EURASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.916 www.in-academy.uz

СИСТЕМЫ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Ахмедов Э.Ю
MSc, ассистент, кафедра программный инжиниринг, Ургенчский филиал Ташкентского университета информационных технологий имени Мухаммада ал-Хорезмий, “05.00.00 - Texnika fanlari”, +998974567912
https://doi.org/10.5281/zenodo.7472035

ARTICLE INFO

Received: 13th December 2022
Accepted: 21th December 2022
Online: 22th December 2022

KEY WORDS

Компьютерная лингвистика, анализ тональности текста, сентимент анализ, бинарная оценка, тернерная оценка, классификация текста.

ABSTRACT

Обработка естественного языка - общее направление искусственного интеллекта и математической лингвистики, изучающее проблемы компьютерного анализа и синтеза естественных языков. Обработка естественного языка можеть дать возможность извлекать различную информацию, представленную в виде текста на естественном языке.

Одной из перспективных направлений компьютерной лингвистики является анализ тональности текста. С помощью анализа тональности текста можно извлечить из текста эмоциональное отношение автора к объектам. Эмоциональное отношение можно оценить бинарной оценкой “хорощо – плохо”.

Современные системы используют бинарную оценку положительный сентимент» или «отрицательный сентимент», но кроме этого можно оценить силу тональности. Сентимент - совокупность чувств и взглядов как основа для действия или суждения, общая эмоциональная установка.

В современном мире наш выбор зависить мнение других людей – мы читаем отзывы о товаре, прежде чем заказать его в интернет-магазине, тщательно выбираем себе ВУЗ, место работы и ресторан, который мы собираемся посетить.

Описание популярных систем анализа тональности текста:

1. SentiStrength – программа оценки силы положительных и отрицательных отношений в текстах, ориентирован работы с краткими текстами на английском языке в социальных сетях (MySpace, Twitter) и комментариями к постам.

Разработчиком данного программного обеспечения является Майкл Фелволл. Систему можно изменить для работы с текстами на ряде других языков. Результат выдается в виде двух оценок – оценка позитивной составляющей текста (по шкале от +1 до +5) и оценка негативной составляющей (по шкале от -1 до -5). Кроме этого можно осуществить предоставления оценок в другом виде:

– Бинарная оценка (позитивный/негативный текст)
– Тернарная оценка (позитивный/негативный/нейтральный)
– Оценка по единой шкале от -4 до +4

Недостатки системы: система может быть сконфигурирована для узбекского языка, но алгоритмы в ней не учитывают морфологию языка, что приводит к ряду проблем. [1]

В большинстве проанализированных исследование применялись подходы на основе правил и базовые подходы к машинному обучению, в нескольких исследованиях использовались нейронные сети. Однако недавние исследования показали, что трансферное обучение предварительно обученных языковых моделей доказало свою эффективность в задаче классификации тональности.

Есть доступные языковых моделей для проведения эксперимента многоязычная версия Bidirectional Encoder Representations from Transformers (M-BERT), RuBERT и Multilingual Universal Sentence Encoder (M-USE). M-BERT, RuBERT и M-USE являются единственными среди наиболее новых языковых моделей. M-BERT уже получила широкое признание среди исследователей, занимающихся контент-анализом неанглоязычных языков. [2]

Алгоритм на основе правил для анализа тональности текста.

Классификатор Наивный Байесовский один из самых известных методов анализа тональности текста на основе лексикона (словарного запаса), основан на упрощенном понимании текста как набора слов (мешка слов), и даже расположение слов в предложении или тексте не имеет значения. Наиболее существенным недостатком является то, что синтаксические связи, которые не учитывает данный классификатор, могут также оказывать значительное влияние на смысл предложения или текста в целом, а значит, и на его оценку.

Определение тональности текста отзыва состоит расчета черновой полярности для каждого предложения, корректировки полярности каждого предложения с помощью правил, применяемых в определенной последовательности, суммирования скорректированных величин и нормализации окончательной величины тональности текста. Этапы определения тональности текста представлены на рисунок 1.

