Kathryn Egan & James Samford

Syracuse university

December 10, 2020

Classification of Anti-Vaccine Articles

ist 664, Natural Language Processing

Table of Contents

Introduction 2

Research Question 2

NLP Investigation:

Data Preparation 2

Data Pre-Processing 3

Analysis at the Semantic Level 4

Analysis at the Syntactic Level 5

Analysis at the Lexical Level 6

Watson Analysis 8

Classification Experiment 9

Discussion and Future Applications 11

References 12

Appendix of Code 13

**Introduction**

In a year marred by easily shared misinformation regarding COVID-19 and the vaccine that is expected to usher our exodus from the pandemic, it is more imperative than ever before to quickly identify and label misinformation before it has time to take hold and be mistaken as truth. Much of the misinformation shared regarding a COVID- 19 vaccine contains rhetoric from the Anti-Vaccine movement that has been gaining ground in the last few decades. The origin of the newest movement is not surprising – misinformation passed off as scientific research. In 1998 a researcher in the UK published findings that linked a common childhood vaccine with rising diagnoses of Autism in children. Though the author later confessed to falsifying his findings, the damage had been done (Omer, 2020).

His blatant lie, published as truth, grew its own legs and has run wild in the minds of fearful parents and conspiratorialists. Today, views regarding vaccine safety and legitimacy is polarizing; it comes down to AntiVaxxers vs. ProVaxxers (Northwestern University, 2020).

While there are misinformation filters being used by social media platforms like Twitter and Facebook to notify people that what they are reading has been flagged as False, the vast majority of articles, blog posts, and YouTube transcripts are readily available outside of social media’s domain(Shu & Shieber, 2020). For the people who make it beyond the reach of these misinformation filters online, how do they know that what they are reading is truth? Is there blind faith that if the author of an article refers to themselves as a doctor, it must be scientifically sound information that can be trusted? Perhaps there is a sentiment that pulls the reader in and builds their trust? Before training models for predicting if an article is Anti-Vaccine or not, a thorough analysis of the similarities and differences between Anti-Vaccine and Pro-Vaccine articles at various levels of the Synchronic Model of Linguistics, needs to be performed.

**Research Question**

Given how polarizing stances on vaccines are in our society, is polarity analysis of vaccine articles the strongest indicator of an article’s vaccine position? Is there a compelling enough partisanship between the two vaccine camps to make subjectivity analysis a greater predictor than polarity analysis? If, these do not appear to be strong predictors, will machine learning models and deep learning models have more success in their predictions?

**Data Preparation**

All articles were taken from websites and compiled in a csv file. Given the time constraints of the course, the corpus of Anti-Vaccine and Pro-Vaccine texts was limited, which only allows for a peak at sentiment differences between texts, but a much larger corpus would provide a stronger statistical analysis of the sentiments.

Five articles were converted to one data frame, using pandas, for Anti-vaccine texts. Then the same steps were taken for Pro-vaccine texts.

Data utilized in the tensorflow experiment was derived both from previously utilized datasets from prior experiments and additional research in order to create balanced models. This data was broken into raw text files stored on disk with a rating of either 0 for pro-vaccine and medical articles, or 1 for antivaccine articles and blog posts.

**Data Preprocessing**

To begin preprocessing the texts, a function was defined to clean the data by converting all text to lower-case, removing special characters and numbers. Next, English stop words were removed using NLTK’s stop words list. The cleaned text was then tokenized to words, followed by lemmatized to their roots, and finally tagged for part of speech. The output of each preprocessing step was saved in the data frame as new columns for ease of access for later analysis.

First rows of the cleaned Anti-Vaccine data frame:

Table

Description automatically generated

First rows of the cleaned Pro-Vaccine data frame:

Graphical user interface

Description automatically generated

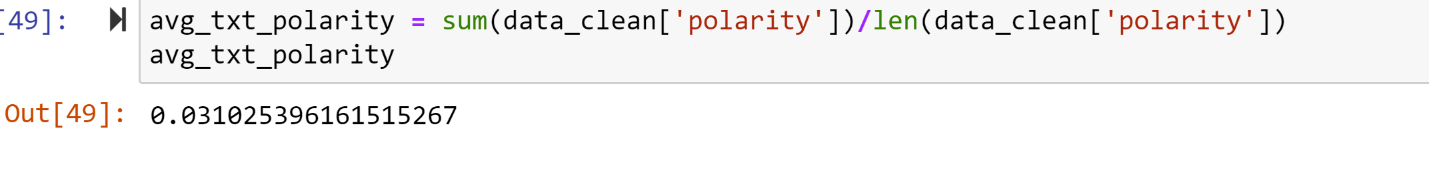
Data for Tensorflow was processed using the Keras library, which was configured to vectorize entire documents for analysis, with any unusual formatting and all punctuation removed, as well as any remnant html excluded.

