Объектно-ориентированный анализ твитов по тональности: результаты и проблемы

© Н. В. Лукашевич
НИВЦ МГУ
Москва
louk_nat@mail.ru

© Ю. В. Рубцова
ИСИ СО РАН,
Новосибирск
yu.rubtsova@gmail.com

Аннотация

В представленной работе описываются результаты решения задачи объектно-ориентированного анализа твитов по тональности, которая была проведена в рамках соревнования SentiRuEval – оценки российских систем классификации сообщений по тональности. В соревновании использовались твиты из двух предметных областей: телекоммуникационных компаний и банков. Задача состояла в определении тональности твита по отношению к упомянутым в твите организациям: положительная, отрицательная или нейтральная. Тональность твита может отражать как положительное или отрицательное мнение автора, так и упоминаемый им положительный или отрицательный факт относительно указанной организации. Основной задачей этой статьи является анализ текущего состояния подходов и алгоритмов, применяемых участниками соревнования.
Работа частично поддержана грантом РФФИ № 14-07-00682

1 Введение

Пользователи Интернета часто используют микроблоговую платформу Твиттер для выражения своего мнения об окружающем их мире. Они могут свободно и открыто писать о том, что им нравится или не нравится относительно происходящих с ними событий, например, обслуживание в кафе или выбор нового устройства. Согласно открытым источникам, количество зарегистрированных пользователей твиттера превышает 500 млн человек и это число продолжает расти. Популярность Твиттера, влияние, которое пользователи оказывают своими высказываниями, доступность данных привлекает исследователей и ведет к постановке различных автоматическому анализу сообщений.

Твиттер отличается от других социальных сетей. В Твиттере ограничена длина сообщений, поэтому можно сказать, что автоматический анализ текстов происходит на уровне фраз или предложений, а не на уровне документов. В Твиттер часто пишут с

мобильных телефонов, поэтому встречается множество опечаток и ошибок, например «зона оьдыха», вместо «зона отдыха». Язык твитов изобилует сленговыми словами, сокращениями, аббревиатурами.

Несмотря на то, что микроблоги достаточно молодое явление, исследователи активно занимаются анализом тональности сообщений блогов в целом и Твиттера в частности [2, 10, 15, 22]. Для того чтобы поддержать исследования в области тонального анализа социальных сетей, проводятся соревнования специализированных программных систем [14, 18]. Так, в 2012–2014 годах в рамках конференции CLEF проходило соревнование систем по оценке репутации в Твиттере [3-4]. В 2014-2015 организовано соревнование анализаторов тональности текстов на русском языке SentiRuEval [11]. Одним из заданий в SentiRuEval было тестирование систем анализа репутации относительно заранее выбранных организаций по текстам твитов. В этой статье мы опишем задачу репутационной оценки постов в Твиттере, исходные данные, результаты участников и их подходы к классификации, а также представим наш анализ проблем существующих подходов.

Статья организована следующим образом. В следующем разделе представлен обзор ранее проведенных соревнований по классификации твитов по тональности. Третий раздел описывает задачу репутационной оценки твита по тональности SentiRuEval, подготовку коллекций, результаты участников. Анализ результатов участников, сложности и проблемы подходов описаны в четвертом разделе. Последний раздел состоит из выводов и заключения.

2 Обзор близких работ

В последние годы было проведено несколько соревнований посвященных анализу тональности

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

сообщений Твитера. В 2013 и 2014 годах, в рамках конференции SemEval проходило соревнование систем автоматического распознавания тональности. Участникам было предложено две задачи: классификация сообщений на уровне фраз и классификация сообщений на уровне целого сообщения. В первой задаче требовалось определить, является ли данная фраза позитивно, негативно или нейтрально окрашенной. Во второй задаче требовалось определить, выражает ли данное сообщение позитивное или негативное мнение автора в противоположность объективной информации [14, 18].

В 2012–2014 годах, в рамках конференции CLEF, проходило соревнование систем оценки репутации (online reputation management systems) RepLab [3-4]. Цель соревнования состояла в том, чтобы определить, положительно или отрицательно твит влияет на репутацию компании.

Организаторы RepLab делают акцент на том, что определение тональности репутации существенно отличается от обычного определения тональности, в котором требуется отличить субъективную информацию от объективной. При определении тональности репутации должны приниматься во внимание как факты, так и субъективные мнения. Совокупность фактов и мнений помогают определить, имеет ли текст позитивные или негативные последствия для репутации выбранного объекта [3-4].

