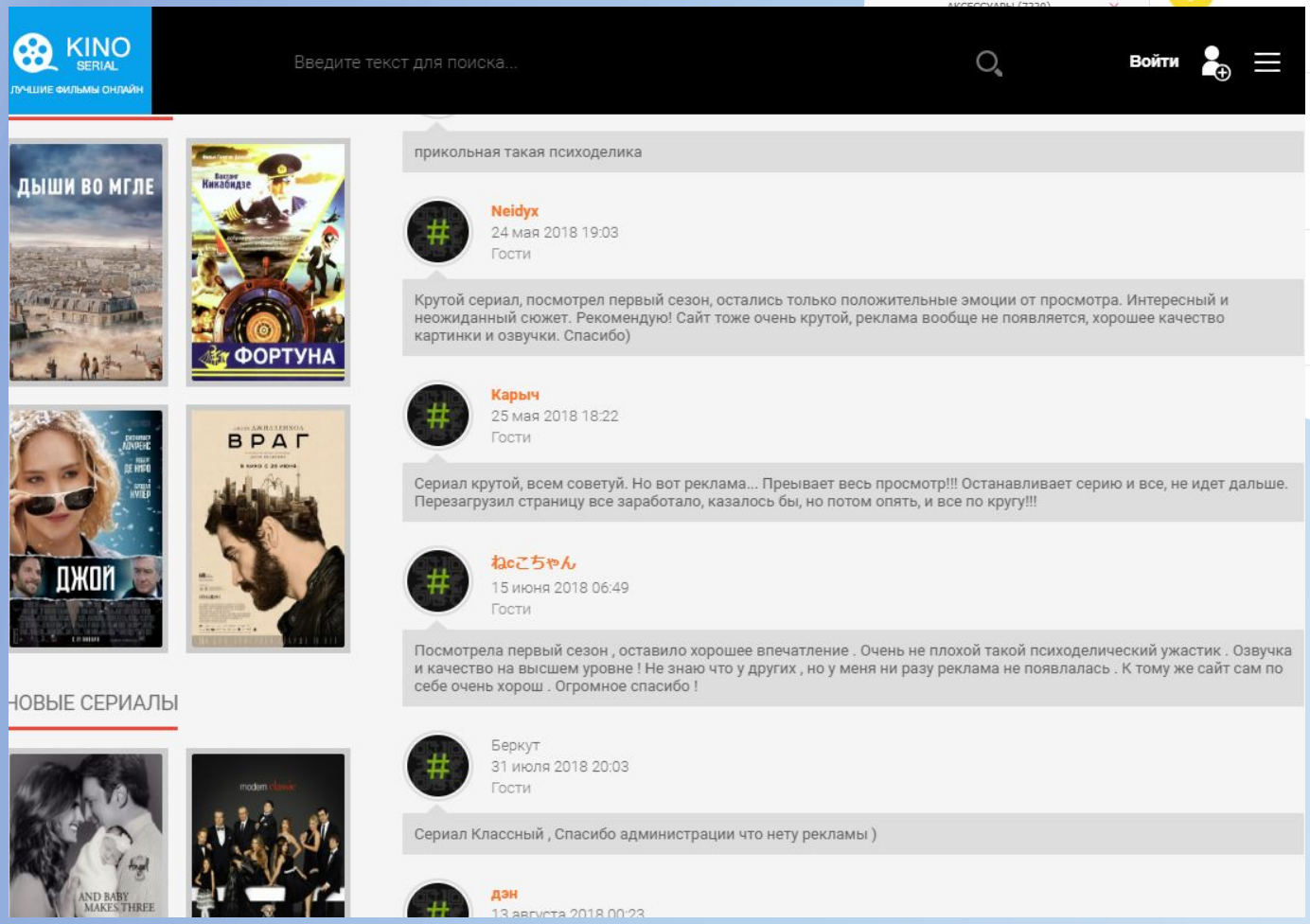
Системы автоматического анализа тональности

