АЛГОРИТМЫ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Баран И. В., Нестеренков С. Н.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: igrbaran@gmail.com, s.nesterenkov@bsuir.by

В последние годы алгоритмы анализа тональности текста привлекают все большее внимание благодаря их применению в различных сферах, таких как маркетинг, социальные исследования и обработка естественного языка. В данной работе рассматриваются два основных подхода к анализу тональности: лексиконные методы и методы машинного обучения, включая наивный байесовский классификатор, метод опорных векторов и сверточные нейронные сети. Проводится сравнение этих подходов с точки зрения их эффективности и применимости к различным типам текстов.

Введение

Анализ тональности текста представляет собой процесс автоматической интерпретации эмоциональной окраски текста, определяя, выражает ли он позитивные, негативные или нейтральные эмоции. В условиях стремительного роста объёмов информации в интернете, таких как отзывы пользователей, комментарии в социальных сетях и статьи, анализ тональности стал важным инструментом для понимания общественного мнения и настроений. Данная технология находит широкое применение в маркетинге, социальной аналитике, политике и других областях, где важно быстро и точно оценивать реакции людей. Алгоритмы, используемые для анализа тональности, основаны на различных методах машинного обучения и обработки естественного языка, которые позволяют анализировать контекст и структуру текста. В этом докладе будут рассмотрены ключевые алгоритмы анализа тональности, их особенности и применимость в различных ситуациях.

I. Лексиконные методы анализа тональности

Лексиконные методы анализа тональности используют заранее созданные словари, в которых каждому слову присвоена определённая эмоциональная окраска (позитивная, негативная или нейтральная). Простейший способ применения таких методов заключается в суммировании полярностей слов, присутствующих в тексте, и определении общей тональности. Этот подход прост в реализации и не требует предварительного обучения модели, что делает его полезным для быстрой оценки.

Однако главный недостаток лексиконных методов заключается в том, что они не учитывают контекст. Одно и то же слово может иметь разные значения в зависимости от ситуации, что лексикон не способен учесть. Лексиконы также не справляются с такими сложными языковыми явлениями, как ирония, сарказм или многозначность слов. Для задач, где важен контекст, лучше подходят методы машинного обучения.

II. Методы машинного обучения для анализа тональности

Методы машинного обучения для анализа тональности основываются на обучении моделей на размеченных данных, где каждый текст заранее классифицирован по тональности (положительная, отрицательная, нейтральная). Машинное обучение позволяет учитывать контекст слов и выявлять сложные закономерности в тексте. Рассмотрим несколько популярных алгоритмов.

Наивный байесовский классификатор (Naive Bayes) предполагает независимость признаков (слов) друг от друга, что является его «наивным» предположением. Вероятность принадлежности текста классу C (например, позитивной или негативной тональности) вычисляется на основе теоремы Байеса:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

где $P(C|X)$ – это вероятность того что текст X принадлежит классу C, $P(X|C)$ – вероятность встретить слова X в текстах класса C, $P(C)$ – априорная вероятность класса C, а $P(X)$ – полная вероятность встретить слова X в любых текстах.

Метод опорных векторов (SVM) – это линейный классификатор, который находит гиперплоскость, разделяющую данные на классы. В контексте анализа тональности SVM пытается найти границу между текстами с разной эмоциональной окраской. Функция для оптимизации в SVM выглядит следующим образом:

$$min_{w,b}\frac{1}{2}||w||^2$$

при условии, что для всех обучающих примеров i:

$$y_i(w * x_i + b) >= 1$$

Где w – это весовой вектор, определяющий гиперплоскость, b – смещение, y_i – метка класса для примера i, x_i – вектор признаков (слов) для текста i. Основное преимущество SVM заключается в его способности эффективно работать с высокоразмерными пространствами признаков,

что делает его популярным в задачах текстовой классификации.

Сверточные нейронные сети (CNN) – это мощный инструмент для решения задач классификации текстовой информации. Задача классификации текстовой информации определяется следующим образом. Пусть существует конечное множество категорий $C = \{c1,c2,\ldots,cm\}$, конечное множество документов $D = \{d1,d2,\ldots,dm\}$ и неизвестная целевая функция Ф, определяющая соответствие для каждой пары <текст, категория> $\Ф : D \times C \to \{0,1\}$. Задача состоит в нахождении функции Ф', которая является максимально близкой к целевой функции Ф. Эта функция называется классификатором.