В качестве наборов данных RepLab были представлены твиты из следующих предметных областей: автомобили, банки, университеты и музыка. Для каждой предметной области было собрано как минимум 2200 твитов: первые 700 твитов формировали обучающую коллекцию, оставшиеся 1500 – тестовую. Обучающие и тестовые коллекции собирались с интервалом в несколько месяцев. Оценка систем была представлена на единой тестовой коллекции, без деления на предметные области.

В России в 2011-2013 годах в рамках семинара ROMIP проводились соревнования систем анализа тональности текстов на русском языке. Участникам была предложена задача классификации по тональности пользовательских отзывов на книги, фильмы и цифровые камеры. Классификация проводилась на уровне документов [5].

Главной задачей проведенных ранее тестирований на русском языке был автоматический анализ тональности небольших текстов на уровне документа – отзывов пользователей (о фильмах, книгах, цифровых фотокамерах) или мнений, выраженных в форме прямой или косвенной речи (новости). Основной целью нового цикла тестирований текстов на русском языке SentiRuEval является автоматическая оценка тональности по отношению к заданному объекту. В данной работе мы представляем подробный анализ достижений и проблем систем-участников в объектно-ориентированном анализе твитов по тональности.

3 Описание задачи

Цель задачи объектно-ориентированного анализа твитов по тональности на SentiRuEval заключается в выявлении твитов, которые оказывают влияние на репутацию организации, упомянутой в твите. Такие твиты могут содержать как положительное или отрицательное мнение автора, так и положительный или отрицательный факт относительно упомянутой организации. В качестве предметных областей были выбраны твиты о телекоммуникационных компаниях (ТКК) и твиты о банках. Важно понимать, что исследуется задача оценки отзыва по отношению к компании, а не текста сообщения в целом.

Задача SentiRuEval похожа на задачу определения тональности репутации на RepLab [3-4]. Разница состоит в том, что для SentiRuEval выбраны твиты из двух предметных областей, и результат работы систем участников для этих предметных областей считается по отдельности, что дает возможность изучить зависимость репутационного анализа твитов от конкретной предметной области. Для проведения тестирования были взяты твиты о восьми банках и семи телекоммуникационных компаниях.

Перед участниками была поставлена задача определить тональность репутации твита по отношению к упомянутой компании: положительная, отрицательная или нейтральная. В обучающих и тестовых коллекциях были выделены поля с перечнем всех организаций выбранной предметной области, по умолчанию поля имеют разметку 0 «нейтральное отношение». Участникам предлагается заменить 0 на «1» (позитивное отношение к компании) или «-1» (негативное отношение к компании) или оставить «0», если отношение автора к указанной в тексте компании нейтральное. Структура одного сообщения представлена в листинге 1.

Листинг1. Структура одного сообщения:

```
<table name="ttk_test">
<column name="id">11</column>
<column name="twitid">40892934798</column> – уникальный идентификатор твита
<column name="date">1386331328</column> – дата твита в unix формате
<column name="name">BboyChapi</column> – имя автора твита
<column name="text">RT @wylsacom: Но в силу поддержки разных частот, работать американские iPhone 5S/5C будут только в МТС/Билайн.</column> – текст твита
<column name="beeline">0</column> – компания, которая упоминается в твите
<column name="mts">0</column> – компания, которая упоминается в твите
<column name="megafon">NULL</column> – компания, которая не упоминается в твите
<column name="tele2">NULL</column>
<column name="rostelecom">NULL</column>
<column name="komstar">NULL</column>
<column name="skylink">NULL</column>
</table>
```

Таблица 3. Распределение сообщений по классам тональности

	Нейтральные	**Позитивные**	**Негативные**	**Общее количество твитов в коллекции**
ТКК				
Обучающая коллекция	2397	973	1667	5000
Тестовая коллекция	2816	413	944	3845
Банки				
Обучающая коллекция	3569	410	2138	5000
Тестовая коллекция	3592	350	670	4549

3.1 Сбор коллекций

Коллекции для тестирования были собраны собирались с помощью Streaming API Twitter. Обучающая коллекция собиралась в промежуток июль – август 2014, тестовая с декабря 2013 по февраль 2014. Специально было сделано так, чтобы по времени сбор обучающей и тестовой коллекциями был отделен значимым промежутком времени. Это важно для моделирования реальной ситуации, когда со временем может меняться язык, а также могут происходить какие-то события, влияющие на восприятие имиджа компании.