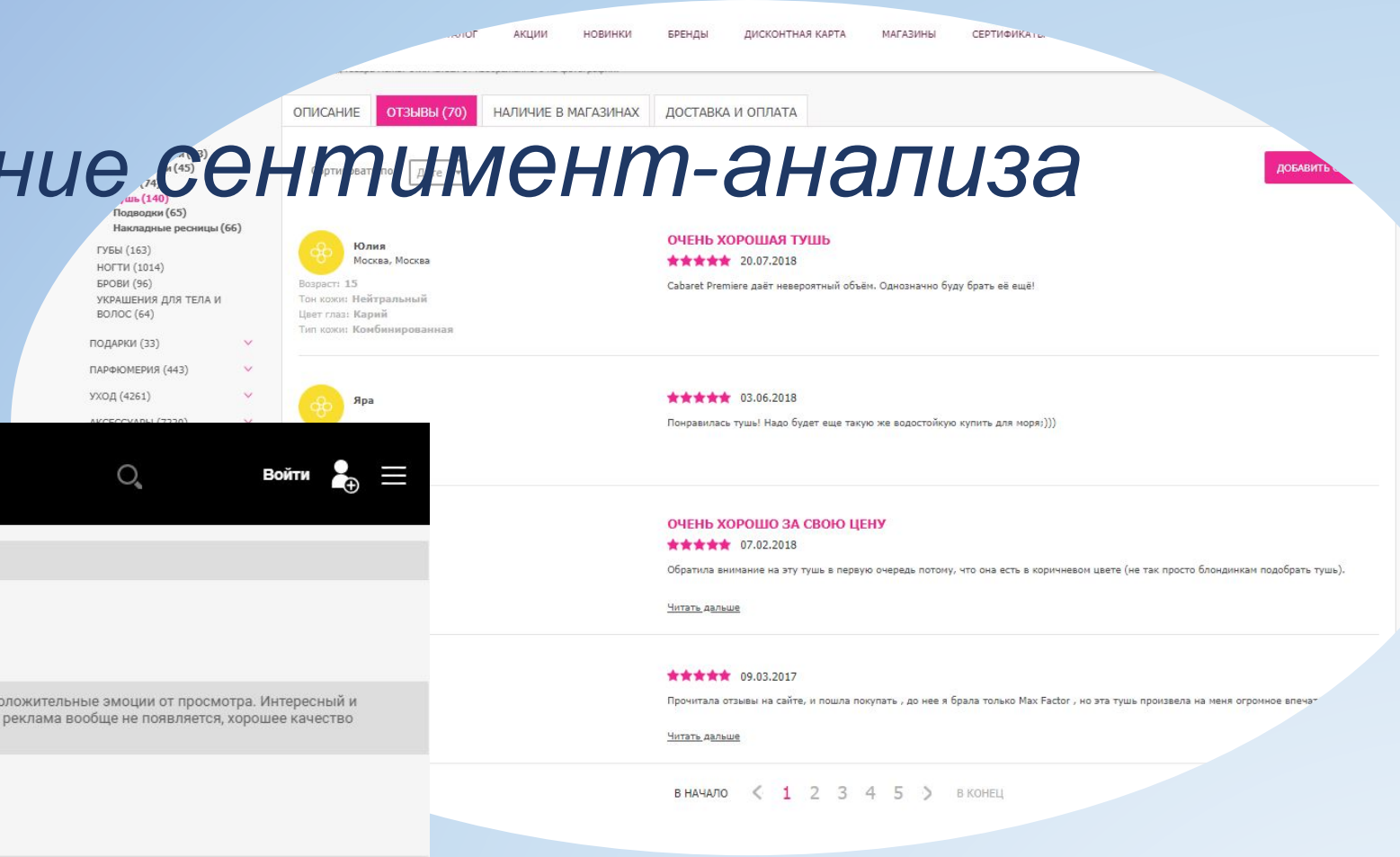
Ильина Александра

324 группа

Определение сентимент-анализа

Анализ тональности (сентимент-анализ) —





это область компьютерной лингвистики, которая занимается изучением мнений и эмоций в текстовых документах.

Зачем нужны системы автоматического анализа тональности?