CNN применяются для решения задачи классификации текста благодаря своей способности выявлять локальные паттерны и особенности в данных. На входе модель получает текст, который предварительно преобразуется в числовые вектора (например, с помощью методов векторизации слов, таких как Word2Vec или TF-IDF). Векторы пропускаются через несколько сверточных слоёв, где каждый фильтр сети выделяет определённые признаки текста, такие как эмоциональные выражения, ключевые слова, или даже сложные грамматические конструкции, связанные с общей тональностью текста или смыслом.

Основным преимуществом использования CNN в задачах анализа текста является их способность выявлять значимые локальные зависимости между словами и фразами. В сверточных сетях используется операция свёртки, которая представляет собой процесс сканирования текста с помощью набора фильтров (сверток). Операция свёртки применяется для обработки последовательностей слов, выделяя локальные зависимости, которые могут быть неочевидны при использовании традиционных методов обработки текста. Это делает CNN особенно полезными для анализа текстов, содержащих сложные структуры, такие как зависимые предложения, многозначные выражения и контекстуальные зависимости.

III. Сравнение лексиконных методов и методов машинного обучения

Лексиконные методы и методы машинного обучения представляют два разных подхода к анализу тональности текста, и каждый из них имеет свои преимущества и недостатки. Лексиконные методы просты, быстры и не требуют предварительного обучения, что делает их удобными для использования в задачах, где точность не критична или для приложений с ограниченными вычислительными ресурсами. Однако они часто не способны точно анализировать сложные тексты, учитывать контекст или выявлять тональность в многозначных выражениях.

Методы машинного обучения, напротив, обладают большей гибкостью и могут справляться с более сложными текстами благодаря обучению на реальных данных. Они лучше работают с многозначными словами, иронией и сарказмом, а также могут учитывать контекст, что делает их более точными в задачах анализа тональности. Однако такие методы требуют значительных вычислительных ресурсов и больших объёмов размеченных данных для обучения, что может ограничивать их применение в некоторых ситуациях.

Выбор между этими методами зависит от конкретной задачи. В простых случаях лексиконные методы могут быть достаточными, тогда как для более сложных текстов и высоких требований к точности лучше использовать методы машинного обучения.

IV. Заключение

Анализ тональности текста – это важная задача в области обработки естественного языка, которая находит широкое применение в маркетинге, социальных сетях и автоматизированной аналитике отзывов. Лексиконные методы и методы машинного обучения представляют собой два ключевых подхода к решению этой задачи, каждый из которых имеет свои сильные и слабые стороны.

Современные тенденции показывают, что для решения сложных задач анализа тональности всё чаще применяются гибридные методы, которые сочетают лексиконы с мощью машинного обучения, обеспечивая баланс между точностью и эффективностью. Выбор подхода всегда зависит от требований задачи, объёма данных и доступных ресурсов. В итоге, развитие методов анализа тональности продолжает расширять горизонты применения этих технологий, делая их неотъемлемой частью анализа больших данных, автоматизации бизнес-процессов и принятия решений.

1. Нестеренков, С. Н. Использование сверточных нейронных сетей для классификации и анализа тональности текстов / С. Н. Нестеренков, П. А. Федоров, В. А. Денисов // Информационные технологии и системы 2019 (ИТС 2019). – Минск, 2019. – С. 248-249.
2. Жалейко, Д. А. Нейросети в анализе эмоционального состояния и развития персонала и его влияния на успех проектов / Д. А. Жалейко, С. Н. Нестеренков, И. Г. Скиба // сб. науч. ст. X Междунар. науч.-практ. конф. – Минск : БГУИР, 2024. – С. 373–376.
3. Yanyan W., Qun C., Jiquan S., Boyi H., Murtadha A, Zhanhuai Li G. Machine Learning for Aspect-level Sentiment Analysis // arXiv:1906.02502 – 2019.
4. Chiranji Lal Chowdhary. Multidisciplinary Applications of Deep Learning-Based Artificial Emotional Intelligence. // IGI Global, 2022. – 324.