УДК 004.912

ИСПОЛЬЗОВАНИЕ СИНТАКСИСА ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТВИТОВ НА РУССКОМ ЯЗЫКЕ

Ю.В. Адаскина[1], П.В. Паничева[2], А.М. Попов[3]

ООО «InfoQubes», Санкт-Петербургский государственный университет
[1]adaskina@gmail.com, [2]p.panicheva@spbu.ru, [3]hedgeonline@gmail.com

Аннотация

Представлен подход к решению задачи анализа тональности в рамках тестирования SentiRuEval – открытого соревнования систем анализа тональности на русском языке. Описанный алгоритм был применен в дорожке по анализу тональности твитов о банках и телекоммуникационных компаниях. Для этих данных была разработана и оценена классификация на три класса: положительный, отрицательный и нейтральный.

Для решения поставленной задачи использовались различные алгоритмы машинного обучения. Признаками для классификатора являлись лингвистические данные, полученные из текста с помощью разработанного нами морфо-синтаксического анализатора. Нормализованные слова, а также синтаксические связи, оказались решающими признаками для достижения наилучшего результата, который был получен с помощью статистического алгоритма опорных векторов.

Оценка, проведенная организаторами конкурса, выявила высокое качество предложенного подхода, который занял первую строчку по трем из четырех мерам качества.

Ключевые слова*: анализ тональности, синтаксические связи, русский язык, статистические методы, классификация текстов.*

ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

Будучи одним из наиболее изученных направлений прикладной лингвистики, анализ тональности остается одной из самых востребованных задач как для теоретических исследований, так и для бизнес-приложений. Анализ тональности применялся на разных уровнях, начиная от документа целиком и постепенно сужаясь к отдельному предложению. Тональность на уровне предложений распознается, исходя из предположения, что одним предложением в языке обычно выражается одно мнение. В последнее время основной фокус сдвинулся на более мелкие единицы внутри предложения, в сферу анализа попадают случаи, когда в предложении есть оценка нескольких сходных объектов (например, нескольких брендов), а также случаи оценки разных аспектов одного и того же объекта (например, таких параметров товара, как прочность, цена, дизайн и т. п.). Основные усилия лингвистов сегодня направлены на создание и развитие высокоточных автоматических методов анализа тональности, что в свою очередь поднимает вопрос о методах оценки качества таких систем. Многие независимые организации проводят тестирования различных методов автоматического анализа естественного языка, самым влиятельным среди российских можно считать соревнование Dialogue Evaluation, проводимое в рамках международной конференции по компьютерной лингвистике «Диалог». В 2015 году состоялось третье тестирование систем анализа тональности SentiRuEval; первые два обсуждаются в [1, 2]. В этом году оценивалось в том числе предметно-ориентированное распознавание тональности на различных типах данных (см. [3]).

В данной статье описан подход к заданиям SentiRuEval, а именно, в двух дорожках, посвященных объектно-ориентированному анализу мнений в твитах о банках и телекоммуникационных компаниях. От участников этих дорожек требовалось произвести трехклассовую классификацию тестовых данных, разделив их на негативный, позитивный и нейтральный классы.

Полученные результаты основаны на классификаторе SVM, хотя предварительные эксперименты показали незначительные различия между ним и классификатором Naïve Bayes. В качестве признаков для обучения использовались нормализованные формы слов в комбинации с синтаксическими связями, где под последними понимаются тройки из нормальных форм двух связанных слов и типа

синтаксического отношения между ними. Для всех предварительных экспериментов добавление синтаксических связей существенно улучшало результаты классификации; в оценке, использованной организаторами соревнования, влияние синтаксиса было чуть менее значительно. Тем не менее, результаты соревнования подтвердили эффективность разработанного метода: по трем из четырех метрик качества созданной системе удалось занять первое место среди участников.

ПРЕДПОСЫЛКИ ИССЛЕДОВАНИЯ

Одним из наиболее распространенных подходов к решению задач анализа тональности является применение машинного обучения. Надо отметить, что анализ тональности хорошо ложится на такую стандартную задачу, часто решаемую с привлечением машинного обучения, как классификация, где документ классифицируется по трем классам тональности: положительному, отрицательному и нейтральному. С одной стороны, машинное обучение, как вероятностный метод, позволяет свести к минимуму лингвистическую составляющую, сохраняя при этом относительно высокие показатели качества [4]. С другой стороны, необходимым условием применения любых алгоритмов машинного обучения является потребность в параметризации обучающих и анализируемых данных. Обычно текст параметризуется как «мешок слов» [4], реже – как n-грамм. В некоторых случаях такой подход оправдывает себя, однако для языков с развитой морфологией число таких признаков может быть очень велико, а их абсолютная встречаемость, наоборот, очень низкой. Это будет препятствовать любым попыткам обобщения, которые проводят алгоритмы машинного обучения. Зачастую, чтобы, с одной стороны, уменьшить число признаков, а с другой, — повысить их встречаемость, применяются различные лингвистические приемы, например, приведение слов к нормальной форме [5], добавление в признаки семантической [6] или синтаксической [7] информации и т. д.

