Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Обзор методов анализа тональности естественно-языковых текстов

Студент: **Маслова Марина Дмитриевна ИУ7-53Б**
Научный руководитель: **Оленев Антон Александрович**

Цель и задачи

Цель: провести классификацию методов анализа естественно-языковых текстов.

Задачи:

- рассмотреть основные подходы к анализу тональности;
- описать методы анализа тональности, относящиеся к каждому из подходов;
- предложить критерии оценки описанных методов;
- сравнить методы по предложенным критериям оценки;
- выделить методы, показывающие лучшие результаты по одному или нескольким критериям.

Основные подходы

- Лингвистический подход:
  - методы на основе правил,
  - методы на основе лексики;
- Подход, основанный на машинном обучении:
  - обучение с учителем,
  - обучение без учителя;
- Гибридный подход.

Лингвистический подход

Идея: анализ лексики в тексте на основе заранее созданных словарей, шаблонов и правил.

К нему относят:

- методы, основанные на правилах;
- методы, основанные на тональных словарях;
- методы, основанные на корпусах.

Достоинства и недостатки:

\+ высокая точность в узких предметных областях;

– ручная подготовка правил/начального набора слов.

Методы на основе машинного обучения

Вероятностные классификаторы:

- Наивный байесовский классификатор (НБК);
- Метод максимума энтропии.

Идея: нахождение наиболее вероятного класса тональности для данного текста.

Достоинства и недостатки:

\+ простота реализации;
\+ малое число обучающих данных;
– ограничение на независимость признаков у НБК.

Методы на основе машинного обучения

Линейные классификаторы:

- Логистическая регрессия;
- Метод опорных векторов;
- Нейронные сети.

Идея: построение разделяющей гиперплоскости в пространстве признаков.

Достоинства и недостатки:

+ высокая точность;
+ способность к масштабированию;
– трудоемкость предварительной обработки.

Методы на основе машинного обучения

Классификаторы на основе решающих деревьев:

- Дерево решений (ДР);
- Случайный лес (СЛ).

Идея: построение дерева/деревьев путем деления текстов обучающей выборки на две группы по каждому признаку.

Достоинства и недостатки:

\+ простота реализации;
– зависимость от обучающей выборки в случае ДР;
– большие временные затраты в случае СЛ.

Методы на основе машинного обучения

- k-ближайших соседей.

Идея: поиск расстояния между вектором анализируемого текста и каждым текстом обущающей выборки, выбор тональности по большиству из k текстов с наименьшими расстояниями.

Достоинства и недостатки:

+ простота реализации;
– большое время выполнения в силу полного перебора.

Гибридные методы

Идея: объединение двух и более методов.

Достоинства и недостатки:

+ преимущества каждого из объединяемых методов;
– недостатки каждого из объединяемых методов.

Иерархия методов

Критерии оценки методов

		Оценка эксперта	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

Точность $$precision = \frac{TP}{TP + FP}$$

Полнота $$recall = \frac{TP}{TP + FN}$$

F-мера $$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Сравнение методов анализа тональности

Метод	К1	К2	К3	К4	К5
На основе правил	74.2%	73.7%	73.9%	ручной	требуется
На основе словарей	97.6%	88.9%	93.1%	смешанный	требуется
На основе корпусов	43.1%	94.3%	59.2%	смешанный	требуется
Наивный Байес	75.5%	52.4%	67.9%	автоматический	не требуется
Логистическая регрессия	76.9%	80.1%	78.5%	автоматический	не требуется
Максимум энтропии	65.2%	98.1%	78.3%	автоматический	требуется
k-ближайших соседей	59.1%	96.1%	73.2%	автоматический	не требуется
Дерево решений	96.7%	64.5%	77.4%	автоматический	требуется
Случайный лес	89.5%	76.5%	82.5%	автоматический	не требуется
Опорные векторы	83.8%	83.9%	83.8%	автоматический	не требуется
Нейронные сети	85.3%	88.4%	86.8%	автоматический	не требуется

К1 — точность;
К2 — полнота;
К3 — F-мера;
К4 — способ подготовки данных;
К5 — повторная настройка.

Заключение

В ходе научно-исследовательской работы выполнены следующие задачи:

- рассмотрены основные подходы к анализу тональности;
- описаны методы анализа тональности, относящиеся к каждому из подходов;
- предложены критерии оценки, по которым проведено сравнение описанных методов;
- выделены методы, показывающие лучшие результаты по по предложенным критериям.

Поставленная цель достигнута.