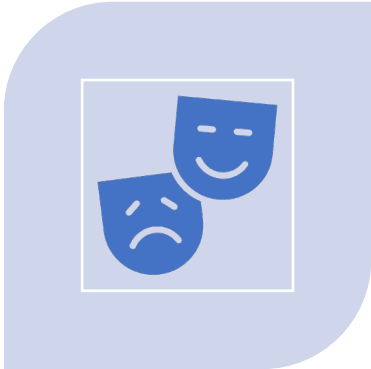
Weekly Wrap-up

Progress Highlights and Insights

Contents

Last Week Recap

Model trained on MBTI 500 dataset and SHAP explanations but not for multi label classification.



Problem with dataset due to presence of personality types in the text.

Prediction

- Cleaning the dataset and retraining the model

Cleaning the dataset and training

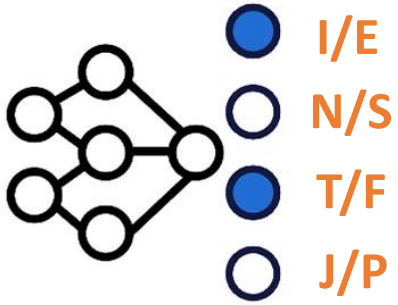
- Deleted the keywords from the dataset.
- Model accuracy went from 94% to 87%.
- Report presents further information on the deleted words and their contribution to the dataset.

Explanations

- SHAP (SHapley Additive exPlanations) for multi-label classifiers

SHAP Multi-Label classifiers

- SHAP depends on the chnages of model output to infer relevance scores.
- 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

Comparison

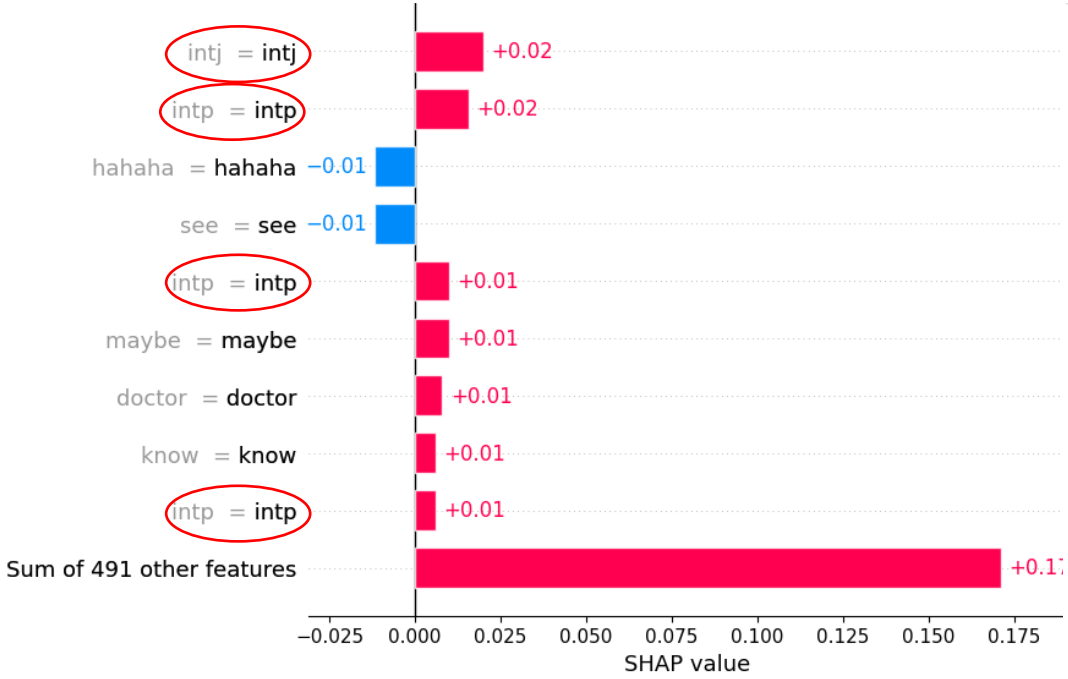
- Using SHAP values to compare the two models

Comparing after and before keyword removal

- A sample of 500 post from the training data.
- Resulting explanations showed high relevance scores assigned to keywords.
- Cleaning pushed to more plausible explanations.

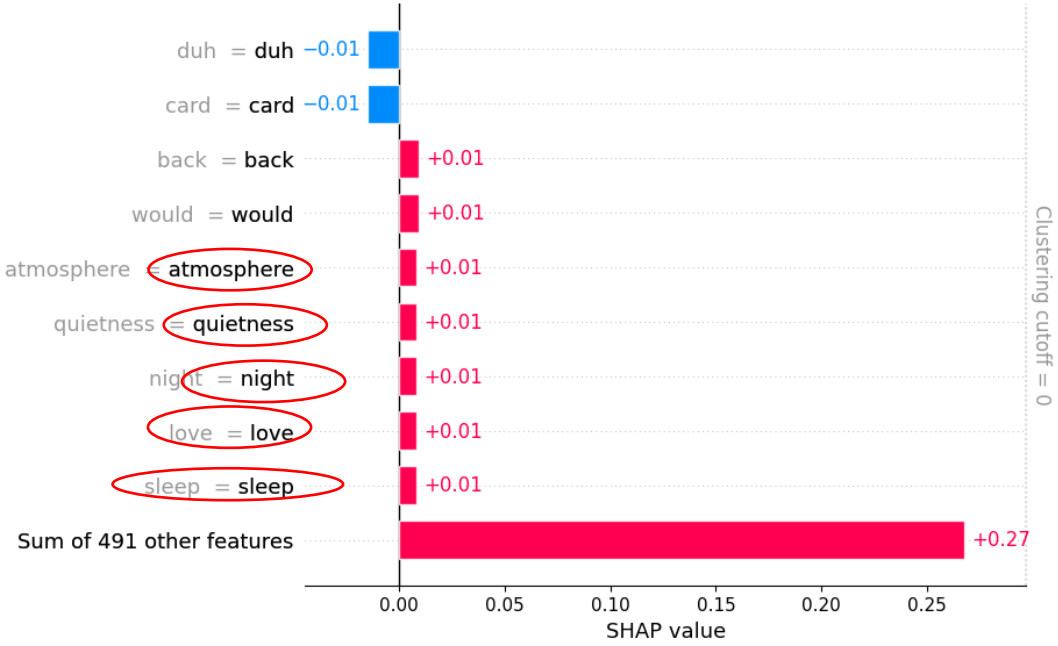
Comparing after and before keyword removal

