Оценка и прогнозирование восприятия решений больших языковых моделей разработчиками программного обеспечения с использованием данных GitHub

Пименов Арсений Егорович,

НИУ ВШЭ Санкт-Петербург, Рафт Диджитал Солюшнс,

канд. техн. наук, Ковальчук Сергей Валерьевич,

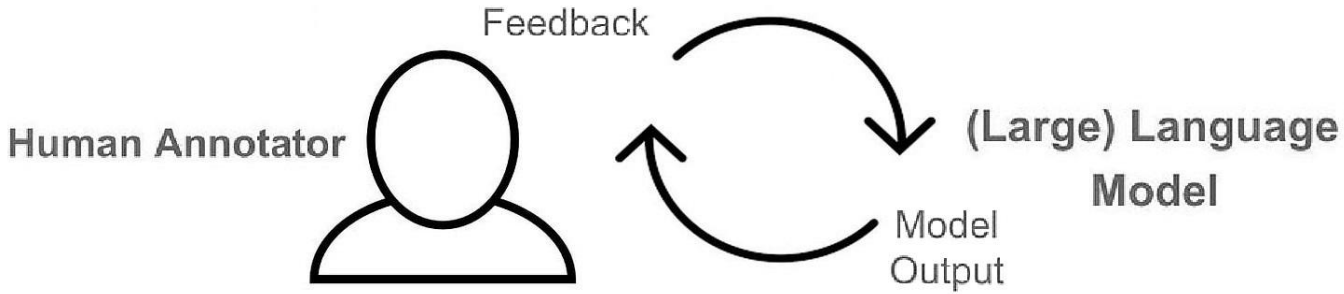
Университет ИТМО, доцент ФЦТ

Введение

- Активное внедрение интеллектуальных ассистентов на базе больших языковых моделей в задачи разработки и поддержки ПО сопровождается недоверием разработчиков.
- Взаимодействие разработчиков и ассистентов происходит в рамках неоднородного и сложного контекста.
- Контекст влияет на восприятие решений и обратную связь разработчиков необходимой для создания более качественных моделей.

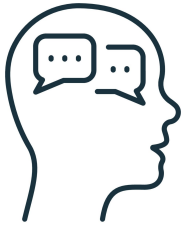
GitHub Copilot

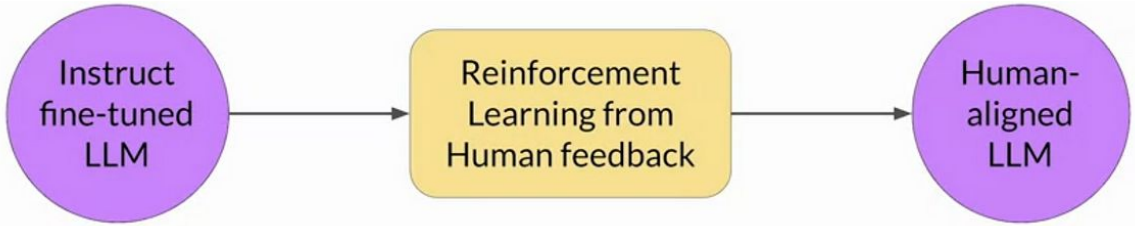
tabnine



Введение

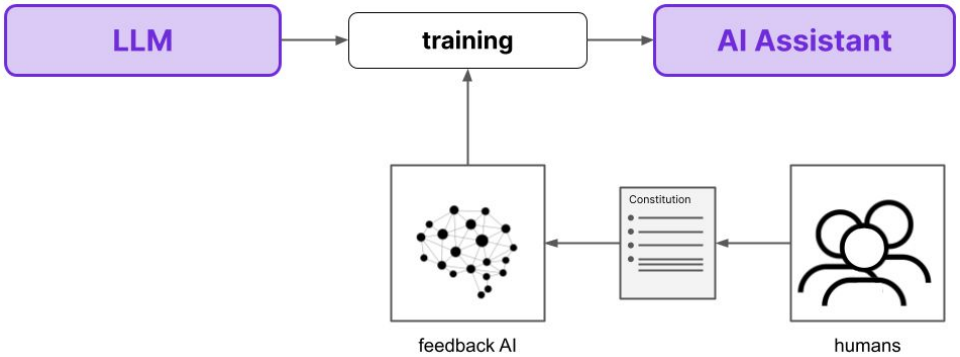
- Использование обратной связи при обучении больших языковых моделей (RLHF) повышает их качество и делает более ориентированными на человека, что улучшает процесс разработки и поддержки ПО.
- Внешний контекст – описание задачи, используемый язык программирования, используемые технологии, уже существующий код проекта
- Внутренний контекст зависит от когнитивной системы: характера, эмоций, настроения, от профессиональных навыков, от решаемой задачи "как разработчик ее держит в голове".





