Jessica Strait
Professor John Yen
DS 200
Mini Project Deliverable #2

# Section A: Important Features

In a decision tree, an easy mathematical measure of important features is the information gain provided by the feature- this can be estimated based on the entropy change between a feature and its children. If the entropy lowers significantly from a node to its child, then there has been significant information gain. Feature importance can also generally be identified by the number of units a node applies to in relation to the total sample size. Through analysis of all the folds across the ideal max_depth of 4, one may also notice recurring keywords that produce very low entropy results. Some of the most important features that yielded low entropy results and were used to identify a large number of tweets included proper noun mentions of Swedish climate activist Greta Thunberg, in fold 1; American Republican politician Mick Mulvaney in fold 4; American Democratic presidential candidate Joe Biden in folds 3, 6, and 7; and "German," referencing criticism of Thunberg by the German media. While proper nouns tended to yield the lowest entropy results in the decision trees, some words such as "is" and "more" also happened to be critical root nodes in several folds; they are reflected in Figure 1 as being some of the most important features regardless of tweet sentiment. These features occasionally did not lead to significant decrease in entropy in the actual decision trees, but they received a large sample of tweets and appeared in multiple folds. However, the 5 features (shown in Figure 1) that were most important across different folds were "is," Biden, "German," "more," and "our." We see the aforementioned phrases in multiple folds, often receiving a large number of nodes and producing a decrease in entropy. There were not many significant differences across the eight folds, but some features continuously yielded high entropy values when an article or otherwise vague keyword was used as the defining feature and was present in a tweet (i.e. "is"). This may be an indication that with the ideal max_depth selected, there are certain folds that may have benefited from a higher max_depth to identify more indicative words than the features selected out of the total data set, though this would have resulted in overfitting for the full model (see Figure 2).

| | is | biden | german | more | our | no | there | or | over | issue | we | but | public | mulvaney | same | than | are | @gretathu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.155558 | 0 | 0 | 0 | 0 | 0.209908 | 0 | 0 | 0 | 0.167154 | 0.162987 | 0.160917 |
| 1 | 0 | 0.170877 | 0.1985 | 0 | 0.217868 | 0.187869 | 0 | 0 | 0 | 0 | 0.224887 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.217375 | 0.213319 | 0 | 0.195923 | 0 | 0 | 0 | 0 | 0.248035 | 0.125348 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.186367 | 0 | 0.210153 | 0 | 0 | 0 | 0.193931 | 0 | 0 | 0 | 0 | 0 | 0.205314 | 0.204235 | 0 | 0 | 0 | 0 |
| 4 | 0.159158 | 0 | 0 | 0.136855 | 0 | 0 | 0 | 0 | 0 | 0.113661 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.204073 | 0.188523 | 0 | 0.258431 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.19037 | 0 | 0 | 0 |
| 6 | 0 | 0.242663 | 0.235684 | 0 | 0.269014 | 0 | 0 | 0.252639 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0.097754 | 0 | 0.163663 | 0.183788 | 0.200598 | 0.354197 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.096081 | 0.090598 | 0.089778 | 0.086111 | 0.076387 | 0.06023 | 0.038832 | 0.028071 | 0.027559 | 0.026557 | 0.024987 | 0.023323 | 0.022813 | 0.022693 | 0.021152 | 0.018573 | 0.01811 | 0.01788 |

*Figure 1: Important feature spreadsheet; row 8 expresses feature importance across all 8 folds*

