Countering Malicious Cyber Actors
A project for Perspecta by Team VeritaZ

Problem:
Foreign actors have been and will interfere with our domestic news in an attempt to shape public opinion

Solution:
Combine various supervised ML techniques and language processing tools to identify these malicious actors from a range of normal reporting.

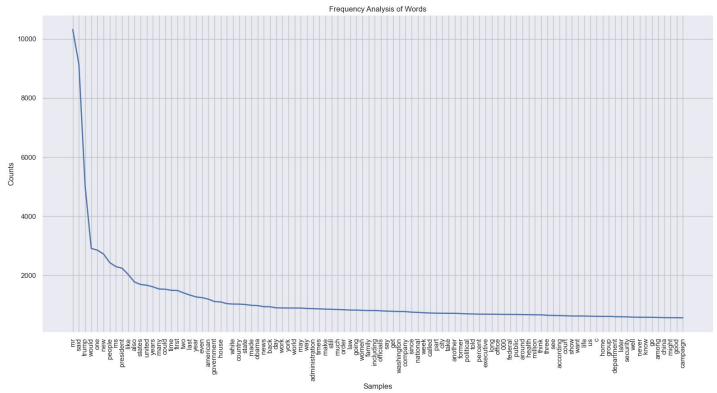
The details:
Our program utilizes a number of ML-centric python packages to process and analyze structured datasets given to the program. This made construction of the program much simpler than other languages and adds a level of user-friendliness for real-world deployment, scanning news and articles.

Methods:

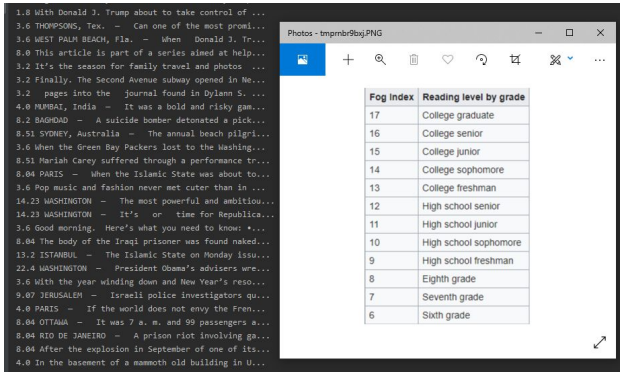
- In our program we utilized natural language processing to lemmatize and segment words as well as locate parts of speech and detect the topic of a given article. Additionally, our program uses frequency analysis to determine what words are used the most in the article to find a pattern of association within the text.
- Reading scales were used in tandem with lemmatization and frequency analysis to determine the necessary reading level of each article. In this case, a lower reading level would indicate a wider audience. Our hypothesis was that if an article was an outlier, being much harder or easier to read, then it was more likely the article is fishing for clicks or searching for attention.
- Semantic analysis was also used to determine the emotion of a given text. If this was against the norm, then we could identify that as an outlier misinforming the public or representing a smaller audience.

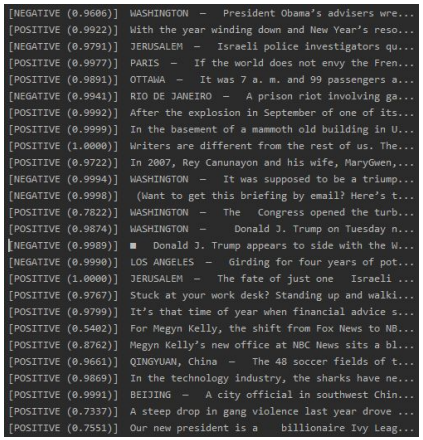
Results:
Our program returns the readability scores, the frequency analysis of a set of articles, giving us reliable indicators as to whether or not the article is real or purposeful misinformation.

Word Frequency in Dataset



Gunning Fog Reading Scores



Fog Index	Reading level by grade
17	College graduate
16	College senior
15	College junior
14	College sophomore
13	College freshman
12	High school senior
11	High school junior
10	High school sophomore
9	High school freshman
8	Eighth grade
7	Seventh grade
6	Sixth grade

Sentiment Analysis of Dataset