Сравнение методов сентимент-анализа

Машинное обучение, словарный и смешанный подходы

Автор: Любовь Беликова

Код: https://github.com/meancatt/CL/tree/master/ProjectFinal

Задача – выявить наиболее эффективный метод сентимент-анализа с учетом особенностей анализируемых данных

Методы:

- Машинное обучение
- Использование тональных словарей
- Использование расширенных тональных словарей с помощью Word2Vec

Данные

Корпус 1

- Русскоязычные новости с сайта Lenta.ru
- 8263 текста
- Три класса: positive, negative, neutral
- Источник: соревнование на Kaggle

Корпус 2

- Отзывы о кино с сайта КиноПоиск
- 46501 текст
- Три класса: positive, negative, neutral
- Источник: д/з по курсу МО

Машинное обучение

- sklearn
- Разные алгоритмы: Naïve Bayes, Logistic Regression, kNN, DecisionTree
- Разные варианты предобработки (с лемматизацией, удалением стоп-слов и пунктуации и без)
- Подбор гиперпараметров с помощью GridSearch
- Разные векторизаторы текста – CountVectorizer и TfidfVectorizer

Словарный подход

- Три словаря:
  - Тональный словарь, размеченный с помощью краудсорсинга. Объем – 11 тыс. слов (источник)
  - РуСентиЛекс, размеченный экспертами. Объем – 10 тыс. слов. (источник)
  - Словарь из исследования Manually Created Sentiment Lexicons: Research and Development (Kotelnikov E. V., 2016), размечен экспертами. Объем – 1114 слов (источник)
- Использование разных методов подсчета тональности текста: сложение точных (например, -0.6, 0.7) и округленных (-1, 1) показателей тональности всех слов в тексте, только позитивно и негативно окрашенных слов; разные числовые «границы» между классами (0 или +-1)

Расширение тонального словаря с помощью Word2Vec

1. Обучение модели Word2Vec на исследуемом корпусе/использование готовой модели, обученной на корпусе текстов по схожей тематике
2. Поиск 5 или 10 ближайших семантических ассоциатов для каждого слова в готовом тональном словаре и его пополнение полученными ассоциатами
3. Классификация текстов с помощью расширенного словаря

Примеры опорных статей: Анализ дистрибутивно-семантических моделей для пополнения словаря оценочной лексики (Пескишева Т. А., 2017); Build a Tourism-Specific Sentiment Lexicon via Word2Vec (Wei L., 2018)

Результаты. Корпус 1

Метод	F-мера	Accuracy
ML (NB, CountVectorizer, без пунктуации и стоп-слов)	**0.70**	**0.70**
Lexicon-based (SentiLex)	*0.46*	*0.48*
Lexicon-based (1114 words)	0.36	0.48
Lexicon-based w/w2v (10 associates, original data)	0.42	0.44
Lexicon-based w/w2v (10 associates, russian_news_model_rusvectores)	0.38	0.42

Результаты. Корпус 2

Метод	F-мера	Accuracy
ML (LR, CountVectorized, без пунктуации)	**0.81**	**0.82**
Lexicon-based (SentiLex)	0.57	0.50
Lexicon-based (1114 words)	0.56	0.49
Lexicon-based w/w2v (10 associates, original data)	*0.69*	*0.70*
Lexicon-based w/w2v (10 associates, web_mystem_skipgram_500_2_2015)	0.60	0.55

Наблюдения и выводы

- Наиболее эффективный метод классификации – машинное обучение
- Лучший словарь – РуСентиЛекс
- При отсутствии разметки в данных и при условии, что корпус достаточно большой, оптимально использовать смешанный подход
- Модель Word2Vec, обученная на «посторонних» данных, даже из схожего домена, расширяет словарь хуже, чем модель, обученная на исследуемом корпусе
- Оптимальное число семантических ассоциатов – 10
- Для метода с Word2Vec лучше использовать небольшой словарь
- Для определения тональности текста в виде «границы» лучше устанавливать не 0, а -1 и 1

Перспективы

- Использование лемматизатора MyStem вместо Pymorphy
- Тестирование метода с Word2Vec на других готовых моделях (более «чистых»)