Report on Symptom Classification & Symptom Severity Dataset Analysis

By Suhaab Shah

Introduction:

This report analyzes two datasets: Symptom Severity and Symptom Classification, each segmented into training, testing, and validation sets. It covers data visualization, text preprocessing, model training, and evaluation using various machine learning techniques. Findings from three detailed analyses of the symptom classification dataset are summarized, incorporating advanced data manipulation, exploratory data analysis, and diverse machine learning methods.

Dataset 1

Data Preprocessing

The preprocessing across the three reports included several consistent steps aimed at cleaning and standardizing the text data:

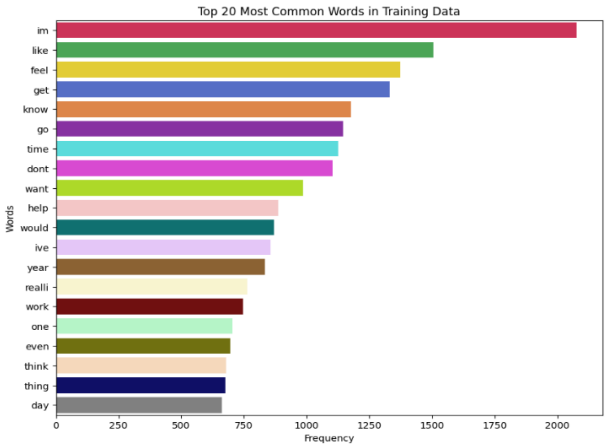
- **Text Normalization:** Conversion of all text to lowercase.
- **Whitespace and URL Removal:** Elimination of extra spaces and removal of web links.
- **Punctuation and Special Characters Removal:** Non-alphanumeric characters were stripped.
- **Stopwords Removal:** Filtering out commonly used words that might not be significant.
- **Tokenization:** Breaking down text into individual terms.
- **Lemmatization and Stemming:** Reducing words to their root forms to standardize variations.
- **Advanced Cleaning:** Included removal of URLs and special sequences.
- **Vectorization:** Employed TF-IDF to convert cleaned text into numerical data suitable for modeling.

The cleaned data was then saved in new Excel files to be used for subsequent analysis and modeling.

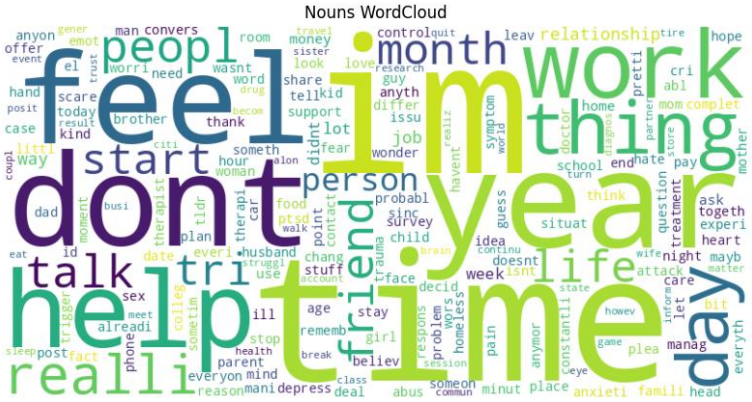
Exploratory Data Analysis (EDA)

The exploratory analysis provided a multi-faceted view of the data:

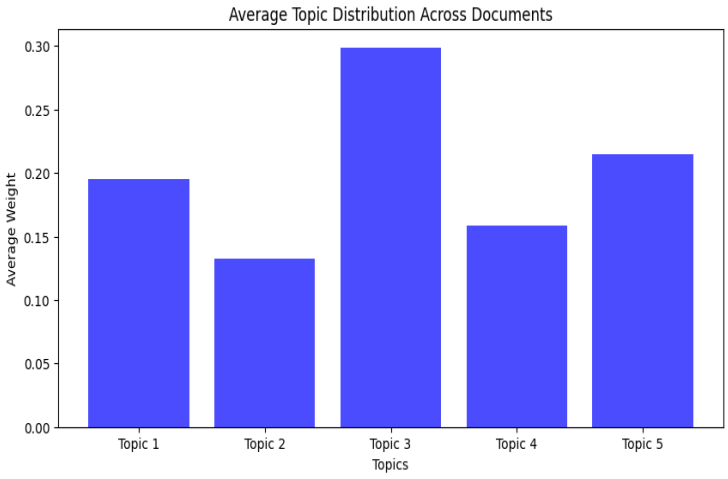
- **Word Clouds and Common Words Visualization:** These visualizations illustrated the most frequent and significant terms across the datasets, revealing dominant themes.

