Итоговое домашнее задание

Дудырев Егор

27.05.2020

1 Задание

Вариант 2.5: Анализ тональности
Анализ тональности текстов или анализ мнений по отношению к какому-то объекту (лицу, продукту, товару и т.п.) путем классификации текстов на 2 класса (положительный/отрицательный или положительный / отрицательный / нейтральный) на основе одного из подходов:

- машинное обучение (например, SVM)
- использование словарей и правил (шаблонов);

2 Используемая модель

На данный момент создано несколько моделей машинного обучения для работы с текстами, в частности - для анализа их тональности. При этом для их функционирования требуется перевести текст в векторное пространство вещественных чисел. Например, Bert переводит каждый текст в вектор из 768 чисел. В результате чего теряется интерпретируемость результатов анализа тональности. Используя Bert мы не можем сказать, почему модель классифицировала конкретный текст как тест с положительной или отрицательной тональностью.

Использовать более простые и интерпретируемые методы для анализа тональности также затруднительно. Логистическую регрессию имеет смысл применять только к векторам вещественных чисел. Для этого текст нужно закодировать. Деревья решений можно использовать и на категориальных признаках (напр. наличие/отсутствие какого-либо слова), но они также требуют определённой обработки текста. При этом любое кодирование либо делает результат неинтерпретируемым (Bert),

либо теряет часть информации о тексте (например TfIdf учитывает частоты слов, но не N-грамм и не коллокаций).

Анализ Формальных Понятий (АФП, Formal Concept Analysis) - область прикладной теории решёток, использующаяся в том числе и в задачах классификации [1]. Особенность АФП состоит в том, что он поддерживает работу со сложными структурами (текстом, изображениями, графами), которые нельзя представить в табличном виде.

Формальное понятие - это пара (A, B), где A - подмножество объектов, B - подмножество бинарных признаков, такое, что B - это максимальное подмножество признаков, которым обладают все объекты из A, A - максимальное подмножество объектов, которые обладают всеми признаками из B. Узорное понятие - обобщение формального понятия на более сложные структуры данных. Оно также состоит из подмножества объектов A, и неких описаний B, таких что B - максимальное описание, которому удовлетворяют все объекты из A, A - максимальное подмножество объектов, удовлетворяющих описанию B. На понятиях задаётся операция включения. Понятие (A_1, B_1) считается включенным в понятие (A_2, B_2), если $A_1 \subset A2, B2 \subset B1$.

Для того, чтобы автоматически получить набор понятий, содержащихся в датасете, можно использовать алгоритм ЗамыкайПоОдному (CloseByOne). В результате его выполнения получаем множество понятий $\{(A_1, B_1)$, (A_2, B_2), (A_3, B_3), ..., $(A_n, B_n)\}$.

Основываясь на данных понятиях можно составить гипотезы о принадлежности объектов к классу C вида:

$$B_i \implies_{p_i} C, p_i = \mathbb{E}[C|A_i]$$

Т.е. если объект удовлетворяет описанию B_i, то он с вероятностью $p_i = \mathbb{E}[C|A_i]$ принадлежит к классу C. Если объект удовлетворяет сразу m описаниям $B_1, B_2, ..., B_m$, при этом описания $B_1, B_2, ..., B_m$ несравнимы (ни одно из них не включено ни в одно другое), то он принадлежит классу C с вероятностью $p = \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[C|A_j]$.

Таким образом для работы с текстами с использованием методов АФП необходимо определить что есть объект, что есть описание объектов и как выполнять операцию включения на описаниях объектов.

Объектами в случае анализа отзывов являются сами отзывы. Описания же можно определить несколькими способами.

