Reddit Miners Final Report

Sridhar Malladi

**Purpose:**

The focus of this project is to classify text samples taken from Reddit to help the city government of San Antonio identify relevant comments from people living in the area.

**Data**:

The dataset used was 2000+ samples, of these 1000 were chosen to work with. The labels used for marking relevant topics in the samples were the following: Technology, Politics, Business, Entertainment, and Miscellaneous.

Data labeling was accomplished by sharing the dataset and category guidelines with groups. Each category was described in the guidelines, specifically the characteristics that show a text sample has relevant material for that category. Each partnered group marked relevant topics for each comment based on provided guidelines. The resulting data showed significant class imbalance with the Entertainment label showing 86% going to the negative class being the most extreme and the Business label at 55% being negative as the best balanced. The other labels showed 75% or greater favorability to one class.

For this project, and for generated models potential future use, each column was handled independently of the others. This means that a text sample could be relevant to technology and business but not politics and would be marked as such. Due to some weaknesses in the guidelines, the cohens kappa for each column was low, with the best being for the political column with a value of 0.404 and the lowest being 0.148 for the miscellaneous column.

**Features:**

Model construction requires information to be either derived from text samples, or have the text samples themselves be tokenized, in a numerical format for the model to recognize. For this project the features generated for model building belonged to four categories. The first being sentiment, this included feature was setup as a one-hot-encoded column with positive, neutral, and negative as the three possibilities. The second feature was the length of the sample or the word count. This was constructed by taking the text sample, splitting along whitespace, and counting the number of word characters found. The third feature was a binary categorization for the presence of a web address. Text samples were scanned using a regular expression to find all web addresses present. The final feature was based on emotional lexicons. The lexicons used were lists of words associated with specific emotions, in this case anger, joy, sadness, surprise, and anticipation. Each emotions column had a count of words found in both the text sample and the emotions lexicon.

Different combinations of these features were fed to different models to compare model performance. One set was the full feature set, the next was only sentiment, the third was only length of sample and web address presence, and the final was the emotional lexicon counts.

**Model Evaluation:**

The data was further prepared for use in models using the train\_test\_split function from sklearn. Each combination of data was split with 70% going to training and 30% going to the corresponding test data. It should be noted that throughout this project data was handled in a pandas dataframe object.

A function was created to run the different combinations of features into the same models. The function also displays the scoring metrics used which were the macro and micro versions of precision, recall, and f1. The models being used are from the sklearn library, they are the SVC, the LinearSVC, and the RandomForestClassifier (max\_depth = 3, random\_state = 30). The SVC and LinearSVC models were set to use their default parameters in the initial run. The models were also evaluated using cross validation score.

Each feature set was run through the function to get a handle on how useful the a given feature set was. The sentiment-based feature set 2 was the least useful, showing the lowest scores relative to the other sets. The scores resulting from this training set struggled in the Macro scoring metrics, with the exception with the business label where the scores were all close at within a few percentage points, the other labels saw 20-30% differences between the Micro and Macro metrics. The best performing feature set was set 0, or the complete feature set. Feature set 1 and 3, emotions and web address/length, performed at a similar level with differences arising with different target labels. Overall, the feature sets followed similar behaviors to the low performer in that the Micro scores did not do as well compared to the macro. This behavior is expected given the data label classes are unbalanced.

The next test was to run the different feature sets against one another for the same label to determine which model, in its given state, and feature set combination would give the best results for the given label. It should be stated that the scores given by the different models were comparable to one another and the following determinations will likely shift between runs given the random nature of splitting data. The best model for the tech label was an SVC trained on feature set 1, for political it was LinearSVC on the complete set, for business it was a Random Forest trained on the complete set, for the entertainment label it was an SVC on the full set, and finally the miscellaneous label top performer was a Random Forest trained on feature set 3.

**Final Predictions:**

Once the best model was determined, each model was given a class weight adjustment for final predictions. After applying class weighting, models using labels that were reasonably balanced, like the business label saw general improvement in both micro and macro. Whereas models looking at labels with more class imbalances, such as the tech focused model, saw increases in the macro performance of up to 5% also decreases in micro performance by up to 10%. This is an example of where its important to understand if its more important to have a balanced outlook or to ensure that the samples flagged as a given label belong. In this instance its more important to be sure that the flagged sample belongs given the large volume of text available and a desire to maximize the value gained vs taxpayer money wasted. Due to the weaknesses in the dataset indicated by each label Cohens kappa, the models were left with their class weighting for the final predictions.