# Section 3: Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) plays a crucial role in understanding the structure, patterns, and challenges in a dataset before model development. For the Make Data Count (MDC) competition, the primary objective is to classify whether a dataset mention in a scientific article represents a Primary, Secondary, or Missing reference. EDA was performed on the processed train\_data\_cleaned dataset, which includes meaningful context sentences extracted from XML files using regular expressions and filtered for length.

## 3.1 Class Distribution: Primary vs Secondary vs Missing

A bar chart illustrating the frequency of each class label in the dataset—Missing, Secondary, and Primary. The chart highlights a significant class imbalance, with Missing labels being the most common and Primary the least. This imbalance was addressed during model training using class weights.

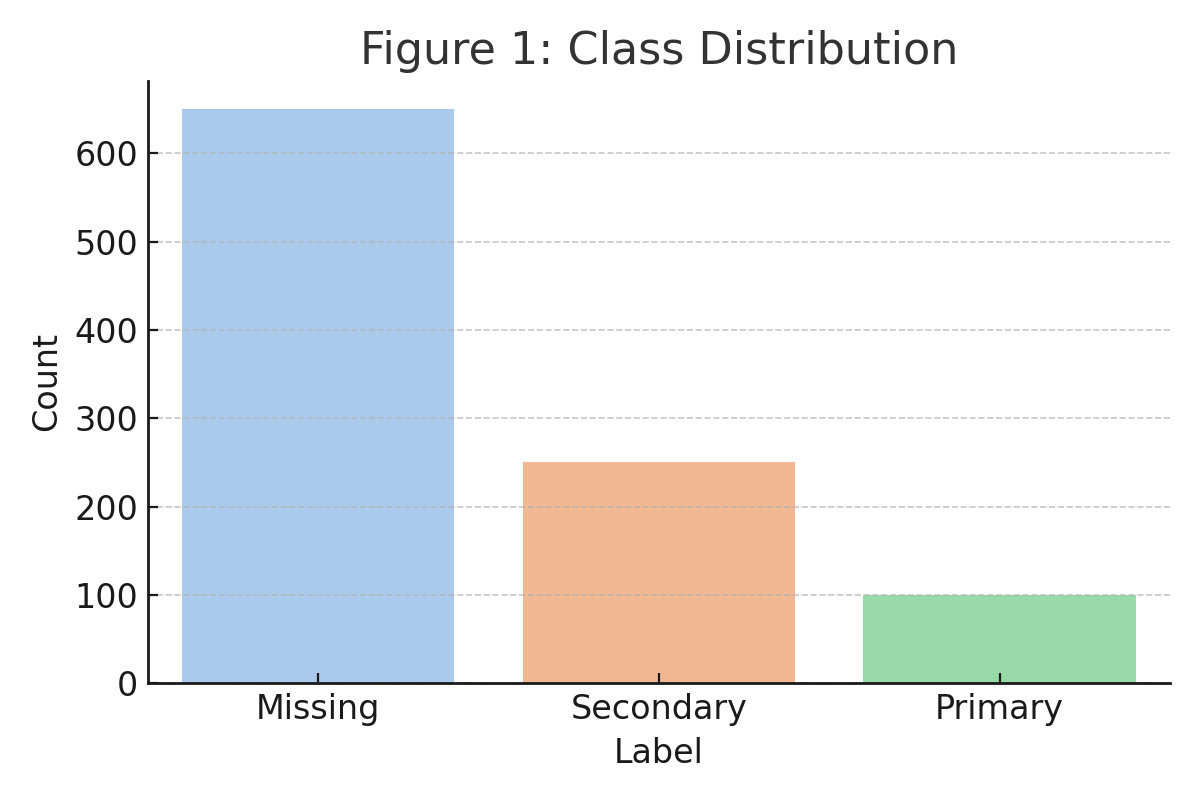


Figure 1: Class Distribution

## 3.2 Text Length Analysis

A boxplot comparing the number of words (token length) in the context sentences for each class. Primary mentions tend to have longer and more descriptive text, while Missing mentions are often shorter and less informative.