2.1 Классификация с помощью фреймов

Один из вариантов описания текстов - это набор фреймов, которые содержатся в текстах. Фреймы здесь указаны в том значении, в каком они упо-

требляются в онтологии FrameNet. Например, в тексте *hello world* содержатся слова, включенные во фреймы {*Being_ born, Attention_ getting, Political_ locales*}. Таким образом, через использование фреймов можно отследить семантику слов в предложении, которая должна кореллировать с тональностью предложения. При этом фреймы удобны для интерпретации человеком, т.к. были сформированы для того, чтобы отображать понимание человеком окружающего мира.

Примеры понятий, которые можно получить с данным подходом:

- (*hello world*, {*Being_ born, Attention_ getting, Political_ locales*})
- (*hello*, {*Attention_ getting*})

Сопоставив такие понятия и тональности отзывов можно вывести гипотезу вида:

$$\{Being\_\ born,\ Attention\_\ getting,\ Political\_\ locales\} \implies_{80\%} Positive$$

Т.о. можно сказать, что если в отзыве находятся слова, соответствующие фреймам {*Being_ born, Attention_ getting, Political_ locales*}, то этот отзыв с вероятностью 80% является положительным. Результаты такого анализа приведены в секции "Результаты эсперимента".

2.2 Классификация через подстроки отзывов

Недостаток работы с фреймами заключается в проблеме омонимии: действительно, в некоторых случаях слово *world* может быть связано с политикой, однако это не относится к тексту hello world. Вторая проблема состоит в том, что словам good и bad, оказывающим противоположный вклад на тональность текста, соответствует один и тот же фрейм *Desirable_ event*. Поэтому с точки зрения фреймов эти два слова неотличимы.

Данные проблемы можно обойти, если считать понятия на исходных текстах без какой-либо обработки.

Например, можно сказать, что описанием текста является множество максимальных N-грамм слов этого текста. А общее описание нескольких текстов A - это множество максимальных N-грамм слов, входящих в каждый из текстов A.

Например, пусть есть тексты:

1. *lorem ipsum dolor sit amet*
2. *ipsum dolor sit amet*

3. *lorem sit*

Тогда,

- Описание текста {1}: {*lorem ipsum dolor sit amet*}
- Общее описание текстов {1,2}: {*ipsum dolor sit amet*}
- Общее описание текстов {1,3}: {*lorem, sit*}

В этом случае, проанализировав отзывы можно составить гипотезы вида:

- {*lorem, sit*} $\Longrightarrow_{70\%}$ *Positive*
- {*ipsum dolor sit amet*} $\Longrightarrow_{90\%}$ *Positive*

Т.е. если в отзыве встречаются слова *lorem* и *sit*, то он с вероятностью 70% окажется положительным. Аналогично, если в отзыве встречается последовательность слов *ipsum dolor sit amet*, то этот отзыв будет положительным с вероятностью 90%. Результаты такого анализа приведены в секции "Результаты экспериментов".

3 Исходные данные

Для создания классификатора были использованы отзывы на сайте Imdb: Large Movie Review Dataset. Датасет содержит два набора отзывов: обучающий (25000 строк) и тестовый (25000) строк. Все тесты проводились только на первом, обучающем наборе, т.к. его объем достаточно большой для анализа.

Датасет с англоязычными отзывами был выбран по трём причинам:

