Finding Commonalities in Misinformative Articles Across Various Topics

By: Maximilian Halvax, Lucas Nguyen, Hwang Min Yu

1. Abstract

In order to combat the large-scale distribution of misinformation online, We wanted to develop a way to flag news articles that are misinformative and could potentially mislead the general public. In addition to flagging news articles, we also wanted to find commonalities between the misinformation that we found. Were some topics in specific containing more misleading information than others? How much overlap do these articles have when we break their content down into TF IDF and see what words carry the most importance when put into various models detecting misinformation. We wanted to narrow down our models to be trained on four different topics: economics, politics, science, and general which is a dataset encompassing the three previous topics. We Found that general included the most overlap overall, while the topics themselves, while mostly different from the other specific topics, had certain models that still put emphasis on similar words, indicating a possible pattern of misinformative language in these articles. We believe, from these results, that we can find a pattern that could direct further investigation into how misinformation is written and distributed online.

2. Introduction to the dataset:

Our data is collected from Simon Fraser University's fake news research where we use the datasets containing Snopes, Politifact, and Emergent.info articles of varying real and fake news from 2010 to 2018. We took articles from each dataset to create a new dataset that contains real and fake news for specific genres of news. We gathered news about 100 data for each economic, political, and scientific topic from the Snopes, Politifact, and Emergent.info datasets to use as our training dataset. The dataset includes both misinformation and non-misinformation. We also created a dataset mixed with the topics we are using as our testing dataset. Our plan with these datasets is to find commonalities of misinformation across different topics. To do this, we are training our models based on set genres and then testing the results on a set of data with varying genres of news.

3. Identify predictive tasks:

For our research, we are using our training dataset to predict whether a random article, regardless of the genre, is misinformative or not. We will train our models so that it learns the commonalities of misinformation for a set topic. Then we will test our findings onto a random article to see if our model can accurately predict whether that article is misinformative or not. We use the scores of the Decision Tree, Logistic Regression, Random Forest Classifier, and SVM to test our models' accuracies. In addition to examining accuracies, we will look at the intersection of a list of words that each model deems most important to determine if an article is misinformative, this will help figure out which topics have common indicators of misinformation. After our models make a prediction on a random genre article, we want to examine differences of misinformation across different genres of news.

4. Describe your models and techniques

We use natural language processing, NLP, to train our models from the texts of articles. We tested out multiple NLP techniques. One technique we used was a bag of words' n-grams. We ended up using a NLP's technique called term frequency and inverse document frequency, TF-IDF, to score the words from articles for the most important words of each topic we were testing. We input the scores from TF-IDF through our models for prediction. As our final result, we get an accuracy, precision, and recall score to determine how well our models predicted articles as misinformative or not.

The following models are the models we used in our replication project in the previous quarter. We use these models since we are familiar with them. Some models have been removed, that we feel are not as useful for our task.

The Decision Tree model utilizes the structure of a tree to classify data. It has branches and leaves which are the classified data path. The Decision Tree model makes a prediction based on the learnings of the decision rules from resulting features of data. We use sklearn for our Decision Tree classifier since it has the option to set the max depth. Having this option allows us to shorten the time for processing this model.

Binary Logistic Regression utilizes linear regression function which is modified to scale any data a value in between 0 and 1. The value assigned is the probability of the prediction belonging to class 1 or 0. We use sklearn implementation of Logistic Regression since linear regression is regularized to prevent overfitting.

The Random Forest Classifier is an estimator. The classifier fits multiple decision trees on smaller sub-samples of the dataset to get a different approach compared to a regular decision tree. Additionally, the Random Forest Classifier averages result to control overfitting and improve the accuracy of predictions.

A Support Vector Machine (SVM) searches a hyperplane in N-dimensional space to classify particular data points from a dataset. The SVM has updatable gradients for the weights when classifying data points. We use sklearn's SVM since it is regularized to prevent overfitting.

5. Literature

For some parts of our project, we relied on some formatting and testing of a covid misinformation report by Sajad Dadgar, titled A COVID-19 misinformation detection system on twitter using network & content mining perspective. We utilized some preprocessing utilities as well as what models to focus on for our news misinformation detection. In addition, we implemented his grid search method for finding optimal parameters for the appropriate models. Finally, we looked at his testing methods and how to display results for finding what model to use. While these were first geared towards twitter posts with mainly lower amounts of text/characters for analysis, we found that it could help with denser articles that contain more details on the subject at hand.

