СИСТЕМА СЕНТИМЕНТНОГО АНАЛИЗА ATEX, ОСНОВАННАЯ НА ПРАВИЛАХ, ПРИ ОБРАБОТКЕ ТЕКСТОВ РАЗЛИЧНЫХ ТЕМАТИК

Паничева П. В. (ppolin86@gmail.com)

EPAM Systems, Санкт-Петербург, Россия

Ключевые слова: сентиментный анализ, анализ тональности, РОМИП

ATEX: A RULE-BASED SENTIMENT ANALYSIS SYSTEM PROCESSING TEXTS IN VARIOUS TOPICS

Panicheva P. V. (ppolin86@gmail.com)

EPAM Systems, Saint-Petersburg, Russia

ATEX is a rule-based sentiment analysis system for texts in the Russian language. It includes full morpho-syntactic analysis of Russian text, and highly elaborated linguistic rules, yielding fine-grained sentiment scores. ATEX is participating in a variety of sentiment analysis tracks at ROMIP 2012. The system was tuned to process news texts in politics and economy. The performance of the system is evaluated in different topics: blogs on movies, books and cameras; news. No additional training is performed: ATEX is tested as a universal 'ready-to-use' system for sentiment analysis of texts in different topics and different classification settings. The system is compared to a number of sentiment analysis algorithms, including statistical ones trained with datasets in respective topics. Overall system performance is very high, which indicates high usability of the system to different topics with no actual training. According to expectations, the results are especially good in the 'native' political and economic news topic, and in the movie blog topic, proving both to share common ways of expressing sentiment. With regard to blog texts, the system demonstrated the best performance in two-class classification tasks, which is a result of the specific algorithm design paying more attention to sentiment polarity than to sentiment/neutral classes. Along these lines areas of future work are suggested, including incorporation of a statistical training algorithm.

Keywords: rule-based sentiment analysis, sentiment classification, Russian language processing, ROMIP

1. Введение

Сентиментный анализ, или анализ тональности — молодой, но быстро развивающийся раздел автоматической обработки текстов. В середине 1990-х гг. исследователи начали проявлять интерес к выражению субъективного отношения автора в тексте [Wiebe], включая в это понятие мнения, настроения, отношение автора, выраженные каким-то образом в тексте[Pang].

С развитием интернета сентиментный анализ привлекает внимание исследователей как один из разделов анализа субъективности, задачей которого является определение значения «тональности» текста, а именно, классификация текста как отражающего позитивное, негативное или нейтральное отношение автора к объектам, явлениям, персонам, упомянутым в тексте.

Важно отметить, что до сих пор не сформулированы четкие теоретические критерии, по которым тот или иной отрезок текста может быть отнесен к позитивному, негативному или нейтральному классам, несмотря на успешные попытки некоторых исследователей теоретически обосновать сентиментный анализ (к примеру, [Balahur]). Таким образом, оценка значения тональности устанавливается опытным путем, с помощью разметки асессорами, которая затем используется в качестве «золотого стандарта» для обучения и оценки результатов сентиментного анализа. Наличие данных, размеченных таким образом, является критическим для развития этой области, в том числе потому, что бóльшая часть исследований сосредоточена на обучаемых методах классификации.

В России сентиментный анализ стал привлекать внимание исследователей в конце 2000-х гг., что отразилось в появлении в 2011 г. в программе семинара РОМИП дорожек по оценке сентиментного анализа на русском языке. Особенность отечественных работ в данной области заключается в большей производственной и коммерческой направленности описываемых систем. В результате оказываются решающими не только численные показатели результатов работы алгоритмов, обученных и проверенных на определенных текстовых выборках, но и более детальная настройка алгоритмов, прозрачная схема определения значения тональности, основанная на явных и четких лингвистических показателях, а также доступность поддержки системы и ее развития для обработки текстов новых жанров/тематик. С этой точки зрения особенно удобными в применении оказываются системы, основанные на правилах ([Kan, Vasilyev]).

