COMPSCI 4NL3: Natural Language Processing
Phase 3 Report

Team 4

Junnan Li
Nawaal Fatima
Rashad Bhuiyan
Sumanya Gulati

17 March 2025

Contents

1 Annotation Analysis

We computed agreement through the use of the Cohen's Kappa agreement metric. This metric is best used for pairwise agreement between two annotators. Since we split the data between 8 annotators with each annotator having some unique sample of a duplicated dataset within their task, this allows for duplicate annotation tasks to exist, making it viable to use Cohen's Kappa to calculate agreement. This is why we chose Cohen's Kappa as our preferred agreement metric.

We calculated the Cohen's Kappa agreement metric using the **cohen_kappa_score** method from the **sklearn.metrics** package. Since the annotations were binary in nature (either having positive-negative sentiment or team-individual focus), we would need to calculate two separate Cohen-Kappa scores to ensure the reliability of our dataset: a Positive-Negative (Sentiment) Score, and a Team-Individual (Focus) Score. We extracted the duplicate data from the combined dataset. Unfortunately, we had only 5% of the duplications available due to a miscalculation of proportions from phase 1, which could affect the reliability of the result as the sample size may seem too small. To combat this, we had our team members personally annotate certain datapoints in the dataset to introduce more duplicate data and included that as part of the full dataset. The Cohen-Kappa scores for each agreement requirement is as follows:

- Sentiment Cohen-Kappa Score: **0.1702838063439066**
- Focus Cohen-Kappa Score: **0.6784420289855073**

Since both of these numbers are positive, this indicates that there is indication of pairwise agreement between two annotators. If the result was negative, that would indicate that the annotators were in disagreement and that annotations are therefore not reliable. In particular, the annotators had stronger agreement when determining the focus of the statement (team-based or individual-based) compared to the sentiment of the statement (positive or negative).

2 Ground Truth Analysis

To decide on the **ground truth labels**, our team chose the 'Majority Vote' method. As per the structure of our datasets and the way our data was annotated, each overlapping data point was annotated twice by two different annotators. In accordance with the requirements of the majority vote method, one of our team members acted as the third annotator and annotated the overlapping datapoints following the same annotation guidelines provided to the other groups.

This method was chosen over the others, namely - 'Weighted Voting' and 'Adjudication' because of the following reasons:

- Since no information was available about the annotators or their experiences, reliability for their annotations could not be satisfactorily determined, thereby, disqualifying the weighted voting method to assess ground truth labels.
- As for adjudication, contacting the annotators, of which there were a total of 8 and conducting discussions till a resolution was reached for each relevant label seemed like an unnecessarily arduous and time-consuming process given our time-constraints.

Taking the aforementioned reasons into consideration, the third annotator's labels acted as the tie-breakers and the most common labels for each category - *positive*, *negative*, *team* and *individual* were then considered as the ground truth labels.

Image 1 shows the amount of data associated with each label. Note that the prefix 'True' for each category name denotes that the category contains calculated ground truth labels.