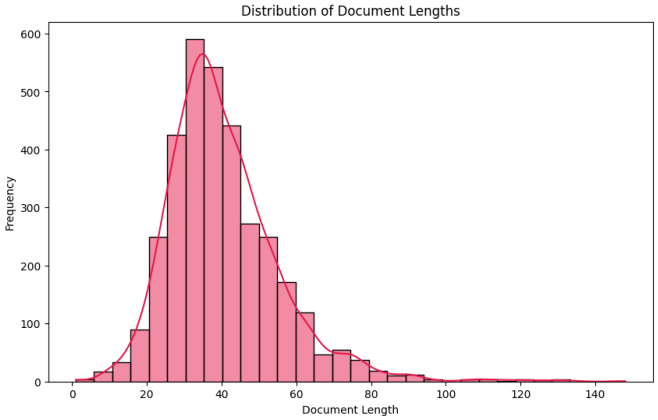


- **N-gram and POS Tagging Analysis:** This analysis showed common phrase structures and grammatical usage within the text, helping to understand the contextual usage of terms.



- **Document Length Distribution and Histogram:** The distribution of document lengths was plotted to understand the variability in the size of text entries, which can affect text analysis and model performance.



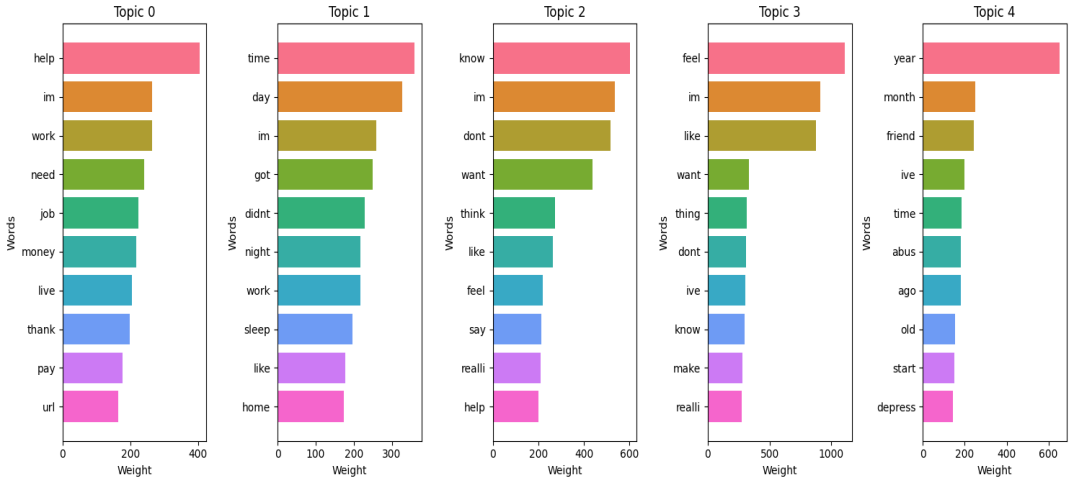


Topic Modelling

Two techniques were highlighted across the reports:

- **BERTopic:** This advanced machine learning technique provided graphical representations of topics, enhancing the understanding of their inter-relationships.

- **Latent Dirichlet Allocation (LDA):** Employed to discover underlying topics in the data, with visualization of each topic's composition to aid in interpreting the text data.