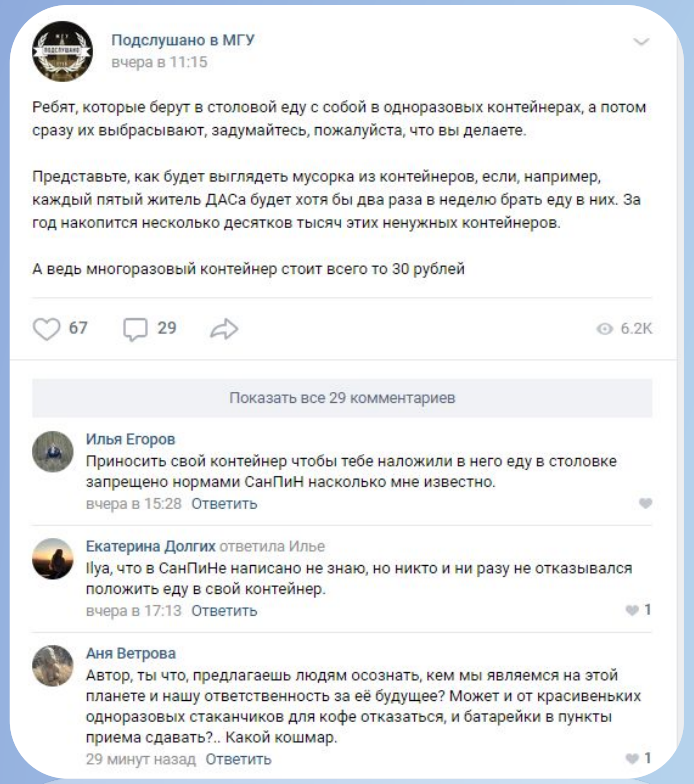
Системы анализа тональности и извлечения мнений находят своё практическое применение в таких областях как:

- **социология**: данные о религиозных взглядах населения;
- **политология**: мониторинг политических взглядов населения;
- **маркетинг**: анализ сообщений в Twitter на предмет того, какая модель ноутбуков пользуется наибольшим спросом;
- **медицина и психология**: сентимент-анализ может использоваться для определения депрессии у пользователей социальных сетей;
- **сфера финансов**: анализ тональности финансовых отчётов и финансовых новостей для определения трендов на фондовом и валютных рынках;
- **поиск спама в отзывах**;

а также в **журналистике**, **бизнесе**, и т.д.

Сбор корпуса данных. Предварительная обработка

Сбор корпуса данных текстов можно делать вручную, а можно использовать для этой цели специальные программы: **Webometric Analyst**, **Datacol**, **VKComment Parser** и др.



Эти программы могут также осуществлять следующую обработку текстов:

- **лемматизацию** - процесс приведения слов к нормальным (словарным) формам (далее будет удобнее искать их в словарях, выделять шаблоны и т. д.);

"такой интересной книги я давно не читала" =>

=>"такой интересный книга я давно не читать"

Предварительная обработка

- **стемминг** - избавление от суффиксов и окончаний:

  *«**мал**~~юсенький~~ экран»*));

- **удаление стоп-слов** - тех, которые часто встречаются, но практически не несут никакой эмоциональной нагрузки:

  - *предлоги (**в**, **на**, **под**);*
  - *некоторые местоимения (притяжательные: **его**, **мой**);*
  - *некоторые наречия (**туда**);*

  и др.

Предварительная обработка

- приведение к нижнему регистру (в некоторых случаях также теряется эмоциональный акцент):

«Приобретение данного товара было ОШИБКОЙ» =>

=> «Приобретение данного товара было ошибкой»

- морфологическая разметка (в текстовый корпус вставляются метаданные для обозначения частей речи и др.);

и т.д.

Выделение сущностей

При сентимент-анализе необходимо выделять следующие составляющие:

1) **субъект тональности** — источник мнения, тот, кто является, автором сообщения;

2) **объект тональности** — то, о чём идёт речь в тексте (фильм, модель ноутбука);

3) **аспект тональности** — характеристика объекта (например, для фильма это может быть: игра актёров, спецэффекты, сюжет, музыкальный ряд и т.д.);

4) **тональная оценка** — тип мнения, отношение автора к отдельному аспекту или к объекту в целом).