6. Exploratory Data Analysis

Since our goal was to see if there were themes to the misinformation of each topic, we looked at what words would be the most important in deciding between informative and misinformative. We used a Word Cloud figure to visualize the important words of informative and misinformative articles across each topic of study. For consistency, we used logistic regression as our base model for determining these clouds.

Figure 1: Word Cloud of Most Important Words in Informative Science Articles

Figure 1 is the word cloud of most important words for informative science articles. The most interesting words of informative science's word cloud are Lexus, chip, honey, and Ukraine. This result is interesting because it shows that informative science articles mainly focus on the subject of the article.

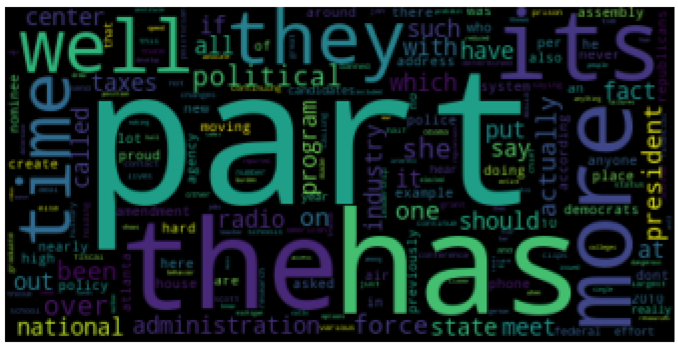

Figure 2: Word Cloud of Most Important Words in Misinformative Science Articles

Figure 2 is the word cloud of most important words for misinformative science articles. The most interesting words that are visible by this word cloud are part, time, well, and more. These words are interesting because they focus on the descriptions to the subject compared to the focus on the subject of the informative science articles.

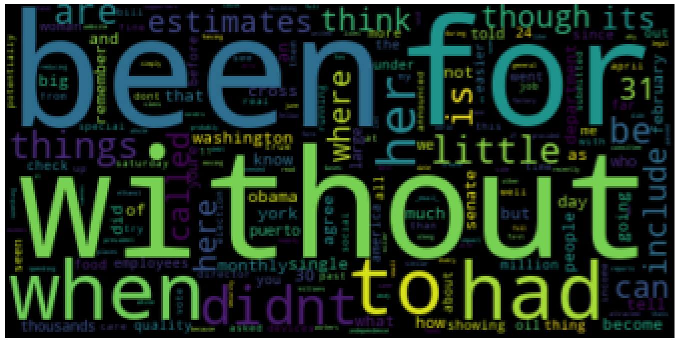

Figure 3: Word Cloud of Most Important Words in Informative Economic Articles

Figure 3 is the word cloud of most important words for informative economic articles. The most interesting words of informative economy's word cloud are been, when, had, and without. This result is different compared to informative science articles where its main focus was the subject of the article. Economic informative article focuses more on occasions.

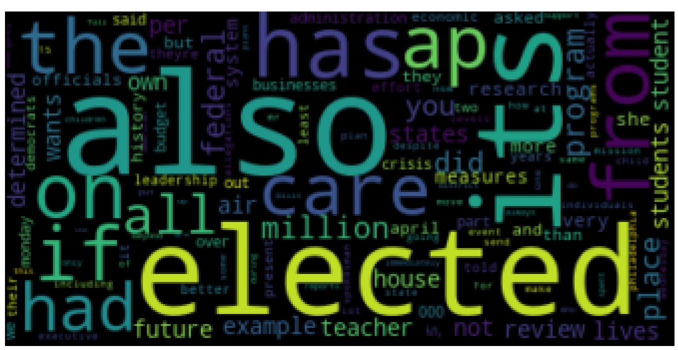

Figure 4: Word Cloud of Most Important Words in Misinformative Economic Articles

Figure 4 is the word cloud of most important words for misinformative economic articles. The most interesting words of misinformative economy's word cloud are care, from, elected, and on. This result focuses a lot more on actions rather than the subject of the article.

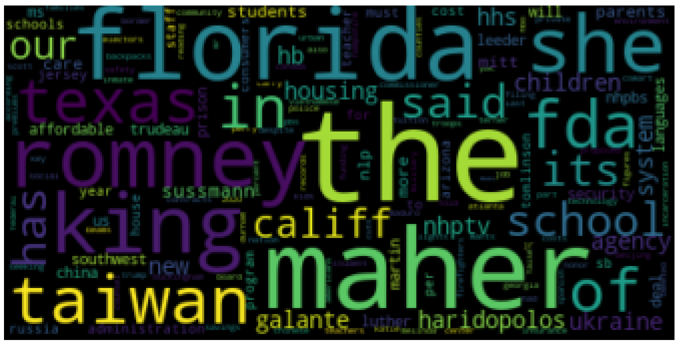

Figure 5: Word Cloud of Most Important Words in Informative Political Articles

Figure 5 is the word cloud of most important words for informative political articles. The most interesting words of informative politics's word cloud are Florida, Maher, Romney, and King. This result is a similar case to informative science article's results. There is more focus on the subject of the article.

