



Выявление мнений и оценка их тональности

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна

Содержание



- Задачи *Opinion Mining* и *Sentiment Analysis*
- Анализ тональности на основе оценочных слов и выражений
- Подходы к оценке тональности
- Аспектный анализ мнений
- Оценка качества анализа тональности

Извлечение мнений и анализ тональностей



- Извлечение мнений (*Opinion Mining*)
мнение – суждение автора по поводу некоторого объекта/услуги/продукта/товара/организации/...
- Оценка (анализ) тональности (*Sentiment Analysis*)
тональность – авторская эмоциональная оценка
- Используется при:
 - извлечении и оценке мнений о персонах, партиях, товарах, услугах, ...
 - анализе новостей, обзоров, сообщений, ... пользователей
- *Opinion Mining* и *Sentiment Analysis* тесно связаны
- Исследования для английского языка с 1997 г.

Способы выражения мнений



- ◆ Использование оценочных слов (с явно положительными/отрицательными ассоциациями)
*Ненавижу эти очереди, а они там всегда
Снято все очень **красиво и качественно***
- ◆ Сравнительные оценки
*Coke tastes **better than Pepsi**
Книга была **лучше***
- ◆ Сарказм, ирония
*Сбербанк – **лучший в мире** производитель
пластиковых карточек для отскабливания льда
"Судьба дарит герою второй шанс. Все считали
его погибшим..." какой **неожиданный поворот**
событий!*

Выражение мнений: сложности



- ◆ Оценочные слова в разных контекстах могут иметь разные смыслы
cool story – cool heater
мягкая мебель – мягкий диван
- ◆ Оценочные слова могут не нести оценки
If I see a good book, I'll tell you
Посоветуйте мне ресторан, в котором вкусно кормят
- ◆ Предложения без оценочных слов могут нести оценку
The car uses a lot of gas
...попался волос в хлебе



Извлечение мнений

- Нахождение текстовых фрагментов, выражающих мнение (разновидность извлечения информации)
 - ✧ распознавание определенных языковых конструкций (прямая и косвенная речь)
 - ✧ поиск эмоционально-окрашенных слов и фраз
- Выявление свойств мнений
 - ✧ объект мнения
 - ✧ эмоциональная оценка (тональность)
 - ✧ метайнформация: автор мнения; время, в которое мнение было высказано
- ◆ Мнением может быть: часть предложения, всё предложение или текст в целом



Анализ тональности

Можно анализировать:

- ☐ Тональность документа
- ☐ Тональность предложения
- ☐ Тональность суждения об объекте и его характеристиках – **аспектный анализ**

Виды оценок:

- ❖ Двоичная: позитивная/негативная
- ❖ Троичная: позитивная/негативная/нейтральная
- ❖ По некоторой шкале

Подходы к анализу:

- Инженерный подход
- Машинное обучение
- Гибридный подход

Анализ тональности: инженерный подход



- Применение методов, основанных на правилах и словарях оценочной лексики
- Оценочная лексика существенно зависит от ПО:
 - ❖ оценочные слова в товарах и услугах отличаются от таковых в новостях
удобный, быстро vs. *катастрофа, успешный*
 - ❖ оценочные слова в одной области могут иметь разный смысл
большой "+" для диска, "-" для «жучка»
- В интернет-текстах много неформальной лексики, которой нет в словарях; ее тональность неясна
- ➔ Важная задача: составление словарей оценочных слов для конкретных ПО

Примеры оценочной лексики мнений о фильмах



Отдельные слова

- прилагательные: *пристойный, отстойный*
- существительные: *творение, нудятина, фигня*
- глаголы: *восторгать, взбесить*
- наречия: *выигрышно, трешово* ("+" или "—"?)

Фразы: *обязателен к просмотру, неожиданный поворот, лишен достоинств, русское кино*

Элементы со специальной функцией

- усиление: *очень, особенно*
- отрицание: *не, ни, никакой*
- ослабление: *вряд ли, не особенно*
- предположение: *если, ли, бы*

Простейший метод анализа тональности



1. Оцениваем каждое слово анализируемого фрагмента (проставляем веса), например: +1 в случае позитивной тональности и -1 в случае негативной
2. Суммируем оценку (веса) всех слов
3. По полученному числу узнаем тональность

Хороший⁺, трешовый[?] фильм, с отличным⁺ чувством юмора. Для любителей гая ритчи самое то, вот только картинка нищенская⁻, но ничего страшного⁺. Это даже колорит⁺ какой-то придает.



Вариации в методах

■ Правила оценивания: можно учитывать словосочетания

Прилагательное	Существительное	Именная группа	Пример
Негативное	Позитивное	Негативное	Неоправданная надежда
Негативное	Негативное	Негативное	Ужасный лгун
Позитивное	Позитивное	Позитивное	Хороший друг
Позитивное	Негативное	Негативное	Безупречные поражение
Позитивное	Нейтральное	Позитивное	Безупречная еда
Негативное	Нейтральное	Негативное	Ужасная еда

■ Значения весов:

- сила тональности (от -5 до +5, 0 – нейтральная)
- вероятность оценочности слова
- частота совместной встречаемости с +/- оценочными словами

