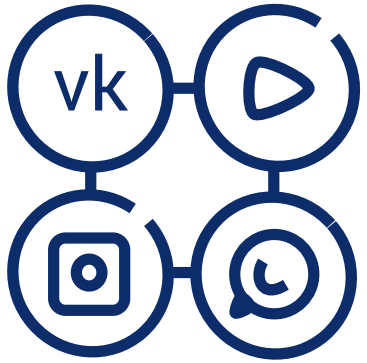
Изучение социального настроения граждан с помощью машинного обучения

Итоговый проект первого курса
магистерской программы ВШЭ
«Машинное обучение и высоконагруженные системы»

Проблема



Гуманитарные науки отстают от развития современных технологий. Большой пласт информации лежит в социальных сетях

Решение

Использование машинного обучения в анализе постов соцсетей выявляет эмоционального настрой населения

Brand Analytics Forum 2023

Аналитика соцмедиа для государства

26 апреля 2023
10:00 – 20:00

Пресс-центр МИА «Россия сегодня»
Москва, Зубовский бульвар, 4

О форуме

Главная тема форума – использование аналитики социальных медиа для решения задач исполнительной власти.

- Что такое клиентоцентричное государство?
- Как выявлять проблемы в социальной сфере?
- Как власти отслеживать эффективность своих решений и реакцию на них в обществе?

Наш форум ответит на эти и многие другие вопросы.

PASHUPATI GUPTA · UPDATED 3 YEARS AGO

Emotion Detection from Text

Predict emotion from textual data : Multi-class text classification

About Dataset

Context

Emotion detection from text is one of the challenging problems in Natural Language Processing. The reason is the unavailability of labeled dataset and the multi-class nature of the problem. Humans have a variety of emotions and it is difficult to collect enough records for each emotion and hence the problem of class imbalance arises. Here we have a labeled data for emotion detection and the objective is to build an efficient model to detect emotion.

Content

The data is basically a collection of tweets annotated with the emotions behind them. We have three columns tweet_id, sentiment, and content. In "content" we have the raw tweet. In "sentiment" we have the emotion behind the tweet. Refer to the starter notebook for more insights.

Acknowledgements

This public domain dataset is collected from data.world platform. Thanks, data.world for releasing it under Public License.

Inspiration

The data that we have is having 13 different emotion 40000 records. So it's challenging to build an efficient multiclass classification model. We may need to logically reduce the number of classes here and use some advanced methods to build efficient model.

Usability

7.65

License

CC0: Public Domain

Expected update frequency

Never

Tags

Text

NLP

Deep Learning

Существующие решения на рынке

Готовые решения	Buzzsumo	Klear	Traackr
Стоимость ($/месяц)	От 119	От 99	От 499
Наличие триального периода	30-дневный	Демо-версия	Демо-версия
Основной покупатель	Малый бизнес	Малый бизнес	Средний бизнес
Наличие кастомизации	Нет	Да	Нет
Отрасль	Ретейл	Ретейл	Поиск инфлюенсеров

Команда проекта

Ксения Лущева

*ML-инженер, тестировщик

Мария Аугуст

*Data Scientist

Павел Егоров

*Аналитик, менеджер проекта


Презентация
Парсинг
шаг 6
шаг 1
Тестирование
шаг 5
Цикл жизни проекта
шаг 2
EDA
шаг 4
шаг 3
Обучение модели
Выбор модели

Цель проекта:

Получен набор данных:

2,7 млн постов vk.com

Цель обучения модели:

определять тексту поста эмоциональную тональность сообщения:

