# АнАлиз тонАльности твитов о телекоммуникАциях и бАнкАх нА основе методаА мАшинного обучения в рАмкАх SentiRuEval

Тутубалина Е. В. (tutubalinaev@gmail.com)1,
Загулова М. А. (mazagulova@stud.kpfu.ru)1,
Иванов В. В. (nomemm@gmail.com)1, 2,
Малых В. А. (валентин.малых@phystech.edu)3

1Казанский (Приволжский) Федеральный Университет (КФУ), Казань, Россия
2Институт информатики, Академия наук Татарстана, Казань, Россия
ЗИСА РАН, Москва, Россия

Ключевые слова: анализ тональности текстов, SentiRuEval, твиттер, классификация твитов

# Контролируемый подход к Задание SentiRuEval на настроения Анализ твитов о телекоммуникационных и финансовых компаниях

Тутубалина Е.В. (tutubalinaev@gmail.com)1,
Загулова М.А. (mazagulova@stud.kpfu.ru)1,
Иванов В.В. (nomemm@gmail.com)1, 2,
Малых В.А. (valentin.malykh@phystech.edu)3

1Казанский федеральный университет (КФУ), Казань, Россия
2Институт информатики Академии наук Татарстана, Казань, Россия
ЗИнститут системного анализа РАН, Москва, Россия

В данной работе описан контролируемый подход к решению задачи анализа тональности твитов о банках и операторах связи. Задача была сформулирована в виде отдельного трека в инициативе «Оценка настроений для русских» (SentiRuEval-2015). Подход, который мы предложили и оценили, основан на

Тутубалина Е.В., Загулова М.А., Иванов В.В., Малых В.А.

Модель векторной машины, которая классифицирует полярность настроений твитов. Набор функций включает функции частотности терминов, функции, характерные для Twitter, и функции на основе лексики. Для заданного домена для извлечения признаков были сгенерированы два типа словарей настроений: (i) созданные вручную словари, построенные на основе обзоров «за» и «против» ; (ii) автоматически генерируемые словари, основанные на точечной взаимной информации между униграммами в обучающем наборе.

В статье мы приводим результаты нашего метода и сравниваем их с результатами других команд, участвовавших в треке. Мы достигли 35,2% макроусредненного F-показателя для банков и 44,77% для твитов об операторах связи.
Метод, описанный в статье, занимает второе и четвертое место среди 7 и 9 команд соответственно. В этой статье также представлены наилучшие настройки SVM после настройки параметров классификатора и анализа ошибок с распространенными типами ошибок.

Ключевые слова: анализ настроений, сентиментальность, твиттер, социальные сети, классификация настроений в твиттере.

## 1. Введение

В последние годы анализу настроений уделяется большое внимание благодаря его способности определять мнения людей о продуктах, названных объектах, фактах (или событиях) и компаниях. Эта область исследований стала важной, особенно из-за быстрого роста сервисов микроблогов, таких как Twitter, в которых люди рассказывают о своем личном опыте.

Цель этой задачи — определить, является ли данный твит положительным, отрицательным или нейтральным в зависимости от его влияния на репутацию телекоммуникационной или финансовой компании. Обычно сложно реализовать традиционный анализ настроений пользователей, поскольку сбор твитов может быть шумным, а каждое сообщение ограничено по длине и может содержать орфографические ошибки, сленг и короткие формы слов. Было проведено большое количество исследований в области классификации тональности коротких неформальных текстов, которые хорошо описаны в (Martí nez-Cá mara, 2014). В современных документах применяются различные наборы функций, от традиционных функций классификации текста (например, энграммы, теги частей речи, основы) до специфических функций Twitter (например, смайликов, хэштегов, аббревиатур) для решения задачи в контролируемом режиме. образом (Киритченко и др., 2014). Поскольку анализ настроений на английском языке был тщательно изучен, исследований по классификации настроений в отзывах пользователей на русском языке не так много. В последних работах основное внимание уделялось решению задачи по анализу настроений в ходе треков анализа настроений ROMIP в 2011–2013 гг. (Четворкин, Лукачевич, 2013; Котельников, Клековкина, 2012; Блинов и др., 2013; Фролов и др., 2013).

