**Applying NLP Techniques to Search for Patterns in Troll Tweets**

**James**

**Abstract:**

Foreign governments are now able to interfere in U.S. elections by making posts on social media that appear to be from an American, when in reality, they are not. The goal of this project was to determine whether or not the detection of “troll tweets” could be automated by applying natural language processing methods such as sentiment analysis to them. The content of the tweets themselves did not provide what would be needed to automate the process of detecting “troll tweets” because of the neutrality produced by sentiment analysis.

**Introduction:**

In 2016, an oligarch-run propaganda factory out of Russia called the Internet Research Agency sent out approximately 3,000,000 tweets in an attempt to influence the U.S. Presidential Election. The effects on the native U.S. population are hard to measure, however, the danger that foreign propaganda poses to the nation as a whole cannot and should not be understated. This project will involve applying natural language processing techniques in order to try to gleam some information from the corpus of the three million tweets.

The dataset has undoubtedly already been examined by others, however, there is no such thing as “too many eyes” when it comes to identifying foreign political attacks. What makes this project worthwhile, then is the varied application of natural language processing techniques such as sentiment analysis which is typically only applied to user reviews. For example, if it was found by a binary classifier that the vast majority of the tweets had negative sentiment, that would be another piece of information that could assist in spotting future attacks.

The aim of this project is specifically to determine whether or not patterns can be found amongst the many tweets that form the dataset that could be helpful in detecting other propaganda attacks in the future. This type of attack is likely to keep happening as long as foreign countries have a vested interest in U.S. politics, and being able to identify it sooner rather than later is quite important. Much debate has taken place regarding the 2016 election, but regardless of someone’s opinions on the matter, surely it is unhelpful to have foreign agents worsening matters.

Throughout the course of the project, small subsets (10,000 sample size out of ~200,000) of the dataset will be preprocessed with spaCy and pandas. SpaCy will be providing the processing pipeline used for this project, because the tweets are all in English, and it has the language-specific rules needed for the preprocessing tasks. Then, sentiment analysis will be done in order to determine what the average polarity/sentiment of the tweets is, as well as how subjective they are.

**Methodology:**

**Pipeline Diagram:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Raw Data** | **Tokenizer** | **Tagger** | **Parser** | **NER** | **SpacyTextBlob** | **Data Collection/Comparison** |

Pipeline:

This project will use spaCy’s medium pipeline. The tokenizer pipe segments the text of the subset into tokens which are needed for further preprocessing. The tagger pipe is capable of assigning parts-of-speech tags to the tokens in a subsets. The parser pipe is necessary as part of the default pipeline from spacy, as it allows everything to run more smoothly down the line. SpacyTextBlob is a pipeline component that enables sentiment analysis by using the TextBlob library. Data will be collected from each run and then collected, to be examined at the end for purposes of gleaming as much information as possible.

In order to apply SpaCy’s powerful tools to it, the subset(s) will first be cleansed of the web addresses that are present in many of the tweets. Regex will be used in this case, proving to be a helpful tool. Text will be restricted to letters, as it is necessary to remove the strange characters like emojis from the tweets. Once the data is ready for entry into the spaCy pipeline, the tokenization can begin. Tokenization enables the stop word removal process, which then passes to the tagger pipe. The tagger pipe will then tag the parts of speech within each sample of the subset(s).

SpacyTextBlob is added as a pipe to the default “en\_core\_web\_md” pipeline of spaCy, and Once the data travels all the way through the pipeline, the results can be compared with a few other similar projects. Tweets present a convenient sample size due to the restriction of 280 characters per tweet.

**Experimental Setup:**

1. Implementation

Python is the programming language of choice for this project. Pandas and the spaCy libraries/tools will be used together in order to accomplish the goals of the project. Pandas provides data analysis and manipulation tools for spaCy. spaCy provides a full suite of tools throughout the pipeline of this project. The project will be carried out on a personal computer, and the medium sized pipeline “en\_core\_web\_md” will be used, instead of the large “en\_core\_web\_trf” pipeline.