■ Формула подсчета суммы

Составление словаря оценочной лексики



- Вручную экспертами
 - ✓ выписываются слова и выражения, указывается их тональность
 - ✧ надежно, **НО** трудоемко и сложно учесть все способы выражения оценки
- Автоматизированно
 - ✓ на основе правил, шаблонов и контекстов
 - ✓ с учетом статистики по корпусу
 - ✓ с использованием словарей
 - ✧ Проблема: как назначить тональность (*sentiment orientation*, *SO*) выявленным единицам?
- Комбинация этих подходов

Автоматизированное построение: идеи

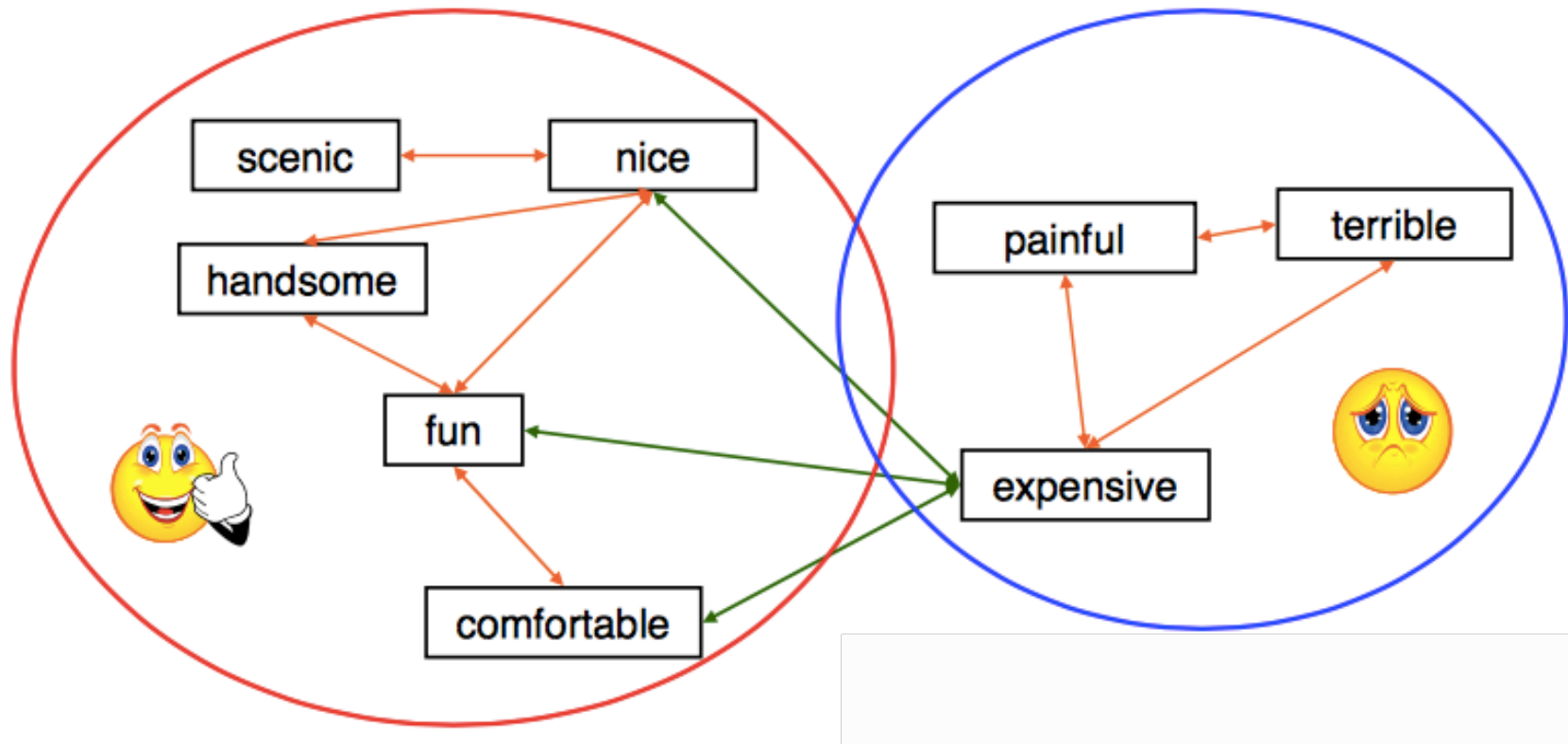


- Учет однородных синтаксических конструкций
радостный и светлый
красивый, но громкий
- Применение методов классификации/
кластеризации
- Учет совместной встречаемости (варианты
MI, log-likelihood)
- Использование словарей со связями:
синонимы, антонимы, однокоренные слова
радостный – безрадостный, радость

Классификация/ кластеризация слов



- Классификация слов: оценочное или нет
- Кластеризация слов: позитивное или негативное



Учет совместной встречаемости



Предположение: слова со схожей семантикой часто встречаются совместно

{*excellent*, *poor*}

Altavista

Seeds: **S+**, **S-**

Separate Corpus

co-occurrence statistics

Pointwise Mutual Information (PMI)

$$PMI(w_1, w_2) = \frac{P(w_1 \text{ NEAR } w_2)}{p(w_1)p(w_2)};$$

$$SO(w) = PMI(w, \textit{excellent}) - PMI(w, \textit{poor})$$

Opinion Words = Seeds + Similar words

- Задаем базовые слова с тональностью
- Тональность (SO) вычисляем по формуле

Извлечение оценочной лексики из WordNet



- Задаем базовые слова с тональностью
- Идем по отношениям
- Для синонимов/гипонимов – такая же тональность, для антонимов – противоположная

