Identification of Quote Themes: A Multi-label Classification Problem

Chandreen Ravihari Liyanage - 1158931
Sara Nazar - 1223124

Outline

- Introduction
- Exploratory Data Analysis
- Model Development
  - Data Pre-processing
  - Feature Extraction and Engineering
  - Implementing Classifiers
- Results and Discussion
- Limitations
- Future Improvements
- Conclusion

Introduction

This is a multi-label classification problem for classifying english quotes into their themes/topics.

The original dataset comprised of 3000 english quotes from goodreads and has three data fields: Quote, tags, author.

Exploratory Data Analysis

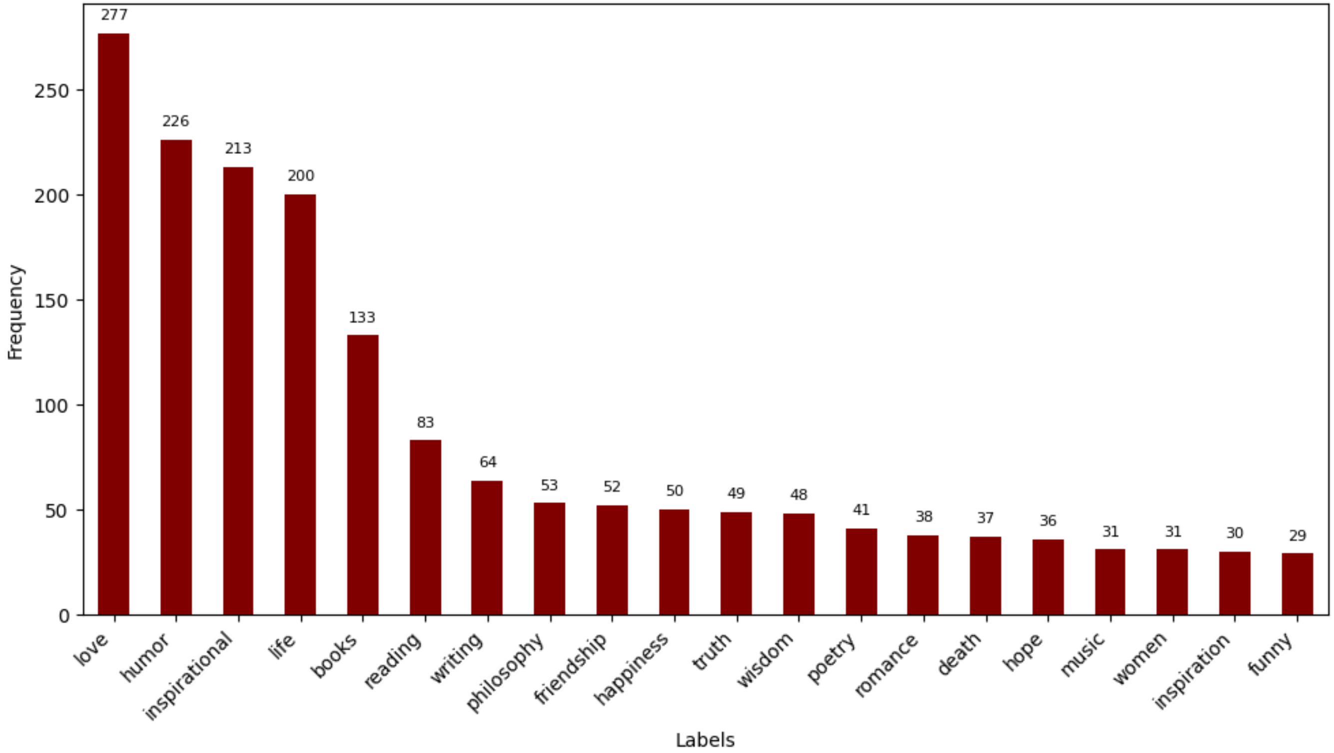
Basic statistics of the complete dataset

Number of records in the dataset	2231
Number of unique labels in the dataset	1592
Maximum number of labels per quote	36
Minimum number of labels per quote	1
Average number of labels per quote	2.36
Average length of quotes	14.34 words