Одним из наиболее известных исследований по анализу тональности с использованием синтаксической информации является работа [8]. В ней описано применение SVM-классификатора для анализа тональности с использованием различных признаков, в том числе, лемматизации и синтаксических поддеревьев. В работе [9] описано использование синтаксической информации в системе ана-

лиза тональности текстов на русском языке. Авторы применяют подход, основанный исключительно на правилах, при котором текст рассматривается не как «мешок слов», а как набор синтаксических деревьев. Данный метод позволяет проводить так называемый «объектно-ориентированный» анализ тональности, когда мнение высказывается относительно какого-то объекта в тексте. В работе [10] используются различные алгоритмы машинного обучения (SVM, Naïve Bayes) для анализа тональности; исследуется влияние лемматизации и применения другого вида лингвистических ресурсов, словарей синонимов, на качество анализа тональности. В частности, сделан вывод о том, что для русского языка лемматизация и словари синонимов оказывают положительное влияние на качество.

Таким образом, список наиболее часто используемых признаков для машинного обучения выглядит следующим образом:

- словоформы (униграммы);
- леммы (нормализованные униграммы);
- n-граммы;
- нормализованные n-граммы;
- бинарная встречаемость слов;
- синтаксические связи/поддеревья.

Следует отметить, что использование синтаксических признаков само по себе подразумевает сложную и длительную процедуру синтаксического анализа. Однако исследования, в которых применяются синтаксические признаки, показывают, что синтаксическая информация позволяет существенно повысить как полноту, так и точность (см., например, [11–13]) алгоритмов классификации текстов. Так, в работе [14], посвященной задаче автоматического извлечения контекста, синтаксические признаки оказывают решающий вклад в достижение F-меры в 70%.

В заключение обзора предпосылок перечислим несколько работ, посвященных анализу тональности на материале твитов — сообщений, представляющих собой отдельный подтип данных [15–19]. Особенности Твиттера — ограничение на длину сообщения и ориентация на жанр мгновенных реакций на происходящее – сказываются на особенностях методов их анализа.

В задачи данного исследования входил поиск ответа на вопрос, как применение синтаксической информации в качестве признаков для машинного обучения повлияет на качество анализа тональности текстов на русском языке.

ДАННЫЕ И ПОСТАНОВКА ЗАДАЧИ

Наша компания принимала участие в дорожках по анализу тональности в твитах, посвященных банкам и телекоммуникационным компаниям. Детальное описание заданий представлено в [3]. Организаторы предоставили участникам обучающие и тестовые выборки размером около 10 тысяч текстов каждая; обе текстовые коллекции, в свою очередь, делились примерно пополам на твиты о двух типах брендов (банки и телекоммуникационные компании). Обучающие данные были вручную размечены экспертами SentiRuEval, каждому тексту было проставлено значение тональности или помечено его отсутствие. Твиты, для которых не было согласия в оценках хотя бы у двух из трех экспертов, исключались из корпуса, в результате чего размер обучающей коллекции для банков составил 4549 документов, для телекоммуникационных компаний – 3845 документов. Тестовый корпус был размечен нейтральными значениями для каждой из компаний, упомянутых в твите, от участников требовалось заменить эти значения на положительные или отрицательные, или же сохранить нейтральное.

АЛГОРИТМ

В основе метода лежит машинное обучение с использованием различных признаков, полученных нашим лингвистическим модулем. Остановимся на этих аспектах подробнее.

Модуль лингвистического анализа

Для анализа текстов использовался морфосинтаксический парсер InfoQubes, который ранее показал свою эффективность для решения задачи полуавтоматического пополнения лексических классов (см. [20]). Эта платформа является коммерческой разработкой нашей компании. Анализатор состоит из нескольких модулей, среди них важно отметить модули: морфологического анализа; распознавания неизвестных слов и слов с опечатками; поверхностного синтаксиса; основного синтаксиса; пост-синтаксической обработки.