1. Dataset description

The dataset being used for this project contains nearly three million tweets from 2012 – 2018, with the vast majority being sent from 2015 to 2017 around the U.S. Presidential Election. This dataset was created by Clemson University researchers Darren Linvill, and Patrick Warren, who used Social Studio to gather the data through custom searches. The U.S. Justice Department filed an indictment against the Internet Research Agency, a Russian “troll factory” who is known to be responsible for the propaganda tweets levied against the U.S. population.

FiveThirtyEighty, a news outlet, posted the dataset onto GitHub in order to provide the public with the opportunity to examine it. It has been split into nine different CSV files, but it is possible to sift through it as a database via BigQuery thanks to GitHub user ‘elithrar’. The content of the data set is inherently limited to 280 characters per document. Many of the tweets contain URLs (website addresses), and some could be harmful and contain malware, so those links are going to be removed during the preprocessing step of the project.

According to the GitHub page where the dataset is shared:

*“The basis for the Twitter handles included in this data are the November 2017 and June 2018 lists of Internet Research Agency-connected handles that Twitter provided to Congress. This data set contains every tweet sent from each of the 2,752 handles on the November 2017 list since May 10, 2015. For the 946 handles newly added on the June 2018 list, this data contains every tweet since June 19, 2015. (For certain handles, the data extends even earlier than these ranges. Some of the listed handles did not tweet during these ranges.) The researchers believe that this includes the overwhelming majority of these handles’ activity. The researchers also removed 19 handles that remained on the June 2018 list but that they deemed very unlikely to be IRA trolls.*

*In total, the nine CSV files include 2,973,371 tweets from 2,848 Twitter handles. Also, as always, caveat emptor -- in this case, tweet-reader beware: In addition to their own content, some of the tweets contain active links, which may lead to adult content or worse.”*

It should be noted that location data of the tweets is irrelevant, as more than half of them have been IP-spoofed to make them appear to be sent from users in the United States. For the purposes of this project, it will be ignored, as the only items of interest are the text contents of the tweets themselves. A smaller section of the original corpus will be used that contains ~200,000 tweets. A csv file containing the 200,000 tweets has been split into 21 csv files, with 20 of them containing 10,000 tweets and the last containing the remainder. The csv file of 200,000 tweets was created by Sagar Chadha, who posted it on a site called Kaggle.

Each document processed by this project contained many strange characters such as emojis that were removed in order to allow the spaCy pipeline to function. Encoding errors when dealing with .csv files and spaCy took significant time and effort to mitigate, but were eventually overcome. Only the contents of the column “tweet\_text\_only” are wanted, so it is isolated when creating each document. Extra whitespace that was created by the encoding\_errors = ‘replace’ line in the read\_csv argument is removed by the split() function in order to prepare the text for regex. Regex is used in order to ensure that only the letters that are wanted appear in the document.

SpacyTextBlob is called to determine the polarity and subjectivity of each chunk{i}.csv file included in the subset.

**Results:**

Sentiment analysis was performed on seven out of twenty-one subsets, processing and handling 70,000 tweets in total. Polarity is a measure of the overall sentiment, and ranges from -1 to 1. A polarity of -1 is negative, 0 is neutral, and 1 is positive. Subjectivity ranges from 0 to 1, where 0 is not subjective at all, 0.5 is neutral and 1 is extremely subjective.

Graphical user interface, application, table, Excel

Description automatically generated

**Discussion and Conclusion**:

As can be seen from the results of this project, processing troll tweets with sentiment analysis is not a helpful metric with which to identify troll tweets. The sentiment returned from this subset of the corpus is neutral, which is inherently not helpful in looking for outliers such as troll tweets that differ from a norm. The publicly available information on Twitter may not be enough to identify troll tweets if further studies proved them to be neutral in polarity and subjectivity. A much better approach may be possible with large scale similarity comparisons included, but the problem is that it is necessary to have something to compare the data to.

The problem is formulating exactly which tweets are troll tweets and which are not from their text content alone, and the approach fell short of expectations. As the group learned more about the project, they learned that the dataset chosen was not a good fit. Future attempts at solving this problem would certainly require more understanding of natural language processing as a whole.

**References:**

[fivethirtyeight/russian-troll-tweets (github.com)](https://github.com/fivethirtyeight/russian-troll-tweets/)

[Linvill\_Warren\_TrollFactory.pdf (clemson.edu)](http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf)

[elithrar (Matt Silverlock) (github.com)](https://github.com/elithrar)

[Sentiment Analysis with Spacy and Scikit-Learn | Engineering Education (EngEd) Program | Section](https://www.section.io/engineering-education/sentiment-analysis-with-spacy-and-scikit-learn/)

<https://www.kaggle.com/code/chadalee/text-analytics-on-russian-troll-tweets-part-1/notebook>

**Related Works**:

***Paper 1:***

***TITLE:*** “Psychometric profiling of individuals using Twitter profiles: A psychological Natural Language Processing based approach”

***SUMMARY:***

In a response to increased internet usage during the covid-19 pandemic, the researchers responsible for this article attempted to measure the semantics of twitter users based on LIWC and SALLEE scores. LIWC and SALLEE are both essentially emotion indicators which seemingly quantify the emotion imparted through a string of text. As they applied this combo, researchers used an MBTI personality dataset and train classifier to try and predict the personality type of the twitter users in question. Their results were more promising than those of other relevant works which primarily relied on the “Big 5” personality types for classification, as they instead generated a psychometric scoring model using an NLP based approach.

***MOTIVATION:***

The group of researchers aimed to gain an understanding of individual twitter user’s personality over time by examining the semantics used within their tweets. Their hope is to be able to group up attributes of people’s personalities in order to establish the basis for a psychometric credit score in relation to social media presence.

***METHODOLOGY:***

In order to accomplish their goals, researchers studied individual twitter user pages for metrics such as tweets, retweets, replies, etc. The textual content of their tweets was classified in terms of anxiety, certainty, negative/positive emotion, and other emotions such as anger or hate. With those features, a predictive model was developed that they then used to train a classifier with in addition to the twitter provided dataset.

***ALGORITHMS DEVELOPED/USED:***

This article references 6 algorithms that were used. The first algorithm was for data collection, and made API calls to twitter in order to form their custom dataset of tweets. The second algorithm was for preprocessing and removed contents such as emojis and stopwords from the tweets. The third algorithm generates LIWC and SALEE scores for the tweets within their custom dataset. The fourth algorithm generates a predictive regression model for the custom dataset. Algorithm 4.5 generates a predictive classification model for the MBTI (Myer-Briggs) personality dataset. The last Algorithm, number 5, predicts a personality for the tweet(s) in their custom dataset, and finds the correlation between the personality type and linguistic scores.

***DATASET(S) USED***:

A custom dataset was generated by an algorithm that made twitter API calls in order to retrieve tweets from users.

***PERFORMANCE***:

The overall performance of this application of NLP was compared with related projects, and it was found to perform better than the others due to not relying entirely on the big five personality types.

***LIMITATIONS:***

The limitations of this project would primarily be that it relies solely on Twitter to form its corpus. Whether or not adding other platforms such as Facebook to the corpus would help is unknown, however, it is a limitation nonetheless.

***COMPARISON OF METHODOLOGIES***:

The methodology of this project differs from the project of our group in that it considers more content from an individuals twitter page than just the content of their tweets. In addition, spaCy was not used in this project, whereas our project relies heavily on its pipeline.

***LINK TO ARTICLE:*** [***https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.7029?casa\_token=AquK3q\_1Ar4AAAAA:UAmdrPN03JfY-MFo-XG1JYJHX0\_8TjHHe7xDdMajN0fBG41uEmfyP20qwgmXjyIDgnvbmxliQtX9qTRD***](https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.7029?casa_token=AquK3q_1Ar4AAAAA:UAmdrPN03JfY-MFo-XG1JYJHX0_8TjHHe7xDdMajN0fBG41uEmfyP20qwgmXjyIDgnvbmxliQtX9qTRD)

***Paper 2:***

***TITLE:*** “Hate Speech Detection in Twitter using Natural Language Processing”

***SUMMARY:***

Modern Natural Language Processing methods are applied to an imbalanced dataset used by previous researchers to train language models on hate speech detection. High accuracy of more than 95% detection is accomplished by this group of researchers, even in their A model which does not use preprocessed data. Ultimately, the researchers determined that more machine learning techniques could be applied to the problem in the future to explore other solutions to the problem.