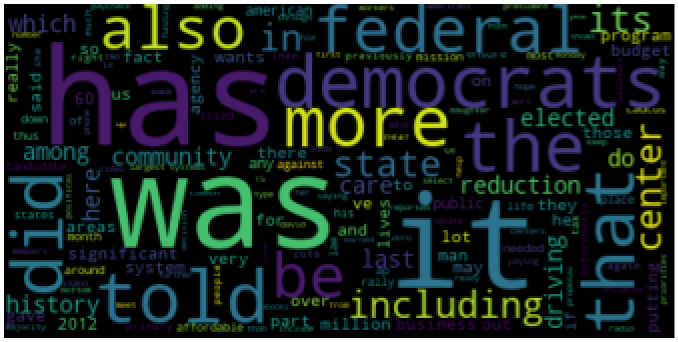

Figure 6: Word Cloud of Most Important Words in Misinformative Political Articles

Figure 6 is the word cloud of most important words for misinformative political articles. The most interesting words of misinformative politics's word cloud are has, was, more, told, and did. This result is similar to misinformative economic article's results. There is more focus on the action rather than the subject of the article.

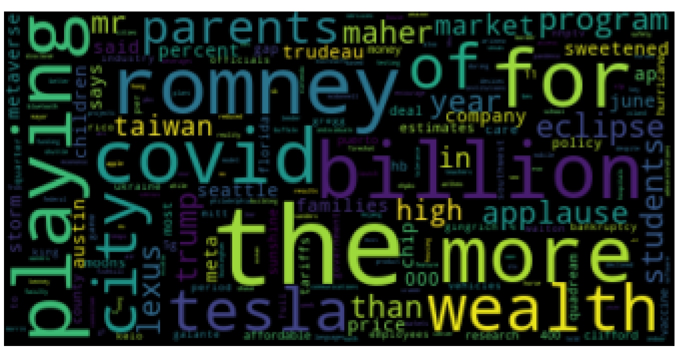

Figure 7: Word Cloud of Most Important Words in Informative Topics Combined Articles

Figure 7 is the word cloud of most important words for informative articles with the topics science, economy, and politics all combined. The most interesting words for this word cloud are Covid, Romney, billion, city, and wealth. The result is interesting since there is a mixture of focus on the subject and description of the subject.

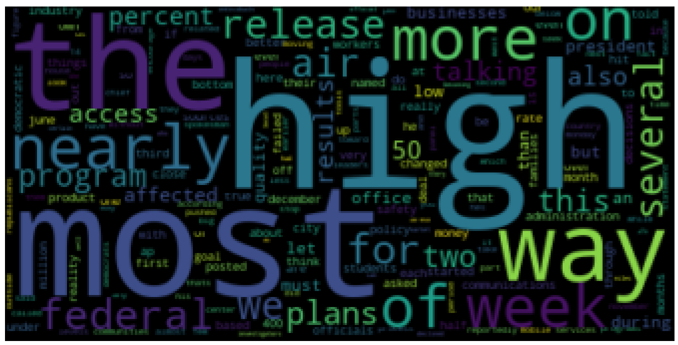

Figure 8: Word Cloud of Most Important Words in Misinformative Topics Combined Articles

Figure 8 is the word cloud of most important words for misinformative articles with the topics science, economy, and politics all combined. The most interesting words for this word cloud are high, most, nearly, several, and more. This result is interesting because these descriptive words are mostly used to describe more of something. Based on the nature of misinformation, it could be hypothesized that misinformative articles use these kinds of descriptive words to exaggerate the subject of its article.

