Семинар №4

Тематическое моделирование и сентимент-анализ

НИУ ВШЭ – Санкт-Петербург

15.02.23

Тональность языка

“Я обожаю курс по молекулярной биологии”

“Я сходил на лекцию по анализу данных”

“Я ненавижу стоять в очереди в Гранолу”

Механизм субъективной оценки эмоциональности нарратива формируется в процессе ранней социализации индивида (Denham, 1986; Ensor & Hughes, 2008). Может ли это делать компьютер?

Формализация

Сентимент-анализ – область компьютерной лингвистики, которая занимается изучением **поляризованной** эмоциональности текста, содержащего мнения и отношения к личностям, событиям и темам.

Область применения:

- Бизнес и управление
- Социальные науки
- Образование
- Государственный сектор

Как провести (...) ?

- Анализ обратной связи потребителей
- Анализ цифровых соц. процессов
- Анализ качества преподавания
- Анализ реакции на законы, реформы и т.д.

Конвенциональные решения:

- Высокоструктурированные инструменты (e.g. опросы)
- Глубинные интервью
- Фокус-группы

Ограничения: сложно, времязатратно, дорого 😩

Операционализация

Как посчитать sentiment score в документе?

1. Абсолютное значение

N (positive terms) – N (negative terms)

2. Нормализация по размеру документа

N (positive terms) – N (negative terms) / N (terms per doc)

Краткий recap 1

Классическое машинное обучение бывает двух видов: с учителем (supervised) и без учителя (unsupervised):

1. Supervised предполагает наличие заранее размеченных данных, на основании которых модель/алгоритм может выполнить предсказание

2. Unsupervised предполагает, что модель/алгоритм самостоятельно выявляет закономерности и зависимости

Краткий recap 2

Supervised ML работает с двумя основными типами задач: регрессия и классификация. В контексте сентимент-анализа может быть как первое, так и второе:

1. Хотим предсказать численное значение сентимента документа => задача регрессии

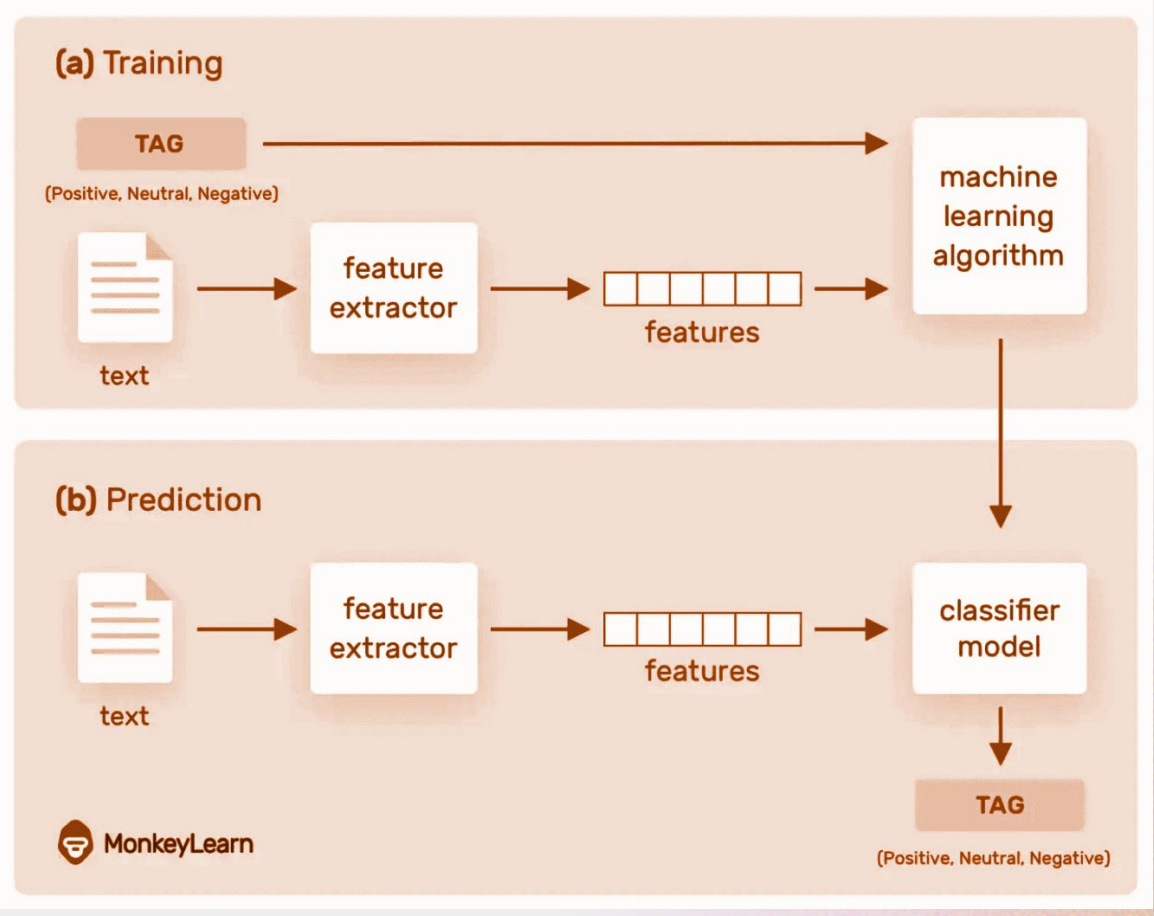
2. Хотим предсказать категориальный класс документа (позитивный, негативный) => задача классификации