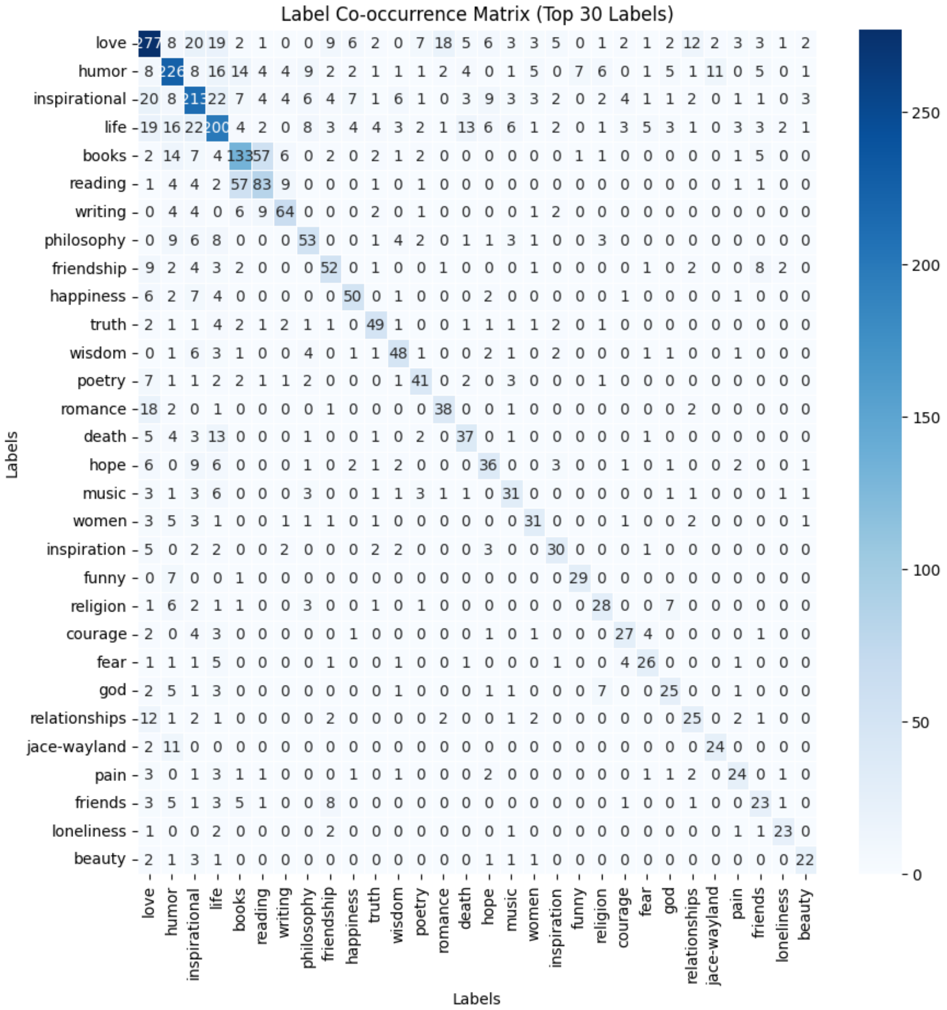
We found that the ratio between the number of records to the number of labels in the dataset is 22:16, which means that the dataset is sparse.

The word cloud for the complete dataset

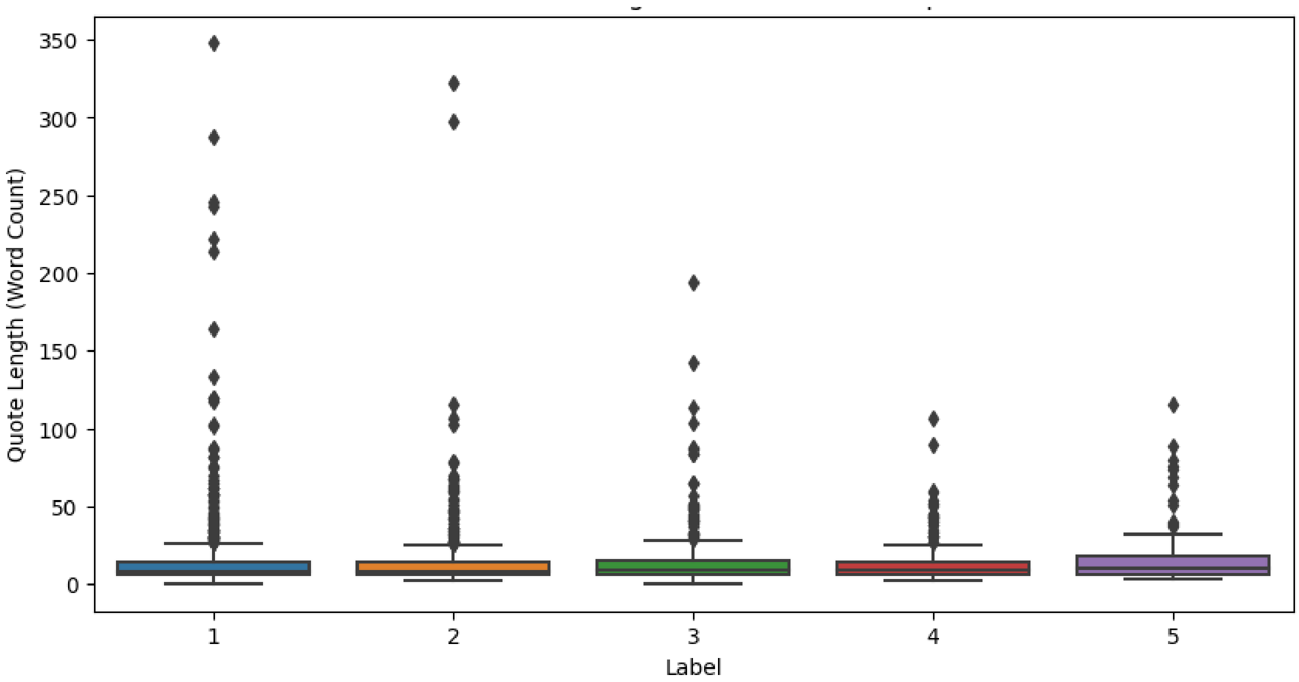
Frequency of the top 20 labels in the dataset.