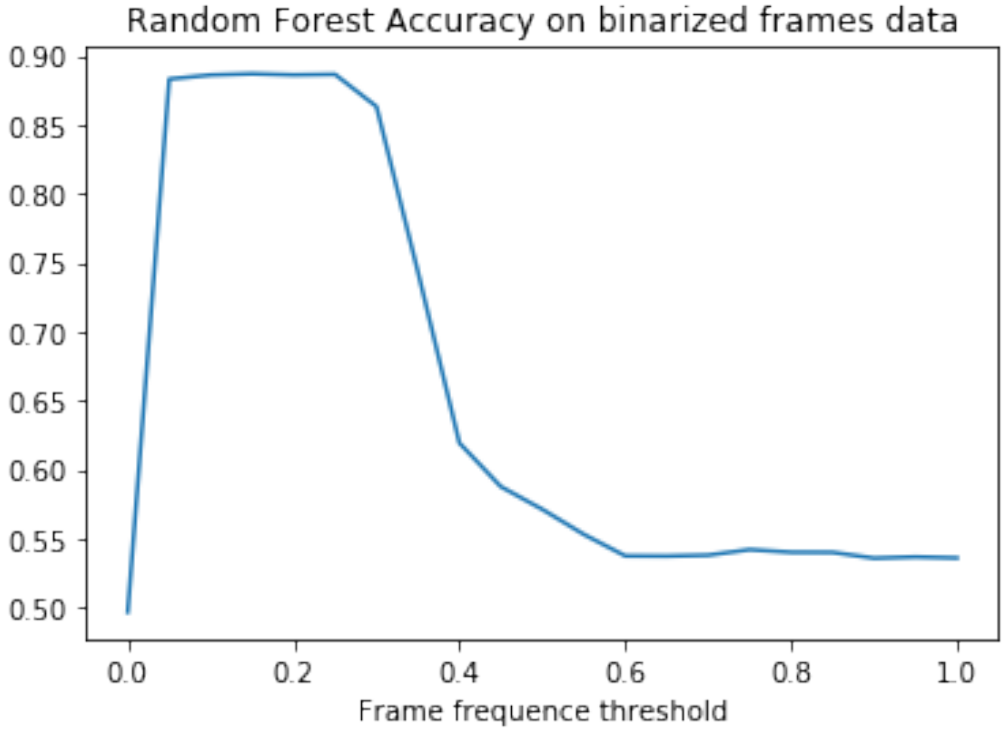
1. наиболее полная база фреймов слов на данный момент FrameNet поддерживает только английский язык.
2. для классификации использовался алгоритм, основанный на АФП. Ввиду новизны подхода, иммет смысл тестировать его на менее изменчивом языке.
3. новый метод классификации лучше сравнивать с лучшими существующими моделями анализа тональности. Например Bert.

4 Эксперименты

4.1 Классификация с помощью фреймов

Как уже говорилось выше, главная проблема связанная с фреймами - омонимия слов. В результате чего в одном тексте могут встречаться слова, относящиеся к 1100 фреймам (из 1221 фрейма, содержащегося во FrameNet). Для того, чтобы уменьшить количество фреймов в тексте, можно сделать следующее предположение: фрейм может содержаться в тексте, только в том случае, если к нему относятся несколько слов этого текста. Порог необходимой доли слов текста, относящихся к фрейму, был найден эмпирически через сравнение качества классификации тональности по оставляемым фреймам методом Random Forest (рис. 1).

Рис. 1: Качество классификации тональности текстов при разном пороге отсечения фреймов



Т.о. будем считать, что фрейм встречается к тексте, только если к нему относятся как минимум 15% слов этого текста.

Получившийся бинарный датасет текстов и содержащихся в них фреймов выглядит следующим образом:

id	Appellations	Personal_relationship	Stimulus_focus	Leadership	Containers
10684	True	True	True	True	True
10611	True	True	True	True	True
16289	True	True	True	True	True
12142	True	True	True	True	True
5367	True	True	True	True	True

Пять текстов, показанных в примере, были выбраны случайным образом. При этом во всех в них встречаются фреймы *Appelations, Personal_relationship* и другие.

Для расчёта понятий был взят датасет из 1000 текстов и 1221 признака, отражающих наличие фрейма в тексте. В результате получилось 96 понятий. Однако все из них содержат в своём описании большое количество фреймов.