Целью данного исследования является тестирование работы системы ATEX, основанной на правилах, настроенной на новостных текстах различного происхождения, без предварительного обучения. Тестирование призвано показать применимость системы к сентиментному анализу текстов различных тематик в сравнении с другими системами сентиментного анализа, в том числе основанных на машинном обучении. Для этого система ATEX была представлена на семинаре РОМИП в наборе дорожек по сентиментному анализу; при этом не проводилось никакого обучения или дополнительной настройки системы.

2. Алгоритм сентиментного анализа на основе лингвистических правил

Система, которую мы представляем на семинаре РОМИП, автоматически реализует сентиментный анализ на основе правил для русскоязычных данных. Правила содержат богатую линвистическую информацию и применяются к структуре текста, полученной в результате работы морфо-синтаксического модуля системы.

2.1. Морфо-синтаксический анализ

Во-первых, на основе морфологического словаря, содержащего для редактирования в текстовом виде парадигмы слов, происходит определение «нормальной формы» каждой из словоформ в тексте и его грамматических атрибутов. Изначально словарь порожден автоматически по базе данных Грамматического словаря А. А. Зализняка [Zaliznyak].

Для словоформ, которые не были найдены в морфологическом словаре, происходит поиск возможной грамматической информации на основе суффикса и окончания, отбрасывания приставки, а также неточный поиск для потенциальных форм с ошибками и опечатками.

Затем грамматическая информация используется для работы синтаксических правил. Синтаксическая обработка представляет собой формальную грамматику, состоящую из нескольких сотен правил, которые разрешают омонимию, объединяют слова в группы, группы, в свою очередь, в более крупные группы, доходя до размера клаузы и сложного предложения, включая предложения с прямой речью. В полученной многоуровневой структуре происходит простановка синтаксических связей ко всем значимым словам.

2.2. Сентиментный анализ

Сентиментный анализ производится на основе ключевых слов, а также сентиментных правил. И в том, и в другом случае результатом сентиментного анализа является значение тональности (+1, −1, 0 или никакое) для одного или нескольких слов[1].

[1] Текущая версия алгоритма не учитывает силу сентимента, т.е. все слова по умолчанию имеют одинаковый вес при вычислении сентимента предложения. Это упрощение оказывается адекватным и не препятствует достижению высоких результатов подготовке системы на основе наших данных, см. «Подготовка системы к сентиментной классификации»

2.2.1. Ключевые слова

В качестве ключевых сентиментных слов выступают слова, которые несут сентиментную окраску в любом контексте, или в подавляющем большинстве контекстов. Ключевые слова хранятся в виде списков нормальных форм в текстовых файлах и содержат, к примеру, такие слова как «хороший», «плохой», «неприятный», «трус», «успех», «провал», «угроза», «позитив», «оперативно», «своевременно», «слишком», и т.п.; всего 1590 негативных и 510 позитивных слов с указанием части речи, что необходимо для правильной обработки омонимичных форм. Если слово из этого списка с соответствующей частью речи встречается в тексте, ему приписывается соответствующее значение тональности.

2.2.2. Сентиментные правила

Сентиментные правила используются для более точного определения сентимента слов и работают на основе более полной морфо-синтаксической информации. Сентиментные правила реализованы на предметно-ориентированном языке программирования и на входе обрабатывают синтаксическую структуру предложения, состоящую из слов и связей между ними, или ее часть, присваивая значение атрибутам отрицания или сентимента определенным словам на выходе.

2.2.2.1. Сентимент словосочетания

В некоторых случаях отдельные слова не несут в себе сентиментного значения, но сентиментом нагружено определенное сочетание некоторых слов или форм слов. Правило, приписывающее сентимент, основано на синтаксической связи определенных слов или словоформ в предложении.

Например, с помощью этих правил обрабатываются такие сочетания, как «пойти навстречу, на лапу, душа компании, по фазе, так себе, промыть мозг, поставить крест, с ума, из ума, ниже плинтуса», и многие другие.