Морфологический модуль (модуль приведения слов к нормальным формам) основан на словоизменительном словаре А.А. Зализняка [21], этот модуль осуществляет лемматизацию и приписывает словоформам наборы значений грамматических категорий. Модуль распознавания неизвестных слов и слов с опечатками анализирует фрагменты текста, которые отсутствуют в морфологических словарях. На основе выделения суффиксов и приставок, а также степени схожести неизвестных слов со словами, имеющимися в словарях, модуль может приписать неизвестному слову грамматические значения. Такая возможность играет особую роль при работе с данными из социальных сетей, особенно короткими текстами Твиттера, которые зачастую пишутся в спешке, что увеличивает вероятность появления опечаток.

Модуль поверхностного синтаксиса собирает основные фразовые категории: имена существительные, имена прилагательные, глаголы и их зависимые. Кроме того, здесь реализованы некоторые вспомогательные функции, например, распознавание именованных сущностей; частично именно этот модуль проставляет маркер отрицания.

Синтаксический модуль представляет собой конечный автомат, который на вход получает текст, обработанный морфологически и поверхностно-синтаксически, а на выходе возвращает синтаксическое дерево. В качестве входной контекстно-свободной грамматики для парсера используется сложная система из 515 синтаксических правил. Обычно синтаксическое правило соединяет два слова или фразовые категории в категорию более высокого уровня и проставляет синтаксическое отношение. Таким образом, из грамматики непосредственных составляющих выводится структура зависимостей. В грамматике разрешены только бинарные связи, каждое синтаксическое отношение характеризуется исходным словом, целевым словом и типом связи между ними. В системе используется 16 синтаксических связей, одна из которых имеет 4 разновидности, поэтому в нашем признаковом пространстве рассматривается как 4 различных типа связи. В Таблице 1 представлены частоты встречаемости 19 типов синтаксических связей в обучающих корпусах дорожек по оценке тональности:

Таблица 1. Синтаксические связи, полученные системой на обучающем корпусе

Название	**Встречаемость в корпусе о телекоме**	**Встречаемость в корпусе о банках**
Argument:DirectObject	2778	2372
Argument:IndirectObject	5748	3585
Argument:PassiveSubject	291	232
Argument:Subject	3148	1805
Attribute	6814	6682
Auxiliary	578	208
Circumstance	3033	1211
Coordinate	1008	1698
Determiner	687	239
Genitive	3963	3355
Identity	2200	4937
Infinitive	772	465
Modifier	707	294
Phrasal	1519	959
Possessive	368	126
Preposition	6582	4554
Quantifier	501	605
Subordinate	226	77
Undefined	1050	1159

Модуль пост-синтаксической обработки анализирует собранные поддеревья и может по необходимости редактировать узлы или связи между ними. На этом этапе устанавливаются недостающие синтаксические связи, например, во фразовых категориях, которые были собраны модулем поверхностного синтаксиса. Кроме того, этот модуль проставляет маркер отрицания и некоторые семантические теги.

Настройка признаков машинного обучения

Эксперименты проводились с единичными леммами (униграммами), сочетаниями лемм (биграммами) и синтаксическими связями в качестве признаков для классификаторов на основе опорных векторов и наивного байесовского (см.

[22]), с использованием трехклассовой классификации (нейтральный, позитивный и негативный классы). В каждом из экспериментов были использованы правила нормализации морфологического модуля. Так как сообщения в Твиттере характеризуются ограниченной длиной, ожидалось, что построение полных деревьев синтаксического разбора будет затруднено. Поэтому в признаковое пространство были включены синтаксические связи как пары связанных слов и тип отношения между ними, иными словами, синтаксические связи представляются как тройки «главное слово – тип связи – зависимое слово». В качестве опциональных параметров также использовалось отрицание, проставляемое нашим морфологическим анализатором: маркер, который получает слово, связанное с одной из отрицательных единиц (в первую очередь, частица «не», кроме того, предлог «без», существительное «отсутствие» и др.). Кроме того, в качестве параметра опционально исключались слова, обозначающие один из исследуемых брендов, так как подразумевалось, что общая направленность на бренды может отрицательно повлиять на результаты. Все использованные признаков и опциональные параметры представлены в Таблице 2.

Таблица 2. Характеристики признаков

№	Пример признака	Тип признака	Опциональные параметры	Расшифровка	Комментарий
1	ВАРИАНТ	Лемма	Маркер отрицания не учитывается	Лемма *ВАРИАНТ*	Нормализованное слово
2	ВАРИАНТ|Argument|НЕТ|PassiveSubject	Синтаксическая связь	Маркер отрицания не учитывается	Связь «субъект в пассивной конструкции» *ВАРИАНТА НЕТ*	Определенный тип синтаксической связи между двумя словами (в данном случае с подтипом, так как связь «аргумент» имеет 4 разновидности)
3	ВАРИАНТ|Attribute|ЭТОТ|	Синтаксическая связь	Маркер отрицания не учитывается	Связь «атрибут» *ЭТОТ*	Определенный тип синтаксической связи между двумя словами