Проблематика

- Мало обратной связи на решения интеллектуальных ассистентов в открытом доступе.
- Прогнозирование восприятия ответов разработчиков с помощью знаний о внутреннем контексте позволит создать искусственную обратную связь.
- Искусственная обратная связь позволит удешевить и ускорить процесс обучения моделей генерации кода.



Обзор литературы

- Ответы моделей не соответствуют требованиям разработчиков. Исследуют взаимодействие разработчиков и интеллектуальных ассистентов[1].
- Важно учитывать человеческое восприятие при оценивании сгенерированных результатов. Есть отличия в метриках и оценках людей[2].
- Применение RLHF для обучения модели решения задач генерации python кода. Возможность замены RLHF на RLAIF из-за стоимости человеческой обратной связи[3].

Цели и задачи

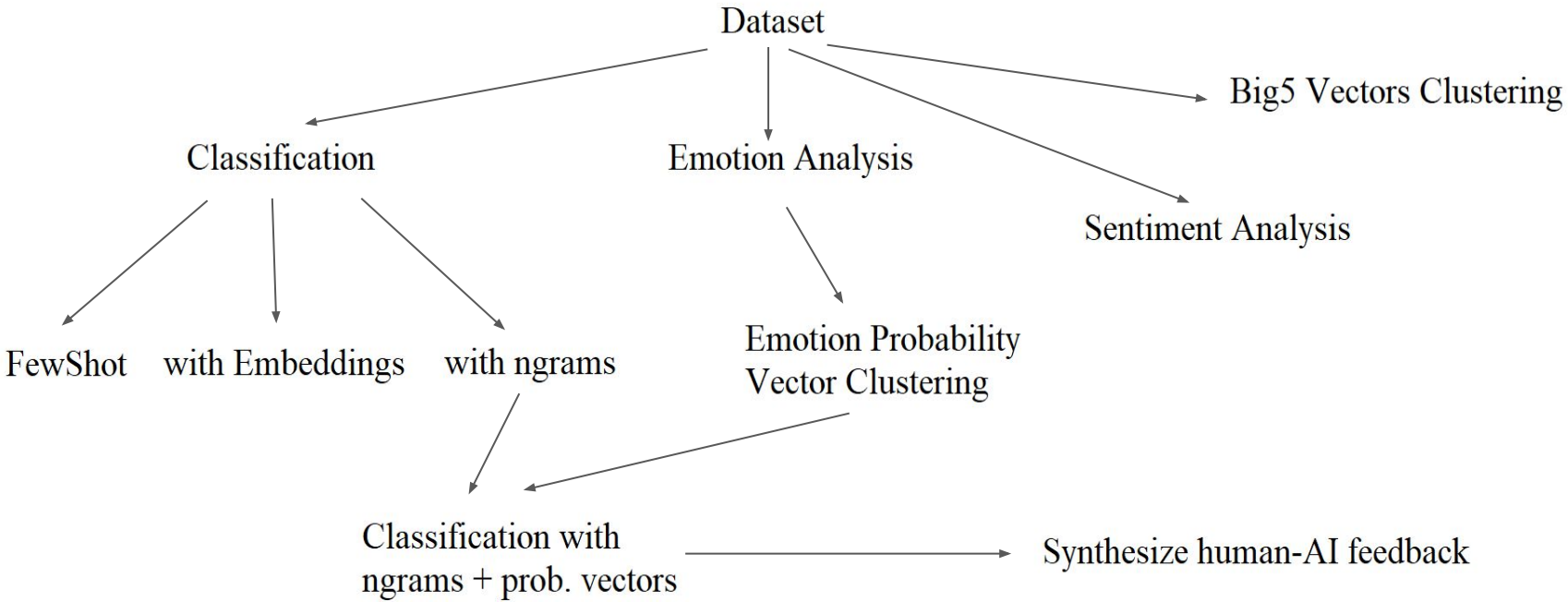
- Оценить восприятие решений, связанных с интеллектуальными ассистентами, на примере ChatGPT.
- Разработать методы синтеза обратной связи к данным решениям из обратной связи к решениям разработчиков с помощью знаний об отличии восприятия разработчиками ChatGPT сгенерированных ответов/кода от восприятия ответов других разработчиков.