Способы реализации

1. Классический ML подход

В большинстве случаев предполагает использование обучения с учителем. Здесь выполняются следующие шаги:

- Очистка данных (пунктуация, особые знаки и т.п.)
- Препроцессинг (e.g. лемматизация)
- Обучение алгоритма на размеченных данных
- Применение алгоритма на тестовых данных
- Оценка качества классификации/регрессии



Источник: monkeylearn.com

Способы реализации

Популярные алгоритмы сентимент-анализа:

- Наивный Байесовский классификатор
- Метод Опорных Векторов (SVM)
- Логистическая регрессия – для бинарного outcome

Где можно реализовать?

- R (e.g. naivebayes, e1071)
- Python (e.g. scikit-learn)
- Иные языка программирования

Способы реализации

2. Подход с использованием тональных словарей

Предполагает использование словаря, состоящего из слов и сентиментов. Базируется на коллекции текстов, объединенных одной темой (но не всегда). Например:

1. **Linis-Crowd** – тональный словарь на базе социо-политических текстов (Alexeeva S., et al. 2015).

2. **РуСентиЛекс** – сборная солянка из твитов, новостных статей и слов из тезауруса русского языка (Loukachevitch N., Levchik A., 2016).

Способы реализации

1. Лемматизация и токенезация
2. Объединение слов из ваших документов со словарем по факту содержания одинаковых значений в обеих таблицах:

Способы реализации

3. Обобщение

- Сколько позитивных/негативных слов в каждом документе?
- Какое значение сентимента для каждого документа?

4. Статистический анализ (опционально)

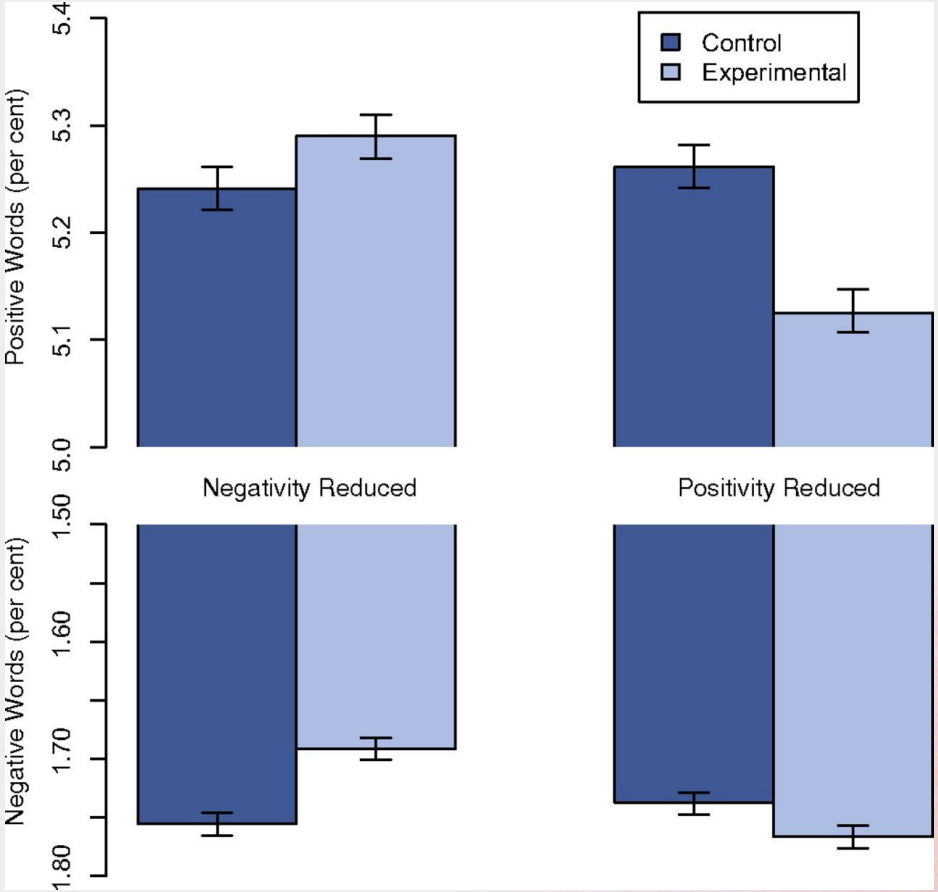
- Тесты на независимость
- Регрессионный анализ (OLS, как правило)

