. 12. С. 73-81.
4. Энциклопедический словарь крылатых слов и выражений / сост. В. Серов. М.: Локид-Пресс, 2003.
5. Annett M. A comparison of sentiment analysis techniques: Polarizing movie blogs / M. Annett, G. Kondrak // *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2008. Pp. 25-35.
6. Brunova E. Parametrical words in the sentiment lexicon / E. Brunova // *(IJCRSEE) International Journal of Cognitive Research in Science, Engineering and Education*. 2013. Vol. 1. No 2. URL: <http://www.ijcrsee.com>
7. Ganapathibhotla M. Mining opinions in comparative sentences / M. Ganapathibhotla, B. Liu // *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, 2008. Pp. 241-248.
8. Hatzivassiloglou V. Predicting the Semantic Orientation of Adjectives / V. Hatzivassiloglou, K. McKeown // *Proc. of the 35th Annual Meeting of ACL*, Madrid, 1997. Pp. 174-181.
9. Liu B. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing* / B. Liu. 2010. URL: <http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf>
10. Manning C. Introduction to information retrieval / C. Manning, P. Raghavan, H. Schütze. Cambridge University Press, Cambridge, 2008.
11. Medhat W. Sentiment analysis algorithms and applications: a survey / W. Medhat, A. Hassan, H. Korashy // *Ain Shams Engineering Journal*. 2014. No 5(4). Pp. 1093-1113.
12. Nasukawa T. Sentiment Analysis: Capturing Favorability Using Natural Language Processing / T. Nasukawa, J. Yi // *Proceedings of the 2nd International Conference on Knowledge Capture*. Florida, 2003. Pp. 70-77.
13. Pal J. Identifying Themes in Social Media and Detecting Sentiments / J. Pal, A. Saha // *International Journal of Statistics and Applications*. 2011. Vol. 1. No 1. Pp. 14-19.
14. Pang B. Thumbs up? Sentiment classification using machine learning techniques / B. Pang, L. Lee, S. Vaithyanathan // *Proceedings of EMNLP*, 2002. URL: <http://www.cs.cornell.edu/people/pabo/papers/sentiment.pdf>
15. Tan Li Im. Rule-based sentiment analysis for financial news / Tan, Phang Wai San, Chin Kim On, Patricia Anthony // *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Hong Kong, 2015. Pp. 1601-1606.
16. Turney P. Thumbs up or thumbs down? Semantic orientation applied to supervised classification of reviews / P. Turney // *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, 2002. Pp. 417-424.
17. Webb G. Not So Naive Bayes: Aggregating One-Dependence Estimators / G. Webb, J. Boughton, Z. Wang // *Machine Learning*. 2005. No 58. Pp. 5-24.

Elena G. BRUNOVA¹
Yulia V. BIDULYA²

**THE CUSTOMER IS ALWAYS RIGHT:
SENTIMENT ANALYSIS
FOR BANK SERVICE QUALITY**

¹ Dr. Sci. (Philol.),
Head of the Department
of Foreign Languages and Cross-Cultural Professional
Communication for Natural Sciences,
Tyumen State University
egbrunova@mail.ru

² Cand. Sci. (Philol.),
Associate Professor,
Department of Information Systems,
Tyumen State University
bidulya@yandex.ru

Abstract

The purpose of this research is to develop a rule-based algorithm for sentiment analysis. The dataset comprises the reviews in Russian on the bank service quality from clients' bank rating, www.banki.ru. Sentiment analysis is considered as the classification task, i.e. matching a text with one of the two classes: with positive or negative polarity. The algorithm is based on the use of certain lexical and syntactic structures, along with the sentiment lexicon consisting of positive and negative lexicons, as well as three service classes. The efficiency of the proposed algorithm is estimated with Precision, Recall and F-measure in comparison with the results of another algorithm widely used for sentiment analysis — the Naive Bayes Classifier. To estimate the efficiency, the dataset of 200 reviews on the bank service quality is used. The values of Precision, Recall and F-measure for the proposed algorithm are 5-8% higher than for the Naïve Bayes Classifier.