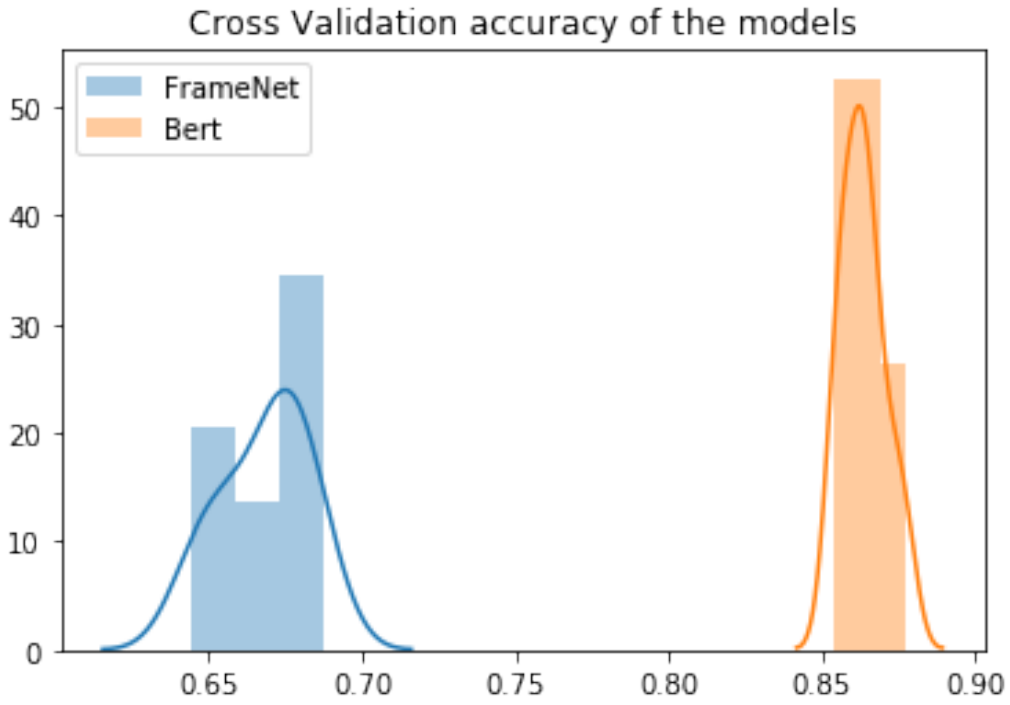
Так, например, понятие, в описании которого содержится наименьшее количество фреймов выглядит следующим образом:

- Понятие 0
- Кол-во текстов: 363
- Примеры текстов:
  №12142, №5367, №1984, ...
- Кол-во фреймов: 60
- Фреймы:
  Appellations, Personal_relationship, Stimulus_focus, Leadership, Containers, Clothing, People_by_vocation, Emotion_directed, Experiencer_obj, Medical_conditions, Kinship, Theft, Killing, Desirability, Body_parts, Buildings, Self_motion, Locative_relation, Frequency, Difficulty, Temporal_collocation, Judgment_direct_address, Placing, Wearing, Weapon, Desiring, Biological_urge, Out_of_existence, Intoxication, Text, Natural_features, Expertise, Locale_by_use, Quantified_mass, Time_vector, Amalgamation, Proportional_quantity, Building_subparts, Preventing_or_letting, Weather, Make_noise, Aggregate, Hostile_encounter, Morality_evaluation, Accoutrements, Scrutiny, People_by_origin, Cause_motion, Come_together, Evidence, Motion, Competition, Process, Activity, Crime_scenario, Criminal_process, Cycle_of_existence_scenario, Cycle_of_life_and_death, Operate_vehicle_scenario, Sleep_wake_cycle
- Вероятность положительной тональности: 0.484

Т.е., если в тексте содержатся 60 вышеперечисленных фреймов, то он будем обладать положительной тональностью с вероятностью 48.4%.

Из-за обилия фреймов такой результат сложно назвать интерпретируемым. Если же сосредоточиться на качестве классификации текстов с помощью фреймов, то можно посчитать частоту встречаемости фреймов в каждом тексте и подать полученные числа на вход логистической регрессии (она даёт лучший результат, чем деревья принятия решений и производные от них методы). Результаты такого подхода представлены на рис. 2.