Synonym **gloss**

(17) **S:** (adj) **excellent**, **first-class**, **fantabulous**, **splendid** (very good; of the highest quality) "made an excellent speech"; "the school has excellent teachers"; "a first-class mind"

Similar word

- **similar to**
 - **S:** (adj) **superior** (of high or superior quality or performance) "superior wisdom derived from experience"; "superior math students"
- **derivationally related form** **Morphologically related word**
 - **W:** (n) **excellency** [Related to: **excellent**] (an outstanding feature; something in which something or someone excels) "a center of manufacturing excellency"; "the use of herbs is one of the excellencies of French cuisine"
 - **W:** (n) **excellence** [Related to: **excellent**] (the quality of excelling; possessing good qualities in high degree)
 - **W:** (v) **excel** [Related to: **excellent**] (distinguish oneself) "She excelled in math"
- **antonym** **antonym**
 - **W:** (adj) **inferior** [Indirect via **superior**] (of low or inferior quality)

antonym

Словари оценочной лексики для английского языка



- Питсбургский словарь эмоционально-окрашенных слов английского языка
- Словари National Research Council (NRC) Canada:
 - ❖ **Word-Emotion Association Lexicon**
 - ❖ **Hashtag Emotion Lexicon**
 - ❖ **Hashtag Sentiment Lexicon** и пр.
- Тезаурусы, построенные на базе WordNet:
 - ❖ **WordNet-Affect** – ручная разработка
 - ❖ **SentiWordNet** – автоматическая разработка
- Тезаурус со связями **SenticNet**

Питсбургский словарь



■ 2005 год

■ Основные поля:

- ✓ тональность: negative, positive и neutral
- ✓ сила тональности: weaksubj и strongsubj
- ✓ часть речи

type=weaksubj word1=**bug** pos1=verb priorpolarity=negative

type=strongsubj word1=**cry** pos1=verb priorpolarity=negative

type=strongsubj word1=**abidance** pos1=adj priorpolarity=positive

type=weaksubj word1=**abilities** pos1=noun priorpolarity=positive

type=weaksubj word1=**analytical** pos1=adj priorpolarity=neutral

type=strongsubj word1=**emotion** pos1=adj priorpolarity=neutra

Word-Emotion Association Lexicon



- 2010 г., 14 182 слов. ПО не ограничена
- Список английских слов и их ассоциаций с
 - ❖ эмоциями (emotions): anger, fear, surprise, trust, sadness, disgust, joy, anticipation
 - ❖ чувствами (senses): positive, negative
- Ручная аннотация, краудсорсинг
- 0 – нет ассоциации, 1 – есть ассоциация

address	anger	0	happy	anger	0
address	anticipation	0	happy	anticipation	1
address	disgust	0	happy	disgust	0
address	fear	0	happy	fear	0
address	joy	0	happy	joy	1
address	negative	0	happy	negative	0
address	positive	0	happy	positive	1
address	sadness	0	happy	sadness	0
address	surprise	0	happy	surprise	0
address	trust	0	happy	trust	1

Словари, автоматически составленные по Twitter



Hashtag Emotion Lexicon, 2013 г.

- Хештеги содержали эмоциональную лексику
- 16 862 слов, указаны emotions и численная сила ассоциации (от 0 до ∞)

anger	#nothappy	1.04368166876984
joy	#happyheart	2.16828383214492
disgust	each	3.18164356645913e-06

Hashtag Sentiment Lexicon, 2013 г.

- Хештеги содержали оценочную лексику
- 316 531 биграмм, значение тональной оценки (от $-\infty$ до $+\infty$), частота +/- связи

. #beauty	6.846	583	1
hurt #bad	-4.999	0	14
2 tickets	0.766	4	3

Тезаурусы на основе WordNet



WordNet-Affect содержит 2,874 синсетов и 4,787 слов

- Синсеты, представляющие эмоции, вручную размечены с помощью меток (A-labels)
- Проставлены дополнительные эмоциональные метки: позитивная, негативная, неоднозначная и нейтральная

A-Labels	Examples
EMOTION	<i>noun anger#1, verb fear#1</i>
MOOD	<i>noun animosity#1, adjective amiable#1</i>
TRAIT	<i>noun aggressiveness#1, adjective competitive#1</i>
COGNITIVE STATE	<i>noun confusion#2, adjective dazed#2</i>
PHYSICAL STATE	<i>noun illness#1, adjective all in#1</i>

SentiWordNet получен путем автоматического проставления каждому синсету оценки

- Каждая из оценок определяет объективную, позитивную или негативную составляющую синсета, принимает значения от 0 до 1 и в сумме дают 1

Тезаурус SenticNet



- С 2010 года, связывает более 14000 понятий на семантическом уровне
- Данные кодируются в RDF-триплеты с использованием синтаксиса XML
- Сопоставляет каждому понятию:
 - ❖ «sentic-вектор» с численными значениями Pleasantness, Attention, Sensitivity и Aptitude
 - ❖ величины основного и дополнительного настроения
 - ❖ набор эмоционально близких понятий
- Если в процессе работы встретится понятие *день рождения*, то SenticNet отнесет его к понятию высокого уровня *события* и свяжет с близкими понятиями *сладкий*, *дружеский сюрприз*, *клоун*, ...