2.2.2.2. Инверсия сентимента

При отрицании в сочетании со словом, которое содержит значение сентимента, его сентимент инвертируется: если слово имеет позитивную тональность, то отрицание модифицирует его на негатив; при отрицании негативной тональности слово в общем случае получает нулевую тональность сентимента.

Важно отметить, что отрицание может выражаться несколькими способами. Основной способ выражения отрицания — частица «не», предикатив «нет». Они приписывают отрицание словам, с которыми они связаны определенными синтаксическими связями. Также при определенных синтаксических связях отрицание ставится за счет группы слов, отличающихся семантикой отрицания, таких как «отсутствие, удаление, лишение, отрицание, устранение, отсутствовать, удалять, лишать, отрицать, устранять», и предлог «без».

При работе с отрицанием также были выделены группы слов — имен существительных, прилагательных, глаголов, — которые получают или меняют значение тональности специфическим образом в сочетании с отрицанием. Эти слова, во-первых, могут не входить в список ключевых сентиментных слов,

но получают значение сентимента при отрицании; во-вторых, могут содержаться в списке ключевых слов с негативным сентиментом, но при сочетании с отрицанием получают, в отличие от общего правила, позитивный сентимент. В первом случае примером могут служить такие слова, как «будущее, дело, желание, мозг, надежда, объяснение, ответ, смысл, ум»; во втором — «вопрос, дефект, конфликт, нарекание, перебой, препятствие, проблема»[2]. Всего в системе порядка 120 таких слов; они хранятся в виде списков в текстовых файлах, обозначенные как «слова, позитивные с отрицанием» и «слова, негативные с отрицанием».

2.2.2.3. Синтаксически связанные слова, входящие в значимые семантические списки

Категория правил, которая заслуживает особого внимания, — правила, приписывающие сентимент на основе синтаксической связи между словами. При этом каждое из слов по отдельности не входит в ключевые сентиментные слова, а связь двух слов не является устойчивым словосочетанием.

К примеру, такие слова как «деньги, доход, зарплата, качество, оборот, отдача, оценка, потенциал, рейтинг, уровень» не содержат позитива сами по себе. С другой стороны, экспериментально подтверждается, что когда явления, обозначенные этими словами, велики, высоки, максимальны, это добавляет положительную тональность, и наоборот — когда они низки, добавляет отрицательную. Ср. «наш рейтинг»/«высокий рейтинг»/«повышение рейтинга»/«низкий рейтинг»/«понижение рейтинга».

Наоборот, такие слова как «издержка, очередь, потеря, расход, риск, урон, ущерб» будут получать позитивный сентимент, когда такие явления минимальны, и негативный, когда максимальны.

Таким образом, если слова из данных списков синтаксически связаны со словами, обозначающими увеличение или уменьшение степени, количества явления или предмета, то главному слову в данной синтаксической связи приписывается соответствующее значение сентимента. Данные правила также включают в себя списки слов, относящиеся к семантике проблем, ситуаций и решения; нехватки; порядка, правил и их гибкости, жесткости, и т. п.

3. Постановка задачи

3.1. Дорожки РОМИП по сентиментному анализу

На семинаре РОМИП были предоставлены 2 вида дорожек по сентиментному анализу: отрывки с цитатами прямой и косвенной речи из новостей, а также тексты блогов, причем последние включали 3 тематики: отзывы

[2] Пример такой инверсии сентимента из данных РОМИП (id отрывков 1049, 1188) см. в разделе «Результаты системы ATEX и их анализ».

о фильмах, книгах и фотокамерах. Тестирование системы проводилось в следующих дорожках:

Новостные фрагменты:

- дорожка по классификации прямой и косвенной речи из новостных лент — 3 класса: положительные, отрицательные, нейтральные (не содержащие оценки).