Важно также отметить, что мы искусственно не завышали количество тонально окрашенных сообщений в обучающих и тестовых коллекциях для построения сбалансированной коллекции. В реальной жизни системы столкнутся с аналогичными наборами данных. Распределение сообщений между имеющими тональность и нейтральными в обучающей и тестовой коллекциях представлено в таблицах 1 и 2.

Таблица 1. Распределение сообщений между имеющими тональность и нейтральными в обучающей коллекции:

	нейтральные, %	тональные, %
ТКК	47,59	52,42
Banks	58,35	41,65

Таблица 2. Распределение сообщений между имеющими тональность и нейтральными в тестовой коллекции:

	нейтральные, %	тональные, %
ТКК	67,7	32,3
Banks	77,9	22,1

Распределение сообщений в обучающей и тестовой коллекциях согласно классам тональности представлено в Таблице 3. Как видно из таблицы, общее количество твитов в коллекции не равно сумме нейтральных, позитивных и негативных упоминаний, так как пользователи в одном сообщении могут упоминать более одной компании. Доля сообщений, в которых упоминается более одной компании составляет 4,12% для обучающей коллекции телекоммуникационных компаний и 16,68% для банков. Пользователи могут перечислять компании, в этом случае, как правило, оценки для всех компаний совпадают, могут сравнивать или противопоставлять компании между собой, в этом случае, оценки у компаний различаются.

Мы заметили, что иногда пользователи не хотят показаться грубыми и добавляют положительные эмотиконы (символы, обозначающие эмоции на письме) в явно негативные или ироничные сообщения. Например: «@VadimSavin неа... мне только мтс впаривает свой смартфон на скидке))». Поэтому простые методы, основанные на извлечении эмотиконов, которые применяют для классификации на уровне всего твита, не всегда дают хорошие результаты [9, 17].

3.2 Разметка коллекций и меры оценки качества классификаторов

Обучающие и тестовые коллекции были размечены асессорами. В общей сложности было размечено 20 000 сообщений, по 10 000 сообщений на каждую предметную область для тестовой и обучающей коллекций. Каждая коллекция была размечена как минимум двумя асессорами. В процессе анализа разметки выяснилось, что некоторые сообщения вызывают сомнения и дискуссии относительно того, к какому классу тональности это сообщение должно быть отнесено. Поэтому, чтобы снизить вероятность ошибок, каждая тестовая коллекция была оценена тремя асессорами, далее была применена процедура

голосования и сформирована финальная тестовая коллекция. Перед асессорами была поставлена задача оценить влияние твита на репутацию упомянутого в нем объекта. Финальный этап подготовки коллекций состоял в фильтрации нерелевантных сообщений из тестовых и обучающих коллекций. Результаты подготовки тестовых коллекций представлены в Таблице 4.

Таблица 4. Результаты процедуры голосования при разметки тестовых коллекций

	Совпадение оценок как минимум двух асессоров	Полное совпадение разметок	Финальное число твитов в тестовой коллекции
Банки	4 915 (98,3%)	3 818 (76,36%)	4 549
ТКК	4 503 (90,06%)	2 233 (44,66%)	3 845

В качестве основного показателя оценки качества классификаторов, использовалось макро-усреднение F-меры, которое рассчитывается как среднее значение между F-мерой положительного класса и F-мерой отрицательного класса, формула 1. Нейтральный класс не участвует в расчётах F-меры, но это не упрощает задачу классификации до классификации твитов на два класса, так как ошибочная разметка нейтральных твитов негативно влияет на F-меру положительного класса и F-меру отрицательного класса.

$$F - macro = \frac{F_+ + F_-}{2}, \quad (1)$$

где F_+ – F-measure для положительного класса, $F\_$ – F-measure для отрицательного класса. F-measure для положительного класса вычисляется по формуле 2. Аналогично вычисляется F-measure для отрицательного класса.

$$F_+ = \frac{2 \times P_+ \times R_+}{P_+ + R_+}, \quad (2)$$

где P_+ и R_+ – Precision и Recall для положительного класса.

Дополнительно для двух классов тональности было рассчитано микро-усреднение F-меры. Микро усреднение получается в результате усреднения Precision и Recall для двух исследуемых классов:

$$F - micro = \frac{2 \times P \times R}{P + R}, \quad (3)$$

3.3 Участники и результаты

В задаче репутационной оценки твитов по тональности на SentiRuEval приняло участие 9 участников, которые предоставили 33 прогона своих систем. Результаты прогонов для телекоммуникационных компаний представлены в таблице 5, для банков – в таблице 6 [1, 16, 19, 20].