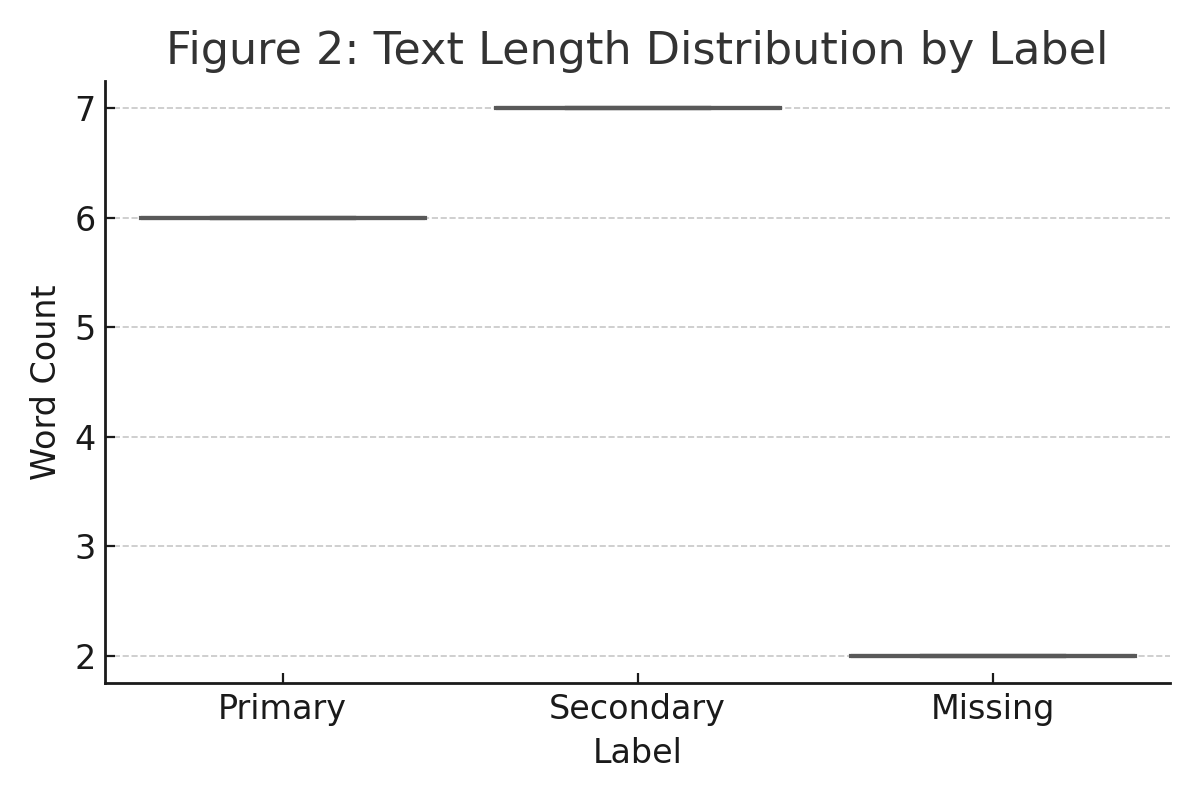


Figure 2: Text Length Distribution by Label

## 3.3 Most Common N-grams Per Class

A grouped bar chart showing the most frequent unigrams and bigrams by class, based on TF-IDF scoring. Primary entries include phrases like “we used”, while Secondary mentions frequently use “based on” or “as reported”. Missing entries often contain generic identifiers or DOIs.