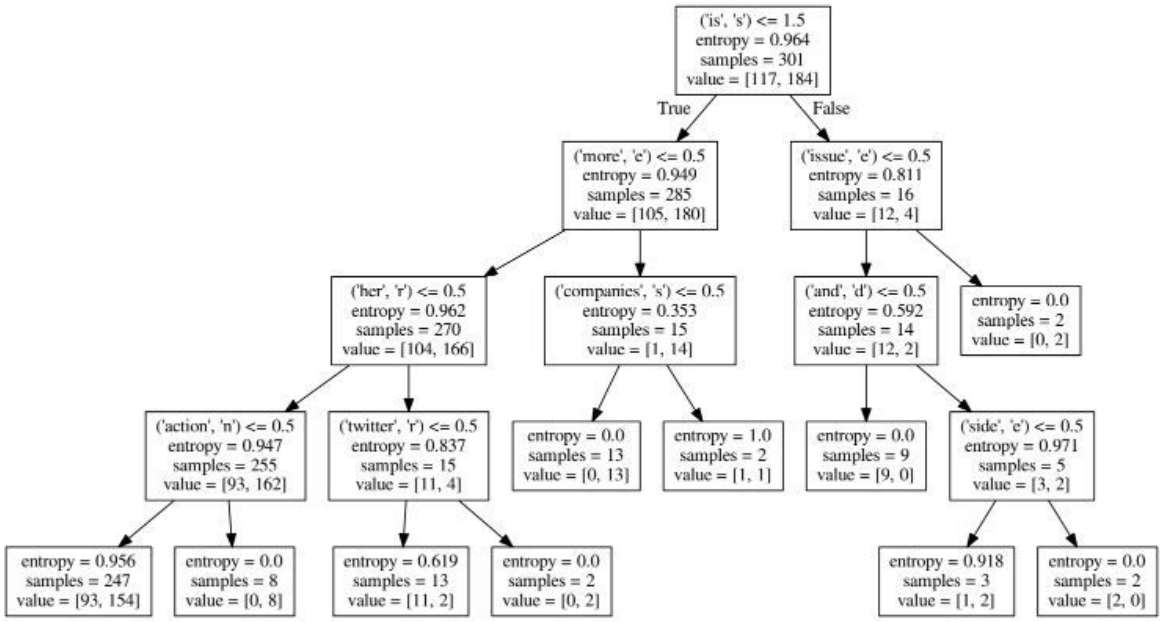


*Figure 2: Decision tree at max_depth 4 for fold 5, which may have benefited from a higher max_depth due to repeated high entropy values when no important feature is in a tweet*

# Section B: Compare Important Features with Word Clouds

When using decision trees to predict the sentiment of tweets gathered, it is important to note that the tree is receiving a relatively equal number of positive and negative sentiments. In contrast, the constructed word clouds are differentiated by the sentiment of the tweets. Important features are highlighted in word clouds through the use of larger text. Visually, this means that words occurring more frequently in a certain sentiment category appear larger in the word cloud. Consider that the positive sentiment word cloud (Figure 3) contains phrases like “Extinction Rebellion.” While this phrase might be extremely indicative of a positive sentiment, if the phrase does not appear in any tweet of the opposite sentiment, it is not an important feature of the entire decision tree. Recall that feature importance is also partially quantified by the proportion of tweets that “reach” and are evaluated by that node. “Extinction Rebellion” does not appear in the negative word cloud (Figure 4), and therefore may not be considered an important feature of the total data set. In the given sample, a small number of tweets of a specific sentiment are not significant enough to be considered important features, even though the word may appear in many positive tweets. For this reason, while important features are highlighted in the word cloud, not every word in a word cloud will be considered an important feature in the overall decision tree that includes tweets of the opposite sentiment.


*Figure 3: Positive sentiment word cloud*

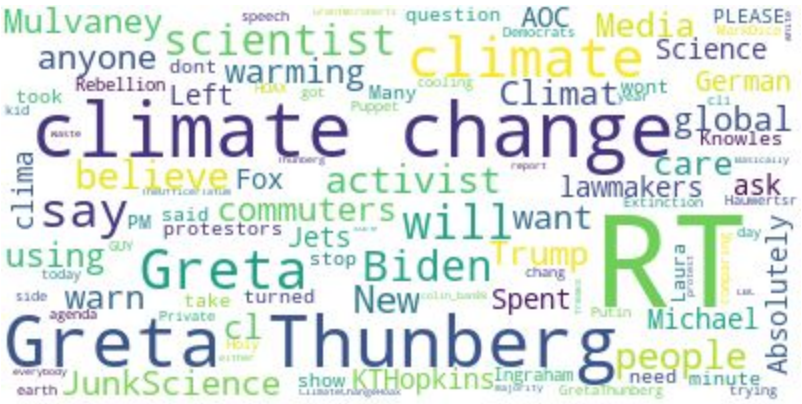


*Figure 4: Negative sentiment word cloud*

## Section C: Interpreting Important Features

While not all features were numerically important across folds, several features were regularly indicative of a certain sentiment regardless of the fold. Mention of German criticism of Thunberg tended to yield a negative sentiment, as did mention of Mulvaney or Biden. However, mention of the phrase "@gretathunberg" tends to yield a positive sentiment, likely because the use of the full username indicates either a direct mention of the activist or a retweet. Use of the word "our," which appeared as a root node three times, was likely used in terms of a collectivist political ideal and was associated with positive sentiment: perhaps referring to "our" planet, action by "our" government, or concern for "our" children. Most sentiments calling for climate action on the part of society as a whole may use "our" as a word to engage readers through pathos, eliciting emotional reactions about climate change. We also saw some consistency in the roles of important features across multiple folds. For the proper noun "German," the role across folds 2, 4, 7, and 8 was identical, and its appearance indicated a negative sentiment; this also holds true for "Biden" in folds 2, 3, 6, and 7 (see Figure 5). This equality of sentiment based on a feature is evident with several other important features such as "Mulvaney." However, for some less

descriptive words (i.e. “is”), they were important features that did not present a certain sentiment and therefore did not hold a consistent role across folds. Generally, for vague keyword important features, the roles across folds were not identical; for highly indicative important features, and all proper nouns included in the decision trees, the roles across folds were identical.