Baseline рассчитывается исходя из наиболее частотного (негативного в обоих случаях) класса из оценочных. Три лучших результата для каждой из коллекций выделены полужирным.

Таблица 5. Результаты участников на коллекции телекоммуникационных твитов

Run_id	Macro F	Micro F
Baseline	**0,1823**	**0,337**
1_01	0,3419	0,38
1_02	0,278	0,3201
1_03	0,2552	0,2944
2_A	**0,4882**	**0,5355**
2_B	**0,4829**	**0,5362**
2_C	0,0659	0,0741
3_02	**0,4804**	**0,5094**
4_1	0,467	0,506
5_2	0,1237	0,1226
6_1	0,1295	0,1906
8_1	0,3324	0,3463
8_2	0,3735	0,4068
8_3	0,3843	0,4283
9_1	0,3158	0,3331
9_2	0,2328	0,2626
9_3	0,3305	0,3371
9_4	0,331	0,3501
9_5	0,3527	0,3765
10_1	0,4477	0,5282

Таблица 6. Результаты систем участников на коллекции твитов о банках

Run_id	Macro F	Micro F
Baseline	**0,1267**	**0,2377**
1_01	0,2986	0,3226
1_02	0,2646	0,2862
1_03	0,2262	0,2592
2_A	0,3345	0,3641
2_B	**0,3354**	**0,3656**
2_C	0,024	0,0194
4_1	**0,3598**	**0,343**
5_1	0,1624	0,1615
5_2	0,2172	0,2141
6_1	0,1469	0,1721
8_1	0,3023	0,3024
8_2	0,3276	0,3432
8_3	0,3197	0,339
10_1	**0,352**	**0,337**

Опишем основные подходы к задаче. Классификатор участника 2 основан на методе SVM, в качестве признаков использовались нормализованные леммы и синтаксические связи (связь = главная лемма, зависимая лемма, тип связи).

Участник 3 использовал подход, основанный на правилах учета синтаксических отношений между тональными словами и целевыми объектами, без применения методов машинного обучения. Участник 4 применил метод максимальной энтропии со следующими признаками: n-граммы слов, символьные n-граммы и результаты тематического моделирования.

Участник 10 использовал метод машинного обучения SVM и следующие признаки: n-граммы слов (ngram_range=(1,4)), буквенные n-граммы (ngram_range=(1,4)), знаки пунктуации, наличие ссылки и ретвита, несколько признаков на основе составленного вручную эмоционально-окрашенного словаря и автоматически созданного лексикона на основе обучающей коллекции (для каждой лексической единицы подсчитан PMI по позитивной/негативной выборке).

Участник 1 выбирал для тестовых твитов из обучающей выборки наиболее близкие твиты, которые сравнивались по метрике Левенштейна (основанного на словах). Этот участник использовал инструмент Word2Vec с разной длиной для трех прогонов: 256 для первого, 1024 для второго и 4096 для третьего. Участник 8 использовал метод SVM со взвешенным словарем униграмм, в качестве весовой схемы была применена схема TF-IDF.

Кроме того, один из участников выполнил независимую экспертную разметку телекоммуникационных твитов и получил Макро-F = 0,703 и Микро-F = 0,749. Эти показатели можно считать максимально возможными показателями качества для автоматических классификаторов.

Как мы видим, лучшие результаты, достигнутые участниками, не высоки, что объясняется сложностью задачи классификации твитов в зависимости от их влияния на репутацию указанной компании. Также было замечено, что участники не воспользовались дополнительными наборами неразмеченных твитов, которые был разосланы вместе с размеченными коллекциями, или какими-либо другими дополнительными коллекциями.

В целом, лучшие результаты сравнимы с результатами RepLab-2012 (F-мера=0,41) [3]. Как указывают организаторы RepLab-2012 относительно низкий уровень результатов RepLab-2012 связан со сложностью задачи и ограниченным объемом обучающих коллекций.

4 Анализ и сравнение результатов участников на коллекциях двух предметных областей

Мы проанализировали полученные участниками результаты с нескольких точек зрения. Во-первых, хотелось разобраться, почему максимальные результаты в двух областях значительно различаются. Во-вторых, были извлечены и классифицированы твиты, которые оказались сложными для подавляющего большинства участников. В-третьих, мы изучили, были ли подходы реально ориентированы на анализ заданных сущностей, или решали общую задачу классификации твитов по тональности. Результаты этого анализа будут рассмотрены в следующих подразделах.

