Аспектный анализ отзывов о ресторанах для рекомендательных систем е-туризма

Е.В. Проноза, Е.В. Ягунова
Санкт-Петербургский государственный университет
{katpronoza, iagounova.elena}@gmail.com

Аннотация

В данной работе представлен аспектно-ориентированный метод анализа тональности отзывов о ресторанах, который может использоваться в рекомендательных системах в сфере е-туризма. Предложенный метод может применяться не только для создания рекомендательных систем в сфере е-туризма, но и для систем автоматического понимания текста вообще, где текстовые данные слабо структурированы и принадлежат разговорному стилю.

В результате экспериментов с различными классификаторами и наборами признаков в данной работе показано, что использование словарей, автоматически или вручную созданных на основе корпуса отзывов, позволяет улучшить качество моделей извлечения ряда аспектов ресторана.

Ключевые слова: анализ тональности; отзывы о ресторанах; машинное обучение; рекомендательная система.

1. Введение

В настоящее время чрезвычайно востребованной является задача автоматического анализа мнений пользователей о товарах или услугах. Анализ пользовательских оценок позволяет оперативно принимать решения о продаже товаров, об изменении планов внедрения услуг и маркетинговой политики компании в целом. В частности, очевидна актуальность автоматического анализа мнений пользователей в сфере е-туризма.

В данной работе представлен аспектно-ориентированный метод анализа тональности отзывов о ресторанах, который может использоваться в рекомендательных системах в сфере е-туризма. Цель работы – разработка метода анализа тональности слабоструктурированных текстов для систем автоматического понимания текста в сфере е-туризма.

Предложенный метод может применяться не только для создания рекомендательных систем в сфере е-туризма, но и для систем автоматического понимания текста вообще, где текстовые данные слабо структурированы и принадлежат разговорному стилю.

Сборник научных статей XVIII Объединенной конференции «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23-25 июня 2015 г.

В данной работе показано, что использование словарей, автоматически или вручную созданных на основе корпуса отзывов, позволяет улучшить качество моделей извлечения ряда аспектов ресторана.

2. Анализ тональности: существующие подходы

Как и в любых других задачах и области автоматической обработки естественного языка, основные используемые методы делятся на методы на основе правил, на основе статистики и гибридные методы. Что касается правил, обычно подобный подход подразумевает использование семантических тезаурусов (например, WordNet-Affect, SentiWordNet, SenticNet), в то время как статистические методы позволяют решить данную задачу даже при полном отсутствии подобных ресурсов.

Так как мы преследуем статистический подход к анализу тональности отзывов, нами было проведено исследование моделей обучения с учителем, которые используются в данной задаче.

2.1. Классификаторы

Как выяснилось, чаще всего для анализа тональности текста используются наивный байесовский классификатор и метод опорных векторов. При этом типичная постановка задачи заключается в том, что необходимо провести классификацию мнений о фильмах или товарах, о политической фигуре или событии (в блоге, твите и т.д.) на положительные и отрицательные. Наивная байесовская модель обычно используется как базовая, простейшая, а метод опорных векторов – как более сложная, так как он считается более эффективным. Однако Бермингем и Смитон [3] и Ванг и Маннинг [25] показали, что наивный байесовский классификатор демонстрирует лучшие показатели по сравнению с методом опорных векторов на коротких текстах, например, на твитах.

Кроме того, в некоторых работах предлагается подход, основанный на применении нейронных сетей с векторными репрезентациями слов в качестве признаков. Данный подход не зависит от языка и не требует семантических тезаурусов (используется только обучающий корпус текстов).

В рамках нашей работы проводились эксперименты с такими моделями, как наивный байесовский классификатор, логистическая регрессия, метод опорных векторов (с различными типами ядра), перцептрон, дерево решений и случайный лес.

2.1. Признаки для машинного обучения

В ходе работы было также проведено исследование признаков, которые обычно используются в задаче извлечения мнений с помощью машинного обучения. Результаты этого исследования представлены в таблице 1.