Model Training and Performance Evaluation

Multiple models were trained to classify the text data based on the processed information:

- **Support Vector Machine (SVM):** Utilized TfidfVectorizer for feature extraction combined with SVM classification. The performance was evaluated using metrics like macro-averaged recall and accuracy.
- **Logistic Regression:** This model also used TF-IDF vectors for training and was evaluated on accuracy and macro-averaged recall metrics.
- **Random Forest Classifier:** Employed on vectorized text data, with performance assessed using accuracy and recall metrics.

Performance Metrics:

Support Vector Machine (SVM):

- **Training Set Macro Average Recall**: 0.9545699761386035
- **Validation Set Macro Average Recall:** 0.8423747276688454

Random Forest

- **Macro-average Recall on Training Data:** 0.999, nearly perfect recall during training.
- **Macro-average Recall on Validation Data:** 0.891, showing high effectiveness on unseen data.

Logistic Regression

- **Training Macro-Averaged Recall:** 0.897, showing the model's capability in identifying all relevant instances in the training set.
- **Validation Macro-Averaged Recall:** 0.832, reflecting the model's effectiveness in identifying relevant instances during validation.

Predictions and Output

Models were used to predict labels for the test data, with the results documented and saved for potential operational use and further analysis.

Dataset 2

Data Preprocessing

Comprehensive text preprocessing steps were implemented across all datasets to enhance model performance:

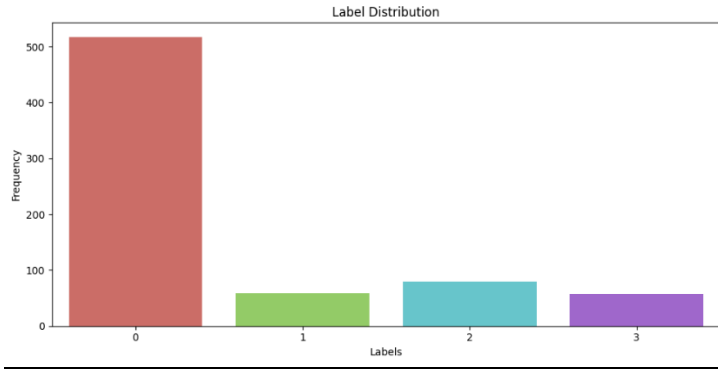
- **Conversion to Lowercase**: Uniformity across texts.
- **Removal of Non-Alphanumeric Characters**: Cleaned texts from irrelevant characters.
- **Tokenization**: Segmented texts into individual tokens.
- **Stopword Removal**: Eliminated common words to focus on valuable terms.

- **Stemming**: Reduced words to their root form for generalization.
- **Advanced Cleaning**: Included removal of URLs and special sequences.
- **Vectorization**: Employed TF-IDF to convert cleaned text into numerical data suitable for modeling.

Exploratory Data Analysis (EDA)

The EDA phase provided visual and statistical insights:

- **N-Gram Analysis**: Utilized unigrams, bigrams, and trigrams to identify frequent terms and phrases.
- **Parts of Speech Tagging**: Analyzed grammatical structures, aiding contextual understanding.
- **Label Distribution**: Visualization of label frequency within the training dataset using bar plots.