4.1 Различие результатов участников в двух предметных областях

Из представленных ранее результатов можно видеть, что максимальные достигнутые результаты участников анализа твитов о банках и телекоммуникационных компаниях значительно различаются (0,36 vs. 0,488 MacroF). Это может быть частично объяснено различиями в уровне представленности в твитах негативного класса, что приводит к существенной разнице в результатах даже для простого классификатора, который всем твитам ставит негативный класс (0,1267 vs. 0,1823 MacroF). Количество негативных твитов в области телекоммуникационных компаний было значительно больше.

Для дальнейшего изучения различий было решено сравнить вероятностные распределения слов в обучающей и тестовой коллекциях для каждой предметной области. Для того, чтобы определить вероятности слов в каждой коллекции и избежать нулевых вероятностей в связи с отсутствием слова в той или иной коллекции, мы применили так называемое аддитивное сглаживание [7], при котором частоты упоминания всех слов в обучающей и тестовой коллекции искусственно повышаются на 1. Таким образом, слова, которые упоминались только в одной коллекции, в другой получали частоту 1. Вероятности слов в коллекциях были вычислены по формуле 4:

$$P(w) = \frac{x_i + 1}{N + d}, (i = 1, \ldots, d). \quad (4)$$

где x_i- это частота упоминания слова w в коллекции, N – количество слов в коллекции, d – количество разных слов в коллекции.

После этого, для подсчета разницы в распределениях слов в обучающей и тестовой коллекции в обеих областях была вычислена дивергенция Кульбака-Лейблера (формула 5), которая представляет собой несимметричную меру удаленности друг от друга двух вероятностных распределений (Таблица 7).

$$D_{KL} = \sum_i test_i \times \ln \frac{test_i}{train_i}. \quad (5)$$

Также мы применили симметричный вариант сравнения вероятностных распределений - так называемую меру Йенсена-Шеннона (формула 6). Результаты распределения слов в обучающей и тестовой коллекции согласно мере Йенсена-Шеннона представлены в таблице 8.

$$D_{JS} = \frac{1}{2}(\sum_i test_i \times \ln \frac{test_i}{M}) + \frac{1}{2}(\sum_i train_i \times \ln \frac{train_i}{M}), \quad (6)$$

$$M = \frac{1}{2}(test + train)$$

Таблица 7. Значения Kullback–Leibler-дивергенции распределения слов в обучающих и тестовых коллекциях.

	Полная дивергенция	Тональные	Позитивные	Негативные
Банки	0,465	**0,505**	0,397	**0,561**
ТКК	0,317	0,287	0,323	0,284

Таблица 8. Значения дивергенции Jensen-Shannon распределения слов в обучающих и тестовых коллекциях.

	Полная дивергенция	Тональные	Позитивные	Негативные
Банки	0,084	**0,123**	0,092	**0,139**
ТКК	0,066	0,066	0,071	0,067

По таблицам 7, 8 можно видеть, что различие в обучающей и тестовой коллекции намного больше в банковской области, и это различие максимально для негативных твитов.

По нашему мнению, такое различие возникает потому, что темы твитов, влияющих на репутацию, значительно зависят от происходящих позитивных или негативных событий с участием целевого объекта, и частично такие события невозможно предсказать заранее**.**

Наше тестирование продемонстрировало это в полной мере. Обучающие коллекции в обеих областях относились к периоду времени июль-август 2014, когда уже начались боевые действия на Украине. Поэтому негативные банковские твиты часто упоминают санкции против российских банков и их последствия. Эти события также частично коснулись и телекоммуникационных компаний, в частности из-за проблем со связью в Крыму.

Для тестирования использовались твиты, относящиеся к декабрю 2013 - февралю 2014, когда события на Украине уже начались, но еще не развернулись в полной мере. Естественно, что в тестовых коллекциях отсутствуют негативные упоминания санкций и проблем со связью в Крыму.

4.2 Анализ сложных твитов

Следующий этап нашего исследования состоял в извлечении и анализе твитов, которые оказались наиболее сложными для участников. Для этого из коллекции были выделены твиты, в которых ошиблось подавляющее большинство участников. Были извлечены: 71 твит в банковской области, в которых ошиблись все участники, и 85 твитов в области телекоммуникаций, в которых правильный ответ дали не более двух из всех участников.

В результате анализа мы разделили эти твиты на две группы, в каждой можно выделить подгруппы, для каждой из подгрупп можно предложить свои методы, которые могут улучшить обработку этих твитов.

К **первой группе** можно отнести твиты, которые неправильно классифицируются участниками из-за ограниченного объема обучающей коллекции, которая не содержала соответствующие обучающие примеры.