Citation: Brunova E. G., Bidulya Yu. V. 2017. "The Customer Is Always Right: Sentiment Analysis for Bank Service Quality". Tyumen State University Herald. Humanities Research. Humanities, vol. 3, no 1, pp. 72-89.

DOI: 10.21684/2411-197X-2017-3-1-72-89

Keywords

Sentiment analysis, natural language processing, content analysis, rule-based classifier, algorithm, user-generated content, lexicon.

DOI: 10.21684/2411-197X-2017-3-1-72-89

REFERENCES

1. Annett M., Kondrak G. 2008. "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs". In: *Advances in Artificial Intelligence*, pp. 25-35. Springer Berlin Heidelberg.
2. Brunova E. 2013. "Parametrical Words in the Sentiment Lexicon". *International Journal of Cognitive Research in Science, Engineering and Education*, vol. 1, no 2. <http://www.ijcrsee.com>
3. Brunova E. G., Bidulya Yu. V. 2014. "Algoritmy elementami formal'noy grammatiki dlya kontent-analiza mneniy" [Algorithm with Formal Grammar Elements for Sentiment Analysis]. *Tyumen State University Herald. Physical and Mathematical Modeling. Oil, Gas, Energy*, no 7, pp. 242-250.
4. Ganapathibhotla M., Liu B. 2008. "Mining Opinions in Comparative Sentences". *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 241-248. Manchester.
5. Hatzivassiloglou V., McKeown K. 1997. "Predicting the Semantic Orientation of Adjectives". *Proceedings of the 35th Annual Meeting of ACL*, pp. 174-181. Madrid.
6. Liu B. 2010. *Sentiment Analysis and Subjectivity*. Handbook of Natural Language Processing. <http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf>
7. Lukashovich N. V., Chetverkin I. I. 2011. "Iz vlechenie i ispol'zovanie otsenochnykh slov v zadache klassifikatsii otzyvov na tri klassa" [Retrieval and Application of Sentiment Lexicon in the Context of Reviews Classifying into Three Classes]. *Vychislitel'nye metody i programmirovaniye*, vol. 12, pp. 73-81.
8. Manning C., Raghavan P., Schütze H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
9. Medhat W., Hassan A., Korashy H. 2014. "Sentiment Analysis Algorithms and Applications: A Survey". *Ain Shams Engineering Journal*, vol. 5, no 4, pp. 1093-1113.
10. Nasukawa T., Yi J. 2003. "Sentiment Analysis: Capturing Favorability Using Natural Language Processing". *Proceedings of the 2nd International Conference on Knowledge Capture*, pp. 70-77. Florida.
11. Pal J., Saha A. 2011. "Identifying Themes in Social Media and Detecting Sentiments". *International Journal of Statistics and Applications*, vol. 1, no. 1, pp. 14-19.
12. Pang B., Lee L., Vaithyanathan S. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques". *Proceedings of EMNLP*. <http://www.cs.cornell.edu/people/pabo/papers/sentiment.pdf>
13. RGGU. 2015. "Kompyuternaya lingvistika i intellektual'nye tehnologii" [Computer Linguistics and Intellectual Technologies]. *Proceedings of the Annual International Conference "Dialogue"*, vol. 2, no 14(21). Moscow: RGGU.

14. Serov V. (comp.). 2003. Entsiklopedicheskiy slovar' krylatykh slov i vyrazheniy [Encyclopedia Dictionary of Popular Expressions]. Moscow: Lokid-Press.
15. Tan Li Im, Phang Wai San, Chin Kim On, Patricia Anthony. 2015. "Rule-Based Sentiment Analysis for Financial News". Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp. 1601-1606. Hong Kong.
16. Turney P. 2002. "Thumbs Up or thumbs Down? Semantic Orientation Applied to Supervised Classification of Reviews". Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424. Philadelphia.
17. Webb G., Boughton J., Wang Z. 2005. Not So Naive Bayes: Aggregating One-Dependence Estimators // Machine Learning, 58, P. 5-24.