**Semantic Analysis: Polarity & Subjectivity of Sentences in the Text**

The cleaned sentences were analyzed for polarity and subjectivity scores using the TextBlob library in Python. The scores were added up and then divided by their length to get the average scores for the sentences in each set.

Polarity is a measure of how positive or negative a sentiment is. The polarity scale runs from -1 being most negative to 1, most positive. As the polarity scores demonstrate below, the Anti-Vaccine sentences were more negative in sentiment than Pro-Vaccine sentences. Surprisingly, the difference is not very profound, but this is most likely due to the small corpus and with more documents added to the corpus, a stronger indication of the polarity differences should be evident

Anti-Vaccine Average Polarity:



Pro-Vaccine Average Polarity:



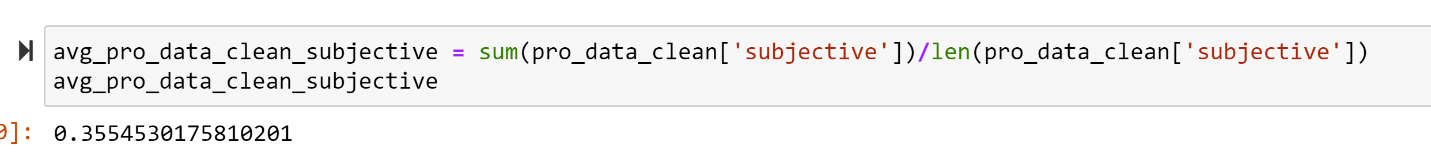
Subjectivity is a measure of prejudice, or lack of objectivity. The Subjectivity scale runs from most objective being 0 and most subjective being 1. There is a minute difference between the subjectivity of sentences from the two stances. Again, with more documents in the corpus, this difference may grow and become more obvious.

Anti-Vaccine Average Subjectivity:

A picture containing graphical user interface

Description automatically generated

Pro-Vaccine Average Subjectivity:



**Syntactic Analysis: Sentiment Analysis of N-Grams**

The cleaned sentences were next broken down to Bigrams and Trigrams. A matrix with frequency of each n-gram was made. Next, the same polarity and subjectivity analysis was done on the bigrams and trigrams as was on the sentences in the previous part.

Anti-Vaccine Most Frequent N-Grams Pro-Vaccine Most Frequent N-Grams

Table

Description automatically generatedTable

Description automatically generated

Looking at polarity of two 2 corpuses at the syntactic level appears to show a much more significant difference in polarity than it showed at the semantic level. Here Anti-Vaccine word combinations are clearly far more negative than Pro-Vaccine word combinations. This is not to imply that the Pro-Vaccine texts are positive, as they are also closer to -1 than 1 on the scale.

Average Anti-Vaccine N-Gram Polarity:

Graphical user interface

Description automatically generated

Average Pro-Vaccine N-Gram Polarity:

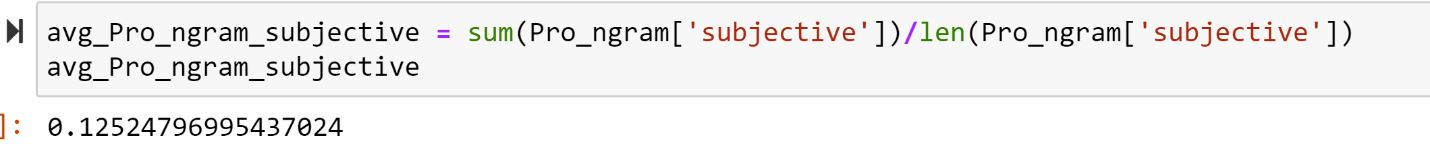


Just as with subjectivity analysis of the sentences in the corpuses, the results of the n-gram subjectivity analysis is not significantly different between the stances. They are both more objective than they are subjective.

Average Anti-Vaccine N-Gram Subjectivity: Graphical user interface, text, application, Word

Description automatically generated

Average Pro-Vaccine N-Gram Subjectivity:



**Lexical Analysis: Frequency Distribution of Words**

While examining word frequencies does not give insight into the importance of the words used in a text, it is a useful tool when comparing documents. For this part of the analysis, the lemmatized words created during the preprocessing phase were used. The NLTK FreqDist library and NLTK Brown Corpus were employed. The most common words in our corpuses were compared against the Brown corpus of most frequent words and the top 100 most common words that matched between our corpus and Brown corpus were removed, like stop words which would create a lot of noise in the data. Next, the frequency distribution was calculated, and the following lists were produced:

Anti-Vaccine Word Frequencies Pro-Vaccine Word Frequencies

Text

Description automatically generatedText

Description automatically generated

To visualize the difference in word choice and frequency, the WordCloud library was imported to create a word cloud for each corpus. The larger the word in the illustration, the more frequently it shows up in the corpus.

Anti-Vaccine Words

A picture containing text

Description automatically generated

Pro-Vaccine Words

A picture containing text

Description automatically generated

In the Anti-Vaccine word cloud, the word *supposed* jumps out as a more frequently occurring than. The word *supposed* has a connotation of distrust. Compare that to frequent Pro-Vaccine words like *tested, licensed*, and *research* which have a more trusting connotation. Further investigation into identifying distrust and trust words and then using those to build a feature for analysis. Perhaps a measure of trust will be a robust tool for classifying vaccine stance.

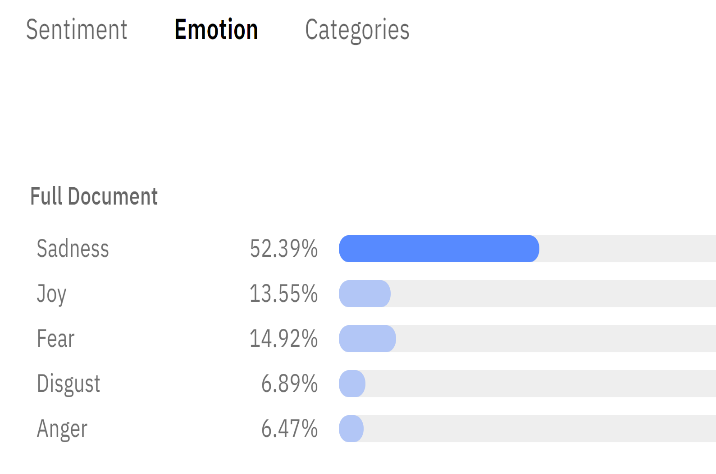
**Watson Analysis: Technology Exploration**

The IBM Watson Natural Language Understanding program is a useful tool for measuring emotion in texts. Each corpus was put into the IBM Watson NLU Demo for analysis of emotion and to compare our sentiment findings to the well-established algorithms of IBM Watson.

At first glance, seeing sadness as the emotion most found in the corpuses, is understandable since much of the discussions discuss death and epidemics. However, a deeper look at the words that IBM Watson tagged with the different emotions, it is extremely evident that this is not a useful tool for analyzing attitudes towards vaccines. For example, joy was ranked highly in the Anti-Vaccine corpus analysis, but it turns out that Watson considers vaccines to be a positive. If the analysis did a better job of identifying the emotion and sentiment of the sentences that the word *vaccine* was in, it would not tag them all as joy. Here is the most glaring example of the mislabeling: *Vaccines are NOT safe*. This is the very first sentence in the Anti-Vaccine corpus and it was wrongly labeled ‘joy’.

Anti-Vaccine Emotion Pro-Vaccine Emotion

Chart

Description automatically generated

Where IBM Watson’s assessment did not fail, was its classification of sentiment polarity. Just as our earlier analysis generated, both Anti-Vaccine and Pro-Vaccine corpuses were negative, with the Pro-Vaccine score a little less negative than Anti-Vaccine.

Anti-Vaccine Polarity Pro-Vaccine Polarity

A picture containing timeline

Description automatically generatedA picture containing chart

Description automatically generated

**Classification Experiment**

After analyzing both vaccine corpuses at different levels of the Synchronic Model of Language, machine learning and deep learning models were built to predict whether a given article is Anti-Vaccine or Pro-Vaccine.

The data used for training and testing the models was the original csv file that all articles were organized after being scrapped from various websites. Each article was labeled pro or anti, depending on the obvious stance after reading the article. The file contained 9 pro and 9 anti articles. After reading the file into python, the text was cleaned using the same code as discussed in the preprocessing phase of the previous corpuses. Once the data was cleaned, the sklearn library was imported. Next the data was separated into a training and a testing set following the 80/20 rule. All text was then transformed to numerical data using Bag of Words.

The first model used for prediction was Multinomial Naïve Bayes. This model classifies by calculating the previous probability of each label, in this case pro and anti, based on how frequent the label occurs in the training set. After using our training data set to train and fit the model, the test data was inputted into the model and the following classification report was generated:

Table

Description automatically generated

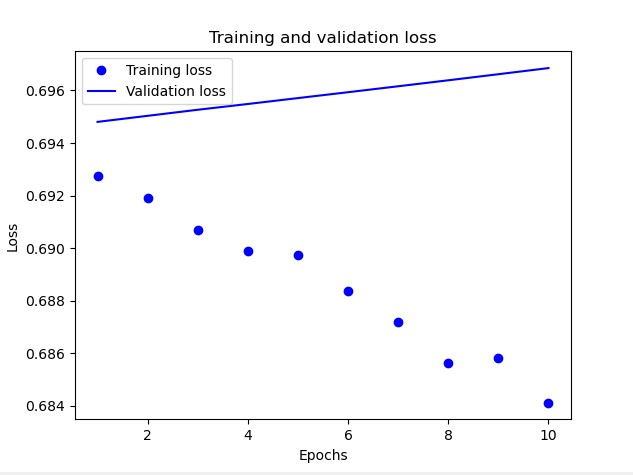
Focusing on the f1-score, which is a measure of accuracy with 1 indicating the highest accuracy, this model’s accuracy is low.

The second classification model used was Random Forest. Simply put, Random Forests classify by creating lots of decision trees with the training data. The steps for preparing data for Random Forest classification is the same as for Naïve Bayes. When training the model, one of the parameters (n\_estimators) had to be tuned multiple times due to overfitting. It is recommended to set the parameter between 500-1000, but this is for larger data sets than being used. Setting n\_estimators = 250 produced the best outcomes, with a f1-score of 0.8.

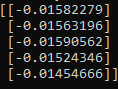
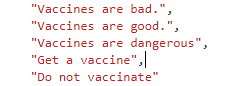
Table

Description automatically generated

Because Deep Learning is where Natural Language Processing is headed, the final model used to predict whether an article is anti-vaccine or pro-vaccine in this study is a Recurrent Neural Network. RNN works by processing sequences of patterns in the training data. Tensor Flow was used to build the RNN. The accuracy report generated by the neural network shows that the more cycles (epochs) that the algorithm ran, the lower the training loss and the higher the validation loss. Note that each iteration delivers different results, as the training model utilized is configured to randomly sample only part of the data. This was done specifically with the intent that larger datasets will become available and compute resources would be limited, or that the datasets would eventually become unbalanced, with antivax-negative texts vastly outweighing antivax positive datasets.



The tensorflow approach has some limitations in this experiment, largely due to the lack of data available for training – the model used would be highly accurate with a library of a hundred or so positive and negative articles, and could be even more accurate with articles being rated on a scale. However, due to the scope and time constraints present, this was not possible. Despite this, the model was able to report an accuracy of about 60%, which is not bad considering that it had only 16 examples to work with. Where the model falls short is with short, test texts, which it will rate close to negative regardless of input. The figures below show each phrase and their rating. Ratings close to 0 would be considered pro vaccine sentiment, where 1 is antivaccine.



**Future Development and Application**

Another evaluation of polarity and subjectivity should be performed on a much larger data set. There is just enough evidence here to suggest polarity is a better indicator of vaccine stance than subjectivity. Also, a deeper look at lexicons that show trust or distrust and their frequencies based on vaccine stance needs to happen. The best indicator of whether an article is anti-vaccine or pro-vaccine will most likely be found through a combination score between polarity and trust word frequencies. This combination should then be used to build a neural network for increased classification accuracy.

Creating an accurate classifier of anti-vaccine articles has a bigger potential application in identifying other misinformation laden or conspiratorial texts passed off as fact.

# **References**

Catherine Shu, J. S. (2020, March 16). *Facebook, Reddit, Google, LinkedIn, Microsoft, Twitter and YouTube issue joint statement on misinformation*. Retrieved from techcrunch.com: https://techcrunch.com/2020/03/16/facebook-reddit-google-linkedin-microsoft-twitter-and-youtube-issue-joint-statement-on-misinformation/

Northwestern University. (2020, December 15). *News Release: Vaccine hesitancy is fueled by polarization and mistrust in science, experts say*. Retrieved from news.northwestern.edu: https://news.northwestern.edu/stories/2020/12/vaccine-hesitancy-is-fueled-by-polarization-and-mistrust-in-science-experts-say/&fj=1

Omer, S. B. (2020, October 27). *The discredited doctor hailed by the anti-vaccine movement: Riveting biography of Andrew Wakefield is a cautionary lesson in the legacy of hubris.* Retrieved from Nature.com: https://www.nature.com/articles/d41586-020-02989-9