Рис. 2: Качество классификации тональности текстов по частоте встречаемости фреймов



Данные метрики были расчитаны на всём наборе данных в 25000 строк с использованием кросс валидации по 10 фолдам. Модель построенная на фреймах даёт качество accuracy в 67%, что лучше чем случайный выбор, однако сильно хуже модели, основанной на кодировании Bert (accuracy = 87%).

4.2 Классификация через подстроки отзывов

Кластеризация с использованием FrameNet не дала ожидаемого интерпретируемого результата, поэтому попробуем поработать с исходным тек-

стом.

Для того, чтобы уменьшить вариативность исходного текста, заменим в нём все словоформы на их леммы. Это позволит избавиться от лишних шумов в данных, при этом не сильно изменив ни смысловую нагрузку текста, ни порядок слов в нём.

Случайным образом выберем 1000 отзывов из датасета и найдём содержащиеся в них понятия. В результате работы алгоритма получаем 410 понятий.

Для выбора наиболее интересных нам понятий будем использовать индекс стабильности понятия [3]. Он показывает насколько вероятно, что рассматриваемое понятие на самом деле существует в реальном мире, а не является шумом в данных.

Ниже приведены наиболее стабильные понятия, показывающие высокую вероятность положительных отзывов:

1. Узорное понятие 14
    - Кол-во объектов: 33
    - Объекты:
      №5364, №5067, №3203, №9630, №3204, ...
    - Описание:
      {'is', 'that', 's', 'the', 'of', 'and', 'a', 'today'}
    - Вероятность положительного отзыва: 80.5
2. Узорное понятие 19
    - Кол-во объектов: 24
    - Объекты: №4789, №12428, №1306, №991, №9068, ...
    - Описание:
      {'is', 'that', 'in', 'him', 'as', 'and', 'hi', 'a', 'live', 'to'}
    - Вероятность положительного отзыва: 66.7
3. Узорное понятие 17
    - Кол-во объектов: 27
    - Объекты:
      '5567', '3111', '1306', '991', '6301', ...
    - Описание:
      {'is', 'befor', 'in', 'movi', 's', 'the', 'wa', 'it', 'but', 'of', 'as', 'film', 'and', 'a', 'to'}

- Вероятность положительного отзыва: 67.5

Из описания узорного понятия №14 следует, что, когда в тексте отзыва содержится слово *today* (остальные леммы являются достаточно распространёнными), то этот отзыв с вероятностью в 80% оказывается положительным. Один из отзывов в этом понятии (№5067) рассказывает про фильм *The running man* с Арнольдом Шварцнеггером. Весь отзыв очень большой, поэтому приведу ту часть, где встречается слово *today*:
*It also features some at the time futuristic digital video editing, allowing the bad guys to change faces in a video to fool their audience. This does not seem futuristic at all today, which is a bit alarming.

If you've seen Arnold movies before then you know when to watch this one.*

Другой обзор из понятия №14 рассказывает о фильме *Dressed To Kill* Брайана Де Пальмы. Отрывок отзыва со словом *today*:
As far as i can tell every scene had a purpose,which i find very rare when compared to many of today's films.

Третий пример: документальный фильм *God's Country* Луи Маля.
*Once a proud community, now no one sees much future at all and parents hope their children will educate them self and do something else than farming.