В этом исследовании мы сообщаем о нашем подчинении задаче SentiRuEval. Подход основан на модели машины опорных векторов. Набор признаков включает признаки частоты терминов, т. е. nграммы слов, nграммы символов; специфические функции Twitter и функции на основе значков lex. Поскольку функции, основанные на лексике, являются наиболее полезными функциями для классификации твитов по тональности на английском языке, мы создали два типа лексики тональности. Вот эти два типа: словари, созданные вручную на основе обзоров «за» и «против» в определенной области; автоматически сгенерированные лексиконы, основанные на точечных

взаимная информация между униграммами в обучающей выборке. Мы достигаем 44,77% среднего макро-показателя F для твитов о телекоммуникационных компаниях и 35,2% для банковского домена, что дает улучшение макро-показателя F1 на 26,54% и 22,53% по сравнению с официальными исходными результатами соответственно.

Оставшаяся часть теста организована следующим образом. В Разделе 2 мы представляем связанную работу по классификации тональности коротких неформальных текстов. В разделе 3 мы описываем предлагаемые классификаторы с набором функций классификации текста и функций, специфичных для Twitter. В разделе 4 представлены результаты экспериментов. Раздел 5 содержит анализ ошибок. Наконец, в разделе 6 мы обсуждаем результаты и будущие расширения нашей работы.

## 2. Сопутствующая работа

Извлечение информации из коротких неформальных текстов, таких как твиты или sms-сообщения, получило большое внимание в анализе настроений (Go, 2009; Киритченко и др., 2014; Сидоров и др., 2013), обнаружении событий (Sakaki et al., 2010), выявление проблем (Гупта, 2013), обнаружение сарказма (Давидов и др., 2010) и отслеживание общественных настроений (О'Коннор и др., 2010). Традиционные подходы к классификации настроений основывались на наличии слов или смайликов, указывающих на положительную или отрицательную полярность (Turney, 2002; Taboada, 2010; O'Connor et al., 2010). В современных работах реализованы гибридные подходы, основанные на использовании методов машинного обучения и лексических ресурсов, таких как словари настроений (Mohammad et al., 2013; Zhu et al., 2014; Kiritchenko et al., 2014; Evert). , 2014). Недавние исследования показали, что важными функциями машинного обучения являются униграммы и биграммы из набора слов, а использование функций синтаксиса твитов (например, хэштегов, ретвитов и ссылок) может улучшить результаты классификации (Barbosa and Feng, 2010). В работе (Киритченко и др., 2014) авторы показали важность определения тональности слов при наличии отрицания. Они использовали отдельные лексиконы для терминов в утвердительном и отрицательном контекстах.

Большая часть работы по анализу тональности связана с использованием существующих словарей тональности и созданием лексических ресурсов, отражающих тональность слов (Martí nez-Cá Mara, 2014). Генерация словарей варьируется от ручных подходов к аннотированию словарей до полностью автоматизированных подходов. В (Evert, 2014) авторы использовали ручное расширение существующих словарей настроений и словарей смайликов и интернет-сленга. В (Mohammad et al., 2013) авторы создали автоматически сгенерированный лексикон хэштегов, оценивающий оценки настроений для терминов на основе точечной взаимной информации между терминами и твитами с полярностями. Вдохновленные этими работами, описывающими контролируемые методы, занимающие первое место в задаче SemEval-2014 по анализу тональности твитов на английском языке, мы решили создать словари тональности аналогичным образом.

