Weekly Wrap-up

Progress Highlights and Insights

Contents

- Last Week Recap
- Explainability
- Better Explanations
- Trait Extraction
- What’s next

Last Week Recap

Model trained on MBTI 500 dataset.

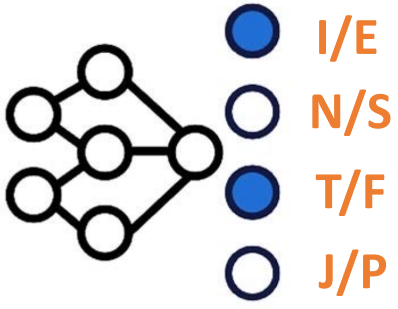
Applied SHAP explanations on the multi label classifier.

Explanations

- Applying LIME explanations

A quick recap

- SHAP depends on the changes of model output to infer relevance scores.
- LIME is no different than SHAP.
- Multi-class classifier output
  - [0.6, 0.35,0.7,0.9]
  - [1,0,0,1] decoded to INTJ



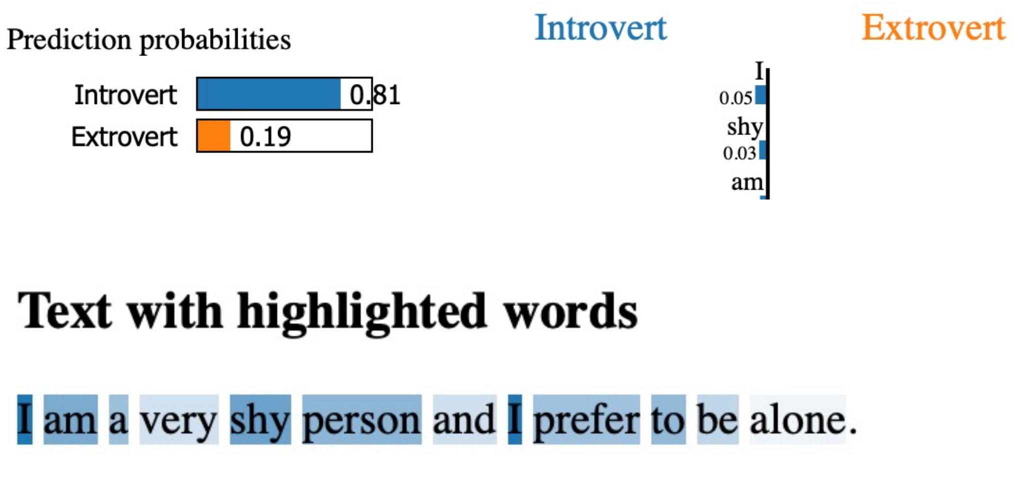
- IE for example:
  - Introvert: 0, Extrovert: 1
  - Focus on a single label 0.6 represents a 60% probability extroversion and 40% Introversion
  - So SHAP explainer receives [0.4,0.6] similar to a binary classifier output

Applying LIME explanations

- From a qualitative point of view LIME explantions were found to be worse than the SHAP explanations.
- LIME would ignore words with high relevance scores in SHAP and assign high relevance scores for others.
- However, the SHAP results seems more reasonable.