Цели и задачи

Задачи:

- Собрать данные о взаимодействии разработчиков с ChatGPT решениями и решениями других разработчиков на Github.
- Применить модели психологического анализа для оценки воспринимаемости разработчиками ответов и получения новых признаков.
- Разработать модели классификации различия между типами обратной связи.
- Применить полученные результаты для предсказания обратной связи.

Схема исследования



Использованные методы

- Анализ сентимента. Модель distilbert-base-multilingual-cased-sentiments-student.
- Анализ текста по психологической модели Большой пятерки и кластеризация (k-means, DBSCAN) полученных векторов личностных характеристик.
- Классификация.
  - FewShot, модель Llama-2-13B-GPTQ, точность 50.2%
  - N-граммы и эмбеддинги с различными моделями классификации.
- Анализ эмоций и кластеризация (k-means, DBSCAN) векторов вероятностей эмоций.
- Обучение с подкреплением моделей перефразирования текста pegasus_paraphraser и chatgpt_paraphraser_on_T5_base.

Источники данных

harshvardhanbarhan commented on Jun 16, 2023 — Author

hi @jeromegamez
May be i am wrong, but it tried the solution from this

https://chat.openai.com/share/58022604-2a05-4ab9-b1dd-8a4c3e8bb471

I want to confirm that is it safe to use Firebase client SDK for phone number authentication
Because keys will remain in front-end in the case of Firebase client sdk

jeromegamez commented on Jun 16, 2023 — Member

ChatGPT is wrong.

Источники данных

- Датасет DevGPT с ссылками на упоминания ChatGPT расширен данными с Github.
- Комментарии с ChatGPT: 281 запись, 253 проекта, 231 автор.
- Обычные комментарии: 2839 записей, 134 проекта, 173 автора.
- В сборном датасете около 600 записей.

	GithubURL	Comments	OldReply	Reply	Label
	https://github.com/app-sre/qontract-reconcile/...	All methods in this class could be properties.	Right, `get_` is a bad name here. fixed	right get_ is a bad name here fixed	0

Предобработка данных

- Обработка текста, удаление ключевых слов (12 слов), связанных с интеллектуальными ассистентами, по типу ‘chatgpt’, ‘llm’, и др.
- Приведение к N-граммам (N ≤ 3) и эмбеддингам (256, 512). К примеру униграммы (576, 2871).
- Класс 0 – обратная связь разработчика на решения разработчика.
- Класс 1 – обратная связь разработчика на решения, связанные с ChatGPT/

Классификация

- Выбор лучшей модели (Логистическая регрессия, Случайный лес, Дерево решений, KNN).
- Осуществлен подбор параметров классификатора и метода векторизации текста.

Вывод:

Униграмы лучше всего улавливают локальный контекст и тональность.
Результаты подтверждают различие между обратными связями.