Классификация при сентимент-анализе

Текст: ***позитивный*** / **негативный**

Комментарий:

- ***грустный***
- ***радостный***
- ***злой***

Отзыв:

- ***положительный***
- ***нейтральный***
- ***отрицательный***

Подходы к автоматическому анализу тональности

Основные подходы к автоматическому определению тональности текста можно разделить на 2 большие группы:

- лингвистические алгоритмы, основанные на правилах, шаблонах и словарях;
- алгоритмы, использующие методы машинного обучения.

Многие коммерческие системы используют первый подход как наиболее точный.

Словари оценочной лексики

Словарь оценочной лексики – база данных, где хранятся слова и n-компонентные цепочки – n-граммы (например, фразеологизмы и различные устойчивые словосочетания («задеть за живое», «с гулькин нос»), при этом каждой такой единице присвоен уровень эмоциональной оценки.

Словари:

- используют различные шкалы оценок;

-с автоматическим пополнением списков и без.

Виды словарей

Словам может быть приписана лишь одна тональная оценка – числовое значение полярности (число большее нуля – позитивный сентимент, число меньшее нуля – негативный сентимент).

Слово/ словосочетание	Уровень эмоциональной оценки
отвратительный	-5
с гулькин нос	-2
модный	+3
усталый	-2

Пример для английского языка:
AFINN
В некоторых других системах (к примеру, SentiStrength) группы слов получают не одну, а две тональные оценки (положительную и отрицательную).

Виды словарей

Существуют лексиконы, в которых словам приписываются разные эмоциональные категории, к ним относится NRC Word-Emotion Association Lexicon. Каждому слову здесь сопоставляются 2 тональные оценки и 8 эмоций: «гнев», «страх», «предчувствие», «вера», «удивление», «грусть», «отвращение», «радость». Списки данного словаря были переведены на несколько десятков языков, среди которых есть и русский.

Слово	Эмоциональная или тональная оценка	Значение (1 – присутствие; 0 – отсутствие)
откровенный	гнев	0
	страх	0
	предчувствие	0
	вера	1
	удивление	0
	грусть	0
	отвращение	0
	радость	0
	положительная	1
	отрицательная	0

Виды словарей

Эмоциональная метка	Пример
Эмоция (emotion)	сущ. гнев#1, гл. бояться#1 (fear)
Настроение (mood)	сущ. враждебность#1 (animosisy), прил. любезный#1J (amiable)
Особенность (trait)	сущ. агрессивность#1 (aggressiveness), прил. соперничающий#1 (competitive)
Когнитивное состояние (cognitive state)	сущ. замешательство#2 (confusion), прил. потрясенный#2 (dazed)
Физическое состояние (physical state)	сущ.хворь#1 (Ilness), прил. выдохнувшийся#1 (all in)
Гедонический сигнал (gedonic signal)	сущ. боль#3(hurt), сущ. страдание#4 (suffering)
Ситуации, вызывающие эмоции (emotion-eliciting situation)	сущ. неловкость#3 (awkwardness), сущ. безопасность#1 (out of danger)
Эмоциональные отклики (emotional responses)	сущ. холодный пот#1 (cold sweat), гл. дрожать#2 (tremble)
Поступки (behaviour)	сущ. преступление#1 (offense), прил. заторможенный#1 (inhibited)
Отношение, позиция (attitude)	сущ. нетерпимость#1 (intolerance), сущ. оборонительная позиция#1 (defensive)
Чувство (sensation)	сущ. холод#1 (coldness), гл. чувствовать#3 (feel)

В тезаурусе WordNet-Affect наряду с метками, указывающими эмоциональную категорию («гнев», «страх», «удивление», «печаль», «отвращение», «радость»), и валентностями (позитивная, негативная, неоднозначная, нейтральная) словарным единицам – синсетам, синонимическим рядам – были сопоставлены метки, описывающие эмоции: «физическое состояние», «настроение», «поведение», «отношение», «чувство» и др. Данный тезаурус был переведен с английского языка на русский и румынский.

