Автоматические определение эмоций в русскоязычных текстовых сообщениях

Александр Бабий

Задача

Разработать мультиклассовый классификатор эмоций, выраженных в коротких неформальных текстах с использованием методов машинного обучения и лингвистических признаков

Разделы доклада

СБОР ДАТАСЕТА

РАЗРАБОТКА МОДЕЛИ

ОЦЕНКА РЕЗУЛЬТАТОВ И ПЕРСПЕКТИВЫ УЛУЧШЕНИЯ

Этап 1. Сбор датасета

- Источники данных: ВКонтакте и Телеграм
- Разработка парсера для VK и TG
- Выборка – 4584 человека
- Нормализация текстовых данных
  - Средняя длина сообщений составляет 4-5 слов
  - Удаление сообщений длиннее 9 слов
  - Удаление различных тегов, хэштегов и сообщений, не содержащих кириллических символов
  - Замена почт и телефонных номеров токенами <email> и <phone>
  - Лемматизация (rnnmorph)
- Размер лемматизированного корпуса – 1.800.000 сообщений

Этап 1. Сбор датасета

Изначальная классификация (по Экману)

- Радость
- Удивление
- Грусть
- Злость
- Страх
- Отвращение
- Презрение

Этап 1. Сбор датасета

Принятая классификация

- Радость
- Грусть
- Злость
- Неуверенность
- Нейтральность

Этап 1. Сбор датасета

Для распределения эмодзи по группам эмоций был создан скрипт, позволяющий проанализировать употребление эмодзи в контексте (поиск в корпусе).

Эмодзи, встретившиеся в корпусе менее 40 раз, были удалены.

Этап 1. Сбор датасета

Создание наборов эмодзи для каждой категории эмоций

Счастье: 😛, 💗, 😊, 🥲, 💚, 🤗, 💘, 💙, 💋, ❤️‍🔥, 😂, 😇, 💝, 💜, 💖, 🧡, 💛, 🖤, ❤️

Грусть: 😩, 😕, 😟, 😒, 😢, 😪, 🥲, 🙄, 😞, 😭, 😔, ☹️, 😓, 😰

Злость: 😡, 😠, 👹, 🥶, 🤦‍♀️, 😐, 😬

Неуверенность: 👀, 🧐, 🤔, 🙏, 🤞

Нейтральные сообщения были размечены вручную, так как они не выражают явных эмоций и достаточно редко содержат эмодзи

Этап 1. Сбор датасета

Размер корпуса сообщений с эмодзи – 4500 документов

Создание бинарных классификаторов для дальнейшей разметки

Размер итогового датасета – 110.000 размеченных сообщений

Этап 1. Сбора датасета

Все сообщения, содержащие нестандартное (ирония, сарказм) использование эмодзи были исключены.

Количество сообщений с использованием эмодзи в прямом смысле составило 4500 сообщений из 11287, содержащих эмодзи.