Применение

Явление эмоциональной контаминации

Эмоциональное заражение может происходить без непосредственного взаимодействия между людьми (достаточно наблюдения того как эмоция выражается), и при полном отсутствии невербальных сигналов (Kramer A., et al. 2014).

Подвержение человека негативной лексике может повлиять на его эмоциональные паттерны использования языка



Источник: (Kramer A., et al. 2014).

Выделение тем корпуса

Ситуация:

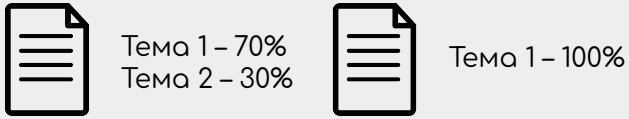
У вас 10 000 текстов, которые вам необходимо содержательно изучить, а затем провести сравнительный анализ, используя их качественные характеристики?

Читать тексты, а затем кодировать их вручную – невыполнимая задача. Как автоматизировать?

Ответ: ✨вероятностные тематические модели✨

LDA

Латентное размещение Дирихле – это модель, которая предполагает, что каждый документ это набор тем, распределение которых соответствует распределению Дирихле.



Каждый документ представляет собой вероятностное распределение **тем**

Каждая тема представляет собой вероятностное распределение **слов**

1. Использует заданное значение количества тем *K*
2. Случайным образом назначает каждое слово в каждом документе одной из *K* тем.
3. Рассматривает каждое слово и тему, к которой это слово было отнесено.
4. Отвечает на вопросы: (1) как частно тема встречается в документе? и (2) как часто данное слово встречается в теме?
5. Основываясь на ответах относит слово к новой теме
6. Повторяет процесс энное количество итераций

Пример:

Документ 1: *я люблю машинное обучение и статистику*
Документ 2: *котята и щенки мягкие и пушистые*

1. Пусть K = 2 : тема №1 (питомцы) и тема №2 (ML)

2. Слово “котята” из документа №2 случайным образом назначается теме №2

3. Тема №1 не часто встречается в документе 1, а слово “котята” не часто встречается в теме №2 => слово “котята” скорее всего принадлежит теме №1

4. Процесс повторяется для каждого слова

Результат:

Тема №1

Слово	Вероятность
котенок	0.8
щенок	0.7
уточка	0.5
...	...

Тема №2

Слово	Вероятность
статистика	1
машинный	0.7
ансамбль	0.5
...	...

Ранжирование по значению вероятности. Интерпретируем по тем терминам, которые наиболее вероятны для той или иной темы.

STM

Структурное тематическое моделирование – объединяет общие тематические модели (LDA, CTM) для создания нового подхода к моделированию тем, который также может включать ковариаты и метаданные при анализе текста (Roberts et al. 2014).

Помогает обойти ограничение LDA и использовать topic proportion для сравнительного анализа. Например, узнать уникальные и общие темы, которые обсуждаются различными соц. группами.

Ограничения

Общие к методам машинного обучения:

- Проблема производительности
- Подверженность переобучению
- Сложность в выборе опт. параметров
- Trash in – Trash out

Общие к сентимент-анализу:

- Идиомы и саркастические выражения
- Чувствительность к размеру текста
- Отсутствие универсальных словарей

References

Denham, S. A. (1986). Social cognition, prosocial behavior, and emotion in preschoolers: contextual validation. *Child Dev.* 57, 194–201. doi: 10.1111/j.1467-8624.1986.tb00020.x

Ensor, R., and Hughes, C. (2008). Content or connectedness? Mother-child talk and early social understanding. *Child Dev.* 79, 201–216. doi: 10.1111/j.1467-8624.2007.01120.x

Alexeeva, S., Kolcov, S., & Koltsova, O. (2015). Linis-crowd.org: A lexical resource for Russian sentiment analysis of social media (in Russian). *Kompyuternaya Lingvistika i Vichislitelnie Ontologii*, 25–34.

Loukachevitch, N., & Levchik, A. (2016). Creating a General Russian Sentiment Lexicon. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1171–1176. https://aclanthology.org/L16-1186

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111