Лексический подход

Шаблоны: ***<им. прил. им. сущ.>, <им. прил. им. прил.>***

По шаблонам из текста извлекаются n-граммы. Их тональность определяется как при помощи словаря, так и посредством правил.

Тональность всего текста складывается из тональности предложений, а тональность предложений — из тональности слов. Для получения итоговой окраски общую сумму весов нужно подсчитать по формуле, которую составляют разработчики конкретного решения, универсальной формулы не существует. Можно, к примеру, просто просуммировать полярности цепочек документа.

Примеры правил при лингвистическом подходе

1. Правила, построенные по модели «если… то…».

Если цепочка содержит глагол из списка («любить», «нравиться», «обожать» и др.) и не содержит глагола из другого списка («ужасать», «отвращать» и др.) или отрицания, то её тональность положительная.

2. Правила, обрабатывающие слова с их модификаторами.

Модификаторы:

- усиливающие (***«очень»***, ***«более»***) исходную тональность;
- снижающие (***«слишком»***, ***«менее»***) исходную тональность;
- преобразующие в обратную (***«не»***, ***«нет»***) исходную тональность.

Модификаторам тональности приписываются некоторые коэффициенты, которые рассматриваются как множители относительно априорной полярности соответствующего оценочного слова.

Примеры правил при лингвистическом подходе

3. Правила обработки слов с коннотациями.

Коннотации — это оценочные ассоциации, связанные со словами. Появление в тексте слов с положительными или отрицательными коннотациями коррелирует с соответствующими оценками, выражаемыми в тексте. Так, в отзывах о фильмах словами с положительными коннотациями обычно являются имена известных актеров. В отзывах о ресторанах на русском языке отрицательными коннотациями обладают такие слова, как «майонез» и «клеенка». Если эти слова появляются в отзыве, обычно в этом месте выражается негативная оценка.

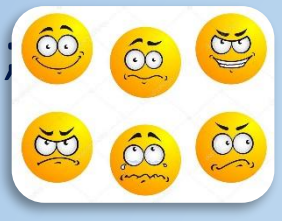
«Вместо нормальных скатертей какие-то клеёнки…»

«В салатах, которые на принесли, было столько майонеза!»

Особенности UGC (user-generated content) текстов

Особенности языка социальных медиа:

- эмотиконы и смайлики;



- опечатки;
- неологизмы (*«пичалька» - опечатка или нет?)*;
- окказионализмы – индивидуально-авторские неологизмы;
- эмоционально окрашенные аббревиатуры (*«omg!..»*);

и много другое.

Смайл	Тональность	Смайл	Тональность
:-)	положительный	:o)	положительный
:-D	положительный	;)	положительный
;-)	положительный	;v)	положительный
xD	положительный	:D	положительный
;-P	положительный	:^D	положительный
:-p	положительный	:/	негативный
8-)	положительный	:-b	положительный
B-)	положительный	=^*	положительный
:-(	негативный	:-x	положительный
;-]	положительный	8-]	положительный
3(	негативный	>:-(	негативный
:'(	негативный	>:-[	негативный
:_(	негативный	:-0	негативный
:((	негативный	:-o	негативный
:o	негативный	:,-(	негативный
3-)	положительный	:'-(	негативный
O:)	положительный	:,-)	положительный
;o	негативный	:<	негативный
<3	положительный	:-||	негативный
:-*	положительный	D;	негативный

Недостатки лингвистического подхода

Плюсы: высокая точность

Минусы:

- составление системы правил очень трудоёмкая задача;
- метод правил и словарей не универсален (есть зависимость от предметной области)

Общие проблемы сентимент-анализа

У любой системы автоматического анализа тональности на данный момент остаются такие проблемы, как:

- обработка саркастических и ироничных отрывков;

«Было скучно. Давно не смотрела фильмов с настолько интригующим сюжетом»

- обработка метафор пользователя;

«школа как второй дом»

- сложности, возникающие, когда в отзыве содержится оценка сразу нескольких объектов, иногда конкурирующих;

«Huawei впервые обогнала Apple по продажам смартфонов в России»

и др.

Спасибо за внимание!