Weekly Wrap-up

Progress Highlights and Insights

Contents

Last Week Recap

Concept transformers and explaining Entrepreneurship through personality traits.

Need to focus on explaining personality traits as phase one

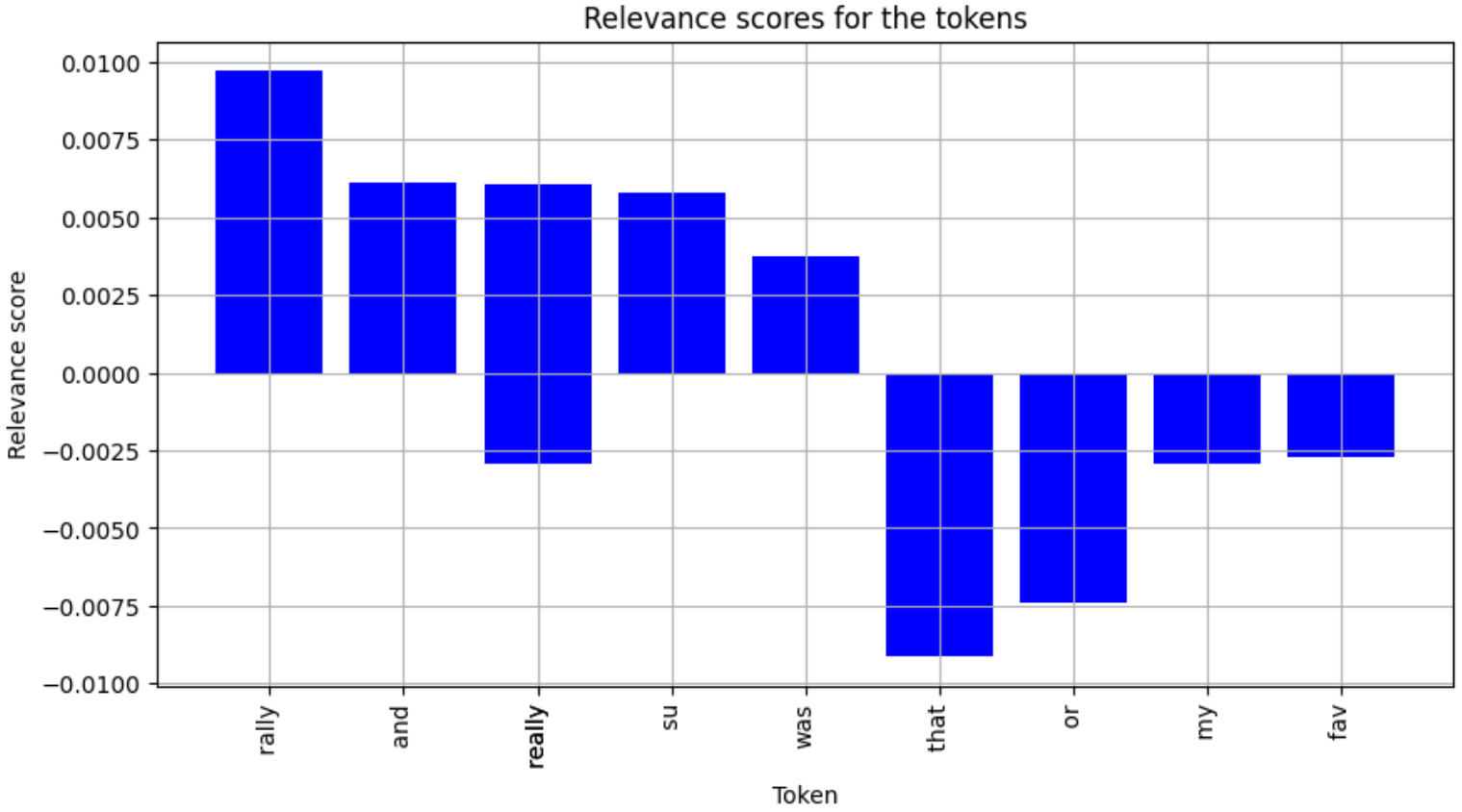
Explanations

- Attention Gradients
- SHAP (SHapley Additive exPlanations)

Attention Gradients

- Mentioned in the literature review AGrad
- Problems:
  - No reference (implemented according to the paper)
  - Not effective with multi-label classification