Comparing after and before keyword removal

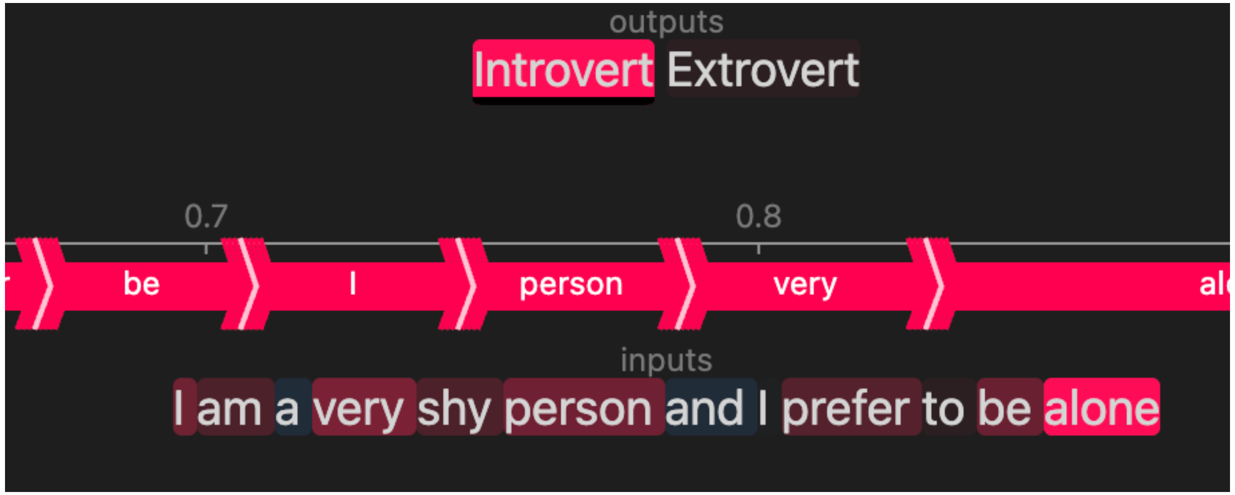
LIME



Text with highlighted words

I am a very shy person and I prefer to be alone.

SHAP



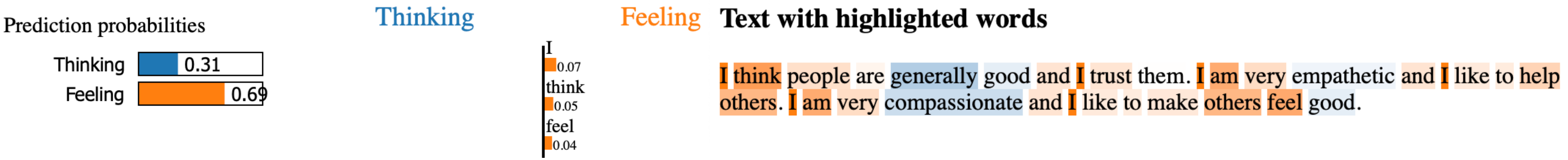
I am a very shy person and I prefer to be alone

LIME VS SHAP

- Notice in the next example how lime assign high relevance for words such as emapthetic and compassionate to the thinking class instead of the feeling class.
- The SHAP explantion seems more plausible, although both agree on some words such as feel.

Another Example on Thinking and Feeling axis

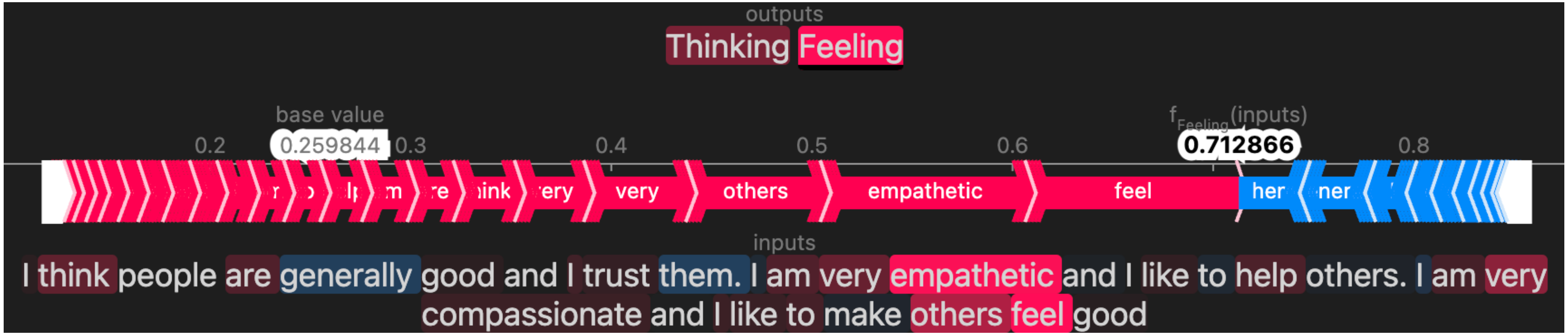
LIME



Text with highlighted words

I think people are generally good and I trust them. I am very empathetic and I like to help others. I am very compassionate and I like to make others feel good.