				ВАРИАНТ, отрицание отсутствует	
4	КРУТОЙ|ВАРИАНТ	Биграмма	Маркер отрицания не учитывается	Биграмма *КРУТОЙ ВАРИАНТ*	Два смежных слова
5	ДРУГОЙ|ВАРИАНТ	Биграмма	Маркер отрицания не учитывается	Биграмма *ДРУГОЙ ВАРИАНТ*	Два смежных слова
6	ВАРИАНТ|0	Лемма	Маркер отрицания учитывается	Лемма *ВАРИАНТ*, на обоих словах нет отрицания	Сочетание нормализованных слов с информацией об отрицании, в данном случае отрицание отсутствует
7	ВАРИАНТ|1	Лемма	Маркер отрицания учитывается	Лемма *ВАРИАНТ*, слово с отрицанием	Сочетание нормализованных слов с информацией об отрицании, в данном случае отрицание присутствует на одном из слов
8	ВАРИАНТ|1|Argument|НЕТ|0|PassiveSubject	Синтаксическая связь	Маркер отрицания учитывается	Связь «субъект в пассивной конструкции» *ВАРИАНТА НЕТ*, слово *ВАРИАНТ* с отрицанием	Сочетание синтаксической связи с информацией об отрицании, в данном случае отрицание присутствует на одном из слов
9	ВАРИАНТ|0|Attribute|ЭТОТ|0|	Синтаксическая связь	Маркер отрицания учитывается	Связь «атрибут» *ЭТОТ ВАРИАНТ*, на обоих словах нет отрицания	Сочетание синтаксической связи с информацией об отрицании, в данном случае отрицание отсутствует

10	КРУТОЙ|0|ВАРИАНТ|0	Биграмма	Маркер отрицания учитывается	Биграмма *КРУТОЙ ВАРИАНТ*, на обоих словах нет отрицания	Сочетание биграммы с информацией об отрицании, в данном случае отрицание отсутствует
11	ДРУГОЙ|0|ВАРИАНТ|1	Биграмма	Маркер отрицания учитывается	Биграмма *ДРУГОЙ ВАРИАНТ*, слово *ВАРИАНТ* с отрицанием	Сочетание биграммы с информацией об отрицании, в данном случае отрицание присутствует на одном из слов

Отметим также, что организаторы ставили перед участниками задачу связывать оценку, содержащуюся в твите, с брендом, к которому она относится. Были проанализированы документы обучающего корпуса, в которых содержатся несовпадающие значения тональности, и выявлено крайне малое их количество: менее 1% для корпусов обеих тематик. Поэтому было принято решение пренебречь этими документам и упростить модель данных, распространяя найденную в документе тональность на все бренды, содержащиеся в тексте.

РЕЗУЛЬТАТЫ ПРЕДВАРИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Предварительные эксперименты были проведены с применением десятикратной кросс-валидации на обучающей текстовой коллекции. Описанный выше алгоритм анализа тональности был совмещен с алгоритмом извлечения названий брендов, основанном на правилах. Для того чтобы оценить результаты, каждому документу были сопоставлены его идентификационный номер, идентификатор бренда и значение тональности. На основании этой информации и предоставленной разметки была подсчитана общая Полнота, Точность и F1-мера. В расчетах учитывался нейтральный класс, а также проверялось качество определения бренда, что отличает полученные нами метрики от метрик, использованных организаторами. Оценки, полученные на основе предварительного эксперимента, представлены в следующих таблицах, наивысший результат выделен жирным

шрифтом. Таблица 3 относится к корпусу твитов о телекоммуникационных компаниях, Таблица 4 – о банках.

Таблица 3. Предварительные результаты для данных «Телеком», SVM

Признаки	**Опциональные параметры эксперимента**		**Оценки**		
	Маркер отрицания	**Удаление названия бренда**	**Точность**	**Полнота**	**F1-мера**
Леммы	–	–	0,7464	0,7482	0,7473
	+	–	0,7549	0,7567	0,7558
	–	+	0,7554	0,7571	0,7563
	+	+	0,7608	0,7625	0,7616
Синтаксические связи	–	–	0,7275	0,5567	0,6308
	+	–	0,7228	0,5532	0,6267
	–	+	0,7196	0,5470	0,6216
	+	+	0,7215	0,5484	0,6231
Леммы + синтаксические связи	–	–	**0,7715**	**0,7734**	**0,7725**
	+	–	0,7692	0,7710	0,7701
	–	+	0,7675	0,7692	0,7684
	+	+	0,7632	0,7648	0,7640
Леммы + синтаксические связи, χ2 распределение для 5000 лучших признаков	–	–	0,5865	0,5879	0,5872
Биграммы	–	–	0,7242	0,7077	0,7158
Биграммы + связи	–	–	0,7204	0,7220	0,7212
Биграммы + леммы	–	–	0,7650	0,7668	0,7659
Биграммы + леммы + синтаксические связи	–	–	0,7684	0,7702	0,7693