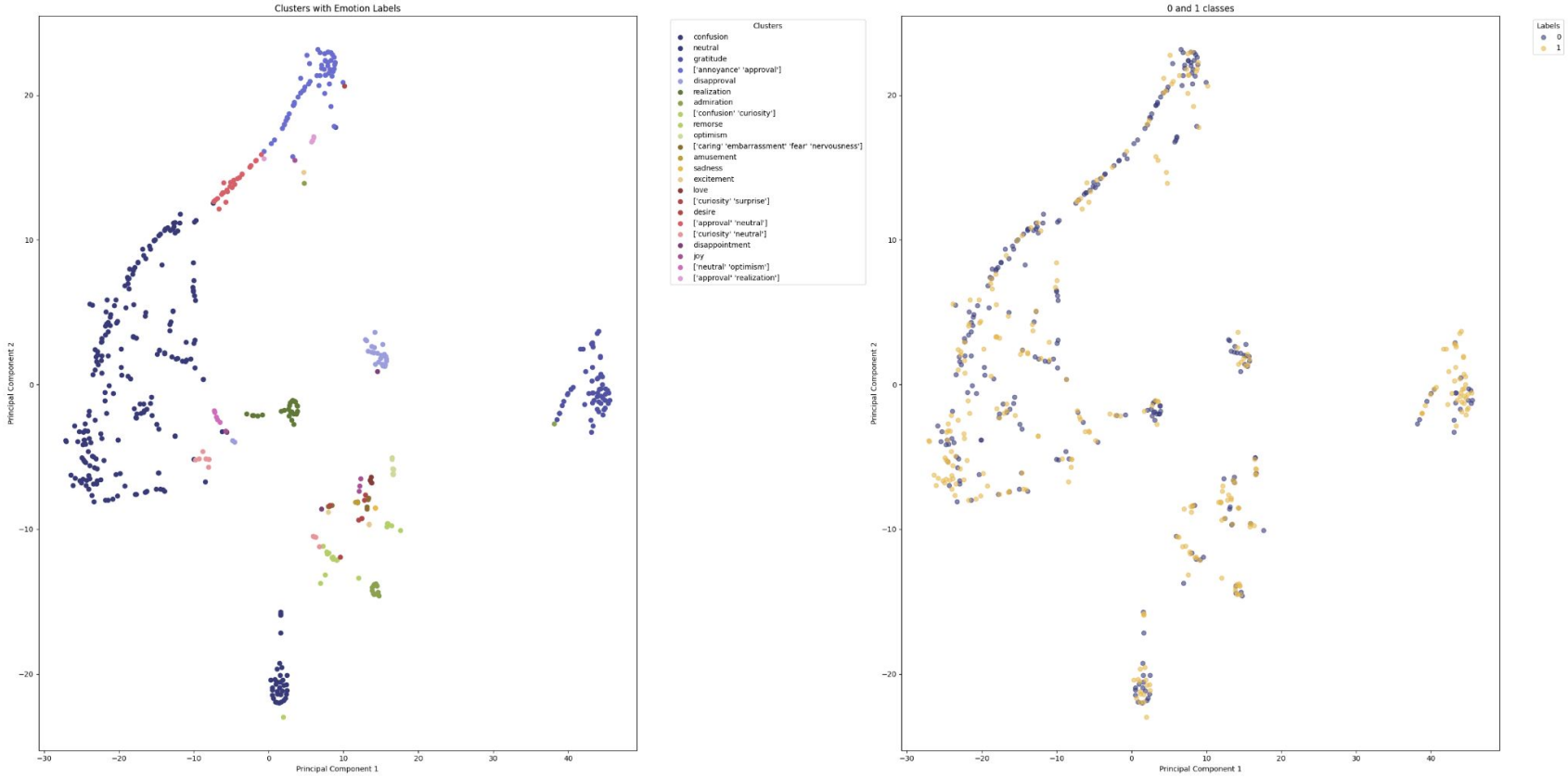
Модель	Данные и размер	CV F1
Логистическая регрессия	Эмбеддинги (576, 32768)	0.58
Логистическая регрессия	Униграммы (576, 2871)	0.61

Анализ эмоций

- Предобученная модель EmoRoBERTa для определения эмоций (22 типа эмоций).
- Для разделения обратной связи по эмоциональным контекстам применена кластеризация.
- Лучшим методом стал Kmeans с применением метода силуэта. Значение индекса силуэта 0.74.

Emotion	Class 0	Class 1
Admiration	8	9
Amusement	2	3
Anger	0	1
Annoyance	1	2
Approval	40	28
Caring	0	2
Confusion	26	22
Curiosity	12	22
Desire	0	1
Disapproval	6	7
Disgust	0	1
Fear	0	1
Gratitude	23	44
Joy	0	3
Love	0	1
Nervousness	1	0
Neutral	149	118
Optimism	2	3
Realization	16	7
Remorse	1	3
Sadness	1	2
Surprise	1	1

Анализ эмоций



Анализ эмоций

- Использование кластеров с высокой мерой разнообразия (<0.69) и явным преобладанием/отсутствием класса 1 (<34%, >66%), как контекст для классификатора.
- Классификатор работает лучше в рамках эмоционального контекста обратной связи.