SHAP



I think people are generally good and I trust them. I am very empathetic and I like to help others. I am very compassionate and I like to make others feel good

Towards better explanations

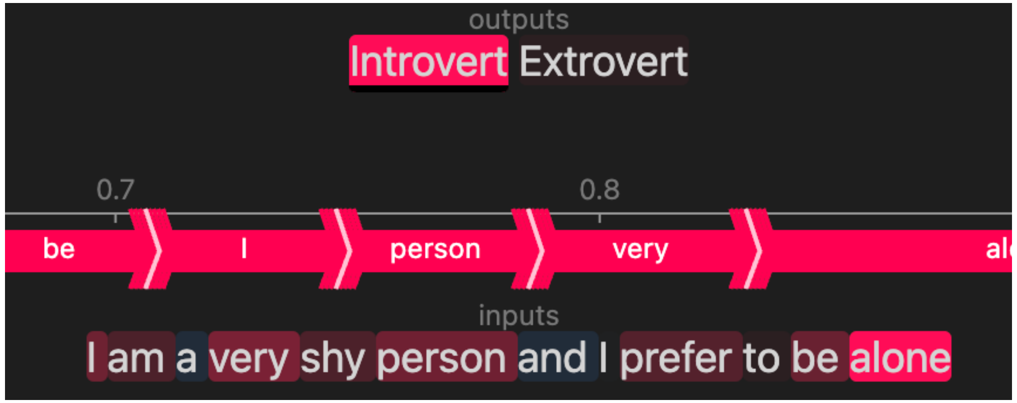
- Apply same cleaning techniques before inference and explanations.

Apply cleaning before inference and explanation

- Clean the input before inference, includes lemmatization and removing
  - Non alphanumeric characters
  - Stop words
  - URLs, hashtags, and @mentions.
- Increase in the explanations quality.
- SHAP and LIME explantions became somehow similar

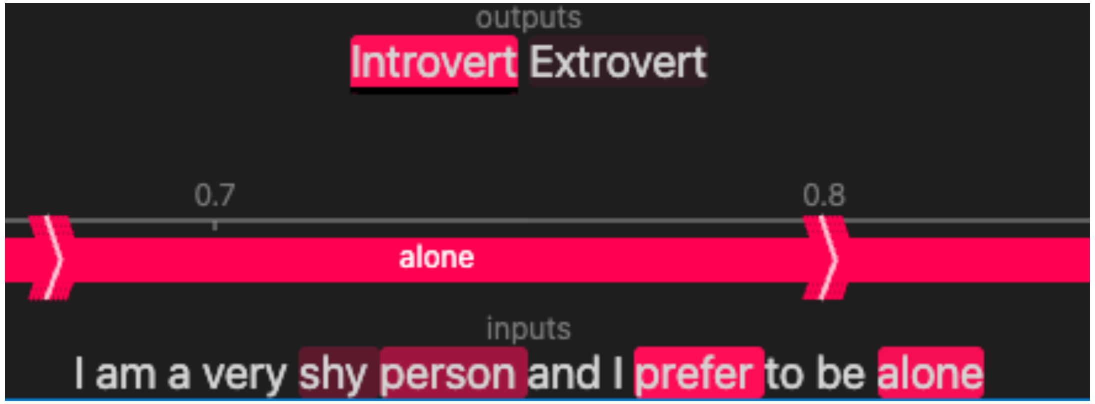
Comparing after and before SHAP explanations