Таблица 4. Предварительные результаты для данных «Банки», SVM

Признаки	**Опциональные параметры эксперимента**		**Оценки**		
	Маркер отрицания	**Удаление названия бренда**	**Точность**	**Полнота**	**F1-мера**
Леммы	–	–	0,9046	0,9061	0,9053
	+	–	0,9021	0,9036	0,9029
	–	+	0,9073	0,9087	0,9080
	+	+	0,9032	0,9046	0,9039
Синтаксические связи	–	–	0,9040	0,8184	0,8591
	+	–	0,9080	0,8220	0,8628
	–	+	0,9040	0,8171	0,8583
	+	+	0,9066	0,8194	0,8608
Леммы + синтаксические связи	–	–	0,9059	0,9074	0,9066
	+	–	0,9047	0,9062	0,9055
	–	+	0,9083	0,9097	0,9090
	+	+	**0,9095**	**0,9108**	**0,9101**
Биграммы	–	–	0,8968	0,8949	0,8959
Биграммы + связи	–	–	0,8957	0,8971	0,8964
Биграммы + леммы	–	–	0,9021	0,9036	0,9029
Биграммы + леммы + синтаксические связи	–	–	0,9026	0,9041	0,9033
Леммы + синтаксические связи, $\chi 2$ распределение для 5000 лучших признаков	–	–	0,8257	0,8269	0,8263

Предварительные результаты показали, что комбинация лемм и синтаксических связей обеспечивает наилучшие результаты для обоих тестовых корпусов, а добавление отрицания и исключение брендов не оказывают существенного влияния на результат. Этот результат подтверждает исходную гипотезу, что синтаксические связи должны улучшить показатели. Биграммы и леммы показывают почти такие же высокие результаты, как леммы и связи. Наивный байесовский классификатор подтвердил эти тенденции с небольшим понижением абсолютных

значений показателей. Также был проведен эксперимент с исключением некоторых низкочастотных признаков, с применением алгоритма отбора признаков (feature selection), но это привело к неудовлетворительным результатам. В таблицы результатов выше были включены значения, полученные при помощи отбора признаков, и можно увидеть значительное падение показателей. Кроме того, применение меры TF-IDF также существенно ухудшило результаты. Представляется, что обучающие данные слишком рассеяны, чтобы на них могли сработать алгоритмы отбора признаков, они, возможно, были бы полезны на большем обучающем корпусе, где была бы выше частотность каждого отдельного признака.

РЕЗУЛЬТАТЫ СОРЕВНОВАНИЯ

Организаторы предоставляли участникам право отправить результаты нескольких прогонов, поэтому нами была выбрана SVM-классификация на основе биграмм и лемм в комбинации с синтаксическими связями. Также из параметрической модели опционально удалялись названия брендов. Для более полного анализа предложенных алгоритмов также был проведен внеконкурсный прогон SVM-классификатора на леммах. Таблица ниже представляет собой несколько модифицированную таблицу лучших результаты участников из обзорной статьи организаторов [3]. В нее добавлены результаты всех наших прогонов, их идентификаторы заменены на названия признаков соответствующих экспериментов в правой колонке. Номера других участников оставлены без изменения. Жирным шрифтом выделен лучший результат, курсивом – наш внеконкурсный прогон. В качестве оценочной метрики организаторы использовали две разновидности F-меры: F-микро и F-макро (подробнее см. [3]). Как видно из таблицы, на основе предложенных алгоритмов были получены лучшие результаты по трем метрикам качества из четырех.

Таблица 5. Результаты соревнования

Область	Мера	Базовый уровень	Результат	Идентификатор участника
Телеком	Macro F	0,182	**0,488**	**леммы+связи**
			0,483	леммы+связи, бренды удалены
			0,480	3
			…	…
			0,469	*леммы*
			0,465	*леммы, бренды удалены*
	Micro F	0,337	**0,536**	**леммы+связи**
			0,536	леммы+связи, бренды удалены
			0,528	10
			…	
			0,512	*леммы*
			0,514	*леммы, бренды удалены*
Банки	Macro F	0,127	**0,360**	**4**
			0,352	10
			0,345	*леммы*
			0,345	леммы, бренды удалены
			0,343	леммы+связи, бренды удалены
	Micro F	0,238	**0,366**	**леммы+связи, бренды удалены**
			0,364	леммы+связи
			0,363	*леммы*
			0,362	*леммы, бренды удалены*
			0,343	8

Результаты организаторов оценки существенно отличаются от наших предварительных результатов, что объясняется различными подходами к оценке: нами использовалась только одна из F-мер (F-микро), а также организаторы исключили из подсчета нейтральный класс документов.