Таблица 1. Признаки в задаче классификации тональности

Признак	Ссылки
Униграммы	[1] [3] [5] [8] [13] [17] [19] [22] [24] [25]
N-граммы, N >= 2	[1] [3] [5] [13] [19] [22] [25]
Униграммы определенной части речи	[1] [5] [22]
N-граммы определенных частей речи, N >= 2	[1] [2]
Позиции токенов в тексте	[13]
Эмотиконы	[6] [10] [18] [19] [24]
Подстроки	[5]
Синтаксические отношения, синтаксические n-граммы	[10] [12] [20] [21]
Модификаторы валентности	[8] [9]
Семантические классы	[5] [17]
Семантические тезаурусы	[4] [18] [22]

Как выяснилось, чаще всего используются такие признаки, как n-граммы . Иногда слова в n-граммах заменяются их семантическими классами [5] [17]. Некоторые исследователи отрицают эффективность использования (смежных) n-грамм высокого порядка [13], в то время как другие утверждают, что для некоторых задач такие признаки могут быть чрезвычайно полезны [5] [25].

Иногда в набор признаков включается и информация о позиции токена в абзаце [13], также проводятся эксперименты с низкоуровневыми признаками, например, с подстроками [5]. Различные способы обработки отрицаний (например, добавление «not_» к слову, к которому относится отрицание) описаны во многих работах [1] [4] [5] [11] [19] [22]. При работе с разговорными текстами некоторые авторы предлагают учитывать эмотиконы [6] [10] [18] [19] [24].

В качестве признаков в задаче анализа тональности также используют высокоуровневые признаки, основанные на лингвистических знаниях: например, в [14] описано применение признаков, отражающих синтаксические связи (в терминах грамматики зависимостей или грамматики составляющих).

У Сидорова [20] рассматривается применение синтаксических n-грамм, состоящих из слов, связанных между собой синтаксическими отношениями. Однако подобный подход требует наличия доступных инструментов синтаксического анализа и потому едва ли может применим для русского языка. С n-граммами с использованием частей речи (и частеречными n-граммами) также проводились эксперименты [1] [19] [22]. Пары «наречие-прилагательное» использовались Бенамарой [2]. В соответствующей работе представлены классификация наречий степени на пять категорий и алгоритм выставления баллов комбинациям наречий и прилагательных. Показано, что использование

наречий позволяет улучшить работу системы в отношении идентификации тональности.

У Кеннеди и Инкпен [9] модификаторы валентности (уменьшительные и увеличительные) включаются в состав биграммных признаков.

В данной работе в качестве базового набора признаков используются смежные n-граммы. Затем этот набор признаков расширяется несмежными n-граммами и признаками, полученными в результате анализа корпуса (а именно, предикативно-атрибутивными словами и модификаторами). Последние обобщают идею комбинации наречий и прилагательных [2] и близки к понятию модификаторов валентности [9]. Несмежные биграммы, в свою очередь, позволяют охватить отрицания, которые в русском языке могут быть выражены различными способами.

Следуя Пэнгу [14], в случае n-грамм мы используем вектора встречаемости вместо векторов частоты, поскольку они являются более эффективными признаками для нашей задачи.

3. Исходные данные

3.1. Корпус отзывов

Исходный корпус отзывов о ресторанах состоит из 32525 отзывов (и 4,2 миллионов словоупотреблений). Корпус представлен текстами на русском разговорном языке с сайта, где пользователи оставляют свои отзывы о различных продуктах, услугах, заведениях и пр. Длина отзыва варьируется от 1 до 96 предложений и в среднем составляет 10 предложений.

Небольшое подмножество исходного корпуса было размечено для последующего машинного обучения. Вначале подкорпус был размечен автоматически на основе ключевых слов, а затем выверен вручную двумя экспертами. Этот подкорпус включает в себя 1025 отзывов о 206 ресторанах в центре Санкт-Петербурга.