Отзывы о товарах:

- дорожка по классификации отзывов пользователей на 2 класса: положительные и отрицательные;
- дорожка по классификации отзывов пользователей на 3 класса: положительные, отрицательные и содержащие достаточно значимые положительные и отрицательные стороны оцениваемой сущности.

3.2. Подготовка системы к сентиментной классификации

Для каждого вида данных на семинаре были предоставлены размеченные выборки для обучения системы. Следует подчеркнуть, что в цели участия в семинаре входило тестирование системы ATEX с исходными настройками, в том числе для новых неисследованных тематик. Поэтому тренировочные выборки не использовались для обучения системы и подготовки к тестовому этапу. Система была настроена заблаговременно в ходе работы над текстами с русскоязычных новостных сайтов русского и казахского доменов на тему политики и экономики — в частности, сентиментно размеченного корпуса, состоящего из 3 тыс. предложений.

Значение сентимента предложения в системе вычисляется как знак среднего арифметического значений сентиментов входящих в него слов. Позитивное или негативное значение сентимента предложения, как и слова, обозначается соответственно как «+1» или «−1». При этом если количество положительно и отрицательно окрашенных слов в предложении одинаково, в том числе и равно нулю, то общий сентимент предложения получается нейтральным. Таким образом, в системе не проводится различие между «нейтральным» сентиментным классом и классом, содержащим достаточно значимые положительные и отрицательные стороны оцениваемой сущности.

Так как большинство данных содержало отрывки, состоящие из более чем одного предложения, система тестировалась в двух режимах:

1. «С предложениями»: вычислялся сентимент каждого предложения в отрывке. Общий сентимент отрывка вычислялся как среднее арифметическое между сентиментами предложений.
2. «Без предложений»: сентимент всего отрывка вычислялся как среднее арифметическое между сентиментами всех входящих в него слов, без учета границ предложений.

3.3. Данные тестирования РОМИП

В Таблице 1 представлена статистика размеченных тестовых данных РОМИП по сентиментным дорожкам. Новостные отрывки были размечены на 3 класса; отрывки из блогов были размечены и оценивались двумя способами: на 2 и 3 класса. Учитывая настройку системы на новостных текстах политической и экономической тематик, именно в ней ожидается получить наиболее высокие результаты сентиментного анализа.

Таблица 1. Статистика тестовых данных РОМИП

Тематика	Всего отрывков	Положительных	Отрицательных	Нейтральных/ содержащих + и −	Процент наибольшего класса во всей выборке, %
Новости	4573	1448	1234	1890	41
Фильмы	408	330	78	—	80
		266	63	79	65
Книги	129	112	17	—	87
		100	9	20	78
Камеры	411	397	14	—	97
			7	53	85

4. Результаты системы ATEX и их анализ

Оценка результатов работы систем сентиментного анализа проводилась на основе четырех показателей: Аккуратность (Accuracy), Полнота (Recall), Точность (Precision), Мера F1 (F-measure) ([Chetviorkin]). В Таблице 1 видно, что тестовые данные не являются сбалансированными относительно итоговых сентиментных классов; поэтому для представления относительных результатов работы систем они были упорядочены по значению F-measure, которое является более подходящей оценкой, чем Accuracy, для несбалансированных данных [van Rijsbergen]. В таблицах ниже приведены результаты работы различных систем; результаты представленной в данном докладе системы выделены жирным. Выведены наилучшие четыре результата, упорядоченные по F-measure.

Таблица 2. Результаты сентиментной классификации отрывков прямой и косвенной речи из новостей

Number	System ID	Object	Classes	Precision Macro	Recall Macro	F_Measure Macro	Accuracy	
1	xxx-4	news	3	0,626	0,616	0,621	0,616	
2	**ATEX**	**news**	**3**	**0,606**	**0,579**	**0,592**	**0,571**	**без предложений**
3	**ATEX**	**news**	**3**	**0,606**	**0,576**	**0.590**	**0,569**	**с предложениями**
4	xxx-5	news	3	0,579	0,568	0,574	0,575	

Согласно ожиданиям, результаты по тематике «новости» оказались высокими и абсолютно, и относительно среди других систем. Это говорит о том, что настройка системы на новостных текстах оказалась полезной, несмотря на различные источники и время появления новостных текстов, используемых для настройки и для тестирования системы.

