МГТУ им. Н.Э.Баумана

Рубежный контроль

По курсу: "Анализ алгоритмов"

Методы анализа тональности текстов

Работу выполнил: Луговой Дмитрий, ИУ7-51Б

Преподаватель: Волкова Л.Л.

Москва, 2019

Оглавление

Введение **3**

1 Аналитическая часть **4**
1.1 Тональность текста 4
1.2 Классификация методов анализа 5
1.3 Анализ методов 6

Заключение **8**

Список литературы **8**

Введение

Цель работы: изучение методов анализа эмоциональной окраски(тональности) текстов.

Задачи работы

1) Изучить существующие методы анализа тональности текстов;

2) Провести сравнительный анализ найденных методов.

1 | Аналитическая часть

В данном разделе содержатся описания методов анализа тональности текста и производится их сравнительный анализ.

1.1 Тональность текста

Тональность — это эмоциональное отношение автора высказывания к некоторому объекту (объекту реального мира, событию, процессу или их свойствам/атрибутам), выраженное в тексте. Эмоциональная составляющая, выраженная на уровне лексемы или коммуникативного фрагмента, называется лексической тональностью (или лексическим сентиментом). Тональность всего текста в целом можно определить как функцию (в простейшем случае сумму) лексических тональностей составляющих его единиц (предложений) и правил их сочетания.

Анализ тональности текста (сентимент-анализ) — класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (мнений) по отношению к объектам, речь о которых идёт в текст.

Основной целью анализа тональности является нахождение мнений в тексте и выявление их свойств. Какие именно свойства будут исследоваться, зависит уже от поставленной задачи.
Виды тональных оценок:

- позитивная;
- негативная;
- нейтральная.

Под «нейтральной» подразумевается, что текст не содержит эмоциональной окраски.
Классификация текстов применяется, в том числе, для:

- Разделения веб страниц и сайтов по тематическим каталогам;
- Борьбы со спамом;
- Определение языка текста;
- Показа более релевантной рекламы;

1.2 Классификация методов анализа

Методы анализа классифицируются на:

- **Методы, основанные на правилах и словарях**

  Эти методы основаны на поиске тональности в тексте по заранее составленным тональным словарям и правилам с применением лингвистического анализа. По совокупности найденной эмотивной лексики текст может быть оценен по шкале, содержащей количество негативной и позитивной лексики. Данный метод может использовать как списки правил, подставляемые в регулярные выражения, так и специальные правила соединения тональной лексики внутри предложения.

  Однако процесс создания этих «фолиантов» очень трудоемкий; основной проблемой является тот факт, что одно и то же слово в разных контекстах может обладать различной тональностью. Это означает, что для адекватной работы системы требуется составить большое количество правил – поэтому чаще всего системы анализа тональности текста создаются с привязкой к определенной предметной области.

- **Методы, основанные на теоретико-графовых моделях**

  В основе этих методов используется предположение о том, что не все слова в текстовом корпусе документа равнозначны. Какие-то слова имеют больший вес и сильнее влияют на тональность текста.
  При использовании этих методов анализ тональности разбивается на несколько этапов:

  1. построение графа на основе исследуемого текста;
  2. ранжирование его вершин;
  3. классификация найденных слов;
  4. вычисление результата.

  После ранжирования вершин графа слова классифицируются в соответствии со словарем тональности, где каждому слову присваивается определенная характеристика («положительное», «отрицательное» или «нейтральное»). Результат вычисляется как соотношение количества слов с положительной оценкой к количеству слов с отрицательной оценкой.

- **Методы, основанные на машинном обучении**

  В наше время наиболее часто используемыми в исследованиях методами являются методы на основе машинного обучения с учителем. Сутью таких методов является то, что на первом этапе обучается машинный классификатор на заранее размеченных текстах, а затем используют полученную модель при анализе новых документов.
  Наиболее популярным является метод с "наивным байессовским"классификатором.
  Краткий алгоритм:

1. Собирается коллекция документов, на основе которой обучается машинный классификатор;
2. Каждый документ раскладывается в виде вектора признаков(аспектов), по которым он будет исследоваться;
3. Указывается правильный тип тональности для каждого документа;
4. Производится выбор алгоритма классификации и метод для обучения классификатора;
5. Полученная модель используется для определения тональности документов новой коллекции.