Эти результаты только отчасти соотносятся с нашими предварительными результатами и нашей исходной гипотезой: на данных о телекоммуникационных компаниях леммы в комбинации с синтаксическими связями работают лучше,

чем одни леммы приблизительно на 2% в микро- и макро- F-мерах. На корпусе о банках результаты неубедительны: добавление синтаксических связей улучшает F-микро на 0,25%, но ухудшает F-макро примерно на 0,2%. В наших предварительных экспериментах результаты на корпусе о банках были выше, чем на корпусе о телекоме, а результат соревнования говорит об обратном. Показатели для банков оказались ниже показателей для телекома у всех участников.

Принятое решение не учитывать документы с упоминанием несовпадающих значений тональности оказалось удачным: в тестовом корпусе их количество тоже было минимальным.

Другие высокие результаты были получены участниками, использовавшими алгоритм, основанный на правилах, классификаторы методом максимальной энтропии и SVM на различных наборах признаков, главным образом, на словах и буквенных n-граммах.

ЗАКЛЮЧЕНИЕ

К задаче анализа тональности на двух предметных областях был применен статистический алгоритм с использованием синтаксических связей, что позволило получить высокие результаты, опередившие другие методы. Использовалась трехклассовая классификация, показатели классификации на леммах в качестве признаков были улучшены за счет добавления синтаксической информации. В некоторых случаях улучшения не происходило за счет высокой разреженности данных, этот вопрос требует дополнительного анализа и исследования. Признаки для классификации получены при помощи нашего морфосинтаксического парсера, уже продемонстрировавшего свою эффективность на другой задаче, связанной с семантикой (см. [20]). Так как число документов, включавших несовпадающие значения тональности, было очень маленьким, модель данных была упрощена до представления «один документ — одна оценка тональности». Для разреженного корпуса небольшого размера классификатор SVM представляется оптимальным методом. Добавление отрицания или удаление брендов не оказывает существенного влияния на результат: можно предположить, что вся информация, которую могут привнести показатели отрицания, уже содержится в синтаксических поддеревьях.

В качестве направлений дальнейших исследований можно отметить несколько пунктов. В процессе подготовки к соревнованию возможности нашего лингвистического анализатора были использованы не до конца, в качестве признаков можно было бы использовать другие результаты обработки, например, семантические теги, синонимические ряды, смайлы, а также более развернутые результаты синтаксического модуля. Для коротких сообщений из Твиттера можно ожидать большое количество ошибок и неполных деревьев в синтаксисе, поэтому применение пар синтаксически связанных слов кажется наиболее перспективным. Для других типов данных возможно расширение синтаксических признаков и улучшение результатов классификации за счет этого. Кроме того, для твитов эффективным может быть анализ смайлов, символов эмодзи и хештегов, что было проделано некоторыми участниками соревнования. Данные также не позволили успешно применить те или иные способы фильтрации признаков, что, как представляется, было бы возможно, если бы частотность признаков была выше, например, за счет замены конкретных слов идентификаторами семантических классов.

СПИСОК ЛИТЕРАТУРЫ

1. *Chetviorkin I., Braslavskiy P., Loukachevich N.* Sentiment analysis track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2012». 2012. P. 1-14.

2. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in Russian // Proceedings of BSNLP workshop, ACL, Prague. 2013. P. 12-17.

3. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Ju., Ivanov V., Tutubalina H.* Sentirueval: testing object-oriented sentiment analysis systems in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». 2015. Issue 14. V. 2. P. 13-24.

4. *Pang B., Lee L., Vaithyanathan S.* Thumbs up? Sentiment classification using machine learning techniques // Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002. V. 10. P. 79-86.

5. *Mullen T., Collier N.* Sentiment analysis using support vector machines with diverse information sources // Proceedings of 9th EMNLP. 2004. P. 412-418.

6. *Turney P.* Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th ACL. 2002. P. 417-424.

7. *Kudo T., Matsumoto Y.* A boosting algorithm for classification of semi-structured text // Proceedings of 9th EMNLP. 2004. P. 301-308.

8. *Matsumoto S., Takamura H., Okumura M.* Sentiment classification using word sub-sequences and dependency sub-trees // Ho T.-B., Cheung D., Liu H. (eds.) PAKDD 2005. V. 3518. P. 301-311.