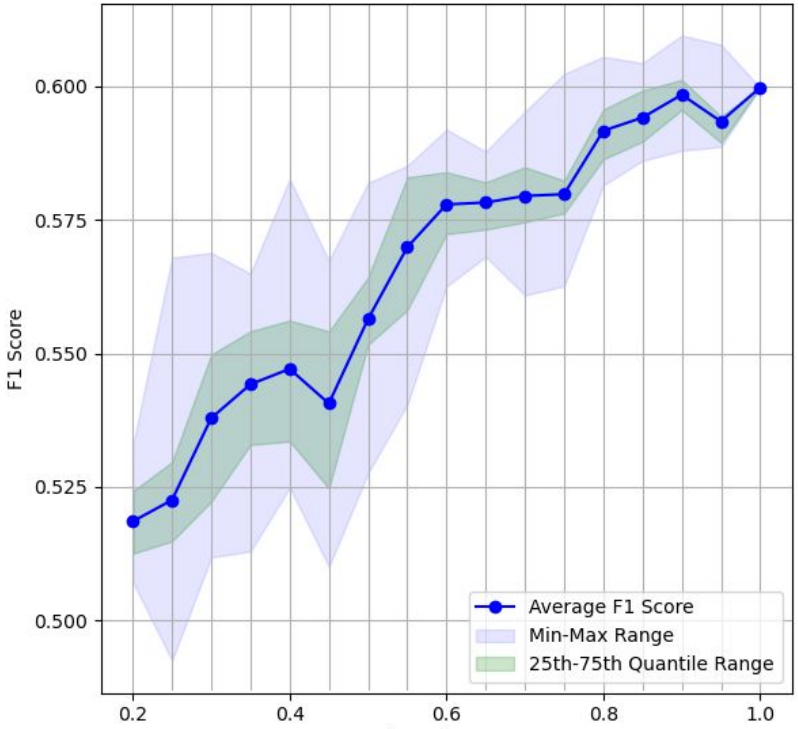
Эмоция	Размер кластера	F1-Score	Precision	Recall	Accuracy
Оптимизм (Optimism)	6	1.0	1.0	1.0	1.0
Одобрение (Approval)	53	0.67	0.76	0.59	0.75
Благодарность (Gratitude)	62	0.85	0.76	0.95	0.77
Неодобрение (Disapproval)	4	0.67	0.67	0.67	0.50
Восхищение (Admiration)	18	0.96	1.0	0.92	0.94
Замешательство (Confusion)	7	0.80	0.67	1.0	0.86

Классификация с дополнительным признаком

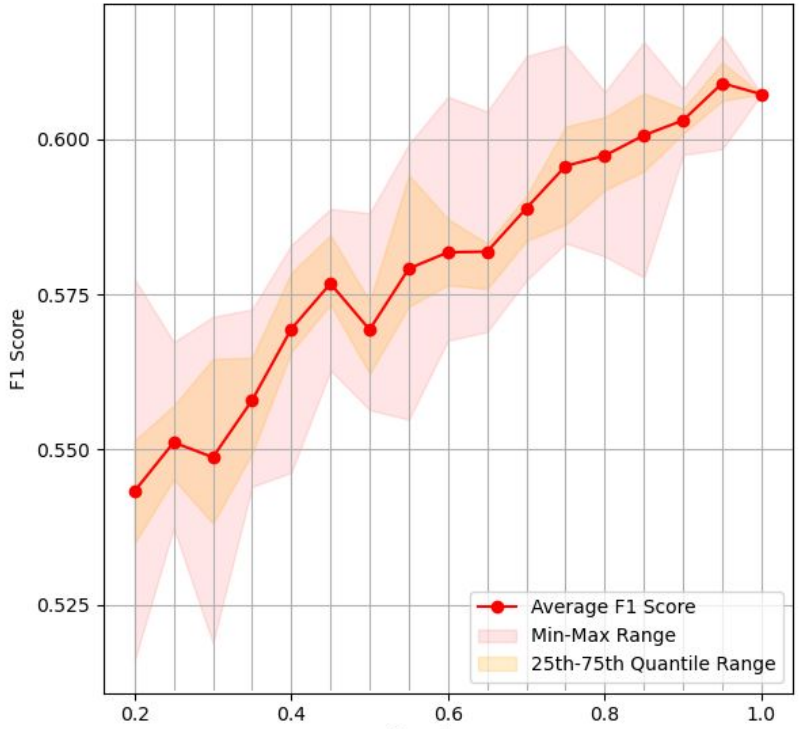
- Добавление вектора вероятностей (28 элементов) к вектору униграмм (2871 элемент).
- Метрика F1 модели:
  - Логистическая регрессия на униграммах с векторами вероятностей: F1=0.63
  - Логистическая регрессия на униграммах: F1=0.61
- Дополнительные признаки в качестве эмоционального контекста позволили улучшить способность модели различать классы.