Также распространены методы без учителя, в основе которых лежит идея, что термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов во всей коллекции, имеют наибольший вес в тексте. Выделив данные термины, а затем определив их тональность, можно сделать вывод о тональности всего текста. Для выделения может быть использована модель TFIDF. Иногда слова могут встречаться во многих документах текстовой коллекции. Следовательно, они не могут характеризовать принадлежность документа тому или иному классу, так как не являются ключевыми. Поэтому вводится так называемая мера IDF (обратная частота документа), которая понижает значимость частотных слов.

$$IDF = \ln(\frac{D}{d}), \tag{1.1}$$

где D — количество всех документов, d — количество документов, в которых содержится данное слово. Таким образом, чем чаще слово встречается, тем меньше его IDF. Мера TF (term frequency) — отношение частоты некоторого слова к общему числу слов в документе. Мера оценивает важность слов в пределах документа. TFIDF равен произведению TF и IDF. Таким образом, TF является повышающим множителем, а IDF — понижающим. Больший вес получат слова, которые часто встречаются в одном документе, но редко в других.

1.3 Анализ методов

Точность и качество системы анализа тональности текста оценивается тем, насколько хорошо она согласуется с мнением человека относительно эмоциональной оценки исследуемого текста. Для этого могут использоваться такие метрики как точность и полнота.
Формула для нахождения полноты:

$$R = \frac{\text{correctly extracted opinions}}{\text{total number of opinions}} \tag{1.2}$$

где correctly extracted opinions — верно определённые мнения, total number of opinions — общее количество мнений(как найденных системой, так и не

найденных).
Точность вычисляется по формуле:

$$P = \frac{\text{correctly extracted opinions}}{\text{total number of opinions found by system}} \tag{1.3}$$

где correctly extracted opinions — верно определённые мнения, total number of opinions found by system — общее количество мнений найденных системой.
Таким образом, точность выражает количество исследуемых текстов, предложений или документов, в оценке которых мнение системы анализа тональности совпало с мнением эксперта. При этом, согласно исследованию, эксперты обычно соглашаются в оценках тональности конкретного текста в 79 % случаев. Следовательно, программа, которая определяет тональность текста с точностью 70 %, делает это почти так же хорошо, как и человек.

Задача определения полярности сообщений, текстов или предложений имеет множество эффективных методов решения, в каждом из которых есть особенности.

Словарные методы не позволяют создать универсальный словарь терминов, так как их вес в разных предметных областях может значительно отличаться или быть противоположным.

Методы обучения с учителем сложны составлением тренировочного словаря для предметной области, в которой используется классификатор, но в тоже время показывают наиболее высокую точность анализа.

На рис. 1.1 представлена сравнительная характеристика наиболее распространенных методов анализа тональности.

	Точность	**Автоматизация**	**Данные для обучения**	**Простота применения**	**Применяемость в коммерческих системах**
Подход на правилах	наиболее точный	подлежит	не требует данных	–	+
Подход со словарем	не универсален	в рамках одной предметной области	требует данные	+	–
Машинное обучение	точный	автоматический	требует данные	+/–	+
Обучение без учителя	низкая точность	автоматический	не требует данных	+	+

Рис. 1.1: Сравнение методов анализа тональности текста

Заключение

На основании проведенного анализа оптимальным является метод машинного обучения с учителем с использованием наивного байесовского классификатора. Выбор метода обусловлен тем, что данный метод позволяет работать с различными исходными массивами данных для обучения системы, что может значительно повлиять на точность анализа. Также этот метод не зависит от специфики текста и может использоваться в очень широком спектре задач.

Список литературы

[1] ИТМО Классификация текстов и анализ тонатльности [Электронный ресурс]. – Режим доступа: URL: http://neerc.ifmo.ru/wiki/index.php?title=Классификация_текстов_и_анализ_тональности. – (Дата обращения: 21.11.2019)

[2] Wikipedia Анализ тональности текста [Электронный ресурс]. – Режим доступа: URL: https://ru.wikipedia.org/wiki/Анализ_тональности_текста. – (Дата обращения: 21.11.2019)

[3] Анализ эмоциональной окраски сообщений в социальных сетях (на примере сети «Вконтакте») [Текст] / И. Е. Воронина, В. А. Гончаров // ВЕСТНИК ВГУ, СЕРИЯ: СИСТЕМНЫЙ АНАЛИЗ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ. – 2015. - № 4. – С. 151-158.