Анализ тональности текстов на русском языке менее изучен. В (Chetviorkin and Lou kachevitch, 2013) авторы описывают первую открытую задачу о классификации настроений в отзывах пользователей на русском языке. Методы с учителем, основанные на SVM-классификаторе в сочетании с ручными или автоматическими словарями или системами, основанными на правилах, занимают первое место по отзывам о фильмах, книгах и цифровых фотоаппаратах в задаче. В работе (Frolov et al., 2013) авторы предложили подход, основанный на использовании специальных словарей и фактосемантических фильтров при анализе настроений пользователей в отзывах о книгах. В (Блинов и др.,

2013) авторы использовали ручные эмоциональные словари для каждой из трех областей и показали преимущества метода машинного обучения по сравнению с лексическим подходом для отзывов пользователей на русском языке. Они сообщили, что было сложно выбрать конкретный метод машинного обучения с лучшими результатами во всех областях обзора.

## 3. Классификация настроений на основе Twitter

Задача определяет, содержит ли каждый твит о телекоммуникационных компаниях (ТТК) или банках положительные, отрицательные или нейтральные настроения. Мы применили подход машинного обучения, основанный на модели мешка слов и наборе специфичных для Twitter функций, основанных на словарном запасе, которые описаны в разделе 3.3.

Следующие примеры иллюстрируют ситуации, в которых в твите появляются различные типы классификационных признаков. Такие твиты, как «Лучи дикой тревогой вашей организации, ГОРИТЕ В АДУ *бешусь*» («Посылая лучи дикой ненависти на вашу организацию, ГОРИТЕ В АДУ *ярость*»), содержат сильные негативные полярности в отношении слов со всеми символами в верхний регистр. Такие твиты, как «Почему у дебетовой карты написаны деньги просто так?!» («Почему с моей дебетовой карты сняли деньги без причины?!») и «Сеть прыгает из E в 3G и примерно примерно 5 минут ((» («Сеть переключается с E на 3G каждые 5 минут ((») не содержат положительных и отрицательных слов. Таким образом, человек-аннотатор обнаруживает отрицательное настроение в каждом твите в отношении контекста твита и того, являются ли последние символы смайликами, восклицательными или вопросительными знаками. Смайлики указывают на положительные или отрицательные настроения в короткие твиты, например, «@sberbank всё спасибо, готово :)» («@sberbank спасибо, дело сделано :)») и «сбербанк продлил рассмотрение дела до 160 дней :(» («Сбербанк продлил рассмотрение дела до 160 days :('). Комплексный анализ настроений в твитах, таких как «Проехать полгорода и узнать карту в другом банке. Всегда записан .__». всегда мечтал .__.») показывает, что некоторые смайлики содержат сарказм, а это означает, что противоположная полярность положительного в твиттере обозначен rd отрезок (приснилось) . Наличие характерных для твиттера функций, таких как URL-адрес или ретвит, указывает на нейтральный контекст твитов о новостях или в официальных сообщениях, например, «mts connect driver for android http://t.co/J3I5SNZuKM» («mts connect driver for android URL ») и «RT @Anna_Anna29: в билайне как узнать свой номер http://t.co/FpDZtLbdMZ» («RT @Anna_Anna2: как узнать свой номер в Билайн URL»).

В следующих примерах мы рассмотрим использование словарей настроений, созданных вручную и автоматически. Словари тональности, созданные вручную, успешно применялись для анализа тональности в традиционных подходах, определяющих, содержит ли сообщение позитивную или негативную тональность (Turney, 2002). Такие твиты, как «хreповый интернет, отвратительная работа с клиентами. Никогда не связывайтесь с этой шайкой» («Паршивый интернет, отвратительная работа с клиентами.
Никогда не связывайтесь с этой бандой») и «МТС пожелали хорошего дня, даже не построились ничего продать. Уверовал в добро» («МТС пожелал мне доброго дня, даже не пытался ничего продать. Я поверил в хорошее») содержат упоминание слов, не зависящих от домена, таких как отвратительный (отвратительный) и хороший .

(хороший). Многие твиты требуют более глубокого анализа тональности из-за сложного контекста сообщений, например негативные твиты «к вашему интернету хочется приложить подорожник» или «Билайн, отдай мне мой интернет». («Билайн, дай мне мой интернет»). По этим причинам для таких случаев автоматически создается другой лексикон настроений.