- Негативное
- нейтральное
- позитивное

План работ

~~Выбор семейства моделей~~

~~Обучение модели~~

~~Тестирование и апробация модели~~

Итоговая презентация проекта

Пример датасета

	id	text	created_at	sentiment	emotion	toxicity	is_congratulation	
0	12963175	Многократно экранизированный и поставленный...	2023-11-01 11:40:28	POSITIVE	no_emotion	TOXIC! Внимание, перед Вами токсичное сообщение!	False	
1	12963451	Продолжает работу Осенний лагерь в нашей школе...	2023-11-01 15:08:04	POSITIVE	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
2	12963452	Сегодня, мы съездили в с\nВоскресенское. Наша ...	2023-11-01 14:52:52	POSITIVE	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Ham
3	12963698	Более 170 лекций запланировали в рамках просве...	2023-11-01 14:40:19	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
4	12963699	1 ноября у нас в гостях были воспитанники подг...	2023-11-01 09:56:06	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
5	12963700	В преддверии Дня народного единства обучающиес...	2023-11-01 09:05:20	POSITIVE	no_emotion	TOXIC! Внимание, перед Вами токсичное сообщение!	False	Spam
6	12963835	Уважаемые жители сельского поселения Орловс...	2023-11-01 17:06:31	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
7	12963836	Вот такая красота на нашем новом мемориале в с...	2023-11-01 07:11:47	POSITIVE	joy	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
8	12963843	В Башкирии отловили алабая, напавшего на людей...	2023-11-01 18:53:58	NEUTRAL	no_emotion	TOXIC! Внимание, перед Вами токсичное сообщение!	False	Spam
9	12966277	Отдам букварь для подготовки к школе.	2023-11-01 10:38:06	POSITIVE	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Ham
10	12966610	Сегодня в службе семьи прошёл "Единый день сем...	2023-11-01 16:53:41	POSITIVE	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
11	12966611	Участник СВО при увольнении в связи с признан...	2023-11-01 11:43:35	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
12	12966996	Сегодня 3 курс Биоэкология на предмете физиоло...	2023-11-01 11:20:17	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
13	12967451	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 18:35:08	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
14	12967452	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 14:33:47	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
15	12967453	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 14:22:27	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
16	12967454	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 14:21:02	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
17	12967455	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 14:20:00	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
18	12967456	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 13:03:04	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
19	12967457	ТСК Оптовик\nсмс рассылка приостановлена \nчто...	2023-11-01 13:02:14	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
20	12967458	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 12:24:42	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam
21	12967459	4-5 ноября\n НОВОЕ ПОСТУПЛЕНИЕ\n -20% на муж...	2023-11-01 11:45:43	NEUTRAL	no_emotion	NOT TOXIC! Это сообщение не является грубым ил...	False	Spam

0 сек. выполнено в 21:40

История не(успеха):
Выбор семейства моделей:

Сначала нами были апробированы:

1. Bag of words
2. TfIdfVectorizer
3. Комбинация TfIdf + дополнительные признаки
4. Комбинация stopwords + Word2Vec

Bag of Words в сочетании с CountVectorizer и Logistic Regression для обучения языковых моделей, является прочным фундаментом в (NLP)

Высокую точность при использовании 1-граммовой модели

Достижение точности 0,94 с помощью 1-граммовой модели впечатляет и указывает на то, что модель очень эффективно классифицирует настроения как положительные или отрицательные на основе отдельных слов. Это говорит о том, что для набора данных наличие или отсутствие определенных ключевых слов является сильным предиктором настроения

Мы получили снижение точности при использовании 3-граммовой модели

Снижение точности до 0,85 при использовании 3-граммовой модели скорее всего говорит о том, что включение контекста окружающих слов (до трех слов вместе) не только не улучшает, но даже может несколько ухудшить работу модели.
- Хотя подход Bag of Words прост и эффективен для многих задач, он не учитывает порядок слов и семантические отношения между словами. Поэтому на следующих этапах мы постарались учесть более сложные модели

TfIdfVectorizer и расширение значений целевых переменных демонстрирует развитие и масштабирование проекта

Использование TfIdfVectorizer с LogisticRegression сохраняет высокую точность бинарной классификации