SenticNet: представление данных



love 1 0 0 1 #joyful #admirable positive 0.667

```
- <rdf:RDF>
- <rdf:Description rdf:about="http://sentic.net/api/en/concept/love">
  <rdf:type rdf:resource="http://sentic.net/api/concept"/>
  <text>love</text>
  <semantics rdf:resource="http://sentic.net/api/en/concept/lust"/>
  <semantics rdf:resource="http://sentic.net/api/en/concept/love_another_person"/>
  <semantics rdf:resource="http://sentic.net/api/en/concept/sexuality"/>
  <semantics rdf:resource="http://sentic.net/api/en/concept/beloved"/>
  <semantics rdf:resource="http://sentic.net/api/en/concept/show_empathy"/>
  <pleasantness rdf:datatype="http://w3.org/2001/XMLSchema#float">1</pleasantness>
  <attention rdf:datatype="http://w3.org/2001/XMLSchema#float">0</attention>
  <sensitivity rdf:datatype="http://w3.org/2001/XMLSchema#float">0</sensitivity>
  <aptitude rdf:datatype="http://w3.org/2001/XMLSchema#float">1</aptitude>
  <moodtag>#joyful</moodtag>
  <moodtag>#admirable</moodtag>
- <polarity>
  <value>positive</value>
  <intensity rdf:datatype="http://w3.org/2001/XMLSchema#float">0.667</intensity>
  </polarity>
</rdf:Description>
</rdf:RDF>
```

Словари для РЯ



- Питсбургский словарь автоматически переведен на РЯ
- Словари NRC Canada автоматически переведены на 39 языков: Arabic, Basque, ..., **Russian**, ..., Zulu
- Тезаурусы на основе WordNet: есть перевод на РЯ
- Словарь Четверкина, РЯ: оценочные слова из отзывов на фильмы, книги, игры, телефоны, камеры

БЕСПОДОБНЫЙ 0.963

ОЖИДАТЬ 0.429

ЖАЖДАТЬ 0.145

- В НКРЯ есть пометы ev:posit (*умница*) и ev:neg (*негодяй*). Их можно учитывать при поиске
- **Linis Crowd** – краудсорсинговый веб-ресурс для РЯ

Linis Crowd: интерфейс



<http://linis-crowd.org>, с 2014 г.

Разметка

затвор

Тональность

Оцените текст:

-2 -1 0 1 2

Оцените слово:

-2 -1 0 1 2

Текст для разметки

Того же завода. Но я взял только заради самого ТТ - все-таки железка 42-го года выпуска :). 7.62x25 мм ТТ - вообще-то, тот самый патрон, под который и был сконструирован этот пистолет-пулемёт... Аутентичность, сэр, АУТЕНТИЧНОСТЬ!) Типа, с дисковым магазином - оно аутентичнее и историчнее?...) Тогда уж и балласт в этот магазин добавить - чтоб и по весу совпадало!) А кто бы спорил :) просто люгеровский 9x19 мальца получше будет, распространённое в нынешние времена и дешевле :) Да-да-да =)Я держал в руках неснаряжённый ППШ - тяжесть та ещё. Да? О как... а я был уверен и даже не интересовался особо =) ну, был - значит, был))) Ну, точно так же, как блокировали автоматический огонь на всех переделках автоматов под гражданское оружие - АКМ, АК-74, АКМС, М16... Но ведь тогда пришлось бы переделывать и ствол, и **затвор**, и магазин... И ещё неясно, как бы себя повело оружие после столь серьёзных переделок. :-/ Патрон-то мощнее ведь... А кстати... вот интересно. С полностью снаряженным дисковым магазином это оружие - ведь гораздо массивнее, так?.. Вот начинаем из него стрелять. По мере опустошения магазина масса оружия уменьшается, так?.. Стрельба - не со станка, оружие ходит . Наверняка же это изменение массы оружия влияет на точность боя не лучшим образом... Так что имхо дисковый магазин, переделанный на 10 патронов и утяжелённый до массы полностью снаряженного - имхо даже лучше, чем натуральный.

Продолжить оценку

Предложить новое слово

Зарегистрироваться

Linis Crowd: статистика



- Размечают добровольцы, проверяют лингвисты
- Тем, кто разметит наибольшее количество текстов и слов, обещан денежный приз
- Размеченные данные можно скачать после регистрации

Оценка тональности проверенных текстов

Всего 9702 слов в базе

Всего 29106 текстов в базе

Статистика оценок

Кол-во оценок	Кол-во слов
0	7136
1	2200
2	328
3	32
4	2
5	1
8	1
13	1
51	1
Итого	9702

Всего 2810 текста с оценками

Всего 26296 текстов без оценок

Всего 3037 оценок слов

Всего 3037 оценок текстов

Анализ тональности: машинное обучение



- Стандартная задача классификации
 - ✓ признаки – слова, части речи, биграммы, позиции в тексте, роль в предложении, семантические классы
 - ✓ разные методы классификации (байесовский подход, SVM, логистическая регрессия, HMM, ...)
- Методы кластеризации используются реже: учитываем слова, которые часто встречаются в данном тексте, но редко в большинстве других

Аспектный анализ мнений



“Комфортный отель”

●●●●○ Отзыв написан 26 июня 2014

Хорошее расположение, весьма уютные номера, на удивление хороший завтрак (если сравнивать в целом с Европой, но может в Швейцарии это везде так). Чай, кофе, чайник в номере, сервис хороший. Вай-фай правда пару раз не работал.