BEFORE



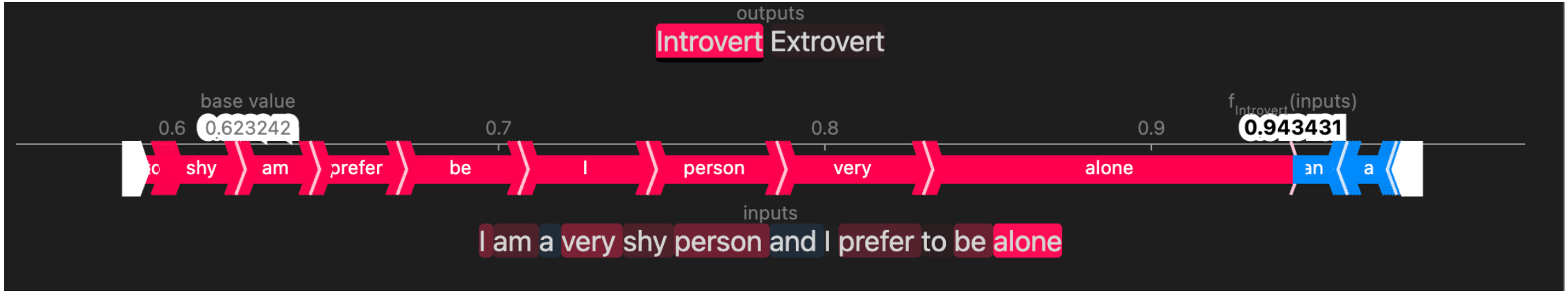
SHAP values IE axis, OUT: Introvert

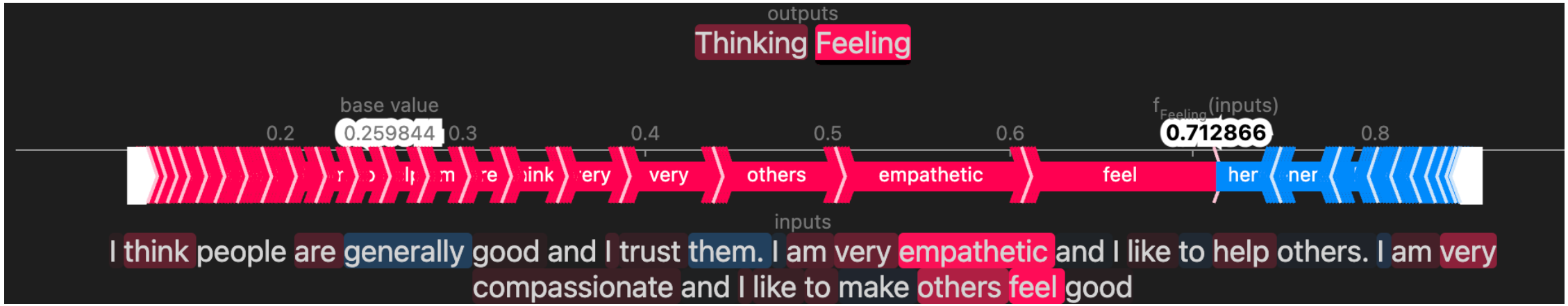
AFTER



SHAP values IE axis, OUT: Introvert

Examples on SHAP output with Text Plot



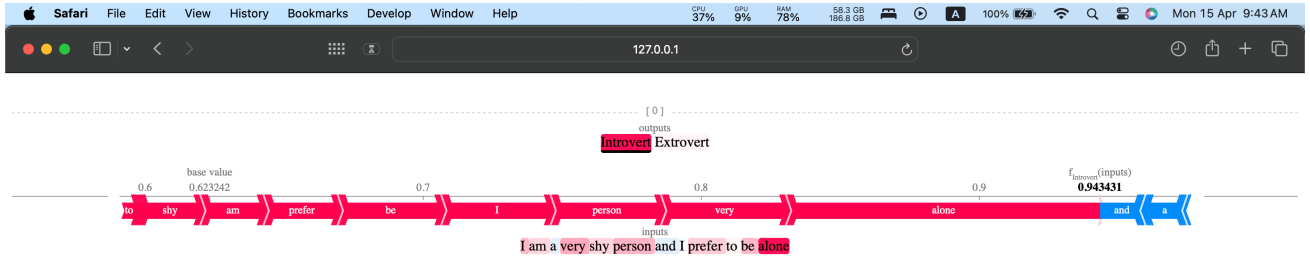


Framing the pipeline

- Deploying the model
- SHAP Outside Python Environment

Towards Model Deployment

- Creating endpoints for prediction and explanations using flask (still on local machine).
- Rendering SHAP plots outside python environment (In the browser).



What’s next

- Creating web interface to present prediction and explanation.
- Containerize the model and deploy the docker image.

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