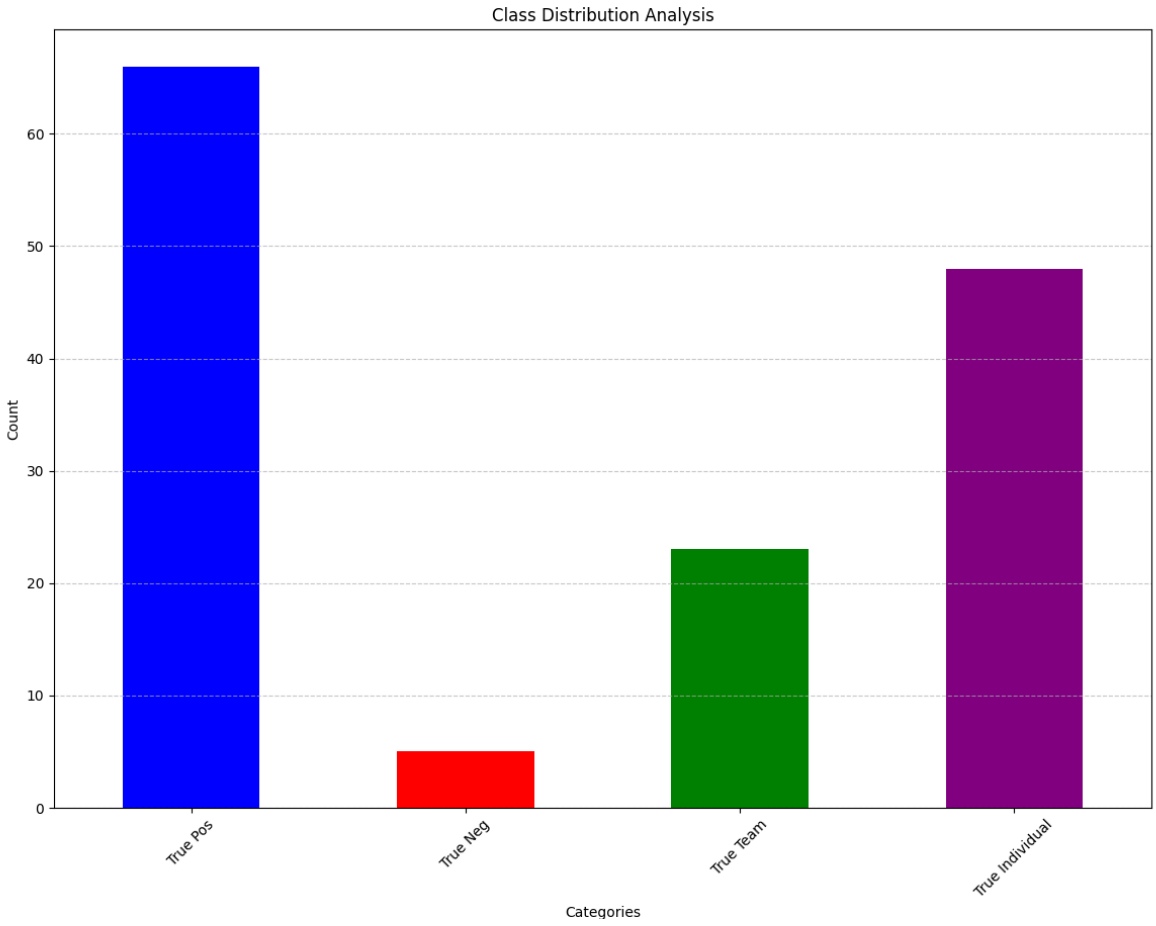

Figure 1: Class Distribution Analysis per Label

The following insights can be gathered from the class distribution analysis:

- *True Pos* is the most frequent labels, significantly higher than the others while its counterpart, *True Neg* has very few occurences. This suggests an imbalance in the sentiment label classes. This large gap between the two might be indicative of a bias towards positive sentiment in the dataset.
- Similarly, *True Individual* appears more frequently than *True Team* meaning individual-focused statements are more common in the dataset. The imbalance between these two label classes suggests that more statements are focused on individuals rather than teams which could indicate bias or be reflective of the nature of the dataset, implying that, the media focuses on the individual over groups.
- A classifier model trained on this dataset may struggle to predict the less frequent classes such as *True Neg*. To avoid potential consequences, the model might need data balancing techniques such as upsampling the minority class or using weighted loss functions in the model in addition to the existing dataset.

3 Baselines

In evaluating the sentiment and focus classification tasks, we chose two primary baselines: the majority baseline and a trained neural network (MLPClassifier). These baselines were selected to provide a clear comparison between a simple, rule-based approach and a more sophisticated machine learning model. The majority baseline predicts the most frequent class in the training data for all instances in the test set. This approach is straightforward but effective, as it serves as a benchmark to measure how much better a trained model performs compared to a naive strategy. For sentiment analysis, the majority baseline predicts the most common sentiment label (e.g., "Positive" or "Negative"), while for focus classification, it predicts the most frequent focus label (e.g., "Team" or "Individual"). This baseline is particularly useful because it reflects the inherent class distribution in the dataset, allowing us to gauge whether a more complex model is adding value or if the task can be adequately addressed by simply predicting the most common class.

The trained neural network, on the other hand, was chosen to evaluate the potential of machine learning models to capture more nuanced patterns in the text data. Neural networks, especially feedforward neural networks like the MLPClassifier, are capable of learning non-linear relationships and can be effective for text classification tasks when combined with appropriate feature extraction techniques like TF-IDF. By training the neural network on the TF-IDF vectorized text data, we aimed to assess whether a more advanced model could outperform the majority baseline and provide better predictions for both sentiment and focus classification. The neural network's ability to learn complex patterns makes it a strong candidate for tasks where the relationships between features and labels are not immediately obvious.

During development, a sample results from the evaluation show that the majority baseline achieved an accuracy of 92.96% for sentiment classification and 67.61% for focus classification. These results indicate that the majority baseline performs well for sentiment analysis, suggesting that the dataset is heavily skewed towards one sentiment class. However, for focus classification, the majority baseline's performance is lower, indicating a more balanced distribution of labels or a more challenging task. In contrast, the neural network achieved an accuracy of 90.14% for sentiment classification and 59.15% for focus classification. While the neural network's performance for sentiment classification is slightly lower than the majority baseline, it still demonstrates strong predictive capabilities. For focus classification, the neural network's performance is lower than the majority baseline, suggesting that the task is more challenging and that the neural network may require further tuning or more sophisticated architectures to improve its performance.

We chose these baselines to provide a balanced view of the classification tasks. The majority baseline helps establish a simple, interpretable benchmark, while the trained neural network evaluates the potential of more advanced models to capture complex patterns in the data. By comparing the performance of these two baselines, we aimed to understand whether the additional complexity of the neural network is justified by improved performance. This approach allows for a comprehensive evaluation of the tasks, highlighting both the simplicity of class distribution and the potential of more advanced models to improve upon it.

4 Relevant Documents