- **Сущность** – объект мнения (продукт, сервис, тема, человек, организация, ...)
- **Аспекты** – составные части, свойства, характерные признаки и черты сущности
- ◆ Мнение касается определенной сущности (*отель*)
- ◆ В нем может высказываться суждение о сущности в целом и/или о ее отдельных аспектах (*расположение, номера, завтрак, сервис, ...*)

Пример сущности *отель* и ее аспектов



Сущность,
аспект GENERAL

Отель

Аспектные категории

Пляж и
бассейн

Питание

Территория

Номер

Сервис

Море
Песок
Лежак
Зонтик

Газировка
Сок
Выпечка
Повар

Аквапарк
Аниматор
Анимация
Бильярд

Балкон
Ванная
Вентиляция
Диван

Администратор
Бронирование
Горничная
Массаж

Аспектные термины

Упоминания аспектов в тексте мнения



- **Явные:** указывают на аспекты, но не содержат в себе оценку
номер, завтрак, wi-fi, СПА
- **Неявные:** сочетают аспект и оценку в одном слове
вкусно = хорошая еда
уютно = хороший интерьер, обстановка
- **Тональные факты:** описывают состояние дел и несут в себе оценку
чайник в номере, одно одеяло, хорошая слышимость

Задачи аспектного анализа



1. Выявление аспектных терминов
 - ❖ опора на частоту существительных и их групп
 - ❖ учет конструкций с оценочной лексикой
 - ❖ машинное обучение
 - ❖ гибридный подход
2. Объединение аспектных терминов, ссылающихся на один аспект
3. Определение тональности:
 - ❖ для мнения
 - ❖ для каждого аспекта
 - ❖ для сущности в целом
4. Обобщение полученной информации

Частотный подход к выявлению аспектов: идея



- ❑ Аспекты выражаются одинаковыми словами, все остальное – разными
- ❑ Малочастотные единицы – не аспекты или маловажные аспекты
- ❑ Наиболее часто встречающиеся существительные и именные группы – аспекты (аспектные термины)



Порог отсеечения – 30

Частотный подход: особенности



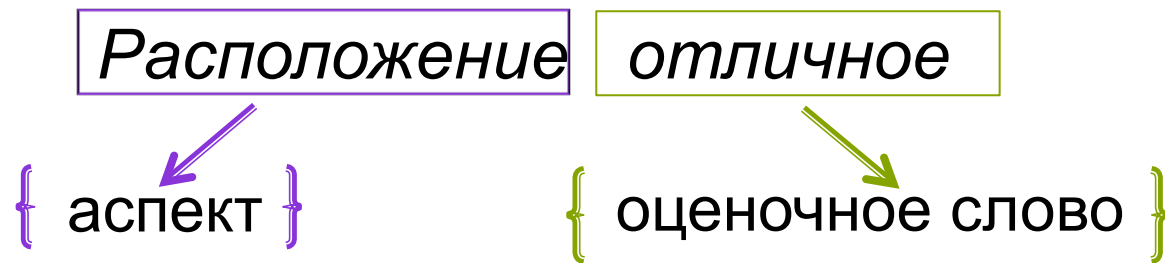
- + Простой, но достаточно эффективный, т.к. от 60 до 70% аспектов выражаются существительными и именными словосочетаниями
- Выделяет много неаспектов
- Пропускает низкочастотные аспекты
- Требуется ручная настройка порога отсечения
- Способы учета частоты:
 - ✓ абсолютное значение
 - ✓ tf-idf на уровне абзаца и документа
 - ✓ tf-idf в рассматриваемом и контрастном (общезыковом) корпусе

Инженерный подход: идея

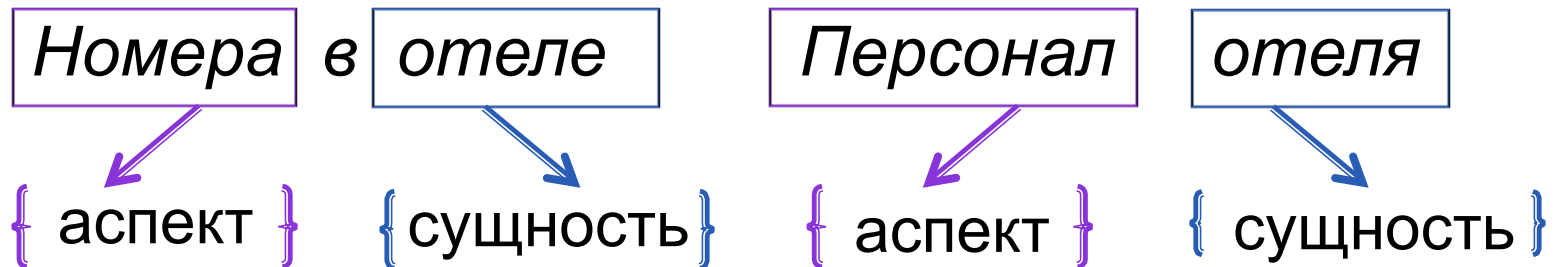


С помощью правил/шаблонов извлекаем существительное или именное словосочетание

- ближайшее к некоторому оценочному слову



- употребляющиеся рядом с названием сущности



Инженерный подход: особенности



- + Опирается на оценочные слова, которые часто известны или их легко найти
- + Выявляет малочастотные аспекты
- Зависит от полноты словаря оценочных слов
- Выделяет много неаспектов
- Требуется вручную размеченной обучающей выборки
- Составление правил/шаблонов:
 - ✓ полностью вручную
 - ✓ с помощью машинного обучения