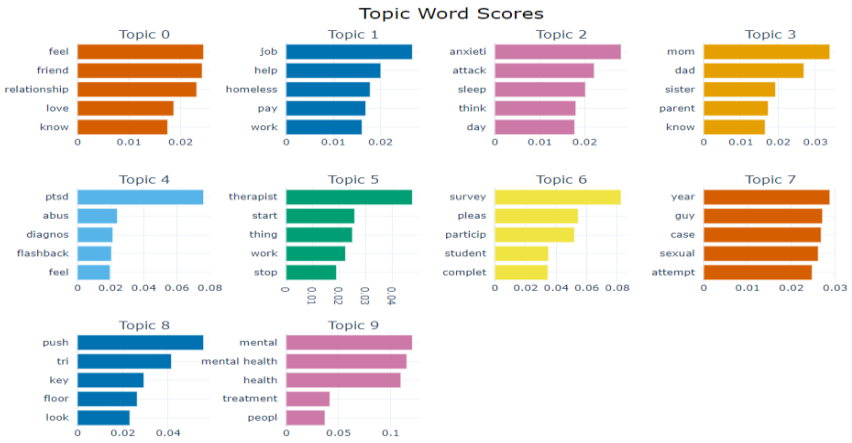


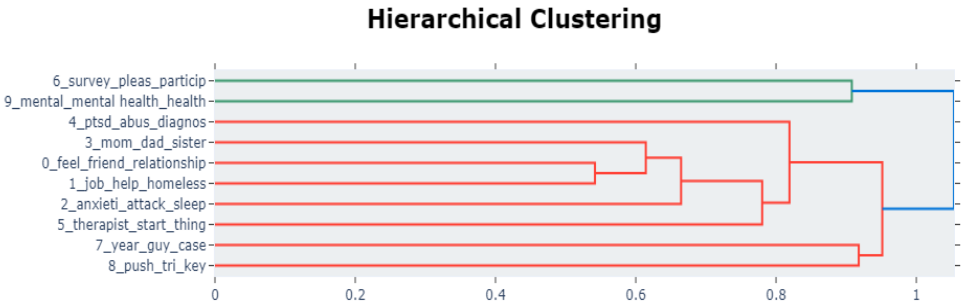
- **Word Clouds**: Generated to highlight common words in the training dataset, providing insights into textual data characteristics.

- **Text Length Distribution**: Histograms displayed text length variability, influencing further preprocessing decisions.

Topic Modelling

- Applied BERTopic to uncover underlying themes or patterns within the text data.





Model Training and Evaluation

Several models were trained and evaluated:

1. **Gradient Boosting and XGBoost Classifiers**: These models handled multi-class classifications and were assessed using accuracy and recall metrics.
2. **Logistic Regression:** Focused on vectorized text data, evaluated through detailed classification reports and accuracy assessments.

Performance Metrics

XGBoost Classifier:

- **Training Macro-Averaged Recall**: 100%
- **Validation Macro-Averaged Recall**: 27.28%

Logistic Regression:

- **Training Macro-Averaged Recall**: 40.53%
- **Validation Macro-Averaged Recall**: 29.73%

Gradient Boosting:

- **Training Macro Average Recall:** 98.9%
- **Validation Macro Average Recall:** 32%

Limitations and Suggestions for Future Work

1. **Lack of Actual Labels for Test Data:**

Without actual labels for the test dataset, comprehensive evaluation metrics like precision, recall, and F1-score couldn't be calculated, limiting the assessment of model performance.

Suggestion: Future studies should aim for a complete dataset with labeled test data to enable a more thorough evaluation and gauge model generalizability.

2. **Dependency on Text Preprocessing:**

Model performance heavily relies on preprocessing steps, risking the loss of context or introduction of noise if preprocessing is inadequate or overly aggressive.

3. **Imbalanced Data:**

The imbalanced distribution of classes in the datasets can bias models towards the majority class, leading to poorer generalization to less frequent classes.

Conclusion:

This report outlines a comprehensive workflow for analyzing textual data, covering preprocessing, modeling, and evaluation stages. Future work should focus on advanced model tuning, additional data integration, and continuous evaluation for improved real-world efficacy. Overall, the analysis underscores the importance of robust text analysis methods and sets the stage for further optimization and exploration in healthcare applications.

Code Links:

Dataset 1:

- https://colab.research.google.com/drive/1xInfnftFbJIEtNYHPSnpePJVxpP_QmhG?usp=sharing
- https://colab.research.google.com/drive/1MyJxkUxEahKNMGmXonW8JQ-imujGmP0_?usp=sharing
- https://colab.research.google.com/drive/1ZJ-Q06CsNLpdDwlvnMfliXR5BlgoTGqJ?usp=sharing

Dataset 2:

- https://colab.research.google.com/drive/1I5GBlTUrg2RkWq_5TcND8K3h2E2Cxy3Z?usp=sharing
- https://colab.research.google.com/drive/1tWJUU8-FPXUMXumC_QLex0P5rggb1DAU?usp=sharing