3.2. Аспекты ресторанов

В данной работе применяется подход, при котором список аспектов объектов (в нашем случае – ресторанов) не извлекается из отзывов автоматически, а составляется вручную исходя из потребностей пользователей системы, в которой будет использоваться модуль анализа тональности.

Список аспектов, извлекаемых из отзывов, включает в себя тип заведения и кухни, качество еды и обслуживания, наличие уюта, романтической обстановки, наличие бара, танцпола, детской комнаты, наличие поблизости гостиницы, парковки и т.д. Аспекты ресторана, которые рассматриваются в данной работе, приведены в таблице 2. Для каждого из аспектов в таблице указано множество значений, которые может принимать этот аспект.

Ресторан имеет как объективные аспекты (например, наличие танцпола, бара, детской комнаты и т.д.), так и субъективные (см. таблицу 2). Для объективных аспектов задача классификации аспекта является задачей извлечения информации, в то время как для субъективных – это задача анализа тональности.

Таблица 2. Аспекты ресторана

Аспект	Множество значений
Качество еды	{-2; -1; 0; 1; 2}
Качество обслуживания	{-2; -1; 0; 1; 2}
Скорость обслуживания	{-2; -1; 0; 1; 2}
Учтивость персонала	{-2; -1; 0; 1; 2}
Приветливость персонала	{-2; -1; 0; 1; 2}
Уровень цен	{-2; -1; 0; 1; 2}
Уровень шума	{-2; -1; 0; 1; 2}

В данной работе рассматриваются субъективные аспекты (это аспекты со значениями из 5-балльной шкалы от -2 до 2, представленные в таблице 2) и решается задача анализа тональности.

Поскольку размеры размеченного подкорпуса невелики, необходимо убедиться в том, что для рассматриваемых аспектов в подкорпусе имеется достаточное количество прецедентов. Так, для каждого из аспектов в таблице 2 была подсчитана доля отзывов в подкорпусе, в которых упоминается данный аспект (и принимает любое из значений). В качестве порогового значения эмпирически была выбрана величина, равная 10%: так, в экспериментах с машинным обучением участвовали только аспекты, упоминающиеся хотя бы в 10% отзывов из размеченного подкорпуса. Как оказалось, все субъективные аспекты (качество еды, приветливость персонала, скорость обслуживания и т.д. – см. таблицу 2) удовлетворяют этому критерию, что подтверждает актуальность данных аспектов с точки зрения пользователей.

4. Метод и этапы исследования

Целью данной работы является разработка аспектно-ориентированного метода анализа тональности отзывов о ресторанах для рекомендательной системы. Поскольку русский язык не отличается большим числом доступных лингвистических ресурсов, мы во многом опираемся на результаты корпусного анализа, пытаясь извлечь максимум из имеющихся данных, и экспериментируем как с правилами, так и с машинным обучением.

В результате анализа классификаторов, которые обычно используются в задаче извлечения мнений, а также возможностей выбранного инструментария для машинного обучения, выбраны бернуллиевский и мультиномиальный наивные байесовские классификаторы, логистическая регрессия, метод опорных векторов (с линейным, полиномиальным, гауссовым и сигмоидным ядрами), перцептрон (с «перемешиванием» обучающего множества), дерево решений и случайный лес. При этом наивный байесовский классификатор демонстрирует лучшие показатели в определении класса мнения для большинства параметров ресторанов. В целом это соответствует выводам, описанным для английского языка в работах [3] и [25].

Выбор признаков также продиктован реалиями русского языка. Отсутствие доступных семантических тезаурусов мы пытаемся компенсировать

полуавтоматическим построением словарей на основе корпуса отзывов, а отсутствие подходящих для нашей задачи инструментов синтаксического анализа – включением несмежных биграмм в состав признаков машинного обучения. В целом, первое созвучно идеям модификаторов валентности и парам «наречие + прилагательное» из работ Кеннеди и Инкпен [9] и Бенамары [2], а второе – синтаксическим n-граммам Сидорова [67]. Различные способы выражения отрицания в русском языке мы также надеемся охватить несмежными биграммами.