Мы протестировали три разных алгоритма обучения: наивный байесовский алгоритм, логистическую регрессию (MaxEnt) и модель машины опорных векторов (SVM). Квадрат евклидовой нормы L2 выбран в качестве стандартного регуляризатора для линейных моделей. На основании результатов, полученных на обучающих наборах, выбираем SVM с параметрами по умолчанию1 для классификации твитов в домене банков.

## 3.1. Два типа лексикона чувств

Мы исследуем два основных метода построения лексикона настроений: ручной и автоматический.

Ручным методом мы собрали отзывы пользователей с сайта otzovik.com: 3357 отзывов. отзывов о банках и 1928 отзывов о телекоммуникационных компаниях. Чтобы сделать корпус более точным, мы включили только положительные отзывы в положительный корпус и отрицательные отзывы в отрицательный корпус. Плюсы (Преимущества) и Минусы (Недостатки) — это части обзора, которые описывают веские причины, по которым автору обзора нравится или не нравится аспект продукта соответственно. Для каждого домена мы отобрали K наречий, прилагательных, глаголов и существительных, наиболее часто встречающихся в каждом корпусе. Затем мы сократили слова-существительные, выражающие эксплицитные аспекты в отзыве пользователя о конкретном домене за счет нейтральной полярности этих аспектов (например, связь (подключение), услуга (услуга), платеж (оплата), скорость (скорость), сотрудник (сотрудник)). ). Кроме того, мы сократили наиболее распространенные прилагательные (например, большой российский ), ( большой ), подключенный (абонент)) и глаголы, выражающие действие (например, использовать (использовать), написать (записать), подключить (подключить)) . Для каждого слова мы добавили другие словоформы. Словарь состоит примерно из 139 положительных и 131 отрицательного слова в домене банков. Словарь состоит примерно из 68 положительных и 168 отрицательных слов в домене телекоммуникационных компаний.

Вслед за Мохаммадом и др. (2013) и другие современные подходы, автоматически генерируемые словари основаны на оценке тональности для каждого термина w в обучающем тесте:

$$\text{оценка}(w) = \text{PMI}(w, pt) \quad \text{PMI}(w, nt)$$

$$\text{PMI}(w,pt) = \log_2 \frac{p(ш, пт)}{p(w) \times p(pt)}$$

где PMI — точечная взаимная информация, pt — положительные твиты, nt — отрицательные твиты, p(w), p(pt) и p(w,pt) — вероятности появления w в положительном корпусе. Слова с сильной полярностью тональности имеют статистически значимую разницу между PMI (w,pt) и PMI (w,nt) в отличие от нейтральных слов. Например, пара значений (PMI (w,pt), PMI (w,nt)) вычисляется по твитам в домене банков.

---

[1] Мы использовали библиотеку scikit-learn в Python.

равно (-0,8016, 0,1450) для нейронного слова еда ( еда); ( 15,2438, 1,5649) для отрицательного слова ущерб (убыток) и (2,1839, 19,2026) для положительного слова выгодный (прибыльный). Поскольку твиты содержат низкочастотные шумные слова, мы игнорировали термины, встречающиеся менее трех раз в обучающей выборке.

### 3.2. Предварительная обработка коротких неформальных текстов

Поскольку необработанные твиты обычно неформальные и очень шумные, выполняются следующие шаги предварительной обработки. Упоминания пользователей нормализуются до @username. Морфосинтаксический анализатор2 применяется для замены слов в твите базовыми формами. Мы определяем отрицательный контекст как часть твита между отрицанием (например, частицей не (нет), предикативным выражением нет (не)) словом и знаком препинания. Слова со связанными отрицаниями (слова после отрицаний) модифицируются вместе с тегом отрицания «neg_». Мы идентифицируем смайлики и заменяем их соответствующими выражениями настроения3 (например, мы заменяем ':-)' на счастливый, 'o_0' на удивление и ' ;-]' на подмигивание).