Attention Gradients

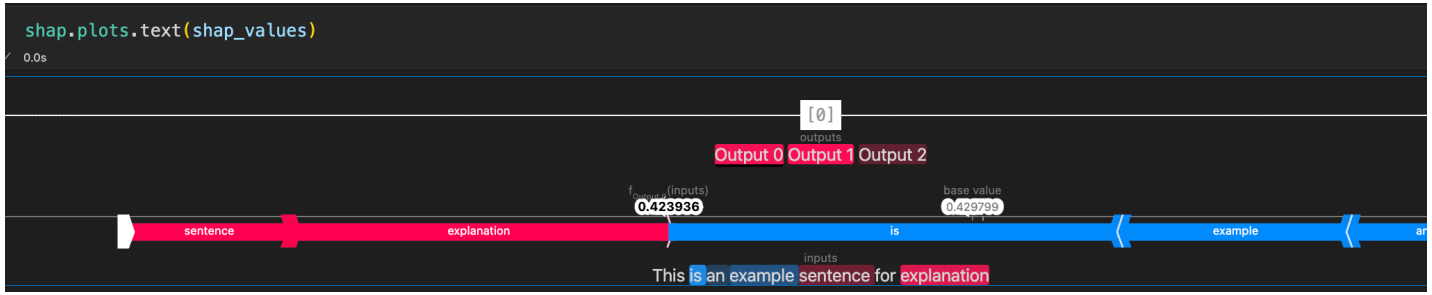


SHAP

- What’s SHAP?
  - Based on shap values from game theory
  - Weigh the importance of each feauture with relevance to the model
  - Make use of permutations and masking the features to observe changes in model’s output

Implementing SHAP explanations

- SHAP API
- Different explainers
- Tabular and textual data



Prediction

- Big five based models
- MBTI based models

Big five based models

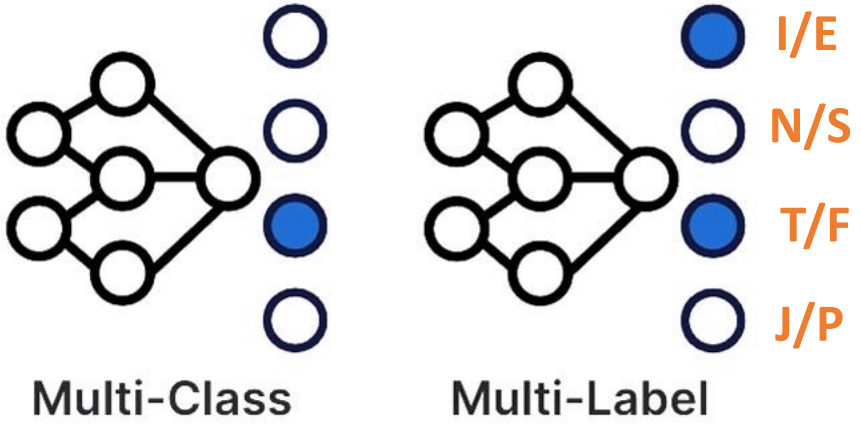
- Used **Stream-of-consciousness Essays** dataset (2.4k)
- Trained different models based on different pre-trained models
  - Bert
  - Roberta
  - Roberta + TF-IDF
  - Ensemble(Bert + Roberta)
- Different Approaches
  - Different epochs count 10, 20, 40
  - L2 regularizations
- Problem
  - Max accuracy 58%
  - Big 5 datasets availability

MBTI based Models

- Dataset:
  - **MBTI 500** (106k instance 500 words each)
- Model
  - Roberta Multi-label classifier
- Training
  - Trained for 3 epochs (6 hours on RTX A4000 16 GB)
- Results
  - 94% accuracy, 90% precision , 90% F1, and 90% recall

MBTI Model

- Further Testing:
  - **MBTI-1** dataset
    - 89% accuracy, 86% precision, 87% recall, 86% F1 score
- Architecture:



What’s next

- Apply SHAP on MBTI Model
- Explore SHAP different visualizations

SHAP with MBTI model

- Applying SHAP on Multi-label classifiers not the usual case
- Idea:
  - Manipulate model output for the 4 MBTI axis
  - Output [0.6, 0.8, 0.8, 0.7] to [0.6, 0.4] for I/E class

References

- A Unified Approach to Interpreting Model Predictions
- Interpretation of multi-label classification models using shapley values

Thank You