Основные этапы проведенного исследования соответствуют решению следующих задач:

- анализ корпуса отзывов (в том числе построение словарей и оценка покрытия корпуса этими словарями);
- выделение признаков для машинного обучения;
- оценка и выбор лучших моделей и лучших наборов признаков машинного обучения.

4.1. Анализ корпуса

Процедура анализа корпуса проводилась в рамках данной работы в несколько шагов. Можно выделить такие этапы, как:

- предварительная обработка корпуса (токенизация, лемматизация, нормализация и разбиение на предложения). Сюда также включается построение частотного словаря и словарей n-грамм;
- построение словарей номинаций и предикативно-атрибутивных словарей. Это процедура проводилась для фреймов кухни и обслуживания как наиболее важных аспектов ресторана. Кроме того, были построены предикативно-атрибутивные словари для фрейма обслуживания, а также для типа кухни и качества еды внутри фрейма кухни. Эти словари в дальнейшем использовались при построении шаблонов и выделении признаков машинного обучения;
- оценка покрытия корпуса словарями номинаций;
- построение шаблонов на основе частеречного распределения слов, соседних со словами-номинациями;
- оценка покрытия корпуса построенными шаблонами.

Все перечисленные этапы подробно описаны в [15]. В данной работе было принято решение использовать предикативно-атрибутивные словари – словари прилагательных и причастий в полной и краткой форме, которые относятся к номинациям ключевых фреймов (т.е. еды и обслуживания). Это обусловлено тем, что анализ распределения частей речи среди несмежных биграмм корпуса отзывов выявил доминирование именных групп при описании аспекта ресторана.

Оценки, полученные на последнем этапе анализа корпуса (покрытие словарями 66% отзывов для фрейма обслуживания и 71% – для качества еды) позволяют надеяться, что для извлечения мнений о качестве еды и обслуживания может быть применена стратегия, основанная на употреблении соответствующих номинаций.

На основе частотного словаря, собранного по корпусу отзывов, был также составлен словарь тональностей. Для этого из частотного словаря были выбраны все прилагательные и причастия (в полной и краткой форме), и каждой паре «лемма + часть речи», в которой лемма могла бы выразить в тексте отношение пользователя к одному из субъективных аспектов ресторана, был поставлен в соответствие один из пяти классов по 5-балльной шкале от -2 до 2. Разметкой словаря занимался эксперт-лингвист.

В данном исследовании словарь тональностей используется только для ключевых аспектов ресторана (фреймы еды и обслуживания):

- качество еды;
- качество обслуживания;
- скорость обслуживания;
- приветливость персонала;
- вежливость персонала.

Баллы из размеченного словаря используются в данном исследовании в качестве одного из признаков в машинном обучении.

Помимо словарей номинаций, предикативно-атрибутивных словарей и словаря тональностей были также построены словари модификаторов и словарь ключевых слов и фраз.

Основной словарь модификаторов представляет собой неразмеченный список наречий степени, частиц и их сочетаний и содержит 94 единицы. Он был получен на основе частотного словаря корпуса. Вначале были применены частеречные фильтры, затем осуществлялась ручная проверка. Кроме того, имеется собранный вручную словарь модификаторов-усилителей, состоящий из наречий степени (например, «очень», «весьма», «чрезвычайно»).

Словарь ключевых фраз состоит из вручную составленных лемматизированных фраз для исследуемых аспектов ресторана.

Все словари ключевых слов и фраз содержат разметку, т.е. помечены значениями параметров (например, лемматизированная фраза «демократичный цена» помечена значением «1» для аспекта «уровень цен»).

4.2. Выбор признаков и моделей машинного обучения

Для всех экспериментов с машинным обучением в данной работе использовалась библиотека scikit-learn для языка python.