9. *Mavljutov R.R., Ostapuk N.A.* Using basic syntactic relations for sentiment analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2013». 2013. P. 91-100.

10. *Yussupova N., Bogdanova D., Boyko M.* Applying of sentiment analysis for texts in russian based on machine learning approach // Proceedings of The Second International Conference on Advances in Information Mining and Management, Italy. 2012. P. 8-14.

11. *Furnkranz J., Mitchell T. M., Rilof E.* A case study in using linguistic phrases for text categorization on the WWW // Proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, US. 2998. P. 5-12.

12. *Caropreso M.F., Matwin S., Sebastiani F.A.* Learner-independent evaluation of the usefulness of statistical phrases for automated text categorization // Amita G. Chin (ed.), Text Databases and Document Management: Theory and Practice. 2006. P. 78-102.

13. *Nastase V., Shirabad J.S., Caropreso M.F.* Using dependency relations for text classification // Proceedings of the 19th Canadian Conference on Artificial Intelligence, Quebec City. 2006. P. 12-25.

14. *Zhao S., Grishman R.* Extracting relations with Integrated Information using kernel methods // Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, US. 2005. P. 419-426.

15. *Jansen B.J., Zhang M., Sobel K., Chowdury A.* Twitter power: tweets as electronic word of mouth // Journal of the American Society for Information Science and Technology. 2009. V. 60, No 11. P. 2169-2188.

16. *Go A., Bhayani R., Huang L.* twitter sentiment classification using distant supervision // Technical report, Stanford. 2009.

17. *Jiang L., Yu M., Zhou M., Liu X., Zhao T.* Target-dependent Twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, US. 2011. P. 151-160.

18. *Kouloumpis E., Wilson, T., Moore J.* Twitter sentiment analysis: the good the bad and the omg! // Artificial Intelligence. 2011. P. 538-541.

19. *Pak A., Paroubek P.* Twitter as a corpus for sentiment analysis and opinion mining // Proceedings of LREC, Valetta. 2010. P. 75-100.

20. *Адаскина Ю.В., Паничева П.В., Попов А.М.* Полуавтоматическое пополнение словарей на основе синтаксических связей // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014), Санкт-Петербург, 19 – 20 ноября 2014 г. 2014. С. 271-276.

21. *Зализняк А.А.* Грамматический словарь русского языка. М.: Русский язык, 1980.

22. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R,, Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* Scikit-learn: machine learning in Python // Journal of Machine Learning Research. 2011. V. 12 (Oct). P. 2825-2830.

USING SYNTAX FOR SENTIMENT ANALYSIS OF RUSSIAN TWEETS

Yu.V. Adaskina[1], P.V. Panicheva[2], A.M. Popov[3]

«InfoQubes», Sanct-Petersburg State University

[1]adaskina@gmail.com, [2]p.panicheva@spbu.ru, [3]hedgeonline@gmail.com

Abstract

The paper describes our approach to the task of sentiment analysis of tweets within SentiRuEval – an open evaluation of sentiment analysis systems for the Russian language. We took part in the task of sentiment analysis of Russian tweets concerning two types of organizations: banks and telecommunications companies. On both datasets, the participants were required to perform a three-way classification of tweets: positive, negative or neutral.

We used various statistical methods as basis for our machine learning algorithms. Linguistic features produced by our morpho-syntactic analyzer are applied to the classification. Syntactic relations proved to be a crucial feature for any statistical method evaluated, and SVM-based classification performed better than the others. Normalized words are another important feature for the algorithm.

The evaluation revealed that our method proved to be rather successful: we scored the first in three out of four evaluation measures.

Keywords: *sentiment analysis, syntactical relations, Russian language, statistical methods, text classification*.

REFERENCES

1. *Chetviorkin I., Braslavskiy P., Loukachevich N.* Sentiment analysis track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2012». 2012. P. 1-14.

2. *Chetviorkin I., Loukachevitch N.* Evaluating sentiment analysis systems in Russian // Proceedings of BSNLP workshop, ACL, Prague. 2013. P. 12-17.

3. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Ju., Ivanov V., Tutubalina H.* Sentirueval: testing object-oriented sentiment analysis systems in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». 2015. Issue 14. V. 2. P. 13-24.

4. *Pang B., Lee L., Vaithyanathan S.* Thumbs up? Sentiment classification using machine learning techniques // Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002. V. 10. P. 79-86.

5. *Mullen T., Collier N.* Sentiment analysis using support vector machines with diverse information sources // Proceedings of 9th EMNLP. 2004. P. 412-418.

6. *Turney P.* Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th ACL. 2002. P. 417-424.