Небольшое увеличение точности 3-граммовой модели с использованием TfIdfVectorizer (с 0,85 до 0,86) по сравнению с подходом BoW позволяет предположить, что взвешивание TfIdf, которое подчеркивает важность менее частотных слов, может быть более эффективным для улавливания нюансов контекста в больших n-граммах

Получили первую проблему с многоклассовой классификацией. Расширение категории до "нейтральный" и обучение ряда классификаторов на векторах (1,2)-грамм показало заметное падение точности во всех моделях

Это может указывать на недостаточно хорошую разметку данных, в связи с чем, одной из потенциально решаемых задач в нашей работе - уточнение сентиментов

Показатели f1 для положительных и отрицательных категорий ниже 0,3 свидетельствуют о значительных проблемах в достижении сбалансированной классификации с помощью текущих моделей

Включение дополнительных признаков в TfIdfVectorizer

На следующем шаге мы постарались включить дополнительные параметры:

- Добавили новые признаки, таких как emotion, toxicity, is_congratulation и spam, чтобы дать модели улавливать более широкий спектр текстовых нюансов
- Применили TruncatedSVD для сжатия матрицы признаков до 100 признаков. Это поможет решить проблемы, связанные с высокой размерностью, такие как проклятие размерности и чрезмерная подгонка, что сделает вашу модель более обобщенной
- Использовали OHE для дополнительных признаков. Это может помочь в точном отражении влияния каждого признака на целевую переменную
- Продолжили эксперименты с TfIdf для кодирования текста с помощью 1-грамм

Использование Word2Vec

Мы предприняли переход к использованию nltk для удаления стоп-слов и модели Word2Vec для векторизации текста

Использование nltk для удаления стоп-слов. Удаление стоп-слов имеет решающее значение для уменьшения шума в текстовых данных

Использование Word2Vec для векторизации текста. Word2Vec учитывает контекстуальные нюансы и семантические связи между словами, в отличие от методов BoW и TfIdf

Используя Word2Vec, мы преобразуем текст в векторы, которые представляют слова в непрерывном векторном пространстве

Этот подход эффективен для улавливания контекста слов, понимания синонимов и для улавливания определенного настроения

Мы рассчитываем, что в нашем случае это приведет к созданию более эффективной модели, особенно для задач, требующих глубокого понимания семантики текста, таких как анализ настроения, классификация текстов и рекомендательные системы

Наконец мы попробовали BERT

Мы дообучили модель DeepPavlov/rubert-base-cased на своей разметке и получили кастомную модель

Для начала мы разметили вручную 20000+ постов.

Затем мы поняли, что получается большой перекос в сторону нейтральных постов

Мы стали прицельно искать позитивные и негативные посты, для баланса выборки

Результаты модели:
Общая точность модели составила 0.94
(подробнее см. следующий слайд)

На финальном этапе мы сделали так, чтобы наша модель работала через FastAPI. Модель работает с локального компьютера, но возможно ее перенесение в облако

Пример работы модели на FastAPI:

FastAPI 0.1.0 OAS 3.1

/openapi.json

default

POST /analyze Analyze Text

Parameters Cancel Reset

No parameters

Request body required application/json

```
{
  "text": "Наш любимый сок отравлен"
}
```

Servers

These operation-level options override the global server options.

/

Execute Clear

Responses

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8000/analyze' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "text": "Наш любимый сок отравлен"
}'
```

Request URL

Результаты нашей модели на тестовых данных:

1

precision:

Negative 0.99
Neutral 0.96
Positive 0.86

2

recall:

Negative 0.97
Neutral 0.95
Positive 0.90

3

f1-score:

Negative 0.98
Neutral 0.95
Positive 0.88

4

Accuracy 0.94

Рефлексия по проекту:

1. Эмоциональная тональность поста наименее точная часть компьютерной лингвистики
2. Использование дополнительных метрик: просмотры, репосты, лайки не дало улучшения модели
3. Основное продвижение дала ручная размета постов, что заняло много времени, но дало результат
4. Определение тональности приемлемо как учебный проект, но нужна опора на компетентного куратора


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •