**Group 3, Text Data Mining of Fake Tweets**

Hasani Jaali, Ada Lazuli, Chukuma Ijioma, Winfrey Johnson

University Maryland Global Campus

Data 630 9041: Machine Learning (2225)

Deanne Larson

August 9 2022

**Introduction**

**Hasani Will expand this section**

Group 3 will mine text data of fake COVID 19 tweets. The team will assume that a fake tweet contains information that the originator of the tweet knows is false and that the tweet is meant to persuade, instill fear, or otherwise disquiet the general population. These tweets are considered *disinformation.* The team will investigate how the proliferation of fake tweets work against our leaders, institutions, and populations, and, as such, should be actively tracked, monitored, and, when possible, contradicted. An example, a fake tweet advocated Ivermectin, a horse drug, as a cure for COVID 19. This widely distributed disinformation caused many people to believe it true, causing a run of the drug at pharmacies, and prompting the Center for Disease Control (CDC) to issue a health advisory (CDC, 2021).

A fake tweet about Ivermectin is only one of many instances of disinformation. Group 3 will model and text mine the tweets with the goal of providing our government, community, corporate, and health leaders with text mined information to combat the negative effect of these tweets on the population.

**Data Preparation & Transformation**

John & Chukuma

Group 3 will use dataset *COVID19 Fake News Dataset NLP* (Kaggle), a data set with 3 variables: id, tweet, label (values = fake or real). Note, each tweet is labeled fake or real. The task here will be to explore and mine the tweets for insights into the composition and trend of these tweets. Below is a sample of the dataset.

Table

Description automatically generated

**Data Preparation**

After examining the structure of the Tweet dataset, the *tweet* field that was of factor type was converted to character data. Since the *id* variable had no meaning for the analysis, that column was removed. Since it is free text, the tweet variables were altered further by converting the characters to lower case, removing English stop words, deleting non-alphabetic characters, and shortening all URLs to simply https or http. This URL shortening was performed as the researchers thought that the fact that URLs were including could be important for the modeling and analysis being performed.

**Exploratory Data Analysis Findings**

**Hasani & Ada**

Exploratory data analysis was conducted in a few phases. mainly focused on word count, sentiment analysis, and word clouds. After initial cleaning It was discovered that both the training and validation files maintained a label split of 47% fake tweets and 52% real tweets, so no imbalances were detected. The data was characterized in several way, Figure – 1 s shows fake tweets tended to have lower character lengths while real tweets tended to have longer character lengths. figure-2 shows that real tweets were found to be twice as likely to have URLs within the tweet text.

**Sentiment Analysis – individual words**

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated | Chart  Description automatically generated |
| Figure -1 Tweet Character length | Figure -2 URL Distribution |

The team explored sever sentiment analysis methods such as: 1) Bing, 2) Afinn, 3) Loughran, 4) NRC. All of the dictionaries appear to indicate that fake tweets are significantly more negative than the real tweets. Figure 3, shows NRC and Loughran dictionaries provided additional sentiment labels that lead to interesting observations. Fake tweets tended to have less anticipation and more anger than real tweets.

Chart, bar chart

Description automatically generated

**Sentiment Analysis – Tweet**

**Hasani**

**Modeling/Analysis**

**Text Mining Technique**

**This section needs to be rewritten and expanded at a high level to introduce the below sections.**

The team will continue with models such as 1) Binary Tree 2) Neural Network, and 3) Naïve Baise. The goal will be to find alternate models with independent variables selected based on text characterization findings. An example, independent variables based on top 10 frequent words, based on several sentiment analysis results, and select words. Also, the team will continue with text mining efforts namely 1) Clustering real & fake tweets, and 2) creating a network of terms.

**TDA**

**Ada**

**Social Network Text Mining**

**Hasani**

**Tree**

**Ada place your stuff here**

**Cluster**

**John place your stuff here.**

**Naive Bais**

**Chukuma place your stuff here**

**Results**

Summary of what we have learned – will be written after we compile the rest of the document.

**Conclusion**

**References**

**Example Reference style -------------------------------------------------------**

Center for Disease Control, 2021. *Rapid Increase in Ivermectin Prescriptions and Reports of Severe Illness Associated with Use of Products Containing Ivermectin to Prevent or Treat*

https://emergency.cdc.gov/han/2021/han00449.asp

Kaggle, 2020. COVID19 Fake News Dataset NLP

https://www.kaggle.com/datasets/elvinagammed/covid19-fake-news-dataset-nlp