Машинное обучение



- Выделение аспектов – задача классификации
- Признаки: конкретные слова, части речи, расстояние между словами и т.д.
- Методы: SVM, HMM, нейронные сети и т.д.
- + Не требует ручной настройки параметров
- + Не пропускает низкочастотные аспекты
- + Редко выделяет лишние аспекты
- Необходимы размеченные данные
- Плохо переносимо на другие ПО
- Результаты работы плохо объяснимы

Гибридный подход



1. С помощью частотных характеристик выделяем именные словосочетания
 2. Отбираем из них аспекты с помощью шаблонов
- + Уменьшает число неаспектов
 - Пропускает редко встречающиеся аспекты
 - Требуется ручное составление правил
 - Требуется ручной настройки порога отсечения
 - Шаги 1 и 2 можно поменять местами

Объединение аспектных терминов



- Один аспект может выражаться разными именными словосочетаниями

Сервис: персонал, администратор, официант, уборщица и т.д.

- Подходы к объединению:

- ☐ учет семантических отношений: синонимия, род-вид
- ☐ использование метрик схожести фраз: расстояние Хэмминга, Левенштейна
- ☐ машинное обучение

Определение тональности



- Определяем тональность:
 - ❖ для мнения
 - ❖ для конкретного аспекта
 - ❖ для сущности в целом
- Подходы: инженерный подход, машинное обучение
- Сложности анализа тональности:
 - для разных областей нужны разные коллекции
 - необходимо отнести оценку к конкретному аспекту

Определение тональности: методы



- **Тональность мнение:** классические подходы
- **Тональность для аспекта:**
 - определяется как объединение тональности всех предложений, описывающих аспект
 - учитывают линейное расстояние, синтаксическое дерево
- **Тональность для сущность в целом:**
 - по аспектам: обобщаем информацию о тональности каждого аспекта
 - по сущности (аспект GENERAL): тональность определяется как объединение тональности всех предложений, описывающих сущность

Обобщение и визуализация информации



Полученная из отзывов информация о сущности, аспектах и их тональности м.б. представлена в:

- структурированном виде: таблицы, диаграммы, облака тегов
- неструктурированном виде: аннотация, краткий реферат, например:
 - ❖ выделение наиболее значимых и содержащих оценку предложений из отзыва
 - ❖ заполнение шаблона

*Расположение отеля _____. Кухня _____. Персонал _____.
Пляж находится _____. Уборка была _____.
Вид из окна _____. Номера _____. Цена за номер _____.*

Структурированный вид: таблица



Отель 1:

Сущность: **отель**

Положительных: 105 <предложения из отзывов>

Отрицательных: 12 <предложения из отзывов>

Аспект: **расположение**

Положительных: 95 <предложения из отзывов>

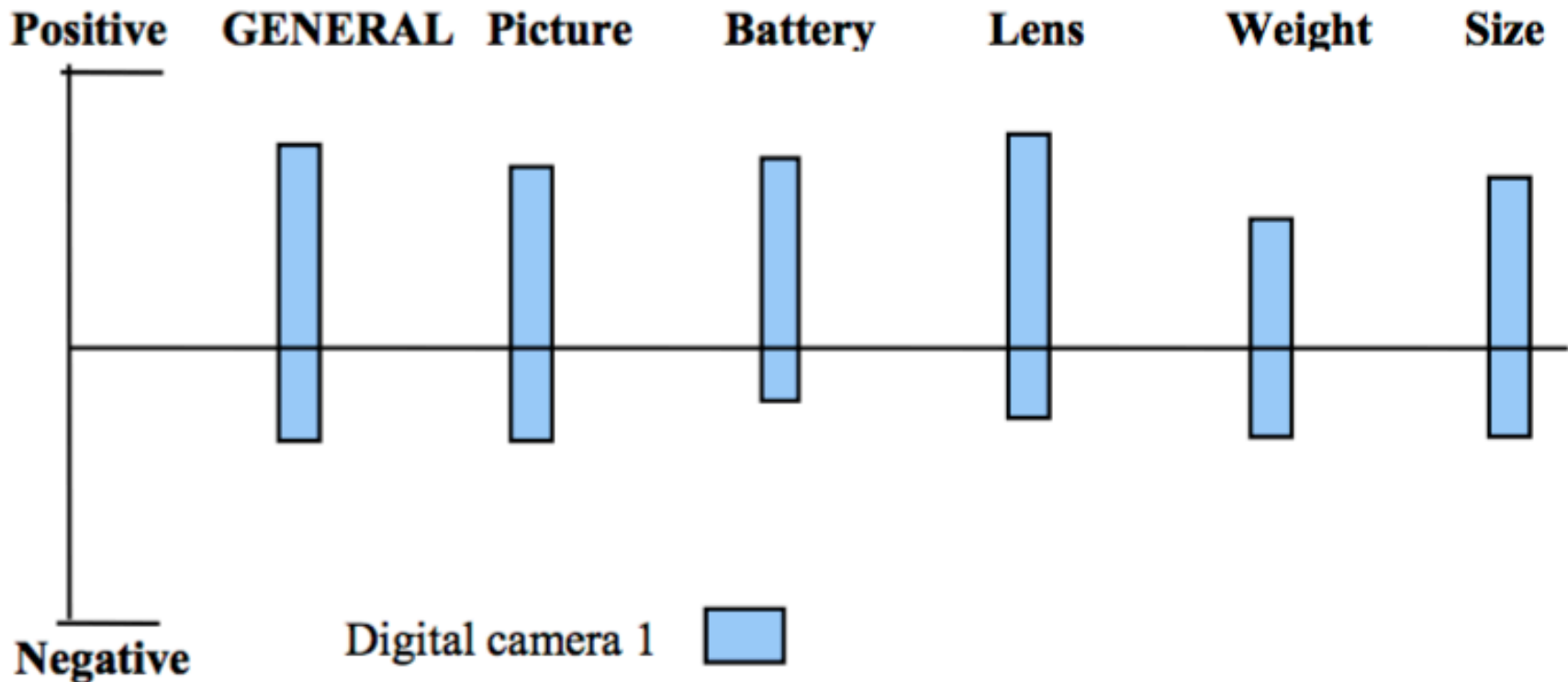
Отрицательных: 10 <предложения из отзывов>