К первой подгруппе этой группы (**подгруппа 1.1**.) неправильно классифицированных твитов относятся твиты, содержащие очевидные оценочные слова или выражения, которых просто не оказалось в обучающей выборке, например:

*Самый **безалаберный** банк по отношению к клиентам - Сбербанк*

*Сбербанк **навязывает** кредитную карту*

***Люблю** сбербанк*

Это означает, что в машинном обучении полезно использовать общеязыковые словари оценочных слов и выражений. Однако пока для русского языка такие словари не опубликованы. Опубликован лишь автоматически извлеченный список оценочных слов, которые не размечены по тональности [6]. Также полезно иметь словарь оценочных жаргонизмов, которые часто используются в Твиттере и других социальных сетях *(жесть, рулед*).

Удачное применение ручных словарей оценочной лексики, а также заранее построенных автоматических словарей оценочной лексики для анализа твитов на английском языке описано в работе [12], в которой представлен метод, показавший лучшие результаты в классификации твитов по тональности в рамках соревнования SemEval 2013. В данном подходе используются три ручных словаря, включая известный словарь MPQA [21] и NRC Word-Emotion Association lexicon, который был получен на основе технологии краудсорсинга и включает слова и выражения, которые ассоциируются у людей с разной тональностью и эмоциями [13].

Вторая подгруппа этой группы твитов (**подгруппа 1.2**) содержит слова, выражающие известные негативные или позитивные ситуации, например, *кража* или *празднование*, которые отсутствовали в обучающей коллекции. Эти слова не являются оценочными, они не выражают

никакого мнения, но имеют позитивные или негативные ассоциации, так называемые коннотации [8]. Например,

*В столице произошло дерзкое **ограбление** Сбербанка*

*В Ижевске нашли **свалку** документов с личными данными клиентов Сбербанка*

*В Башкирии инкассатор "Сбербанка" **смертельно ранен** в голову*

Для решения проблемы обработки таких слов, отсутствующих в обучающей выборке, необходим общий словарь слов с коннотациями, поскольку появление этих слов в твитах в связи с целевой компанией может быть сильным сигналом, что твит соответствующим образом влияет на репутацию. Для английского языка эти слова включены в вышеупомянутый словарь NRC Word-Emotion Association lexicon, поскольку вызывают у людей соответствующие эмоции.

Третья подгруппа **(группа 1.3)** проблемных твитов содержит отсылки к текущим позитивным или негативным новостям. Частично такие события и их влияние на репутацию конкретной компании трудно предсказать, их упоминания естественно отсутствуют в обучающей выборке, если она относится к другому времени, например,

У нас в городе в одном из филиалов Сбербанка сделали детский уголок - для нашего города это вновинку - чтобы детишк..

Сбербанк с 2016 года начнет сокращать банкоматную сеть

В этом случае полезным является параллельный анализ текущего потока новостей, целью которого является извлечение корреляций между словами в твитах и выражаемыми эмоциями в новостях.

Вторая группа неправильно классифицированных твитов содержит твиты, которые на самом деле являются сложными для анализа, поскольку они включают оценочные слова разной полярности, иногда относящиеся к разным упоминаемым компаниям, а также иронию. Таким образом, сложные твиты можно подразделить на следующие подгруппы:

Подгруппа 2.1 включает твиты, в которых упоминаются два объекта, с разным к ним отношением. Таких твитов особенно много оказалось в области телекоммуникаций (25 из 85). Например:

@ru_mts МТС не работает! Вечно вне зоны доступа. Связь постоянно прерывается. Всю семью переводим на Билайн.

Здесь также можно выделить дополнительную подгруппу, связанную со сравнением объекта самого с собой в прошлом. Таких твитов оказалось по одному в каждой из областей, например,

@Beeline_RUS Спасибо, раньше просто всегда приходили и было очень удобно, а сейчас не понятно когда кончается трафик.

Вторая подгруппа сложных твитов (**подгруппа 2.2**) содержит сложно сформулированные твиты с несколькими оценочными словами разной тональности:

Новый интерфейс сбербанк онл@йн: Внимательно соотносим иконки и подписи, впадаем в когнитивный диссонанс.

Наконец, последнюю подгруппу твитов (**подгруппа 2.3**.) составляют иронические твиты. Такие твиты характеризуются тем, что имеют превалирующее количество позитивных слов или равное количество позитивных и негативных слов, однако фактически эти твиты явно негативные, например:

Сбербанк - самая крупная сеть неработающих банкоматов в России

Сбербанк - "мы идиоты и гордимся этим!" / / Не важно что написано в заявлении клиента, мы истолкуем все по-своему.

