Final Project

University of Texas – San Antonio
IS-6713-901: Data Foundations
Professor Anthony Rios
05 December 2023

San Annotations:
Emily Bates XZZ320
Bella Lin PVZ339
Diego Mejia OOL750
Richard Wycklendt NWR023

Task Design

Our group, The San Annotations, were responsible to create and write-up guidelines that would instruct the annotators on how we wanted the dataset annotated. We created a list of categories that may or may not pertain to each comment in the data set to help us create a machine learning model. These guidelines would be the basis for each annotation. We outlined the categories and defined them to guide the annotators. The categories were all annotated if the text was related or not (binary). Our categories were Technology related, Internet Cost, Internet Speed, and Internet Accessibility.

Our guideline specified two tasks: identify technology-related comments and annotate the second category for each comment. In the guidelines, we provided both positive and negative examples for each category. A positive example would be a text that would match one of the categories we defined. A negative example was a text that had similar information of the category defined but did not match the category. In addition to the positive and negative examples, we gave examples of what might be a corner case or tricky to decide if it matched the criteria.

We then sent the first draft of our guideline along with 25 comments to two annotator groups: the Annotation Avengers and the Predictors. We instructed them to annotate the first 25 comments as a test and get an idea of how our guidelines faired.

After receiving negative guideline feedback along with 25 annotated comments, we have decided to modify the tasks. We kept the first task as identifying technology-related comments, and changed the second task to determine whether the comment is discussing internet cost, internet speed, internet accessibility, or others.

Annotation Modeling

Before modeling our project, we reviewed the annotations from the Annotation Avengers and The Predictors, creating a Golden Standard annotation file. We then calculated Cohen's Kappa Score to evaluate the agreement between the two annotation groups. When looking at the scores, the result below showed that there is generally low agreement between The Predictors and The Annotation Avengers across the different categories. The negative Kappa value for the "other" category is unusual and might indicate that there is systematic disagreement or a lack of agreement beyond what would be expected by chance.

```
The Predictors vs. The Annotation Avengers:

Cohen's Kappa for 'Technology-related': 0.013781254838923784
Cohen's Kappa for 'Cost': 0.22097696570528136
Cohen's Kappa for 'Speed': 0.0541946467417006
Cohen's Kappa for 'Accessibility': 0.04101995565410199
Cohen's Kappa for 'Other': -0.005963929007688495
```

Next, we compared each annotator group with our golden standard and calculated Cohen's Kappa Score for each pair as seen below. We can see from the

result that The Annotation Avengers had low levels of agreement with our gold standard compared to The Predictors who had higher Kappa values. Therefore, we believe the Predictors performed the annotation task reasonably well since they have high agreement scores with our golden standard.

```
The Predictors vs. Gold Standard Annotation:

Cohen's Kappa for 'Technology-related': 0.6225136915304315
Cohen's Kappa for 'Cost': 0.5610514698505067
Cohen's Kappa for 'Speed': 0.6635050158783571
Cohen's Kappa for 'Accessibility': 0.3659420289855072
Cohen's Kappa for 'Other': 0.4959390342533633

The Annotation Avengers vs. Gold Standard Annotation:

Cohen's Kappa for 'Technology-related': 0.05656522885721815
Cohen's Kappa for 'Cost': 0.14064441315331733
Cohen's Kappa for 'Speed': 0.05650262678245943
Cohen's Kappa for 'Accessibility': 0.07063197026022305
Cohen's Kappa for 'Other': 0.005334355014588343
```

For the modeling process, we decided to use four features: the number of capital words, the number of exclamation points, positive word counts, and negative word counts. These features can contribute to understanding the style and tone of the text, which may be helpful as input features for machine learning models. Since our project focuses on classification, we have explored two machine learning models: LinearSVC and Random Forest. From our class, we have grown more familiar with LinearSVC and we believe it is more suitable for binary and multiclass classification, therefore we chose to use LinearSVC as our training model.

Evaluation Metrics and Error Analysis

Our machine learning task is classification, therefore we chose to use Precision, Recall, and F1 score as evaluation metrics for both macro and micro to assess the performance of our model.

Validation Macro Metrics:

Validation Precision: 0.6505102040816326
Validation Recall: 0.27154471544715447
Validation F1 Score: 0.31865079365079363

Validation Micro Metrics:

Validation Precision: 0.61
Validation Recall: 0.6559139784946236
Validation F1 Score: 0.6321243523316062

As seen above, the validation set metrics show us better results for micro metrics. We see a relatively low F1 score for macro, meaning less balance compared to the higher F1 score for micro. The micro metrics also indicate a higher overall precision but slightly lower recall, suggesting that the model is better at making accurate positive predictions but may miss some positive instances.

Test Macro Metrics:

Test Precision: 0.2669753086419753
Test Recall: 0.1802615193026152
Test F1 Score: 0.1878121878121878

Test Micro Metrics:

Test Precision: 0.5595238095238095
Test Recall: 0.5
Test F1 Score: 0.5280898876404494

From the test set, we can see lower macro metrics compared to the micro metrics. The macro results show that on average about 26.7% of the positive predictions made by the model are correct across all classes, with the model capturing 18.0% of the actual positive. However, the low F1 score indicates low performance balance between precision and recall. Looking at the micro result, we can see that the micro metrics are generally higher than the macro-average metrics. This suggests that the model is performing well when training on classes with larger sample sizes (like tech-related) but may not generalize as well to smaller classes when considering the macro metrics.

In conclusion, we believe that if we were to look at each classification individually, the model might have better performance for some columns compared to others. The model might be better at identifying technology-related comments compared to identifying internet feature-related comments since we do not have a large sample for the training model. Having a larger dataset would allow us to better train the model, and would subsequently increase our precision, recall, and F1 scores.