7. Results and conclusions

Model	**Accuracy %**	**Precision %**	**Recall %**
Logistic Regression	**61.6**	**84.6**	**29.8**
SVM	**61.6**	**76.5**	**35.1**

Decision Tree	**53.4**	**54.3**	**51.4**
Random Forest	**50.7**	**52.0**	**35.1**

Figure 9: Evaluation Results For “General” Models

Model	**Accuracy %**	**Precision %**	**Recall %**
Logistic Regression	**56.5**	**100**	**10.0**
SVM	**56.5**	**66.7**	**18.1**
Decision Tree	**73.9**	**69.2**	**81.8**
Random Forest	**69.6**	**75.0**	**54.5**

Figure:10 Evaluation Results for “Science” Models

Model	**Accuracy %**	**Precision %**	**Recall %**
Logistic Regression	**60.0**	**75.0**	**42.9**
SVM	**56.0**	**61.5**	**57.1**
Decision Tree	**52.0**	**55.5**	**71.4**
Random Forest	**60.0**	**83.3**	**35.7**

Figure 11: Evaluation Results for “Politics” Models

Model	**Accuracy %**	**Precision %**	**Recall %**
Logistic Regression	**56.0**	**47.4**	**90.0**
SVM	**56.0**	**47.4**	**90.0**
Decision Tree	**60.0**	**50.0**	**50.0**
Random Forest	**48.0**	**42.1**	**80.0**

Figure 12: Evaluation Results For “Economics” Models

When examining the accuracies of the models, we decided to show both ends of performance for each topic, the best and the worst. The best performing model for the general classifier was SVM with an APR[1] score of: 61.4%, 76.5%, and 35.1% respectively. The worst

[1] APR stands for Accuracy Precision and Recall, these are the scores of which the model is evaluated by in the project

model was Random Forest with an APR score of 50.7%, 52%, and 35.1%. The best performing model for Science was Decision Tree with an APR score of 73.9%, 69.2%, and 81.8%. The worst performing model was Logistic Regression with an APR score of 56.5%, 1.0, and 9.1%. For politics our best model was Random Forest with an APR score of 60%, 83.3%, and 35.7%. The worst model was Decision Tree with an APR score of 52%, 55.5%, and 71.4%. For economics our best model was Decision tree with an APR score of 60%, 50%, and 50%. The worst model was Random Forest with an APR score of 48%, 42%, and 80%.

General	Politics	Science	Economics
D.T./D.T. Politics (53.8%)	D.T./D.T. General (53.8%)	D.T./D.T General(37.8%)	D.T./D.T. General(50.8%)
D.T./D.T. Economics (50.8%)	D.T./D.T. Economics (41.8%)	D.T/D.T Politics(37.4%)	D.T./D.T. Politics (41.8%)
L.R/L.R Economics (38.6%)	D.T./D.T Science (37.4%)	D.T/D.T Economics(31.6%)	L.R/L.R General(38.6%)
D.T./D.T. Science (37.8%)	SVM/SVM General(33.2%)	L.R./L.R General(31.4%)	SVM/L.R. General (37.8%)
L.R/SVM Economics (37.8%)	SVM/L.R. General(33.2%)	L.R./SVM General(31%)	L.R./SVM General(37.4%)

Figure 13: Table of Top 5 Intersections by Topics and models

Next we wanted to examine the overlap of sets of important words between models. We did this by creating a set of the keys returned by our models as being the 500 most important words in determining if an article is misinformative or not. While turning these keys and coefficients into simple sets of keys removes some of the magnitude of these words, it still lets us examine which articles have similar “queues” as to whether they are misinformative or not. Figure 13 shows the intersections. Expectantly, we found the general models had the most overlap. What was interesting was that Decision Tree models maintained very similar word coefficients across topics. For specific topics, it seems that Politics and Economics have higher intersection rates with Science being lower overall. This is likely due to the unique wording of science documents that might make it easy to mislead the reader, whereas economic and political articles will use well known, common words to mislead the reader.

With our results, we can not make a definite conclusion. There were limitations to the project that we could not handle. The major limitation for our project is the nature of human language. There are many connotations and hidden meanings behind sentences in the human language that computers have a hard time processing. It is difficult for the computer to process the human language that is constantly evolving everyday. This limitation makes it difficult to get an accurate score for perfect predictions of articles being misinformative. But, we found a good direction for the project to go off of. Our process and methods led us to results that are satisfactory despite the low scores. Some good ideas for similar future projects are usage of deep learning models, usage of structure of text instead of words, and more data for computation. Hopefully, our research will help similar future projects improve the predictions of misinformative articles.

Works Cited

Dadgar, Sajad. "Sajaddadgar/A-Covid-19-Misinformation-Detection-System-on-Twitter-Using-Network-Content-Mining-Perspective." *GitHub*, https://github.com/sajaddadgar/A-COVID-19-misinformation-detection-system-on-Twitter-using-network-content-mining-perspective.