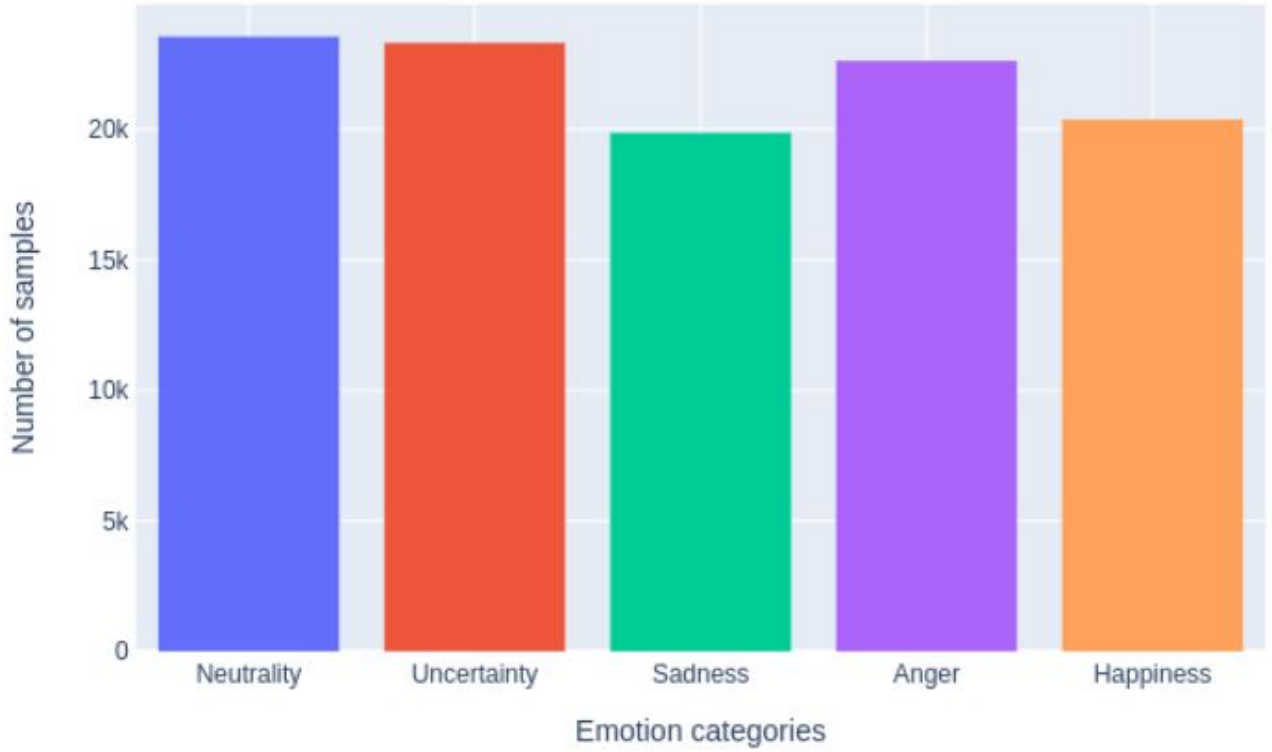
Этап 1. Сбор датасета

Для создания бинарных классификаторов логистическая регрессия и наивный байесовский классификатор были использованы.

Для этого сообщения с нужной эмоцией были отмечены как «1», а остальные как «0» (1-vs-all strategy).

F1-Score классификаторов составляет 90-93%. Наконец, были отобраны только те сообщения, которые были помечены как «1» только одним из классификаторов.

Этап 1. Сбор датасета



Этап 2. Разработка модели

Сравнение традиционных моделей машинного обучения

	Precision	Recall	F1
Logistic Regression	0.72	0.69	0.70
Naïve Bayes	0.73	0.68	0.69
Random Forest	0.67	0.65	0.65

Этап 1. Сбора датасета

С помощью *scikit-learn* имплементации TF-IDF vectorizer были извлечены 29714 признаков.

Логистическая регрессия показала лучший результат.

Этап 2. Разработка модели

На этапе конструирования признаков были проведены POS-tagging на сообщениях (до нормализации) и эксперименты со следующими признаками:

- Коэффициент Трейгера (отношение количества глаголов к количеству прилагательных в единице текста); связан с уровнем эмоциональной стабильности и указывает на соотношение у субъекта высказывания наклонности к активности
- Коэффициент агрессии (отношение количества глаголов к общему количеству слов)
- Коэффициент определённости действия (отношение количества глаголов к количеству существительных)
- Количество цифр, восклицательных знаков и т.д.

Но это не дало результатов.