Предварительная обработка включает следующие шаги:

Шаг 1. Разбить текст на предложения, пометить начало и конец каждого предложения.

Шаг 2. Разбить каждое предложение на фрагменты, заменить знаки препинания внутри предложения, кавычки и союзы специальными символами.

Шаг 3. Разбить текст на слова, используя пробелы.

Шаг 4. Подсчитать количество слов в тексте.

Шаг 5. Пометить слова, набранные прописными буквами.

Шаг 6. Заменить все прописные буквы строчными.

Шаг 7. Подвергнуть все слова стеммингу.

Рис 1. Этапы определение тональности текста

Черновая полярность каждого предложения рассчитывается после заменены все слова, входящие в положительный или отрицательный класс оценочного лексикона на служебные символы (POS или NEG соответственно).

Он включает три класса: модификаторы полярности, усилители полярности и антимодификаторы полярности. Модификаторы полярности – это изменяющие слова полярность предложения на противоположную. Усилители полярности - это слова, увеличивающие счет полярности предложения. Антимодификаторы полярности - это слова, отменяющие изменение полярности предложения на противоположную.

Правило 1.1. <ALT>(<POS>){n}→<NEG>{n}

Если в промежутке от начала предложения или знака препинания, или союза «и, или» до следующего знака препинания или союза «и, или» имеется модификатор полярности, то положительная полярность всех слов, входящих в оценочный лексикон, в данном промежутке изменяется на отрицательную. «Специалисты не могут дать внятного ответа на простой вопрос». Count = −3.

Правило 1.2. <ALT>(<NEG>){n}→<POS>{n}

Если в промежутке от начала предложения или знака препинания, или союза «и, или» до следующего знака препинания или союза «и, или» имеется модификатор полярности, то отрицательная полярность всех слов, входящих в оценочный лексикон, в данном промежутке изменяется на положительную. «Никто не заставил меня ждать». Count = +1.

Правило 2.1. (<INC><NEG>) →<NEG><NEG>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеются слова отрицательный лексикон и усилитель, то каждый усилитель засчитывается за одно слово отрицательного лексикона. «Банк совершенно не в состоянии провести поиск счетов в течение месяца». Count = −2.

Правило 2.2. (<INC><POS>) →<POS><POS>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеются слова в положительный лексикон и усилитель, то каждый усилитель засчитывается за одно слово положительного лексикона. «Тут могу отметить очень удобную возможность погашения кредита». Count = +2.

Правило 3.1. (<TH><ALT><NEG>)→<NEG>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеется модификатор полярности (ALT) и антимодификатор полярности (TH), то полярность слов в данном промежутке не изменяется. Правило 1 не применяется.

«Таких проблем с ипотекой я не ожидал». Count = −1.

Правило 3.2. (<TH><ALT><POS>)→ <POS>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеется модификатор полярности (ALT) и антимодификатор полярности (TH), то полярность слов в данном промежутке не изменяется. Правило 1 не применяется. «Такого чуткого отношения к клиентам я нигде не встречал». Count = +1.

Правило 4. <POS>{1}<QM>→<NEG>

Если предложение состоит из одного слова с положительной полярностью и заканчивается ? или ?!, то полярность данного слова изменяется на отрицательную: «Оперативно ?». Count = −1

Правило 5. <QM>→<NEG>

Если предложение заканчивается вопросительным знаком или вопросительным знаком с восклицательным знаком, при этом в предложении нет слов в оценочном лексиконе, то полярность данного предложения равна −1. «Но на каком основании мне была подключена данная услуга?». Count = −1.

Правило 6. <Q><POS><Q>→<NEG>

Если слово с положительной полярностью заключено в кавычки, то полярность данного слова изменяется на отрицательную. Кавычки являются модификатором полярности: «“Забота” о клиенте». Count = −1.

Правило 7.1. <WT>(w|INC){0,}<POS>→<NEG>

Если перед словом с положительной полярностью имеется слово «без» тогда полярность слова изменяется на отрицательную. Слово «без» является модификатором полярности. Между словом «без» и словом с полярностью может быть усилитель (особый, всякий, единый): «Банк начислил пени без всяких разъяснений». Count = −1.

Правило 7.2. <WT>(w|INC){0,}<NEG>→<POS>

Если перед словом с отрицательной полярностью имеется слово «без», то полярность данного слова изменяется на положительную. Слово без является модификатором полярности. Между словом «без» и словом с полярностью может быть усилитель (особый, всякий, единый и т. п.):

«Оплата с моей стороны происходила без задержек». Count = +1.

«Я смог без особых проблем снять деньги со счета». Count = +2.

Правило 8.1. <POS><EM>→<POS><POS>

Если предложение заканчивается восклицательным знаком, а счет предложения положительный, то восклицательный знак приравнивается к одному слову с положительной полярностью: «Научились, делают, молодцы!». Count = 2. «Великолепное отношение к клиентам!». Count = 2.

Правило 8.2. <NEG><EM>→<NEG><NEG>

Если предложение заканчивается восклицательным знаком, или вопросительным знаком, или вопросительным знаком с восклицательным знаком и счет предложения отрицательный, то восклицательный знак, или вопросительный знак, или вопросительный знак с восклицательным знаком приравнивается к одному слову с отрицательной полярностью: «Умопомрачительная халатность сотрудников!». Count = −2. «Почему сотрудники данного филиала не знают об этом?!». Count= −2. «И зачем полгода мурыжили?» Count = −2.

Правило 9.1. <POS><CAP>→<POS><POS>

Если в предложении имеются слова из прописных букв и счет предложения положительный, то каждое слово из прописных букв приравнивается к одному слову с положительной полярностью: «Уровень сервиса — ВАУ!». Count = +3.

Правило 9.2. <NEG><CAP>→<NEG><NEG>

Если в предложении имеются слова из прописных букв, а счет предложения отрицательный, то каждое слово из прописных букв приравнивается к одному слову с отрицательной полярностью: «Сотрудники НЕ ЗНАЮТ свою работу». Count = −2.

Результатом алгоритма является суммирование полярностей всех предложений, нормализация данной величины относительно количества слов в тексте отзыва и вывод системы (отзыв положительный и отрицательный). [3]

References:

1. Меньшиков И.Л., Кудрявцев А.Г. Обзор систем анализа тональности текста на русском языке Молодой учёный №12 (47) 2012 г.
2. Сметанин С.И., Анализ тональности текстов из социальных сетей на основе методов машинного обучения для мониторинга общественных настроений. Резюме диссертации Москва 2022.
3. Брунова Е. Г. Клиент всегда прав: анализ тональности текста в отзывах о качестве банковского обслуживания / Е. Г. Брунова, Ю. В. Бидуля // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2017. Том 3. № 1. С. 72-89. DOI: 10.21684/2411-197X-2017-3-1-72-89
4. Гималетдинова Г.К., Довтаева Э.Х. Сентимент-анализ читательского комментария: автоматизированная vs ручная обработка текста // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. – 2021.
5. В. В. Гаршина, К. С. Калабухов, В. А. Степанцов, С. В. Смотров «Разработка системы анализа тональности текстовой информации» «Вестник ВГУ. Серия: Системный анализ и информационные технологии» 2017, №3
6. Мельчук И. А. Опыт теории лингвистических моделей «Смысл ⇔ Текст».2-е изд. М.: Яз. рус. культуры, 1999. 368 с.
7. В.В. Воронович. “Машинный перевод” МИНСК – 2013
8. Усталов Дмитрий Алексеевич. Модели, методы и алгоритмы построения семантической сети слов для задач обработки естественного языка. Екатеринбург-2017