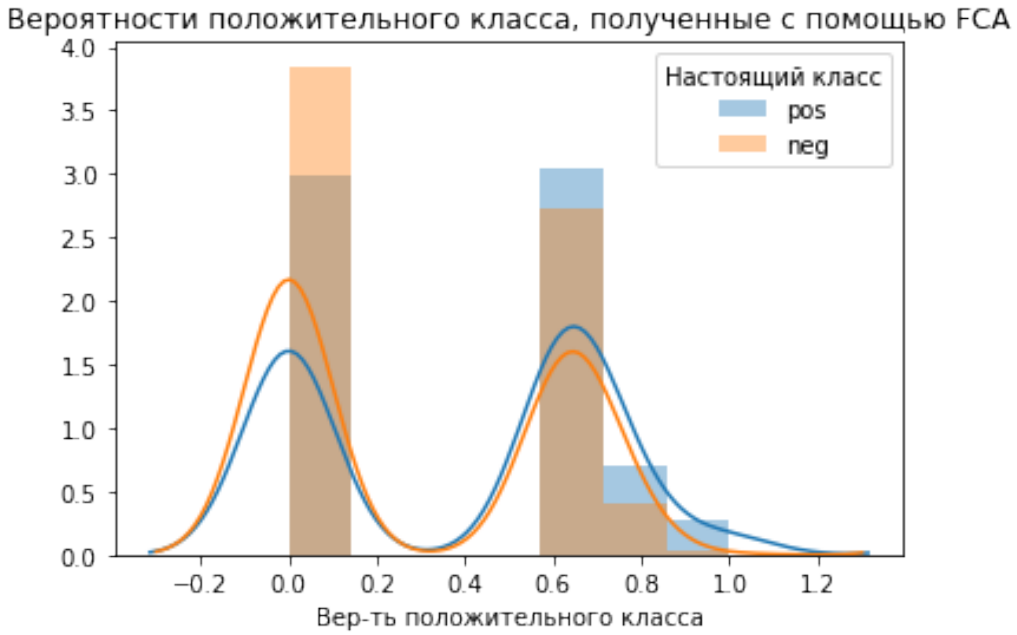
This documentary is quite relevant today. Our financial crises today started because of what was happening then.*

Можно предположить, что люди часто употребляют слово *today* в отзыве на киноленту, когда рассказывают об актуальности фильма, или когда испытывают ностальгию по давним временам. Поэтому актуальные или ностальгические картины обычно получают положительные оценки.

Остальные понятия можно найти в формате json здесь.

Однако если смотреть на качество такой модели на тестовой выборке из 1000 отзывов, то её accuracy составляет 56.2%. Это чуть лучше чем случайный выбор, чуть больше чем средняя доля положительных отзывов (50.7%). Но сильно хуже модели, основанной на Bert (86.6%).

Рис. 3: Вероятности принадлежности к положительному классу от алгоритма на АФП



На рис. 3 видно, что АФП может разделять положительные и отрицательные отзывы, но на данный момент делает это не очень хорошо.

5 Вывод

Данная работа была посвящена созданию интерпретируемого алгоритма по классификации тональности отзывов. В качестве метода, который может дать интересные интерпретируемые результаты, был выбран Анализ Формальных Понятий. В проведённых экспериментах он не дал какого-либо хорошего качества кластеризации в сравнении с Bert. Главная вероятная причина такого низкого качества - обучение лишь на 1000 примеров, ввиду низкой скорости работы алгоритма, а также его экспоненциальной вычислительной сложности от размера датасета.

Однако АФП позволяет высветить интересные особенности данных. Например то, что обзоры на актуальные или ностальгические фильмы (если в них встречается слово *today*) часто оказываются положительными.

Также было показано, что для качественного применения онтологии FrameNet необходимо использовать более сложные алгоритмы опреде-

ления фреймов в предложении. Т.к. подход, основанный на доле слов в тексте, которые относятся к тому или иному фрейму, даёт достаточно плохое качество предсказания.

6 Ссылки на код

- Ноутбуки с экспериментами
- Библиотека для работы с FCA
- NLTK версия онтологии FrameNet

7 Список источников

1. Ю.С. Кашницкий, Методы замкнутых описаний в задаче классификации данных со сложной структурой. Электронный доступ.
2. Aleksey Buzmakov, Sergei O. Kuznetsov, Amedeo Napoli. Concept Stability as a Tool for Pattern Selection. FCA4AI 2014. What can FCA do for Artificial Intelligence?, Aug 2014, Praque, Czech Republic. ffhal-01095903