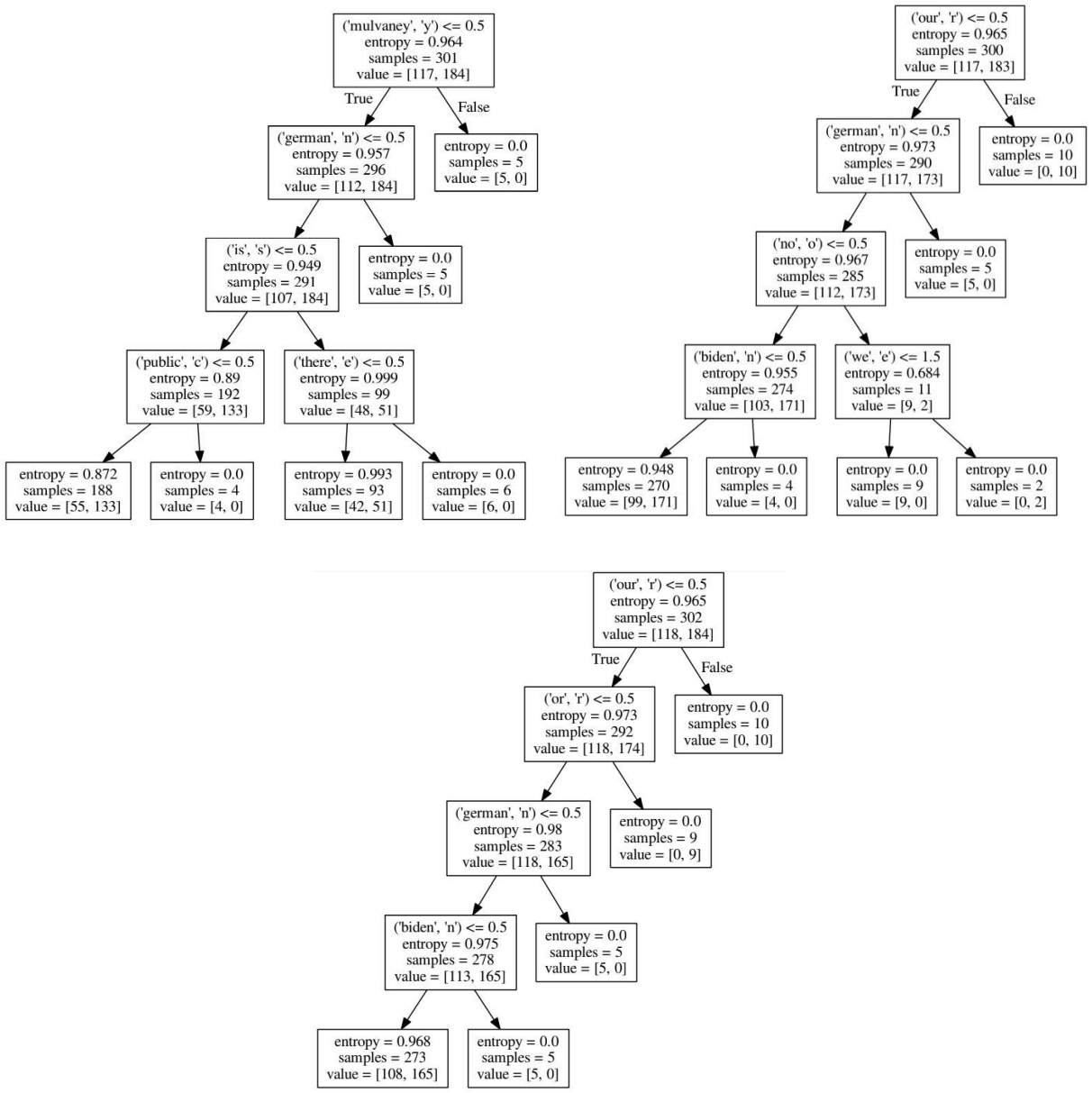


*Figure 5: Decision trees at max_depth 4 for folds 4, 2, and 7 respectively*

# Section D: Findings, Questions, and Hypotheses

Several patterns across tweets related to climate change, either denying or encouraging action, became clear when analyzing the hundreds of data samples gathered. There were several words

and phrases that were repetitive across tweets and prominent in word clouds, even if they were not considered important features in all of the decision trees. Just by analyzing the decision trees at the ideal max_depth of 4, phrases like “Biden,” “German,” and “Mulvaney” consistently indicated a negative sentiment in a tweet. Similarly, the root node of “@gretathunberg” in fold 1 gave a clear picture of the types of interactions positive tweets are seeking by directly tagging or retweeting the young climate activist; this type of connection was also evident by the feature importance of “our,” and the positive sentiment it regularly yielded. It would be appropriate to hypothesize that tweets containing direct tags or personal connections typically indicate a positive sentiment; supportive tweets are perhaps more likely to be intentionally communicating with the individual in question. Alternatively, through negative tweets mentioned proper nouns to criticize individuals, tags or personal connections do not occur as frequently in these tweets. This may pose a broader question about the style of communication within tweets- meaning, are positive tweets in general more likely to form a personal connection with the object of the tweet than a negative tweet? Overall, analyzing decision trees and word clouds gave a picture of the features not only existing in tweets about climate change, but also in the underlying political argument about the issue.