***MOTIVATION:***

The motivation of this project is to automate the detection of hate speech online, as manual detection on a platform such as twitter is not possible.

***METHODOLOGY:***

Machine learning is used in binary classification (either a tweet is hate speech, or it is not). Preprocessing including tokenization, removal of stopwords, stemming, and case folding was applied to the dataset in order to prepare it. Feature Extraction is then performed, in order to convert every tweet in the corpus into a set of fixed attributes so that they are interpretable by the ML models.

***ALGORITHMS DEVELOPED/USED:***

Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest algorithms(RF) were used in this project.

***DATASET(S) USED***:

A dataset composed of tweets that was posted on GitHub by researchers from another project was used.

***PERFORMANCE***:

The performance with and without preprocessing was measured for this project. The model was tested in two different sections, A and B, having different parameters. Model A was done without preprocessing, while model B did use preprocessing. The group of researchers concluded that certain measures needed to be taken due to their imbalanced data set.

***LIMITATIONS:***

Due to their specific implementation, this approach can only be used to datasets from twitter. In addition, their data set was imbalanced, which requires implementation of certain methods in order to be handled properly.

***COMPARISON OF METHODOLOGIES***:

Whereas our group uses a corpus composed solely of propaganda/troll tweets from Russian agents around the 2016 U.S. Election, this group of researchers used a corpus of tweets containing a mixture of tweets that contain hate speech, or not. The goals are not identical, but similar in the sense that they are both steps being taken to try and reduce negative interactions on twitter.

***LINK TO ARTICLE:***

<https://ieeexplore.ieee.org/abstract/document/9388496?casa_token=Lt8ES9MFJWQAAAAA:ary3m9VOX9lS-_m_hpySK8HNBtthaRskG5VnoIq_uVokMCaJ3yQmzMiQtib8FgmWw3DyuqyAz30>

***Paper 3:***

***TITLE:*** “Twitter Sentiments Analysis Using Machine Learninig Methods” (sic)

***SUMMARY:***

The textual data from Twitter makes a good target for sentiment analysis due to the limit of 280 characters per tweet. This group of researchers tested three different machine learning methods with sentiment analysis tasks. Out of all three methods, the Naïve Bayes Classifier was found to the most accurate, although the Support Vector Machine showed promise as well.

***MOTIVATION:***

The limited character count of tweets on Twitter serve as good source for sentiment analysis. This group of researchers believes that Twitter represents people of all age groups and demographics as well, so they decided to compare various machine learning methods to see how accurate and precise they are.

***METHODOLOGY:***

The methodology for this project is applying different machine learning methods to analyze the sentiment of tweets. They began with data gathering, where they obtained developer credentials on Twitter in order to access the API. With their textual data gained, the group pre-processed it using python. The latter steps of their project were feature extraction, feature selection, and classification.

***ALGORITHMS DEVELOPED/USED:***

The algorithms they chose are as follows: Naïve Bayes Classifier, Support Vector Machine, and Maximum Entropy Method.

***DATASET(S) USED***:

The data set used for this groups project is a pre-processed Twitter corpus.

***PERFORMANCE***:

The Naïve Bayes Classifier presented the group of researchers with the highest accuracy of sentiment analysis on their created data set. They determined that improvements to their performance could be made with more work in the future.

***LIMITATIONS:***

This group limited the scope of their project to only using textual data from Twitter, as opposed to using other platforms such as Facebook. This is due to the appealing characteristics of tweets, such as their limit of 280 characters.

***COMPARISON OF METHODOLOGIES***:

This group of researchers also chose Twitter as the sole source of textual data for their dataset, although our group did not make our own custom dataset as they did. Our group aims to determine topic classification in addition to performing sentiment analysis, so the motivation of our project differs from theirs as well.

***LINK TO ARTICLE:*** [***https://ieeexplore.ieee.org/abstract/document/9154183***](https://ieeexplore.ieee.org/abstract/document/9154183)