### 3.3. Особенности классификации для классификации твитов по настроению

Каждый твит представлен в виде вектора признаков; краткое описание функций, которые мы используем, представлено ниже: • словесные n-граммы:

в качестве функций используются униграммы (отдельные слова) и биграммы (выражения из нескольких слов), извлеченные из твита. Выбираются объекты с частотой документов больше двух. • символьные n-граммы: для отбора признаков рассматривались

строчные n-граммы символов для n=2,...,4 с частотой документа больше двух. • слова, написанные заглавными буквами: функция подсчитывает количество слов, содержащих все заглавные буквы. Аббревиатуры компаний (например, МТС (MTS), ВТБ (VTB)) исключены.

- пунктуация: функции подсчитывают количество знаков в последовательностях восклицательных знаков, вопросительных знаков или комбинации этих знаков и количество знаков в смежных последовательностях точек. Последовательности, состоящие из более чем одной метки, рассматриваются для выбора признаков.
- последний символ: двоичная функция указывает, является ли последний символ твита бывшим символом. восклицательный знак или скобка.
- смайлики: извлекаются четыре признака: количество положительных смайликов; количество отрицательных смайликов; две двоичные функции, которые указывают, является ли последний символ твита положительным или отрицательным смайликом соответственно.
- Специфические для Twitter функции: три бинарных функции, которые указывают, содержит ли твит упоминание пользователя Twitter, ретвит и наличие URL-адреса.

---

2 Мы использовали инструмент Mystem, url: https://tech.yandex.ru/mystem/

3 Мы использовали некоторые выражения настроения из http://en.wikipedia.org/wiki/List_of_emoticons.

- функции на основе лексикона: для каждого из двух сгенерированных словарей функции рассчитывается следующим

  образом: – для словаря, созданного вручную, мы подсчитываем количество слов с положительной тональностью, слов с отрицательной тональностью. Слова настроения с отрицанием меняют полярность настроения, например, положительное слово с суффиксом отрицания рассматривается как отрицательное слово.

  – для автоматически создаваемого словаря добавлены четыре функции: подсчет слов с ненулевыми баллами; сумма оценок тональности слов, нормированная по количеству слов; максимальная оценка тональности и минимальная оценка тональности в твите. Слова настроения с отрицанием сдвигают оценку настроения в сторону противоположной полярности.

## 4. Экспериментальные результаты.

Мы использовали обучающий набор из 5000 аннотированных твитов для каждого домена, предоставленного для задачи SentiRuEval. Итоговое количество твитов в тестовой коллекции — 4549 твитов о банках и 3845 твитов о телекоммуникационных компаниях.

Официальные результаты, полученные нашими классификаторами на тестовой выборке, представлены в таблице 1. В таблице представлены официальные исходные результаты и результаты метода, занимающего первое место по макросредней F-мере как основному показателю качества в задаче ( Лукашевич и др., 2015). Макросредняя F-мера рассчитывается как среднее значение между F-мерой положительного класса и F-мерой отрицательного класса. Классификатор был обучен предсказывать все три класса (положительные, отрицательные и нейтральные), но эта макроусредненная мера не учитывает правильно классифицирующие нейтральные твиты. Наш метод занимает второе место среди 7 команд с 14 прогонами в домене банков. Метод занимает четвертое место среди 9 команд и пятое среди 19 прогонов в области телекоммуникационных компаний. Лучший подход имеет улучшение макроэкономического показателя F1 на 0,007% по сравнению с нашим подходом в области банков.

