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

3 Baselines

4 Relevant Documents