Таблица 3. Результаты сентиментной классификации на 2 класса блогов по тематике «Фильмы»

Number	System ID	Object	Classes	Precision Macro	Recall Macro	F_Measure Macro	Accuracy	
1	**ATEX**	**film**	**2**	**0,695**	**0,719**	**0,707**	**0,806**	**с предложениями**
2	xxx-23	film	2	0,731	0,641	0,683	0,831	
3	xxx-2	film	2	0,667	0,687	0,677	0,787	
4	xxx-12	film	2	0,759	0,586	0,661	0,828	

Несмотря на различие в тематиках, система показала наилучший результат в дорожке по классификации отзывов о фильмах на два класса. Следует отметить, что в классификации на 2 класса отзывов о книгах и о фотокамерах система занимает третью и пятую строки соответственно относительно других систем. Предположительно, язык выражения сентимента в описании фильмов оказывается наиболее близким к языку выражения сентимента в политике и экономике, и по-видимому, наиболее общим, не обладающим большим количеством специфических сентиментных слов и выражений. В действительности, для сравнения, описания фотокамер содержат большое количество подробностей о функциональных качествах, свойствах самих камер, которые не могут быть освоены без знакомства с самой тематикой и создания специфических правил; что делает такие тексты специфическими и близкими к техническим описаниям.

Таблица 4. Результаты сентиментной классификации на 3 класса блогов по тематике «Фильмы»

Number	System ID	Object	Classes	Precision Macro	Recall Macro	F_Measure Macro	Accuracy	
1	xxx-11	film	3	0,569	0,479	0,520	0,694	
2	**ATEX**	**film**	**3**	**0,486**	**0,521**	**0,503**	**0,596**	**с предложениями**
3	xxx-0	film	3	0,505	0,477	0,491	0,627	
4	xxx-7	film	3	0,566	0,429	0,488	0,360	

Предположение о более общих сентиментных моделях в тематиках фильмов, политики и экономики подтверждается также в результатах классификации на 3 класса: система занимает вторую строку в тематике фильмов, и пятую и шестую строки соответственно для тематик камер и книг.

Важно отметить, что система, основанная на правилах в сравнительном анализе значений F-measure и Accuracy результатов, получает высокий показатель F-measure при относительно более низком значении Accuracy. Это говорит о более равномерном механизме классификации такой системы относительно других систем, особенно при условиях несбалансированной обучающей выборки вероятностных систем, которые при тестировании, предположительно, демонстрируют «перекос» результатов в сторону наиболее частотного класса, что проявляется в их высоком значении Accuracy, но низком значении F-measure относительно системы, основанной на правилах. С другой стороны, это характеризует относительно более высокую воспроизводимость результатов последней на различных данных.

Для иллюстрации работы сентиментных правил приводятся примеры работы системы на цитатах из новостных лент. Подчеркиванием выделены слова, получившие соответствующую тональность в результате работы всей последовательности сентиментных правил и повлиявшие на правильный конечный результат.