Этап 2. Разработка модели

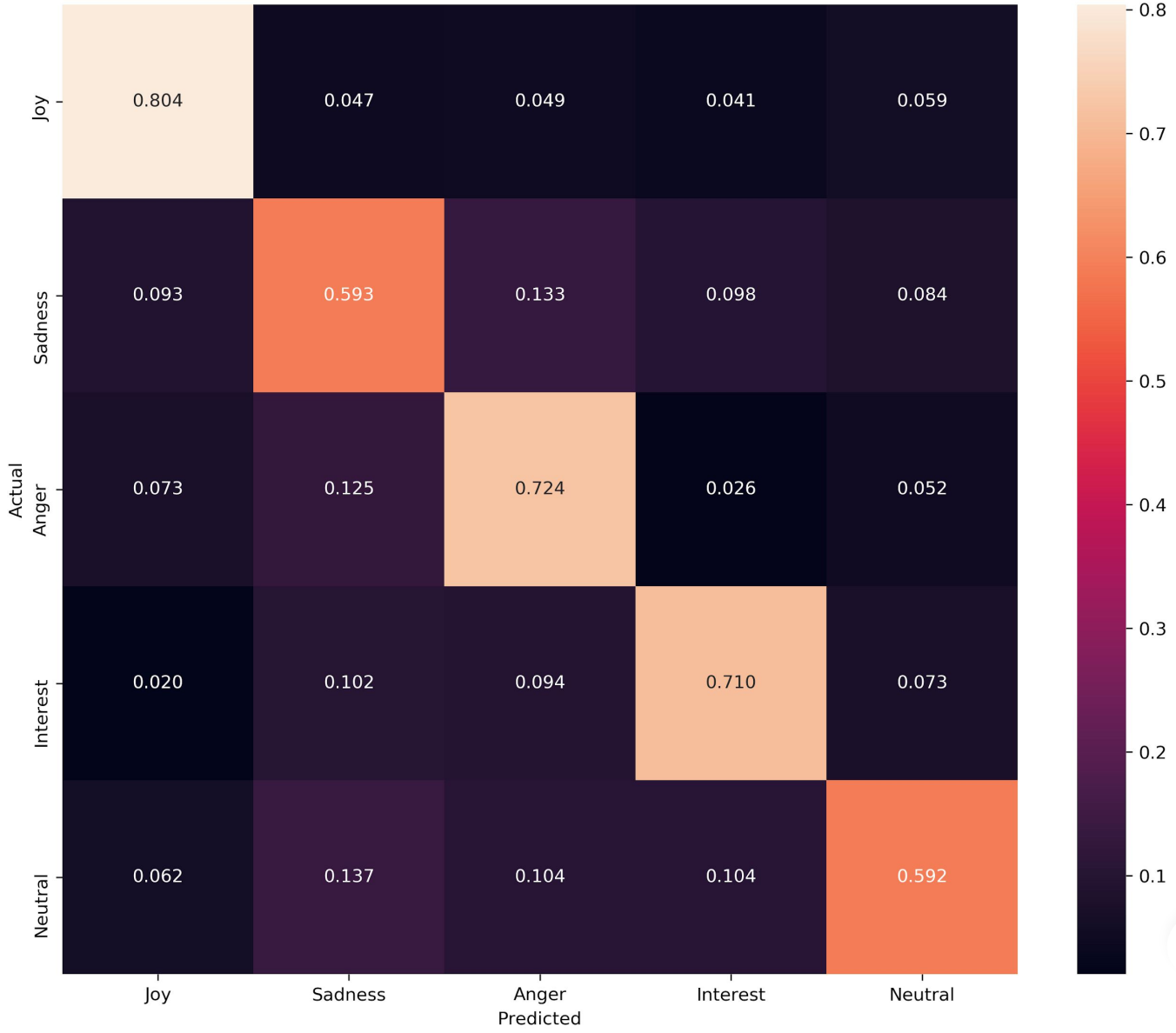
Калибровка гиперпараметров с помощью *scikit-learn GridSearchCV*

	Precision	Recall	F1-Score
Joy	0.90	0.80	0.85
Sadness	0.72	0.59	0.65
Anger	0.50	0.72	0.59
Interest	0.59	0.71	0.64
Neutrality	0.48	0.59	0.53
Weighted average	0.74	0.71	0.718

