MoI Classification: Three Approaches

Presentation and code by Aussie Frost
ausdfrost@gmail.com | github.com/ausdfrost

Overview

Objective

Using a labeled subset of CAHOOTS case narratives, devise pipelines of combinations of TF-IDF and DistilBERT preprocessing and Random Forest and SVM classification methods to determine the best combination of methods to classify the entire CAHOOTS case narrative dataset.

We wish to classify Modes of Intervention (MoI) that are taken for each case that exists in the CAHOOTS case narrative dataset.

Three approaches to MoI classification

- Random Forest with TF-IDF preprocessing
- Random Forest with DistilBERT preprocessing
- Linear SVM with TF-IDF preprocessing

Methods

1. Build pipelines for respective classification approaches
    a. Preprocess data
    b. Classify data
    c. Output meaningful results, metrics, visualizations
2. Iterate and fine-tune models*
3. Run each pipeline 100 times and average results*
4. Understand best approaches*

* not completed

Preprocessing

TF-IDF Text Vectorization

- Text is vectorized based on most and least frequently occurring words
- Results are easily interpreted as the original text is preserved when corresponding vectors are created
- Low computational effort required

Source: https://www.deepset.ai/blog/what-is-text-vectorization-in-nlp

DistilBERT Embedding

- Text is transformed by a pre trained model into high-dimensional embeddings of vectors
- Results aren’t as easily interpretable as the original text is not easily traced back
- Much higher computational effort required
- Performance should improve when using a custom model trained on your data
- Flexible hyperparameters allow for advanced customization

Source: https://huggingface.co/docs/transformers/en/model_doc/distilbert

Classification

Random Forest (SKLearn.ensemble.RandomForestClassifier)

- Random Forest classification builds an ensemble of decision trees that can visually show branches of features that might lead to one outcome or another (or in our case, MoIs)
- Each tree is trained on a subset of features, and then the final classification is made based on the aggregate of predictions across all 100 trees in my implementation
- When coupled with TF-IDF Text Vectorization, I found that Random Forest was able to use the term-frequency vectors created from the text to understand patterns in the data, leading to fast, scalable, and interpretable predictions

Classification

Linear SVM (SKLearn.svm.LinearSVC)

- Linear SVM attempts to find the best hyperplane to separate the MoI classes that are present in the case narrative data
- The nature of using SKLearn’s LinearSVC model allows for implementing hyperparameter tuning
- Linear SVM does not produce interpretable results in the same way that Random Forest does, results are simply metrics and the predicted classifications for the test dataset

Output

All results can be found in the output directory of the respective pipeline.

Results Log

- A log file is generated to understand the number of features present in the data
- Accuracy metrics are output here (and more metrics can be added)

Confusion Matrix

- A confusion matrix is generated from the TP/TN/FN/FP of each classification

Tree Graphs

- A Random Forest tree is generated to get an idea of what features are driving predictions

Replication

Accessing this code

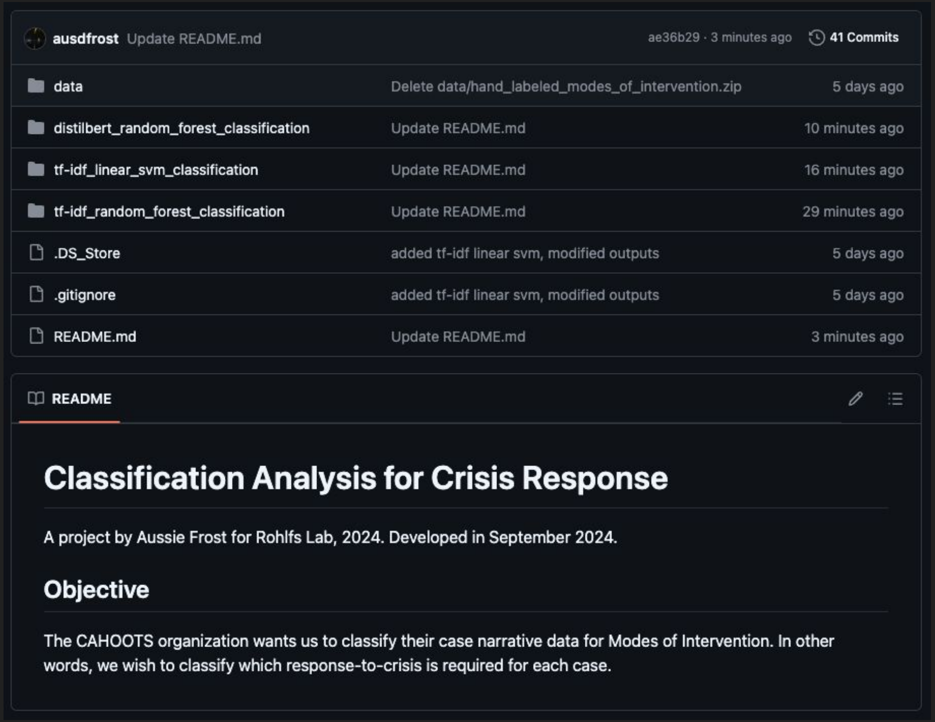
The code for this exploration on MoI classification is available on GitHub.

Please feel free to clone the repository to further develop these analyses.

If you would like to make direct contributions to the repository, please contact me or make a pull request!

Important notes

- Data repository only contains a fake dataset for testing
- For my set of hand labeled CAHOOTS data, please contact Rori
- The three non-data directories are the three classifiers used in this analysis



ausdfrost Update README.md ae36b29 · 3 minutes ago 41 Commits

data	Delete data/hand_labeled_modes_of_intervention.zip	5 days ago
distilbert_random_forest_classification	Update README.md	10 minutes ago
tf-idf_linear_svm_classification	Update README.md	16 minutes ago
tf-idf_random_forest_classification	Update README.md	29 minutes ago
.DS_Store	added tf-idf linear svm, modified outputs	5 days ago
.gitignore	added tf-idf linear svm, modified outputs	5 days ago
README.md	Update README.md	3 minutes ago

README

Classification Analysis for Crisis Response

A project by Aussie Frost for Rohlfs Lab, 2024. Developed in September 2024.

Objective

The CAHOOTS organization wants us to classify their case narrative data for Modes of Intervention. In other words, we wish to classify which response-to-crisis is required for each case.

Running these pipelines

- To run each respective pipeline, first access it’s main directory
- Ensure to specify the data path you would like to use
- Run the ‘case_narrative_classification.py’ file
- Observe model output in its respective subdirectory

In summation

Conclusion

During this exploratory analysis process, I found that classifying Modes of Intervention for CAHOOTS case narrative data was indeed feasible. Even with a small dataset, we can begin to build a scalable pipeline based on the models tried (and new ones). These pipelines all have their benefits and drawbacks, and I am excited to see where the study goes.

Next steps

- Each pipeline should be run many times and averaged to gain clear benchmarks for comparison
- Visualizations of model metrics in comparison to one another could be created
- Hyperparameter tuning could be implemented in various areas
- Larger case narrative datasets could be leveraged to train a custom DistilBERT model, as well as to gain more accurate performance metrics

Thanks DS4SJ team!

- Aussie Frost