Проверка устойчивости модели

Классификатор без доп. признаков



Классификатор с доп. признаков



Синтез обратной связи

Цель: изменять обратную связь разработчика другому разработчику, чтобы получилась обратная связь на ChatGPT решение.

- Применение модели pegasus_paraphrase и chatgpt_paraphraser_on_T5_base.
- Обучение с подкреплением, подсчет награды с обученным классификатором на перефразированном тексте.
- Проверка качества с помощью классификатора. Успех, если на перефразированном тексте предсказан новый класс.

Синтез обратной связи

```
original_proba = cls.predict_proba(original_vector)[0][0]
paraphrased_proba = cls.predict_proba(paraphrased_vector)[0][0]
cos_sim_reward = 1.001 - cosine_similarity(original_vector,
↪  paraphrased_vector)[0][0]
reward = (original_proba - paraphrased_proba) * cos_sim_reward
```

Done with this commit. Not sure about the new failing tests. They seem to only occur with methods midpoint or lower and python version lower.

Finished with this commit. Want to know the new failing tests for this commit. Methods midpoint or lower and python version lower seem to be the only ones that occur with them.

Результаты

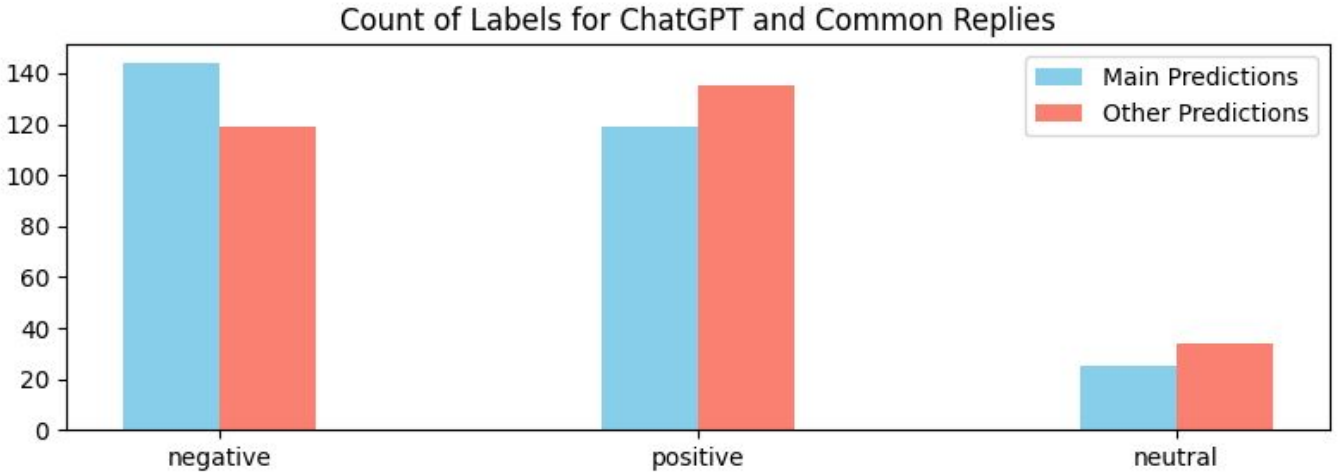
- Собран датасет двух типов обратной связи. Почти 300 ответов на решения, связанные с ChatGPT, и около 3000 ответов на решения без них. Для обучения использовалось всего около 600 записей.
- Разработана модель классификации, различающая обратную связь разработчика ChatGPT и разработчика разработчику в рамках общения на Github. Это указывает на отличие типов обратной связи.
- Работа модели классификации проверена в сложном и неоднородном контексте – эмоциональном, и улучшена с помощью дополнительных признаков.
- Модели синтеза обратной связи на базе моделей перефразирования текста после обучения с подкреплением повысили процент успешно синтезированных ответов до 16% и 20% случаев на тестовой выборке, показав прирост с 2% и 6%.