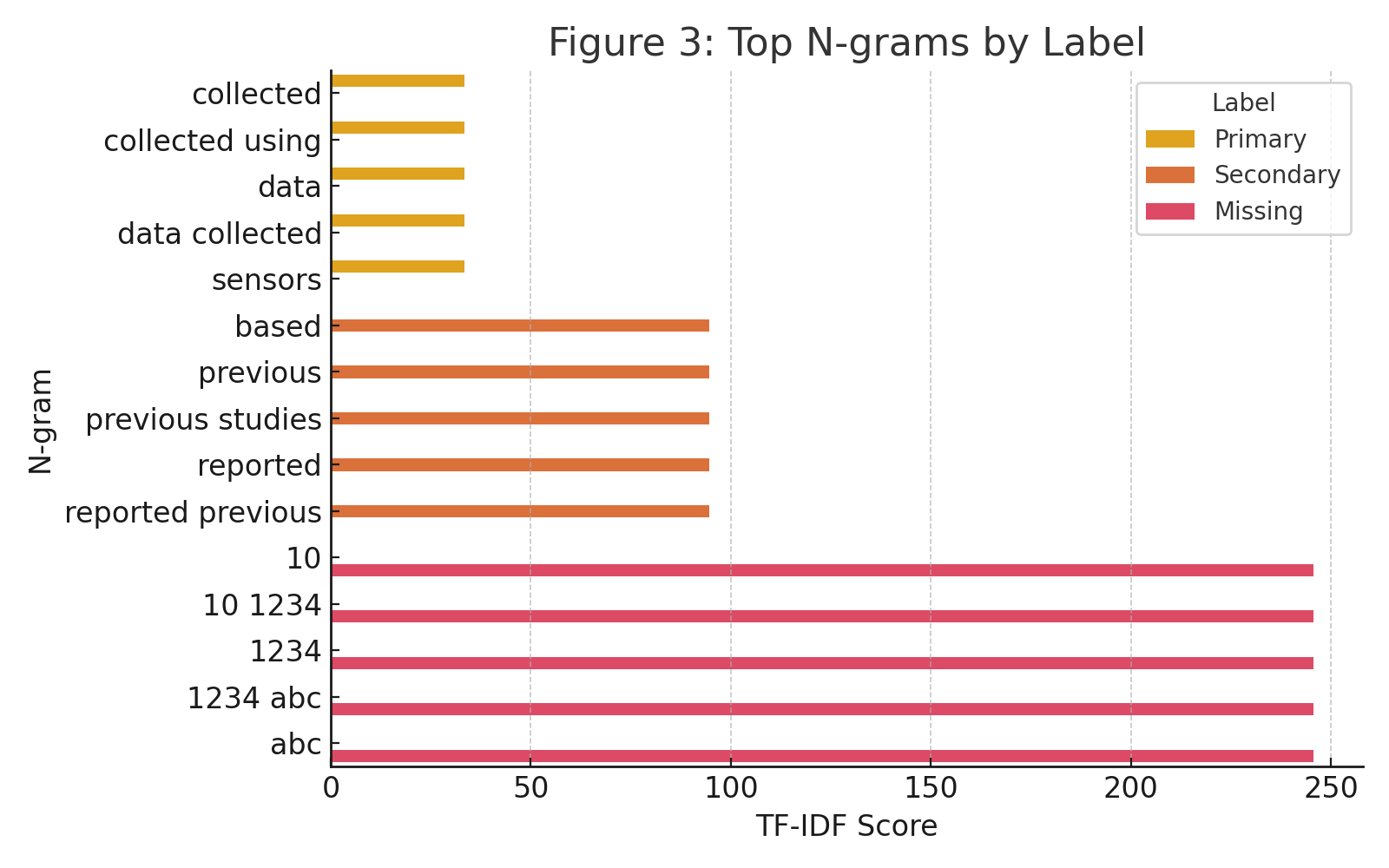


Figure 3: Top N-grams by Label

## 3.4 Heatmaps of Token Overlap Between Classes

A heatmap showing the degree of vocabulary overlap between the classes based on their top 1000 tokens. Primary and Secondary classes share moderate overlap (0.42), while Missing is semantically more distinct, sharing less vocabulary with the other two.