BEFORE



I am a very shy person and I prefer to be alone

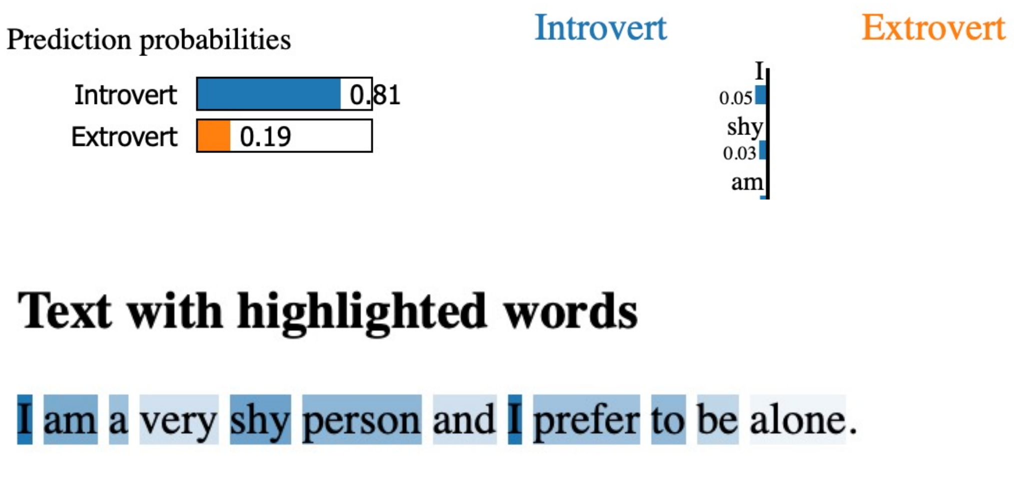
AFTER



I am a very shy person and I prefer to be alone

Comparing after and before LIME explanations

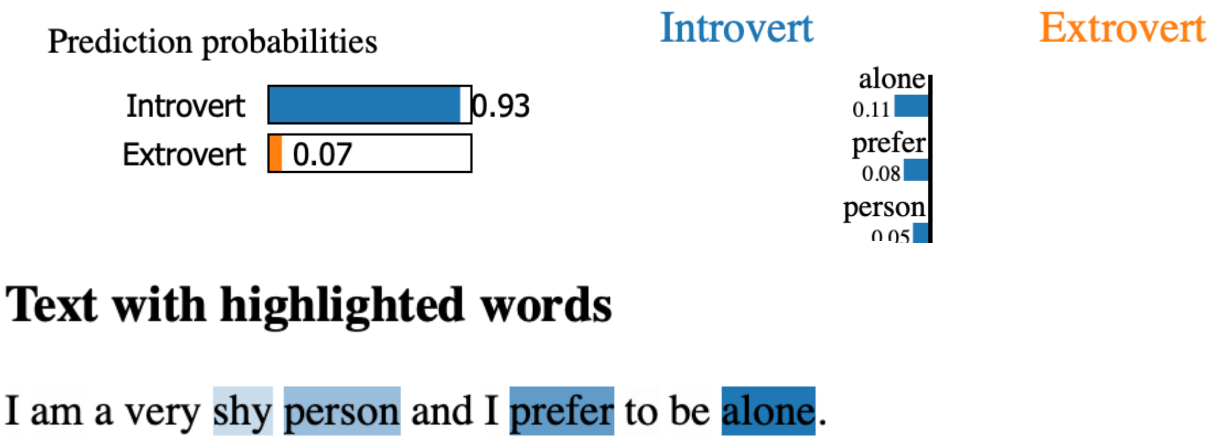
BEFORE



Text with highlighted words

I am a very shy person and I prefer to be alone.

AFTER

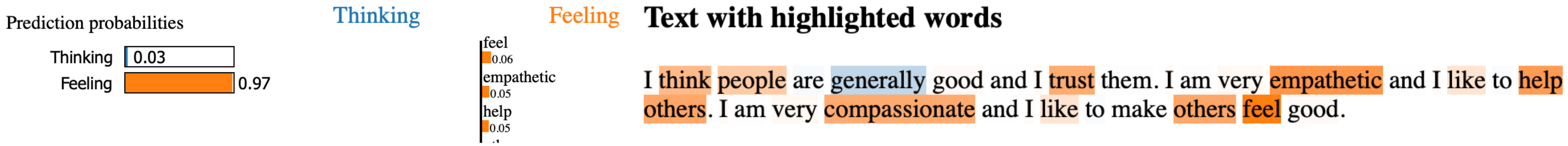


Text with highlighted words

I am a very shy person and I prefer to be alone.