Таблица 1. Показатели производительности в задаче классификации твитов в двух доменах: телекоммуникационные компании и банки

| | телекоммуникационные компании | | банки | |
|---|---|---|---|---|
| | микро Ф | макрос Ф | микро Ф | макрос Ф |
| Лучший | 0,536 | 0,488 | 0,343 | 0,359 |
| Наш подход | 0,528 | 0,448 | 0,337 | 0,352 |
| Официальный базовый уровень | 0,337 | 0,182 | 0,238 | 0,127 |

Мы также представляем эксперименты по абляции признаков в тестовом наборе, удаляя по одному элементу каждой отдельной категории из полного набора. В табл. 2 представлены результаты экспериментов по абляции, в каждой строке указаны макросредняя точность, макросредняя полнота и макросредняя F-мера, рассчитанная как среднее значение между соответствующими мерами положительного и отрицательного классов. Наиболее эффективными являются словесные n-граммы для твитов о телекоммуникационных компаниях. Наиболее эффективными функциями являются

на основе символьных n-грамм и смайликов в банковском домене. Метод также показывает улучшение F-меры на 0,021% после сокращения n-грамм слов в области банков и улучшение на 0,041% F-меры после сокращения автоматических словарей слов в области ttk. Эти улучшения могут быть вызваны динамическим контекстом твитов-сообщений о компаниях. Твиты обучающей выборки были опубликованы в 2014 году, твиты тестовой выборки — в 2013 году.

Таблица 2. Результаты экспериментов по абляции в двух областях

| | телекоммуникационные компании (ттк) | | | банки | | |
|---|---|---|---|---|---|---|
| | макрос P | макрос R | макрос F | макрос P | макрос R | макрос F |
| Все функции | 0,443 | 0,471 | 0,447 | 0,538 | 0,279 | 0,352 |
| без персонажа n-граммы | 0,447 | 0,413 | 0,405 | 0,444 | 0,233 | 0,301 |
| без смайлов без | 0,413 | 0,450 | 0,406 | 0,489 | 0,274 | 0,335 |
| обоих словарей без | 0,419 | 0,553 | 0,475 | 0,496 | 0,276 | 0,337 |
| последнего символа | 0,458 | 0,379 | 0,390 | 0,509 | 0,274 | 0,340 |
| без словаря (ручная версия) | 0,379 | 0,505 | 0,432 | 0,516 | 0,270 | 0,340 |
| без словаря (автоматическая | 0,427 | 0,569 | 0,488 | 0,426 | 0,292 | 0,343 |
| версия) без слов, | 0,446 | 0,447 | 0,436 | 0,498 | 0,293 | 0,349 |
| написанных заглавными | 0,429 | 0,429 | 0,412 | 0,522 | 0,286 | 0,350 |
| буквами без знаков препинания | 0,447 | 0,441 | 0,443 | 0,491 | 0,289 | 0,351 |
| без твиттера особенности синтаксиса без словесных n-грамм | 0,390 | 0,412 | 0,373 | 0,507 | 0,316 | 0,373 |

Мы также проанализировали значение настройки SVM для нашего метода. После переноса метода регуляризованной регрессии SVM на эластичную сеть, которая линейно сочетает штрафы L1 и L2 и альфа члена регуляризации до 0,0001, классификатор продемонстрировал улучшения на 4–5% в макро-мерах F1 по сравнению с нашими результатами с параметрами SVM по умолчанию в обоих случаях. домены. Настроенный классификатор достигает средней макроэкономической F-меры 39,46% для домена банков и 50,6% для твитов о телекоммуникационных компаниях. Результаты показывают, что тщательная настройка алгоритма машинного обучения может дать гораздо лучшие результаты.

## 5. Анализ ошибок

После анализа ошибок мы выделяем следующие типы наиболее частых ошибок в классификации твитов:

- •орфографические ошибки и трудности с транслитерацией английского текста на русский
- •множество хэштегов
- •эмоциональное обсуждение нейтральных тем
- •недостаточный размер лексикона настроений (наличие нелексических слов в тексте тестовый набор)

Из таблицы 3 видно, что большинство ошибок вызвано недостаточной информацией о контексте в положительных или отрицательных твитах о компаниях.

