Group 15: Identifying Misinformation

Team Members: Theodoros Pateros, Saloni Sharma, Brenden Moore, Yukti Handa

Business Understanding and Societal Implications:

Misinformation and fake news are now commonplace in our society. In the last two decades, online news sources and social media have risen and become the main source for news for many individuals, thus, it is easier than ever for false information to spread. Negative consequences that have arisen from this issue range from the spread of conspiracy theories, erosion of public trust in institutions, and even more recently, the promotion of dangerous health and safety practices with regard to the COVID-19 pandemic. Also, fake news has provided credibility to people's political beliefs ultimately leading to further polarization and a population that is unwilling to consider alternative perspectives.

The term "fake news" was popularized during the 2016 election cycle when thousands of biased or false articles were posted on social media in pursuit of advertising revenue. The then presidential candidate, Donald Trump, used the term as a way to combat negative stories against him attempting to highlight fundamental flaws in the way news was consumed.

While many probably think of the pervasiveness of fake news more on a societal or political level, it is also important for companies and corporations to mitigate the spread of fake news regarding their practices. For a company, if they were to be falsely accused of wrongdoing or unethical behavior, their overall reputation could be damaged thus eroding trust with customers, shareholders, and stakeholders. While the possible fallout of fake news on a company is endless, lawsuits taken against a company would be costly in terms of time and money. A large scandal not rooted in truth would take company resources away from the core operations jeopardizing a firm's long-term well-being.

With regard to how high fake news has spread throughout society, 80 percent of consumers in the United States have reported having seen fake news on the coronavirus outbreak. In addition, 64% of adult Americans in a study conducted by the Pew Research Center believe that fake news has caused a great deal of confusion regarding the basic facts of current

events. Our group identified individual instances of when the dissemination of fake news warped our perspective of the political landscape and current events.

As a group, we recognized the wide-spread implications of this issue and decided to create a classifier model with the goal of detecting possible fake news in mind. For the sake of simplicity, we decided to create a model that focuses on identifying the words that are most commonly used in both true and false articles of news. This model will serve as the foundation for a possibly stronger and more effective tool that can be applied on a greater scale. It is conceivable to believe that all companies could use some sort of fake news detection software to maintain their brand's reputation, it could be of particular usefulness for news outlets, social media firms, and even government agencies.

Dataset:

Our dataset can be found on Kaggle and contains 79,000 articles of misinformation, fake news, or propaganda. Articles are labeled as either true or false, and we are also given the article's text, title, and the date of publication. The data is separated into three different CSV files with the first containing the TRUE articles, the second being FALSE, and the third containing articles that have been titled as Russian propaganda. Fake articles are taken from American right wing extremist websites and disinformation or propaganda cases. True articles are sourced from a variety of entities such as the Washington Post, Reuters, and the New York Times. For simplicity's sake, we focused only on the TRUE and FALSE articles as the third Russian propaganda set could complicate the model.

Data Exploration/Analysis:

As mentioned in the previous section, we utilized two datasets to construct our model. The first dataset pertained to accurate text information while the second dataset pertained to misinformation. Both of these datasets contained empty rows which we decided to drop entirely. The resulting datasets without null values contained 43,642 and 34,946 rows for the fake and true datasets respectively.

Due to the large size of both datasets, we decreased the number of rows to 2,500 for each dataset. This process was conducted by randomly selecting 2,500 rows from each dataset and dropping the remaining ones. The next step was to add a column to both datasets labeled as "true(0)/fake(1)" which would serve as the binary classifier for our analysis. We populated this column with 0 values in every row for the dataset that contained true information and 1 in every row for the dataset that contained fake information. We then concatenated the two datasets to create a merged dataset that contained the text and the binary classifier pertaining to each text. The resulting dataset contained 5,000 rows where 2,500 rows scored 0 in the binary classifier and 2,500 rows scored 1 in the same column.

Data Modeling:

In order to predict whether a text pertains to accurate information or misinformation, our group elected to utilize a Linear Support Vector Machine (SVM) Classifier. This is because our research suggested that this model performs best in the task of text classification. SVM models perform by creating a hyperplane that divides a space into two subspaces. One subspace contains tags that belong to one group while the other contains the vectors that do not belong to that group.

Taking the above into consideration, the first step in creating our model was to remove all the non-alphanumeric characters from the text and tokenize the text into individual words using the function 'word_tokenize' from the 'Natural Language Toolkit (NLTK)' module. We then converted all the words to lowercase in order to make the words more suitable for feature extraction.

After tokenizing the text we were able to identify the most common words in our concatenated dataset. In doing so, we excluded stop words and special characters in order to provide meaningful results. The top-10 most common words that resulted from this process are pasted below along with their respective count of occurrence.

1. **said: 14248**
2. **trump: 13203**

3. **would: 6285**
4. **people: 5760**
5. **president: 5444**
6. **one: 5360**
7. **u: 4537**
8. **clinton: 4420**
9. **mr: 4411**
10. **new: 4265**

Next, we conducted feature extraction in order to make the data suitable for a classifier model. We did so by mapping each word from our vocabulary to an index in our feature vector. This process allowed us to develop a feature matrix that served as a numerical representation of the text where each row corresponds to a text sample and each column corresponds to a word in the vocabulary. The value that was inserted into each cell represents the frequency of occurrence of a word in the corresponding text. A visual representation of our feature matrix is attached below:

Out[293]:

	feature_0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	...	feature_53283	feature_53284	feature_53285
0	1.0	3.0	1.0	9.0	5.0	4.0	1.0	1.0	5.0	1.0	...	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

5000 rows × 53293 columns

Completing this step enabled our team to fit a linear SVM model on our concatenated dataset. We split the data into 70% train and 30% test in order to develop a prediction model for the classification of a text as misinformation.