7. *Kudo T., Matsumoto Y.* A boosting algorithm for classification of semi-structured text // Proceedings of 9th EMNLP. 2004. P. 301-308.

8. *Matsumoto S., Takamura H., Okumura M.* Sentiment classification using word sub-sequences and dependency sub-trees // Ho T.-B., Cheung D., Liu H. (eds.) PAKDD 2005. V. 3518. P. 301-311.

9. *Mavljutov R.R., Ostapuk N.A.* Using basic syntactic relations for sentiment analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2013». 2013. P. 91-100.

10. *Yussupova N., Bogdanova D., Boyko M.* Applying of sentiment analysis for texts in russian based on machine learning approach // Proceedings of The Second International Conference on Advances in Information Mining and Management, Italy. 2012. P. 8-14.

11. *Furnkranz J., Mitchell T. M., Rilof E.* A case study in using linguistic phrases for text categorization on the WWW // Proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, US. 2998. P. 5-12.

12. *Caropreso M.F., Matwin S., Sebastiani F.A.* Learner-independent evaluation of the usefulness of statistical phrases for automated text categorization // Amita G. Chin (ed.), Text Databases and Document Management: Theory and Practice. 2006. P. 78-102.

13. *Nastase V., Shirabad J.S., Caropreso M.F.* Using dependency relations for text classification // Proceedings of the 19th Canadian Conference on Artificial Intelligence, Quebec City. 2006. P. 12-25.

14. *Zhao S., Grishman R.* Extracting relations with Integrated Information using kernel methods // Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, US. 2005. P. 419-426.

15. *Jansen B.J., Zhang M., Sobel K., Chowdury A.* Twitter power: tweets as electronic word of mouth // Journal of the American Society for Information Science and Technology. 2009. V. 60, No 11. P. 2169-2188.

16. *Go A., Bhayani R., Huang L.* twitter sentiment classification using distant supervision // Technical report, Stanford. 2009.

17. *Jiang L., Yu M., Zhou M., Liu X., Zhao T.* Target-dependent Twitter sentiment classification // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, US. 2011. P. 151-160.

18. *Kouloumpis E., Wilson, T., Moore J.* Twitter sentiment analysis: the good the bad and the omg! // Artificial Intelligence. 2011. P. 538-541.

19. *Pak A., Paroubek P.* Twitter as a corpus for sentiment analysis and opinion mining // Proceedings of LREC, Valetta. 2010. P. 75-100.

20. *Adaskina Yu.V., Panicheva P.V., Popov A.M.* Poluavtomaticheskoie popolnenie slovarei na osnove sintaksicheskih svyazei // Tehnologii informacionnogo obtschestva v nauke, obrazovanii i kul'ture. Trudy XVII Vserossiiskoi ob'edinennoi konferencii «Internet i sovremennoe obschestvo» (IMS-2014), Sankt-Petersburg. 2014. S. 271-276.

21. *Zaliznyak A.A.* Grammaticheskii slovar russkogo yazika. M.: Russkii yazik, 1980.

22. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R,, Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* Scikit-learn: machine learning in Python // Journal of Machine Learning Research. 2011. V. 12 (Oct). P. 2825-2830.

СВЕДЕНИЯ ОБ АВТОРАХ

АДАСКИНА Юлия Владимировна – кандидат филологических наук, лингвист-эксперт компании «Инфо-Кьюбс».

Yulia ADASKINA received her Masters and PhD degrees in Theoretical and Applied Linguistics from Moscow State University. Currently is an expert linguist at InfoQubes, Moscow. Her scientific interests include syntax, data mining and distributional semantics.

email: adaskina@gmail.com.

ПАНИЧЕВА Полина Вадимовна – аспирант кафедры теоретической и прикладной лингвистики Санкт-Петербургского государственного университета.

Polina Vadimovna PANICHEVA, received her MSc degree in Information Technology from ITT Tallaght, Dublin, Ireland (2011). Currently is a PhD student at the Department of Theoretical and Applied Linguistics of St. Petersburg State University, Russia. Current scientific interests: distributional semantics, cognitive semantics, affective language, linguistic psychological profiling.

email: p.panicheva@spbu.ru

ПОПОВ Андрей Михайлович – аспирант кафедры математической лингвистики Филологического факультета Санкт-Петербургского государственного университета.

Andrei Mikhailovich POPOV, received MS degree in linguistics from Saint-Petersburg State University (2014). Currently is a graduate student at the Saint-Petersburg State University. Current scientific interests: machine learning, syntax parsing, fact extraction, sentiment analysis.

email: hedgeonline@gmail.com

Материал поступил в редакцию 15 июля 2015 года