Аспект: **номер**

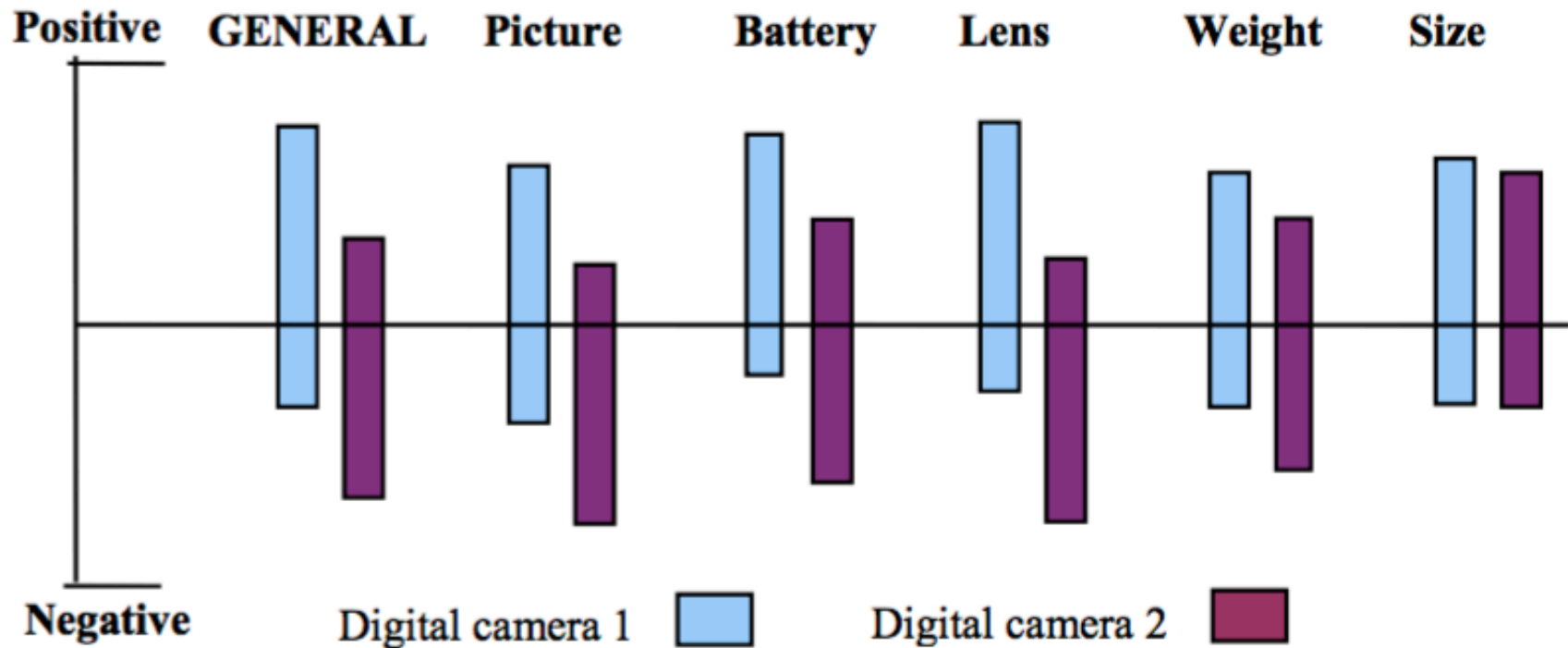
Положительных: 50 <предложения из отзывов>

Отрицательных: 9 <предложения из отзывов>

Структурированный вид: диаграмма



Структурированный вид: сравнительная диаграмма



Структурированный вид: диаграмма с оценками



Canon PowerShot SX40 HS 12.1 MP Digital Camera

\$355 online

★★★★★ 482 reviews  +25 Recommend this on Google

#1 in Digital Cameras

September 2011 - Canon - Point & Shoot - 12.1 megapixel - Electronic Viewfinder - Compact Sensor - 35 x optical zoom
- CMOS - Pop-up Flash - ISO 3200

[← Back to overview](#)

Reviews

Summary - Based on 482 reviews



What people are saying

pictures		"Picture clarity is great."
features		"Great images, great features, easy to use."
zoom/lens		"This is a great low end camera for Canon DSLR users."
design		"Another point is the overall camera speed."
video		"Great video quality."
value		"Amazing product with excellent price!!"
size		"The SX 40 is very easy to use.small learning curve."