The relationships between the most popular 30 labels.



Distribution of text length in the top 5 labels.



Model Development

Data Preprocessing

- Non-alphanumeric characters were removed, text was converted to lowercase, and extra spaces, punctuations, and stopwords were removed to make the text more readable.
- One-hot encoding was used to convert labels into columns.
- Redundant columns and rows were removed
- Quotes with a high number of labels were also deleted to reduce the size of the dataset.

Feature Extraction and Engineering

- TF-IDF (Term Frequency-Inverse Document Frequency), and text embedding methods, including Word2Vec, and SentenceTransformers were used to identify features
- Dimensional space was reduced using PCA.
- The feature extraction and engineering techniques were performed separately on the training and testing datasets.

Implementing classifiers

K-Nearest Neighbour (KNN), Random Forest (RF) and Multi-Layer Perceptron (MLP) were developed as classifiers using MultiOutputClassifier (MOC) and BinaryRelevance (BR) multi-label classification strategies.

24 result sets were expected as 4 feature types, including combined TF-IDF SentenceTransformers, 3 classifiers and 2 multi-label constructors were used, but due to RAM and GPU limitations 17 result sets were obtained.

Accuracy, Hamming loss, micro Precision, micro Recall, and micro F1-Score were generated to evaluate the performance of the model.

Results

- Ovearll, SentenceTransformers perform better than TF-IDF across all the matrices.
- When combined with TF-IDF, the classification performance of SentenceTransformer features decreases in terms of recall and F1-score, while it increases in terms of precision.
- RF demonstrates notable performance superiority over both KNN and MLP across all metrics.
- No significant difference observed in the results of the two multi-label classification constructor strategies.

Multi-label classification results

Model	Feature vector	Multi-label constructor	Classifier	Accuracy	Hamming loss	Precision	Recall	F1-score
M1	TF-IDF	MOC	KNN	0.0000	0.0026	0.5039	0.0759	0.1320
M2			RF	0.0000	0.0025	0.7915	0.0947	0.1692
M3			MLP	0.0000	0.0027	0.3622	0.0667	0.1126
M4		BR	KNN	0.0000	0.0026	0.5039	0.0759	0.1320
M5			RF	0.0000	0.0026	0.8286	0.0832	0.1513
M6			MLP	0.0000	0.0027	0.3856	0.0694	0.1176
M7	SentenceTransformers	MOC	KNN	0.0045	0.0026	0.5489	0.1660	0.2549
M8			RF	0.0000	0.0025	0.9795	0.1075	0.1938
M9			MLP	0.0030	0.0026	0.5492	0.1488	0.2341
M10		BR	KNN	0.0045	0.0026	0.5489	0.1660	0.2549
M11			RF	0.0030	0.0025	0.9841	0.1048	0.1895
M12			MLP	0.0030	0.0025	0.5646	0.1530	0.2408
M13	Combined	MOC	KNN	0.0000	0.0026	0.5368	0.0888	0.1524
M14			RF	0.0000	0.0025	0.9941	0.0952	0.1738
M15			MLP	0.0045	0.0026	0.5547	0.0917	0.1574
M16		BR	KNN	0.0000	0.0026	0.5368	0.0888	0.1524
M17			RF	0.0000	0.0025	0.9947	0.1044	0.1889
M18			MLP					

Limitations

Overall, multi-label classification models perform poorly due to higher label space.

Higher feature space posed challenges related to computational efficiency and increased resource requirements during training and inference.

Future Scope

Using label space feature selection may improve performance as it would reduce dimensionality.

Sentiment analysis tools like VADER and EmoTFDF may also improve performance

Experiments on class distribution could also improve performance.

Conclusion

- Various feature extraction and classification techniques were employed to carry out multi-label text classification of english quotes.
- Random Forest with Binary relevance and multi-output classifier strategies yielded the highest precision which is 99.47% and 99.41% in label prediction.
- Large label sets create a more diverse range of concepts which makes it harder for the model to interpret and understand label relationships, thus lowering accuracy, recall and F1 score.

Thank you

We would like to express our special thanks to Dr. Saad Bin Ahmed for his dedication towards this course and providing us with thorough knowledge in machine learning concepts.