Количественная оценка состава этих групп показала, что до 30% трудных для участников твитов в банковской области связано со словарными ресурсами (общие оценочные словари, словари коннотаций, разговорной лексики, анализ новостей). Для телекоммуникационных компаний доля таких твитов ниже - около 15%, здесь значительна доля сравнений разных компаниями с противоположными оценками. Это означает, что появление разных типов оценочных словарей, а также их внедрение в алгоритмы машинного обучения может значительно улучшить качество классификации твитов в рамках репутационного анализа.

4.3 Являются ли системы участников, ориентированными на анализ тональности объектов

На последнем шаге исследования было проверено, решали ли участники на самом деле задачу анализа тональности, ориентированную на конкретную сущность, или они классифицировали твит как целое. Для этого были извлечены твиты, которые упоминали более одной сущности: 58 твитов в банковской области (15 твитов с разной полярностью для разных банков) и 232 твита в телекоммуникациях (71 твит с разной полярностью для разных компаний).

Качество анализа твитов с несколькими объектами оказалось существенно ниже, чем качество анализа всех твитов. Лучший результат для телекоммуникационных компаний составил 0,3463 Macro F (0,4882 для всех твитов – Таблица 5), для банков 0,3095 Macro F (0,3598 для всех твитов – Таблица 6). В области телекоммуникаций падение результатов значительно больше, что объясняется большим количеством твитов с разными оценками для упоминаемых компаний, а также их значительным количеством среди сложных твитов, в которых ошиблись подавляющее большинство участников (см. раздел 4.2.).

Используя эти твиты, было проверено, классифицируют ли их участники всегда одинаково относительно упоминаемых организаций или по-разному. Мы выяснили, что только трое участников из девяти рассматривали задачу как ориентированную, на конкретную сущность. Остальные участники классифицировали твит в целом, т.е. упомянутые компании всегда получали одинаковые тональности.

В этой ситуации возникает вопрос, какая стратегия была более эффективной с точки зрения полученного результата: пытаться разобраться с упомянутыми организациями или ставить им одинаковую тональность. Анализ качества работы разных участников на твитах, упоминающих разные организации, показал, что лучшие результаты получили участники, которые классифицировали твит в в целом. Таким образом, можно заключить, что пока возможности систем к анализу тональности твитов по отношению к заданным сущностям остаются очень ограниченными.

5 Заключение

В данной статье мы описали задачу анализа репутации по твитам на русском языке, которая была поставлена в рамках открытого тестирования систем анализа тональности на русском языке SentiRuEval. Целью такой обработки твитов является классификация твитов в соответствии с их влиянием на репутацию упоминаемой в твите компании. Твиты, затрагивающие репутацию компании, могут выражать как личное мнение автора, так и приводить некоторый позитивный или негативный факт об этой компании.

В тестировании приняли участие 9 участников, которые в основном применяли методы машинного обучения, среди которых наиболее популярным был метод SVM.

Мы проанализировали результаты участников и выяснили, что:

- уровень лучших достигнутых результатов в классификации твитов в некоторой области коррелирует с различием между обучающим и тестовым множеством: чем больше такое различие, тем хуже достигнутые результаты. Характерной особенностью задачи анализа репутации является то, что сильное различие между обучающей и тестовой выборкой может возникнуть в любой момент из-за каких-то важных происходящих в мире событий, которые невозможно отразить в обучащей выборке,

- в связи с ограниченным размером обучающей выборки и динамичностью твитов, значительный рост качества обработки твитов может базироваться на интеграции в методы машинного обучения дополнительных ресурсов, включая нескольких видов словарей: словарь общезначимых слов оценочных слов, словарь слов с положительными и отрицательными коннтоациями,

- большинство участников не решало задачу классификации твитов, ориентированную на конкретную упоминаемую сущность, а решали общую задачу классификации твитов; объектно-ориентированным подходам не удалось достичь лучших результатов на твитах, упоминающих несколько компаний.

Все подготовленные коллекции доступны для исследовательских целей по адресу http://goo.gl/qHeAVo.

Литература

[1] Adaskina Yu. V., Panicheva P. V., Popov A. M. Syntax-based Sentiment Analysis of Tweets in Russian // In Proceedings of International Conference Dialog. – 2015. – T. 2. – C. 25-35.

[2] Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau, R. Sentiment analysis of twitter data //Proceedings of the Workshop on Languages in Social Media. – Association for Computational Linguistics, 2011. – C. 30-38.

