Biv Hack Challenge

Автоматическая детекция цели платежа

DataDisasters

15-17 ноября 2024

Roadmap проекта

	1-й день	2-й день	3-й день
Natural Language Processing	**EDA** баланс классов, анализ зависимостей, графики частотности, анализ ошибок	**PREPROCESSOR** класс-предобработчик: устранение ошибок, добавление тэгов; очистка, токенизация и лемматизация текстов	
Machine Learning		**ПОДБОР МОДЕЛИ** подбор классификатора на размеченных данных **SSL** предсказание меток на неразмеченных данных (классификатор + SelfTraining)	**КЛАССИФИКАЦИЯ** fine-tuning предобученного классификатора на всех данных
Docker		**ФОРМИРОВАНИЕ DOCKER IMAGE**	

Техническая информация

КОД, ФРЕЙМВОРК ТЕХНОЛОГИИ, БИБЛИОТЕКИ

Предобработка текстов:

RegEx (re)

- приведение текстов к единому формату
- добавление тэгов

NLTK

- выявление специфичных стоп-слов
- очистка от стоп-слов

SpaCy

- лемматизация (ru_core_news_sm)

Модели:

Sklearn

- Метод опорных векторов (SVM) + TF-IDF векторы текстов + PCA
  accuracy 0.97
- SelfTrainingClassifier для разметки

Huggingface Transformers

- Дообучение модели bank-transactions-statements-classification
  accuracy 0.969

Описание продукта

Ценность

Позволяет **быстро** определить назначение платежа

Реализация

Python code + NLP models + ML approach = **Docker Container**

Проблема

Ограниченное количество данных для обучения, **человеческий фактор** в текстах

Продукт

Репозиторий GitHub, в котором содержится код предобработки и обученная на тренировочных данных модель, классифицирующая платежи в одну из 9 категорий + **Docker Container**, в котором реализован весь процесс от предобработки до сохранения меток классов

Функционал и фичи

Предобработка текстов:

собственный класс-препроцессор

- очистка данных на основе особенностей датасета (разные форматы, опечатки, шум в данных)
- выбор стоп-слов, специфичных для датасета
- сохранение label-specific частотных слов и маркеров

Модели:

использование технологий Semi-Supervised Learning

- решение проблемы малого объёма обучающей выборки (размеченных данных)

классификация предобученным трансформером

- использование модели классификации текстов транзакций

Демонстрация проекта

Ссылка на репозиторий GitHub с реализацией продукта

Ссылка на fine-tuned модель на HuggingFace (использована для предсказаний)

Конкурентные преимущества

01

Предобработка текстов реализована разработчиками с **лингвистическим образованием**, гарантируя **хорошее качество** очищенного материала

02

Предобработка дополнена функционалом устранения **ошибок в данных и шумов** для **максимальной очистки** датасета.

03

Используется **трансформер**, заточенный **под данную задачу** классификации, **учитываются особенности текстов**, что помогает достичь **лучших результатов**.

Выводы

1. Удалось решить задачу классификации при малом объёме размеченных данных.
2. Были тщательно изучены классы и выявлены особенности данных в зависимости от класса.
   a. наличие class-specific частотных слов (*лизинг* для LEASING, *услуга* для SERVICE, *займ* для LOAN и т. д.)
   b. взаимосвязь классов и наличия/отсутствия НДС
3. Модели классификации показали высокие результаты accuracy на тестовой выборке.