Таблица 3. Распределение типов ошибок

| | Опечатки и транслитерация | Несколько хэштегов | Эмоциональный обсуждение | Недостаточный размер лексикон настроений |
|---|---|---|---|---|
| телекоммуникации компании | 20,40% | 8% | 14,90% | 43% |
| банки | 9% | 1% | 11% | 64% |

Такие твиты, как «Билайну труба короче» («Игра Билайна окончена»), содержат скрытый отрицательный смысл, например, «игра окончена» со словом «труба» («труба»). Негативные твиты, такие как «Самый безалаберный банк!» («Самый неорганизованный банк!») неправильно классифицируются из-за низкочастотных слов, таких как «безалаберный», которые не содержатся ни в обучающем наборе, ни в созданных лексиконах.

Мы не применяли корреляцию ошибок для случаев орфографических ошибок, таких как ацтой (вздор) и аккорд (черт), а правильное написание этих слов включено в созданные вручную лексиконы. Такие твиты, как «Билайн. Дисконнектинг пипл.
("Билайн. Отключение людей") с транслитерированными словами с ярко выраженной отрицательной полярностью на английском языке были ошибочно классифицированы как нейтральные. Анализ показывает, что орфографические ошибки вызывали меньше ошибок в твитах, чем удлиненные, транслитерированные слова и наличие звездочки (звездочки) в нецензурных словах.

Такие хэштеги, как #отстойсвязь (#yourconnectionsucks), #мтсумри (#mtsdie), #люблюего (#loveit), содержат сильную ориентацию на чувства. 8% ошибок в телекоммуникациях можно устранить, если разбить хэштеги на слова, а затем рассчитать оценки тональности хэштегов.

Ошибки четвертого типа связаны с нейтральными твитами о телекоммуникационных компаниях или банках, которые содержат положительную или отрицательную полярность по другим темам (например, твиты о дресс-коде компании , дружеском разговоре или флирте с сотрудником компании). Другим типом таких твитов является твит, описывающий какое-то ежедневное мероприятие компании: «Матч штаб квартиры Вымпелком — Сибирь. Пока ведем!!! :)» («Матч штаб-квартиры «Вымпелкома» — «Сибирь». Мы побеждаем!!! :)»). Во всех этих случаях твит о компании нейтрален. Наши классификаторы не учитывали такие случаи, на которые приходится до 11% ошибок о банковских твитах и 14,9% ошибок о телекоммуникационных твитах.

## 6. Заключение

В этой статье мы описали контролируемый метод классификации настроений финансовых или телекоммуникационных данных Twitter с упором на потребительский опыт. Предлагаемый метод использует машины опорных векторов с функциями частоты терминов, функциями, специфичными для Twitter, и функциями на основе лексики. Для твита функции на основе лексикона были сгенерированы путем проверки того, входит ли слово в лексикон настроений, которые были созданы как автоматически, так и вручную на основе отзывов пользователей. Для того чтобы произвести автоматически

Создав лексикон, мы использовали точечную взаимную информацию для расчета оценки тональности и связывания каждого слова из обучающего набора с соответствующим классом тональности.

Мы продемонстрировали, что при использовании этих признаков эффективность классификации повышается с базовых макроусредненных F-мер с 0,265 до 0,447 для телекоммуникационных компаний и с 0,225 до 0,352 для банков. Мы планируем создать большие корпуса положительных и отрицательных твитов для улучшения классификаторов с автоматически создаваемыми лексиконами.

## Благодарности

Работа выполнена за счет субсидии Правительства РФ на поддержку Программы повышения конкурентоспособности Казанского федерального университета и при поддержке РФФИ (проект РФФИ 13-07-00773).

## Рекомендации

1. Барбоса Л., Фэн Дж. (2010), Надежное обнаружение тональностей в Твиттере по необъективным и зашумленным данным, Материалы 23-й Международной конференции по компьютерной лингвистике, Пекин, стр. 38–42. 2. Блинов П., Клековкина М. .,
Котельников Е., Пестов О. (2013), Исследование лексического подхода и методов машинного обучения для анализа настроений, Вычислительная лингвистика и интеллектуальные технологии, Vol. 2(12), стр. 48–58.
3. Четверкин И., Лукашевич Н. (2013), Оценка систем анализа настроений.
на русском языке, ACL 2013, с. 14.
4. Давыдов Д., Цур О., Раппопорт А. (2010), Полуконтролируемое распознавание саркастических предложений в Twitter и Amazon, Материалы четырнадцатой конференции по компьютерному изучению естественного языка, Ассоциация компьютерной лингвистики, стр. 107. –116.