Таблица 5. Примеры работы системы для цитат из новостных лент

Id отрывка	Текст	Общий сентимент
1049	«На данный момент не вижу перспективы(−1) никаких военных действий за исключением мер по защите дипломатических представителей, а также справедливого наказания ответственных за эту ужасную акцию(−1)», — сказал Терци.	−1
1068	«Льоренте все еще принадлежит Атлетику и, похоже, готов играть. Впрочем, мы все равно потеряли(−1) одного отличного(0) футболиста(0) и хорошего(0) человека(0)», — сказал Бьелса, намекая на уход Хави Мартинеса в мюнхенскую «Баварию».	−1
1108	«В период после нашей предыдущей встречи мировая экономика по-прежнему испытывала немалые трудности и продолжает подвергаться рискам падения; финансовые рынки остаются(−1) нестабильными, тогда как высокий уровень(−1) дефицита госсектора и государственной задолженности в некоторых развитых экономиках в значительной мере сдерживает процесс восстановления экономики», — отмечается в документе.	−1
1151	«Еще одна трата(+1) на проведение саммита — обеспечение безопасности. Но деньги пошли на обеспечение спецслужб, оборудование не будет выброшено(0), но будет использовано для проведения Универсиады в Казани, Олимпиады в Сочи, на форумы «восьмерки» и «двадцатки». Ничего не пропадает(0). Все траты в целом абсолютно обоснованы», — подчеркнул Путин.	+1
1188	Конкуренция — не проблема(+1) для меня<…>.	+1
7943	«Принять данный документ позволило повышение(+1) возможностей медицинских учреждений по диагностике и лечению заболеваний», — констатируют в оборонном ведомстве.	+1

5. Выводы и дальнейшая работа

Система, основанная на правилах, настроенная на новостных текстах без дополнительной настройки и обучения, в сентиментной классификации на 2 и 3 класса для различных тематик демонстрирует хорошие результаты, сравнимые с результатами систем, в том числе обученных на текстах соответствующих тематик. Особенно высокие показатели полноты, точности и F-measure система демонстрирует, как и ожидалось, в тематике новостей, а также в тематике отзывов о фильмах, что характеризует особенности выражения сентимента в последней.

Числовые показатели оценки говорят о высокой воспроизводимости результатов системы на различных текстах в различных тематиках, при отсутствии тренировочной размеченной выборки и связанных с ней ограничений.

При более детальном исследовании результатов в дальнейших работах наиболее полезной представляется информация о сработавших в ходе сентиментного анализа правилах. Это позволило бы, во-первых, сформировать статистику наиболее частотных моделей выражения сентимента; во-вторых, охарактеризовать различные тематики исследования с точки зрения специфических присущих им моделей, правил и лексики; наконец, это создало бы основу для автоматического создания и лексического наполнения недостающих правил.

В дальнейшем будет полезно, учитывая значительный объем текстов в некоторых тематиках и важность понятия «нейтрального» сентиментного класса, настраивать систему с помощью машинного обучения. В качестве параметров следует использовать количество и, возможно, качество положительных, отрицательных и нейтральных слов в тексте, обработанном системой. В результате следует с помощью алгоритма обучения настраивать общий сентиментный класс, соответствующий всему тексту. Такое дополнение позволило бы, во-первых, более четко определять границу между сентиментным и нейтральным текстом; во-вторых, разграничивать действительно нейтральные тексты как не содержащие сентимент от текстов, в которых указываются достаточно значимые положительные и отрицательные стороны оцениваемой сущности.

Литература

1. *Balahur A., Montoyo A.* Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification, Proc. AISB Convention Comm., Interaction and Social Intelligence. 2008.
2. *Chetviorkin I., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2012. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2013" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013"]. Bekasovo, 2013.
3. *Kan D.* Rule-based approach to sentiment analysis at ROMIP 2011. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2012" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"]. Bekasovo, 2012.
4. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval, no. 2, pp. 1–135. 2008
5. *van Rijsbergen C. J.* Information Retrieval, Butterworths, London, (1979)
6. *Sokolova M., Lapalme G.* A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4, pp. 427–437. Jul. 2009
7. *Vasilyev V. G., Khudyakova M. B., Davydov S.* Sentiment classification by fragment rules. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2012" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"]. Bekasovo, 2012.
8. *Wiebe J. M.* Tracking point of view in narrative. Computational Linguistics 20 (2), pp. 233–287. 1994.
9. *Zaliznyak A.* Grammaticheskij slovar' russkogo jazyka. Moskva, 1977, (further editions are 1980, 1987, 2003).