Compare the after for SHAP and LIME

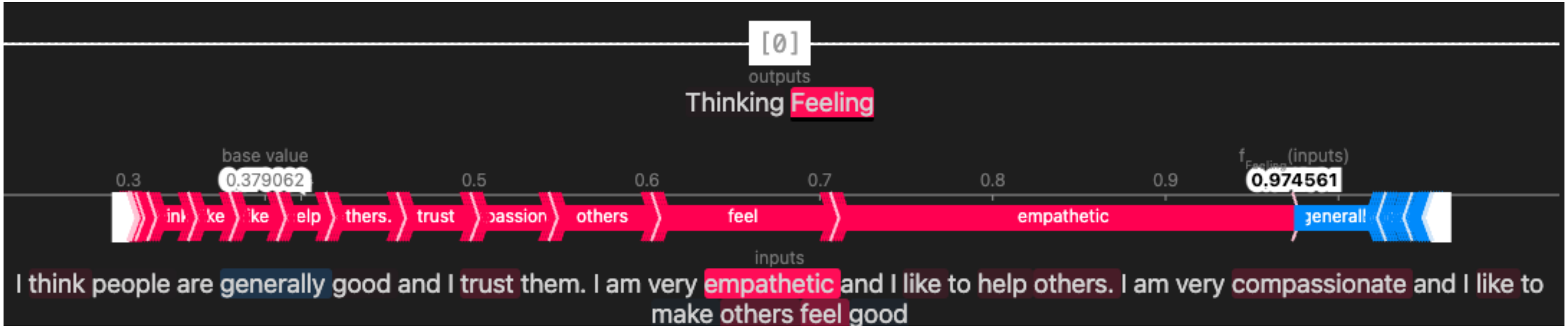
LIME



Text with highlighted words

I think people are generally good and I trust them. I am very empathetic and I like to help others. I am very compassionate and I like to make others feel good.

SHAP



I think people are generally good and I trust them. I am very empathetic and I like to help others. I am very compassionate and I like to make others feel good

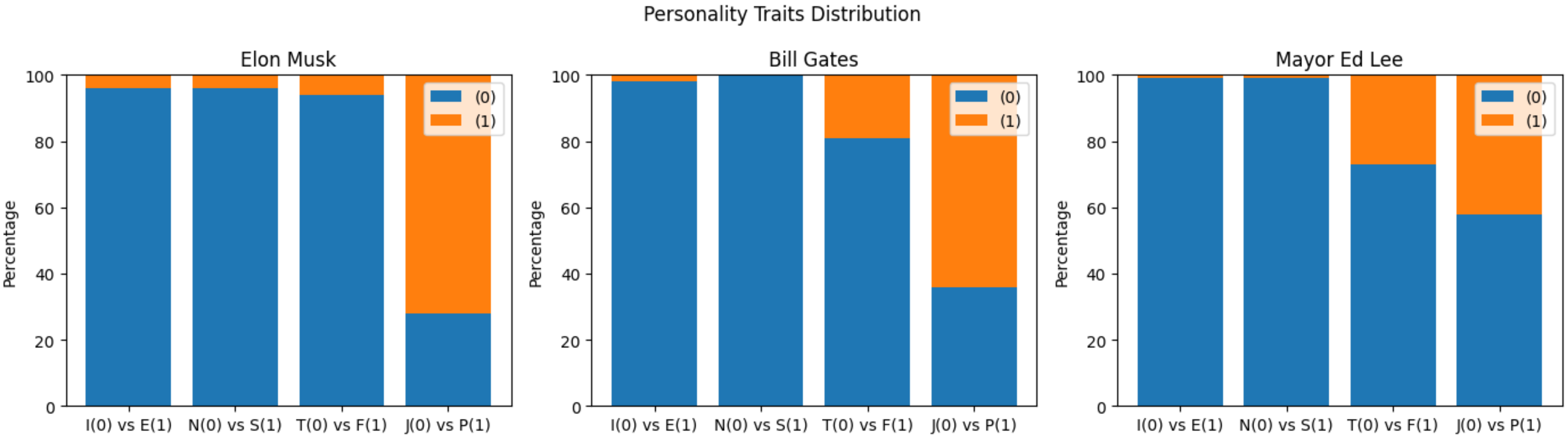
Trait Extraction

- Extracting traits for 2 famous entrepreneurs

Trait Extraction

- Dataset
  - ～2k tweets per user
  - 3 users (Elon Musk, Bill gates, Mayor Ed Lee)
  - Cleaned from tweets with no content other than URLs, hashtags and mentions.
- Results, a percentage for each class through the 4 MBTI axis
  - Elon Musk : INTP
  - Bill Gates : INTP
  - Ed Lee : INTJ

Traits across the 3 users



What’s next

- Write the implementation chapters
- Add SHAP to the literature chapter

References

- A Unified Approach to Interpreting Model Predictions
- Interpretation of multi-label classification models using shapley values
- https://flask.palletsprojects.com/en/3.0.x/quickstart/
- https://shap.readthedocs.io/en/latest/

Thank You