Для каждого из аспектов, наиболее часто встречающихся в отзывах, проводились эксперименты с такими классификаторами, как наивный байесовский классификатор (бернуллиевский (NB) и мультиномиальный (MNB)), логистическая регрессия (LogReg), метод опорных векторов (SVM) (с линейным, полиномиальным, гауссовым и сигмодиным ядрами), перцептрон (с «перемешиванием» обучающего множества), дерево решений и случайный лес. Однако все виды SVM, кроме линейного, показали во время предварительных экспериментов результаты на порядок хуже остальных классификаторов (а именно, на десятки процентов хуже), поэтому они не были включены в процедуру кросс-валидации и серию тестов для выбора лучшей модели.

Поскольку размеченный подкорпус содержит большое количество пустых значений, мы разделили задачу классификации на две: вначале необходимо

оценить релевантность отзыва в отношении аспекта, а затем, если отзыв релевантен, нужно определить значение аспекта (т.е. тональность).

При выборе используемых признаков мы постарались объединить результаты анализа корпуса отзывов и идеи, предлагаемые в работах по извлечению мнений.

Было выделено 11 различных наборов признаков. Базовый набор состоит только из униграмм и смежных биграмм (он используется в первой задаче определения релевантности), а следующие 10 наборов расширяют его теми или иными признаками: несмежные биграммы, эмотиконы и восклицания, предикативно-атрибутивные слова, ключевые фразы, словарь тональностей. Эти признаки в различных комбинациях используются во второй задаче определения класса.

Все признаки, за исключением эмотиконов и восклицаний, являются бинарными и отражают наличие определенного слова или комбинации слов в отзыве. Под несмежными биграммами подразумеваются такие биграммы, компоненты которых необязательно являются соседними словами в тексте (в данном случае допускается наличие 1-2 слов между компонентами.

Предикативно-атрибутивные признаки, как и признаки словаря тональностей, представлены для каждого слова из соответствующих словарей. Что касается эмотиконов, позитивные и негативные эмотиконы сгруппированы в две разные категории.

Перед обучением классификаторов на описанных наборах признаков применялась процедура автоматического отбора наиболее важных признаков. Для этого использовались встроенные возможности библиотеки scikit-learn: отбор признаков осуществлялся с помощью рандомизированной логистической регрессии.

5. Результаты

Для выбора оптимальной комбинации классификатора и набора признаков, были проведены следующие 2 процедуры: кросс-валидация (с делением размеченного корпуса на 10 частей и с предварительным «перемешиванием» отзывов) и серия статистических тестов (для проверки того, что выбранные комбинации классификатора и признаков значительно лучше остальных).

Модели оценивались с помощью показателей точности и средней взвешенной F-меры. Серия статистических тестов Холма-Бонферрони проводилась на основе рекомендаций, приведенных в [7], и с помощью метода, описанного в [16]. Она позволяет выбрать комбинацию классификатора и набора признаков с показателями, статистически лучшими, чем у остальных комбинаций (или одними из лучших), см. табл. 3.

Для большинства исследуемых аспектов LogReg и линейный SVM оказались оптимальнымы классификаторами при решении задачи определения релевантности отзыва. Таким образом, эти классификаторы могут быть рекомендованы в задаче выделения текстов (из массива разговорных неструктурированных текстов), которые содержат мнение о некотором объекте.

Для задачи выбора определенного значения аспекта (т.е. класса), независимо от параметра, оптимальными оказались наивные байесовские классификаторы.

Поэтому можно предположить, что в задаче определения тональности наивный байесовский классификатор (мультиномиальный) является оптимальным выбором (из семи рассматриваемых типов классификаторов) для данного типа текстов.

Таблица 3. Примеры определения класса и релевантности для аспектов

Аспект	Классификатор:Набор признаков	Точность, %	Средняя F1, %
Определение класса			
учтивость персонала	MNB:extended_KWs	80,34	79,75
приветливость персонала	MNB:extended_All	78,38	77,72
уровень шума	MNB:baseline	80,67	76,17
качество еды	MNB:extended_All	74,95	73,59
качество обслуживания	MNB:extended_Distant_Emotions_Lex	73,73	72,70
скорость обслуживания	MNB:extended_Distant_Emotions	72,00	71,35
уровень цен	MNB:baseline	64,75	60,95
Определение релевантности			
качество еды	SVM:baseline	93,61	93,51
уровень шума	SVM:baseline	93,61	93,45
уровень цен	LogReg:baseline	91,39	91,38
скорость обслуживания	LogReg:baseline	90,37	90,39
приветливость персонала	LogReg:baseline	85,93	85,77
качество обслуживания	LogReg:baseline	85,65	85,61
учтивость персонала	MNB:baseline	85,83	85,12