Список литературы

1. Liang и др. A large-scale survey on the usability of AI programming assistants: Successes and challenges, ICSE, 2024.
2. Kovalchuk и др. Human perceiving behavior modeling in evaluation of code generation model, GEM Workshop 2022.
3. Chen и др. Improving Code Generation by Training with Natural Language Feedback. 2024.

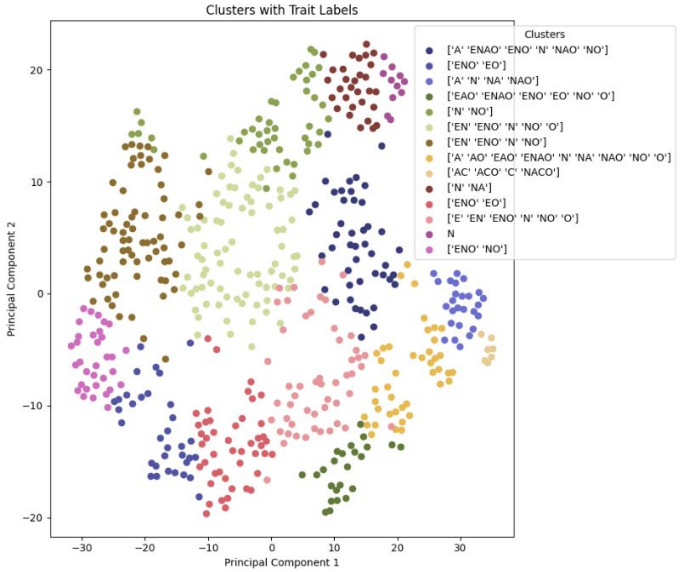
Анализ сентимента

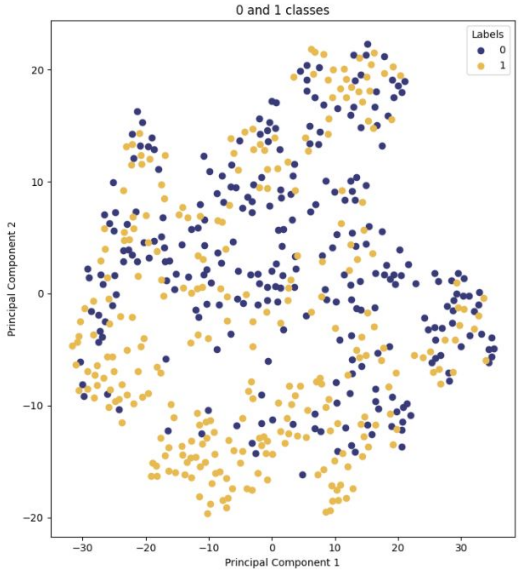
Использование предобученной модели lxyuan/distilbert-base-multilingual-cased-sentiments-student не дало значимых результатов. По результатам проверки распределения вероятностей каждого из классов для обеих групп ответов оказались разных типов.