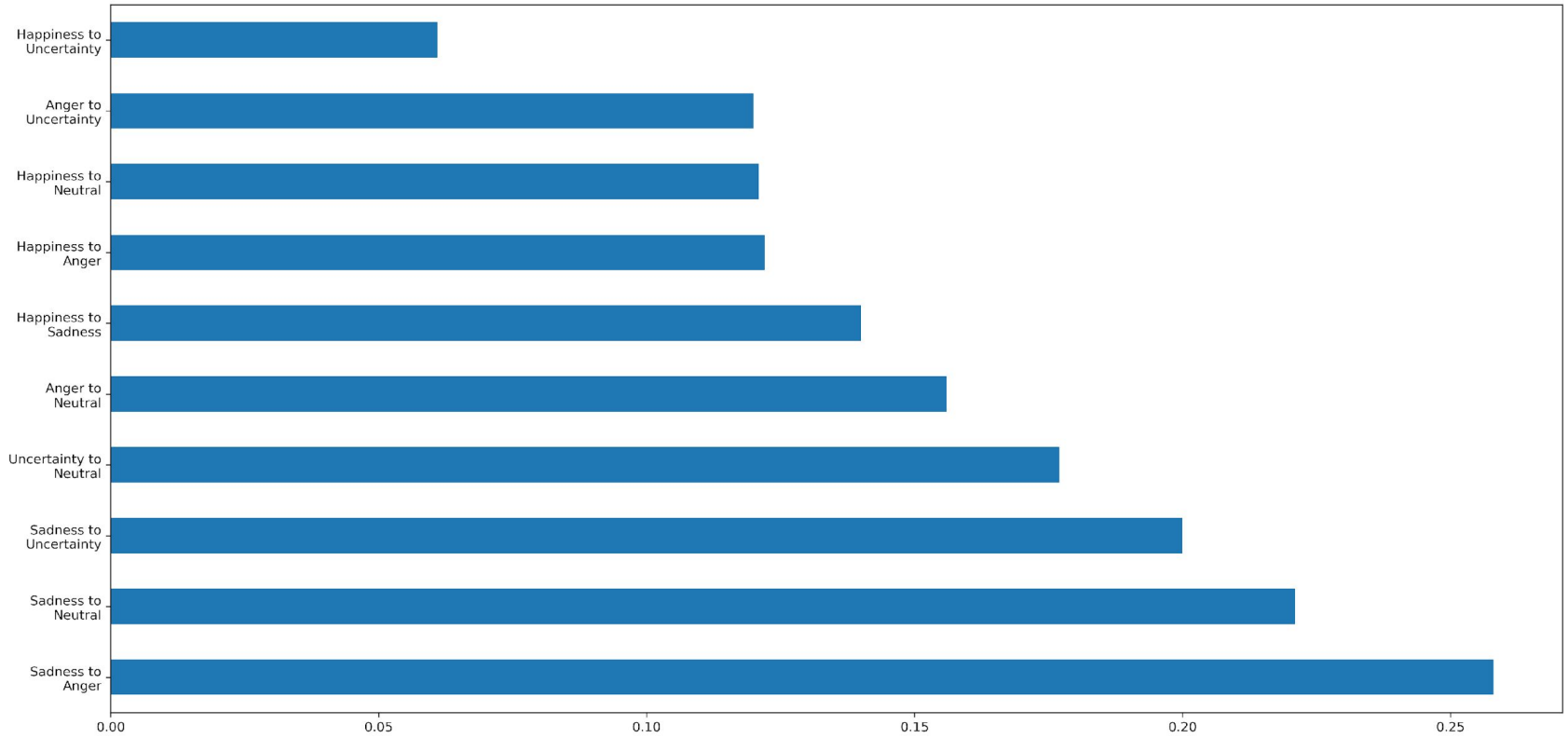
Этап 3. Оценка результатов

- Разметка части валидационного набора данных людьми
  - 4 человека возрастом 18-19 лет разметили 250 сообщений
  - Средняя точность – 74%
  - Согласие между разметчиками – 70%
- Разметка модели
  - Средняя точность – 71.8%

Этап 3. Оценка результатов



Этап 3. Оценка результатов



Перспективы

- Провести эксперименты с новыми признаками
- Проанализировать информацию о важности признаков с помощью *eli5*
- Переработать классификацию эмоций
- Рассмотреть использование метаданных сообщений
- Провести эксперименты с другими типами ML моделей
- Провести эксперименты с DL

Ссылки

Парсер: https://github.com/bac03704-byte/AIST/blob/master/Parser.py

Лемматизатор: https://github.com/IlyaGusev/rnnmorph

Маркеры эмоций: http://www.vestnik.vsu.ru/pdf/lingvo/2015/03/2015-03-16.pdf