[3] Amigo E., Corujo A., Gonzalo J., Meij E., de Rijke M. Overview of RepLab 2012: Evaluating Online Reputation Management Systems // CLEF 2012 Evaluation Labs and Workshop Notebook Papers. – 2012.

[4] Amigo E., Albornoz J.C., Chugur I., Corujo A., Gonzalo J., Martin T., Meij E., de Rijke M, Spina D. Overview of RepLab 2013: Evaluating online reputation monitoring systems //Information Access Evaluation. Multilinguality, Multimodality, and Visualization. – Springer Berlin Heidelberg, 2013. – C. 333-352.

[5] Chetviorkin I., Braslavskiy P., Loukachevich N. Sentiment analysis track at romip 2011 //Dialog. – 2012.

[6] Chetviorkin I., Loukachevitch N. V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain //COLING. – 2012. – C. 593-610.

[7] Chen S. F., Goodman J. An empirical study of smoothing techniques for language modeling //Proceedings of the 34th annual meeting on Association for Computational Linguistics. – Association for Computational Linguistics, 1996. – C. 310-318.

[8] Feng, S., Kang, J.S., Kuznetsova, P., Choi, Y.: Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning //ACL (1). – 2013. – C. 1774-1784.

[9] Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. Exploiting emoticons in sentiment analysis //Proceedings of the 28th Annual ACM Symposium on Applied Computing. – ACM, 2013. – C. 703-710.

[10] Kouloumpis E., Wilson T., Moore J. Twitter sentiment analysis: The good the bad and the omg! //ICWSM. – 2011. – T. 11. – C. 538-541.

[11] Loukachevitch, N., Blinov, P., Kotelnikov, E., Rubtsova Yu, V., Ivanov, V. V., Tutubalina, E.

SentiRuEval: testing object-oriented sentiment analysis systems in Russian. // In Proceedings of International Conference Dialog. – 2015. – T. 2. – C. 12-24.

[12] Mohammad S. M., Kiritchenko S., Zhu X. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets //Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR'13). – 2013.

[13] Mohammad S. M., Turney P. D. Crowdsourcing a word–emotion association lexicon //Computational Intelligence. – 2013. – T. 29. – №. 3. – C. 436-465.

[14] Nakov P. et al. Semeval-2013 task 2: Sentiment analysis in twitter // Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013), – 2013. – C. 312–320.

[15] Pak A., Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining //LREC. – 2010. – T. 10. – C. 1320-1326.

[16] Polyakov P. Yu., Kalinina M. V., Pleshko V. V., Automatic Object-oriented Sentiment Analysis by Means of Semantic Templates and Sentiment Lexicon Dictionaries // In Proceedings of International Conference Dialog. – 2015. – T. 2. – C. 68-76.

[17] Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification //Proceedings of the ACL Student Research Workshop. – Association for Computational Linguistics, 2005. – C. 43-48.

[18] Rosenthal S., Ritter A., Nakov P., Stoyanov V. SemEval-2014 Task 9: Sentiment Analysis in Twitter // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). – 2014. – C. 73–80.

[19] Tutubalina E. V., Zagulova M. A., Ivanov V. V., Malykh V. A., A Supervised Approach for SentiRuEval Task on Sentiment Analysis of Tweets about Telecom and Financial Companies // In Proceedings of International Conference Dialog. – 2015. – T. 2. – C. 89-99.

[20] Vasilyev V. G., Denisenko A. A., Solovyev D. A., Aspect Extraction and Twitter Sentiment Classification by Fragment Rules // In Proceedings of International Conference Dialog. – 2015. – T. 2. – C. 100–110.

[21] Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis //Proceedings of the conference on human language technology and empirical methods in natural language processing. – Association for Computational Linguistics, 2005. – C. 347-354.

[22] Рубцова Ю. В. Разработка и исследование предметно независимого классификатора текстов по тональности //Труды СПИИРАН. – 2014. – Т. 5. – №. 36. – С. 59-77.

Entity-Oriented Sentiment Analysis of Tweets: Results and Problems

Natalia Loukachevitch, Yuliya Rubtsova

This paper summarizes the results of the reputation-oriented Twitter task, which was held as a part of SentiRuEval evaluation of Russian sentiment-analysis systems. The tweets in two domains: telecom companies and banks – were included in the evaluation. The task was to determine if an author of a tweet has a positive or negative attitude to a company mentioned in the message. The main issue of this paper is to analyze the current state and problems of approaches applied by the participants.