Кроме того, как оказалось, для аспектов «уровень шума» и «уровень цен» достаточно использовать базовый набор признаков и MNB классификатор, чтобы получить максимальные по этим параметрам результаты.

Включение словарей в состав признаков (для фреймов обслуживания и кухни) улучшает показатели для всех аспектов, кроме скорости. Аспект скорости лучше всего извлекается с использованием эмотиконов и несмежных биграмм, которые покрывают большое число способов выражения мнения. В

настоящее время мы заняты более детальной проработкой используемых словарей.

6. Заключение

В результате проведенного исследования предложен и апробирован аспектно-ориентированный метод анализа тональности отзывов о ресторанах, который может использоваться в рекомендательных системах в сфере е-туризма. Сфера его применения не ограничена данной предметной областью. Предложенные идеи могут учитываться при решении задач автоматического понимания слабо структурированных текстов разговорного стиля на флективном языке со свободным порядком слов (особенно в том случае, когда для данного языка имеется малое количество доступных лингвистических ресурсов).

Для каждого из аспектов ресторана в работе выбирается оптимальная (с точки зрения качества работы и вычислительной эффективности) комбинация классификатора и набора признаков. Особое внимание при этом уделяется проверке статистической значимости полученных результатов.

Предложенный в работе подход основан, главным образом, на максимальном использовании данных, полученных во время анализа корпуса, при построении моделей. Показано, что с помощью этих данных может быть повышена эффективность работы системы в отношении некоторых аспектов ресторанов. Так, например, использование размеченного словаря тональностей, словарей номинаций, предикативно-атрибутивных словарей и словарей модификаторов в целом приводит к повышению точности и средней взвешенной F1-меры при извлечении мнений о качестве еды и обслуживания.

В результате экспериментов с машинным обучением были выделены классификаторы, показавшие себя наиболее эффективными для анализа тональности отзывов о ресторанах. Для большинства аспектов ресторанов логистическая регрессия и линейный метод опорных векторов оказываются оптимальными методами классификации отзывов на релевантные и нерелевантные относительно исследуемого аспекта. Для классификации отзывов на категории по отношению к исследуемому аспекту для большинства аспектов ресторанов наиболее эффективным становится мультиномиальный наивный байесовский классификатор.

Направления будущей работы, помимо расширения размеченного корпуса, включают в себя более детальную проработку используемых семантических ресурсов, в частности, словаря тональностей. Планируется проведение экспертной разметки словаря в разрезе определенных аспектов ресторанов и с добавлением особой глагольной разметки.

Работа выполнена при поддержке гранта СПбГУ 30.38.305.2014/

Литература

[1] Bakliwal A., Patil A., Arora P., Varma V. Towards Enhanced Opinion Classification using NLP Techniques // Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP, pp. 101–107, 2011.