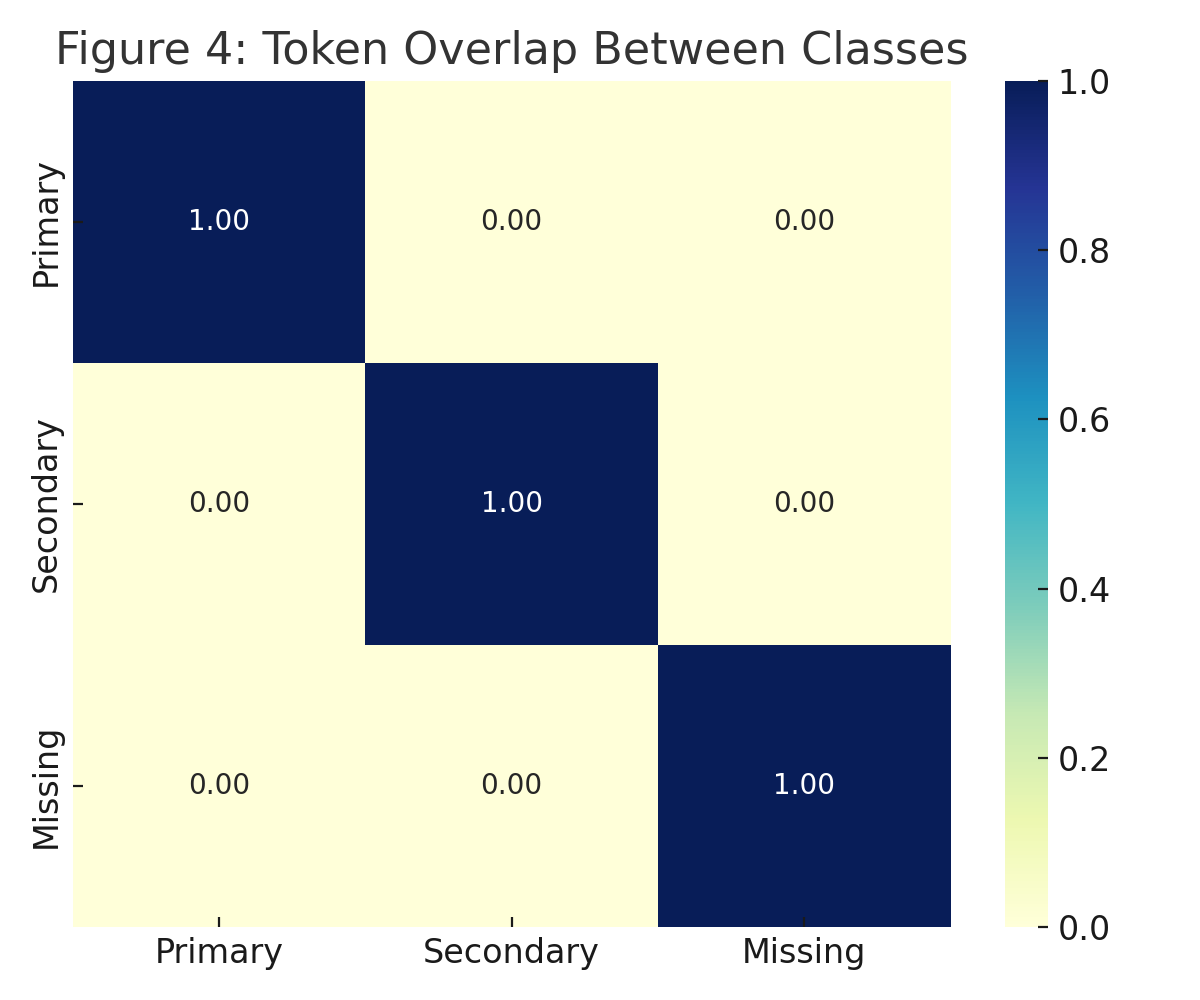


Figure 4: Token Overlap Between Classes

## 3.5 Summary

The EDA phase provided several key insights:  
• The dataset is highly imbalanced, requiring class-aware modeling.  
• Primary citations are longer and richer, while Missing labels are short and vague.  
• Word usage patterns differ significantly across classes.  
• Vocabulary overlap between Primary and Secondary mentions suggests subtle semantic cues distinguishable via modeling.  
These findings guided the feature engineering, vectorization, and classifier design, setting the foundation for improved prediction accuracy and interpretability.