5. Эверт С., Пройсл Т., Грейнер П., Кабаши Б. (2014), SentiKLUE: Обновление полярности Классификатор за 48 часов, SemEval 2014, Дублин, с. 551.
6. Фролов А. В., Поляков П. Ю., Плешко В. В. (2013), Использование семантических фильтров в приложениях. ссылка на анализ настроений рецензий на книги, доступно по адресу: www.dialog-21.ru/digests/dialog2013/materials/pdf/FrolovAV.pdf 7. Го А., Бхайани Р., Хуанг
Л. (2009), классификация настроений в дистанционный контроль, Отчет о проекте CS224N, Стэнфорд, стр. 1–12.
8. Гупта Н.К. (2013), Извлечение фраз, описывающих проблемы с продуктами и услугами, из сообщений Twitter, Computació n y Sistemas, Vol. 17(2), стр. 197–206.
9. Кириченко С., Чжу С., Мохаммад С.М. (2014), Анализ тональности коротких неформальных текстов, Журнал исследований искусственного интеллекта, Vol. 50, стр. 723–762.
10. Котельников Е.В., Клековкина М.В. (2013), Анализ тональности текстов на основе методов машинного обучения // Автоматический анализ тональности текстов. Вычислительная лингвистика и интеллектуальные технологии: Материалы международной конференции «Диалог», стр. 753–762.

11. Лукашевич Н., Блинов П., Котельников Е., Рубцова Ю., Иванов В., Тутубалина Е.
(2015), SentiRuEval: тестирование систем объектно-ориентированного анализа настроений на русском языке, Материалы международной конференции «Диалог-2015», Москва, стр. 3–9.
12. Мартинес-Камара Э., Мартин-Вальдивия М.Т., Урена-Лопес Л.А., Монтехо-Раес А.Р.
(2014), Анализ настроений в твиттере, Natural Language Engineering, Vol. 20(01), стр. 1–28.

13. Мохаммад С.М., Кириченко С., Чжу С. (2013), NRC-Канада: создание современного анализа тональности твитов, Материалы второй совместной конференции по лексической и вычислительной семантике (SEMSTAR'13). ), Атланта, с. 2 14. O'Коннор Б., Баласубраманян Р.,
Рутледж Б.Р., Смит Н.А. (2010), От твитов к опросам: связывание настроений текста с временными рядами общественного мнения, ICWSM-11, Барселона, стр. 122–129.

15. Сакаки Т., Окадзаки М., Мацуо Ю. (2010), Землетрясение потрясло пользователей Twitter: обнаружение событий в реальном времени с помощью социальных датчиков. Материалы 19-й международной конференции по всемирной паутине, ACM, стр. 851–860.
16. Сидоров Г., Миранда-Хименес С., Виверос-Хименес Ф., Гелбух А., Кастро-Санчес Н., Веласкес Ф., Гордон Дж. ( 2013), Эмпирическое исследование подхода на основе машинного обучения для сбора мнений в твиты, Достижения в области искусственного интеллекта, Vol. 7629, стр. 1–14.
17. Табоада М., Брук Дж., Тофилоски М., Фолл К., Стеде М. (2011), Методы анализа настроений на основе лексикона, Вычислительная лингвистика, Том 37(2), стр. 267–307.
18. Turney PD (2002), Большой палец вверх или большой палец вниз?: применение семантической ориентации к неконтролируемой классификации обзоров, Proceedings of the 40th Annual Meeting on Association for Computer Languages, Филадельфия, стр. 417–424.
19. Уилсон Т., Вибе Дж., Хоффманн П. (2005), Признание контекстуальной полярности в анализе настроений на уровне фраз, Материалы конференции по технологиям человеческого языка и эмпирическим методам обработки естественного языка, стр. 347–354.