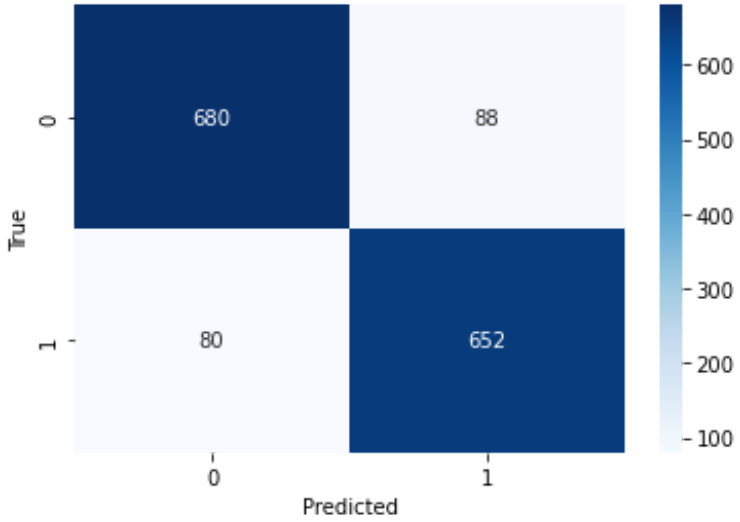
Model Evaluation & Performance

Our model was tested across four different evaluation metrics: Accuracy, Precision, Recal, and F1-score. The results for each metric are summarized below along with a description of the metric:

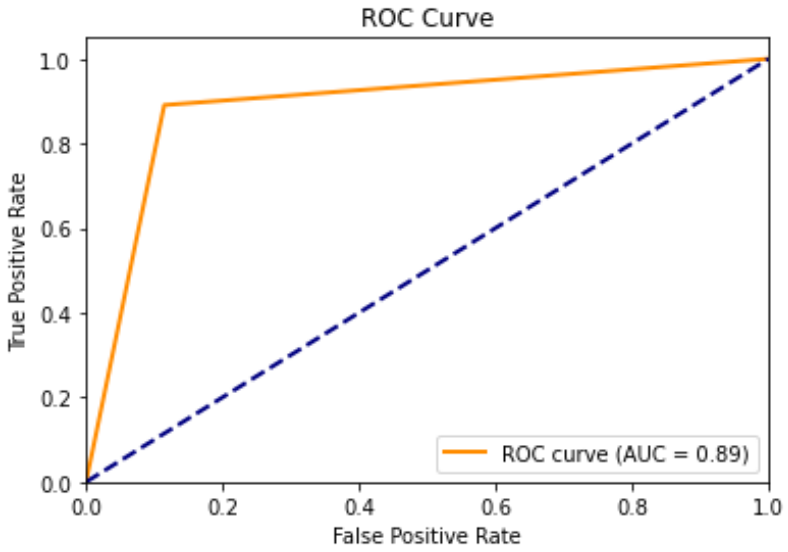
- **Accuracy: 0.888**
    - Accuracy is a metric that measures the proportion of correct predictions among all predictions. In other words, it is calculated by dividing the number of True Positives(TP) and True Negatives(TN) by the total number of samples. The Accuracy Score of 0.888 means that our model is expected to classify between true and false accurately 88.8% of the time.
- **Precision: 0.881**
    - Precision is used to measure the proportion of True Positives among all the predicted positives. Its calculation is derived from dividing the number of True Positives(TP) by the sum of True Positives(TP) and False Positives(FP). The Precision Score of 0.881 means that 88.1% of the samples that we predicted as positive were actually positive. This shows that our model identified the majority of the positive samples (fake news).
- **Recall: 0.891**
    - Recall measures the proportion of True Positives(TP) among all the actual positives(TP+FN). We calculate Recall by dividing the number of True Positives (TP) by the sum of True Positives and False Negatives(TP+FN). The Recall Score of 0.891 means that our model correctly identified 89.1% of all the positive samples in the dataset.
- **F1-Score: 0.886**
    - The F1-Score is a combination of Precision and Recall as it utilizes the harmonic mean of the two. It is used to estimate the model's balanced ability to both capture positive cases (Recall) and be accurate with the cases that it does capture (Precision). We calculate the F1-Score as 2 times the product of Precision and Recall divided by their sum: (2*Precision*Recall)/(Precision+Recall). The

F1-Score of 0.886 means that the model made the correct prediction 88.6% of the time across the entire dataset.

Attached below, is the confusion matrix that our model developed:



To conclude the model evaluation performance, our team plotted the Receiver Operating Characteristic Curve (ROC Curve) for the model. The ROC curve is a graphical representation of the performance of a classification model at all classification thresholds. We also calculated the Area Under the ROC Curve (AUC) at 0.89 which indicates that our model does a good job of correctly classifying instances from both classes. The graphical representation of the ROC curve is attached below along with the model's AUC score:



Recommendation

In conclusion, the ability to identify between true and false information has become of key importance in an era when the spread of misinformation and propaganda has skyrocketed to levels that we have not seen before. Thus, a tool such as the one that we are proposing in this analysis is critical for both individual and organizational use. We have developed this tool in an attempt to distinguish between news that are accurate and news that are likely to contain false information in an attempt to tackle the spread of misinformation around the world. This tool can be utilized by social media companies to aid their efforts to flag and censor misinformation on their platforms. By implementing our tool into their platforms, social media companies are likely to see an increase in their user base since customers will begin to trust their platforms as a news source. This is estimated to increase online advertisement in their platforms which is the leading source of revenue for social media companies. Therefore, we are estimating that the implementation of our tool in social media platforms will have an important impact in the generation of revenue for such social media companies.