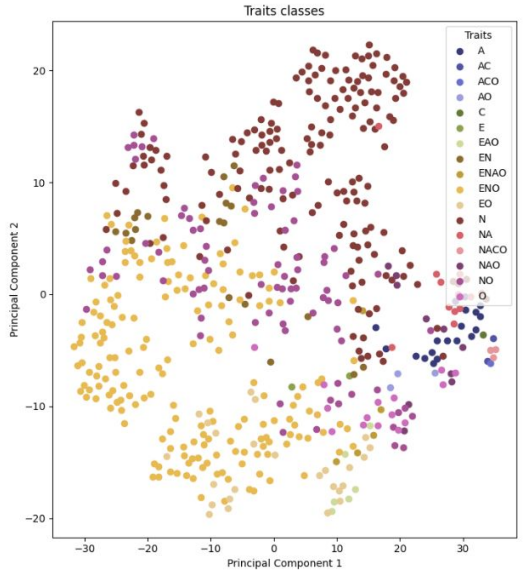


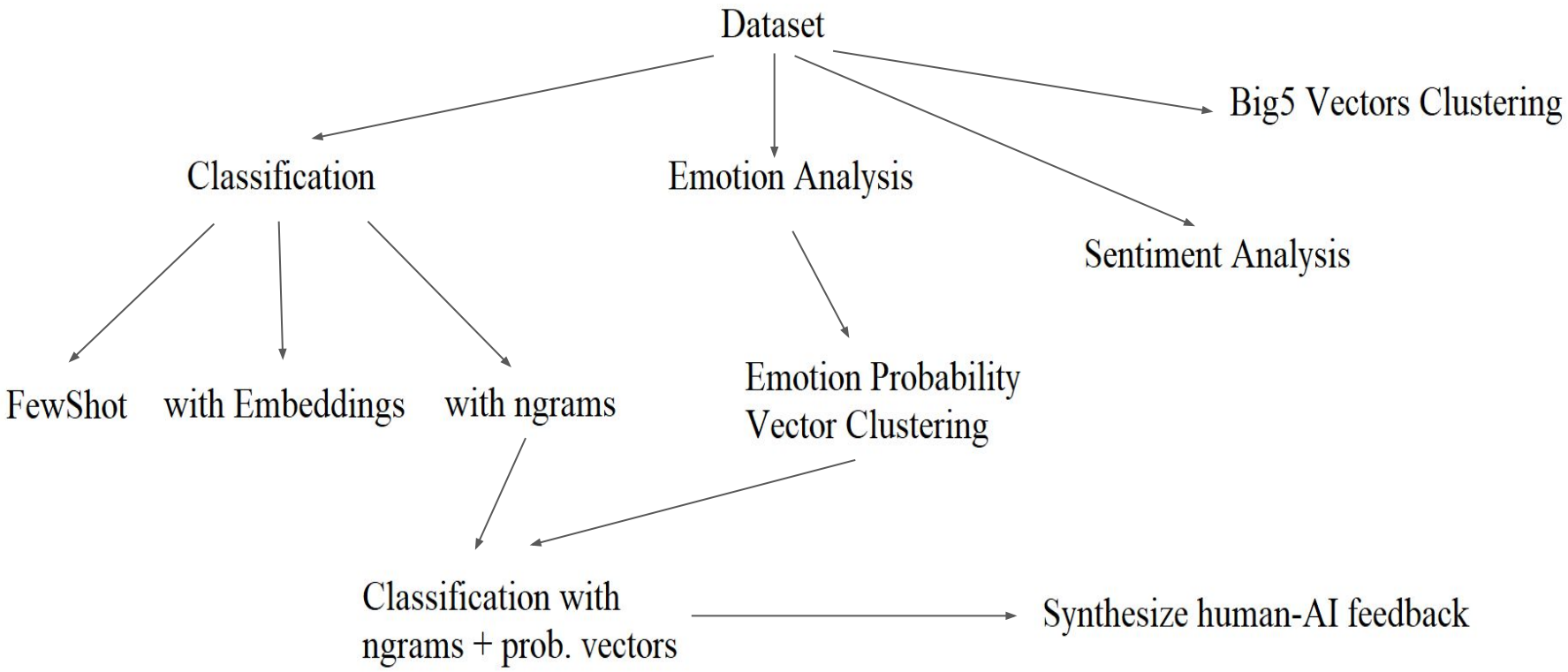
Кластеризация векторов Big5

Вектора Big Five -- экстраверсию, доброжелательность, добросовестность, нейротизм, открытость опыту. k-means, DBSCAN кластеризация. Четко разграничить не получилось








Dataset
Big5 Vectors Clustering
Classification
Emotion Analysis
Sentiment Analysis
FewShot
with Embeddings
with ngrams
Emotion Probability
Vector Clustering
Classification with
ngrams + prob. vectors
Synthesize human-AI feedback

| Column 1 | Column 2 | Column 3 |

|----------|----------|----------|

| Row 1 | Cell 2 | Cell 3 |

| Row 2 | Cell 5 | Cell 6 |

| Row 3 | Cell 8 | Cell 9 |