Showing reviews that mention: Size - [Show all reviews](#)

Структурированный вид: облако тегов



CHI Papers and Notes are archival publications of original research in the field of Human-Computer Interaction (HCI). CHI Papers and Notes are read and cited worldwide. They have broad impact on the development of HCI theory, method and practice. Papers present significant contributions to research and practice in all aspects of Human Computer Interaction. Notes present brief and focused research contributions that are noteworthy. See Papers versus Notes: What is the difference?

Оценка качества систем анализа тональности



- ◆ Измеряют точность, полноту, F-меру, аккуратность
- ◆ Сравнивают с оценками экспертов
- ◆ Особенности задачи:
 - ✓ эксперты сходятся в оценке тональности в 79% случаев → для программы приемлемо 70% точность
 - ✓ сущность, оцененная одним автором как 5, другим может быть оценена как 7
 - ✓ при настройке на определенный тип текстов (ПО/тема/автор) качество выше

Аккуратность



Результаты работы системы можно разделить

эксперт система	правильные (positive – P)	неправильные (negative – N)
правильные (P)	True P = TP	False P = FP
неправильные (N)	FN	TN

- Аккуратность (*Accuracy*) – отношение правильно принятых решений к общему числу решений
$$A = (TP + TN) / (TP + FP + TN + FN)$$
- Возможности человека различать отзывы с близкими оценками по некоторой шкале:
 - если разница между отзывами 3 и более баллов – аккуратность 100%
 - два балла – 83%, один балл – 69%, ноль – 55%

Соревнования для русского языка: РОМИП 2011 И 2012



- Анализ тональности:
 - ❖ отзывов о фильмах, книгах, цифровых камерах; классификация на 2, 3 и 5 классов
 - ❖ в 2012 г. были добавлены цитаты из новостной статьи; классификация на 3 класса
- Наибольшая аккуратность:
 - для 2-х классов $A=96,6\%$ (камеры)
 - для 3-х классов $A=85,4\%$ (камеры)
 - для 5-х классов $A=62,6\%$ (камеры)
- Лучший метод машинного обучения – SVM
- Для анализа цитат лучше работают инженерные методы, наибольшая $A=61,6\%$

SentiRuEval 2015: определение аспектов



Лучший метод:

- ✓ слова w_j представляются в виде векторов (W2V)
- ✓ для каждого аспекта Ai из обучающих данных определяется начальный набор называющих его аспектных терминов
- ✓ для каждого w_j подсчитывается его схожесть (косинусная мера) с Ai
- ✓ если похожесть выше заданного порога, получаем новый аспектный термин

F=67,6% (точное совпадение, автомобили)

SentiRuEval 2015: определение тональности



Лучший метод:

- ✓ каждому w_j ставится в соответствие:
 - ✧ максимум похожести между w_j и эталонными позитивными/негативными словами
 - ✧ число (PMI), отражающее, насколько w_j чаще встречается рядом с позитивными словами чем с негативными
- ✓ слова w_j представляются в виде векторов, учитывается их контекст
- ✓ для классификации используется дерево принятия решений

F=56,8% (рестораны)

Заключение (1)



- Существует множество прикладных областей, для которых необходимо решать задачу извлечения мнений и оценки их тональности
 - определение субъективности/объективности текста
 - поиск цитат и их оценка
 - ответы на вопросы типа "Какова реакция Вануату на победу Трампа?"
 - мониторинг изменения мнений
- В зависимости от области постановка задачи меняется

Заключение (2)



Существует множество методов для ее решения:

- ◆ методы, основанные на машинном обучении: становятся все более сложными и громоздкими
- ◆ инженерный подход: активно используется поверхностный/полный синтаксический и семантический анализ
- ✓ есть методы, опирающиеся на теоретико-графовую модель



СПАСИБО ЗА ВНИМАНИЕ!

-

Система SentiStrength



- Анализирует короткие неструктурированные неформальные тексты
- Может работать с русским языком
- Результат – 2 оценки: позитивная (от +1 до +5) и негативная составляющая (от -1 до -5)
- Метод: поиск максимального значения тональности в тексте для каждой шкалы

The text 'We are the champions, my friends, and we'll keep on fighting 'til the end.' has positive strength **2** and negative strength **-3**

Approximate classification rationale: We are the champions ,my friends[2] ,and well keep on fighting[-3] til the end .[sentence: 2,-3] [result: max + and - of any sentence][overall result = -1 as pos<-neg] (Detect Sentiment)

Программа LIWC



- LIWC (Linguistic Inquiry and Word Count) разработана исследователями, интересующимися психологией
- В тексте подсчитывается процент слов, которые отражают различные эмоции, стили мышления и пр. и сравнивается с текстами из разных областей

TRADITIONAL LIWC DIMENSION	YOUR DATA	AVERAGE FOR SOCIAL MEDIA: TWITTER, FACEBOOK, BLOG
I-WORDS (I, ME, MY)	6.7	5.51
SOCIAL WORDS	26.7	9.71
POSITIVE EMOTIONS	6.7	4.57
NEGATIVE EMOTIONS	6.7	2.10
COGNITIVE PROCESSES	0.0	10.77
SUMMARY VARIABLES		
ANALYTIC	51.4	55.92
CLOUT	99.0	55.45
AUTHENTICITY	43.4	55.66
EMOTIONAL TONE	25.8	63.35