[2] Benamara F., Cesarano C., Picariello A., Reforgiato D., Subrahmanian V. S. Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone // Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007.
[3] Bermingham A., Smeaton A. Classifying Sentiment in Microblogs: Is Brevity an Advantage? // Proceedings of the International Conference on Information and Knowledge Management (CIKM), 2010.
[4] Das S. R., Chen M. Y. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web // Management Science, vol. 53 (9), pp. 1375–1388, 2007.
[5] Dave K., Lawrence S., Pennock D. M. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews // Proceedings of the 12th International Conference on World Wide Web, pp. 519–528, 2003.
[6] Davidov D., Tsur O., Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys // Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, pp. 241–249, 2010.
[7] Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets // Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.
[8] Gatterbauer W., Bohunsky P., Herzog M., Krüpl B., Pollak B. Towards Domain-Independent Information Extraction from Web Tables // Proceedings of the 16th International Conference on World Wide Web, pp. 71–80, 2007.
[9] Kennedy A., Inkpen D. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters // Computational Intelligence, 2006.
[10] Marchand M., Ginsca A. L., Besançon R., Mesnard O. [LVIC-LIMSI]: Using Syntactic Features and Multi-polarity Words for Sentiment Analysis in Twitter // Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 418–424, 2013.
[11] Narayanan V., Arora I., Bhatia A. Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model, 2013.
[12] Pak A., Paroubek P. Language Independent Approach to Sentiment Analysis (LIMSI Participation in ROMIP'11) // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог», № 11 (18), Изд-во РГГУ, М.: 2012, с. 37–50.
[13] Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002.
[14] Pang B., Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Infomation Retrieval, vol. 2 (1–2), pp. 1–135, 2008.
[15] Pronoza E., Yagunova E., Lyashin A. Restaurant Information Extraction for the Recommendation System // Proceedings of the 6th Language Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 2nd Workshop on Social and Algorithmic Issues in Business Support: "Knowledge Hidden in Text", 2013.
[16] Pronoza E., Volskaya S., Yagunova E. Corpus-based Information Extraction and Opinion Mining for the Restaurant Recommendation System. Proceedings of the 2nd Statistical Language and Speech Processing. L. Besacier et al. (Eds.): SLSP LNAI 8791, pp. 272–284, 2014.

[17] Saif H. Sentiment Analysis of Microblogs. Mining the New World. Technical Report KMI-12-2, 2012.
[18] Shah K., Munshi N., Reddy P. Sentiment Analysis and Opinion Mining of Microblogs. 2013. Retrieved from: http://www.cs.uic.edu/~preddy/dm1.pdf, Access date: 28 May 2014.
[19] Sidorov G., Miranda-Jiménez S., Viveros-Jiménez F., Gelbukh A., Castro-Sánchez N., Velásquez N., Díaz-Rangel I., Suárez-Guerra S., Treviño A., Gordon J. Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets // Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence, Springer-Verlag, Berlin, Heidelberg, pp. 1–14, 2013.
[20] Sidorov G., Velasquez F., Stamatatos E., Gelbukh A., Chanona-Hernández L. Syntactic N-grams as Machine Learning Features for Natural Language Processing // Expert Systems with Applications, vol. 41 (3), 2014, pp. 853–860, 2014.
[21] Socher R., Perelygin A., Wy J. Y., Chuang J., Manning Ch. D., Ng A. Y., Potts Ch. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank // Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013.
[22] Sui H., Khpp Ch., Chan S. Sentiment Classification of Product Reviews Using SVM and Decision Tree Induction // Proceedings of the 14th Annual SIG CR Workshop, 2003.
[23] Wakade S., Shekar Ch., Liszka K. L., Chan Ch.-Ch. Text Mining for Sentiment Analysis of Twitter Data. Retrieved from: http://worldcomp-proceedings.com/proc/p2012/IKE3997.pdf, Access date: 28 May 2014.
[24] Wang S., Manning Ch. D. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers – vol. 2, pp. 90–94, 2012.

Aspect-based Restaurant Sentiment Analysis for E-Tourism Recommender Systems

E. Pronoza[1], E. Yagunova[1]
[1]Saint-Petersburg State University

This paper presents an aspect-based approach towards sentiment analysis of reviews about restaurants for e-tourism recommender systems. The proposed method can be applied not only in the recommender systems but in the other text understanding systems where the data is unstructured and the texts belong to colloquial style. We conduct a series of experiments with various feature sets and classifiers, and it is shown that the use of resources learnt from the corpus of reviews leads to the improvement of the restaurant aspects classification models. Naïve Bayes appears to be the best choice for identifying sentiment class, while Logistic Regression and SVM are best at deciding on the relevance of a review with respect to the particular restaurant aspect.

Keywords: sentiment analysis, reviews about restaurants, machine learning, recommender system.