PolitiTweet 2020: A Twitter Data Set Encompassing the 2020 Presidential Election

Eric Cao
University of Illinois at Urbana-Champaign
ejcao2@illinois.edu

Mark Cockburn
University of Illinois at Urbana-Champaign
markc2@illinois.edu

Hunter Dyer
University of Illinois at Urbana-Champaign
hadyer2@illinois.edu

ABSTRACT

In this work, we attempt to develop a misinformation classification data set of text-based social media posts through both an unsupervised labeling and manual labeling processes. We assess our methodology and results for the unsupervised labeling approach and discuss areas for improvement and future work. Additionally, we contribute PolitiTweet 2020, a data set that may aid future research in this space, and others. Our data set was collected amidst notable world events that were a catalyst for the spread of misinformation on social media.

The data set introduced in this work consists of a set of tweets in the time leading up to, and following, the 2020 United States presidential election. In addition, the data collected included information shared at the height of the COVID-19 pandemic. The data set was constructed in a manner similar to Ma et. al [5] in which stories from PolitiFact are used to locate related tweets, with benign tweets mixed in. The content of our data set and the results of manual labeling are presented and discussed.

1 BACKGROUND

Social media platforms' success has established a new ecosystem, allowing the easy creation and sharing of misinformation and other content [7]. As more adults rely on social media for their news [9], the impact of fake news and misinformation grows proportionally. The spread of misinformation seems to become more prevalent during transitions of power, such as during the 2016 and 2020 United States presidential elections. Aside from politically motivated misinformation, we have also seen misinformation come into the public discourse regarding the COVID-19 pandemic. This misinformation includes conspiracy theories about the virus being created as a biological weapon, to faux treatment remedies claiming to cure the virus. The impact of these claims range from merely attempting a non-effective treatment to engaging in dangerous behavior, such as failing to follow safety protocols and inadvertently spreading the virus [6].

Misinformation can occur in almost any media type, whether it be videos, text posts, images, or memes. In particular, if it is present in the last form, incorporating humor with the misinformation can help it spread easier. In order to combat misinformation, many have employed deep learning in order to detect misinformation in text-based posts.

As the topics of misinformation changes over time, it might render previously viable approaches to misinformation detection obsolete due to concept drift and other factors. In order to aid in the combat of misinformation, we hope to contribute a data set that can capture the historical essence of 2020, along with a sliver of the misinformation present in it. Our hope is that this data set will provide current and up-to-date data so that future work in this area can adapt to the constantly changing ecosystem of misinformation.

In this paper, we present a collection of tweets collected during the span of the 2020 United States presidential election and amidst the COVID-19 pandemic. We explore the labeling of these tweets through unsupervised means with ground truths extracted from PolitiFact, as well as an anecdotal discussion of the trends that we observed while manually labeling these tweets.

2 RELATED WORKS

Our literature review helped us locate relevant and popular data sets, which helped us develop a plan for our own data collection. Particularly, we found a Twitter data set from Ma et al. [5] to be particularly relevant and useful. Since the main contribution of this paper is a modern data set, we were inspired by the methodology used in their approach. In particular, this work only included tweets that contained keywords from different Snopes articles. They located the tweets through the Twitter search function, and manually curated their search terms so that the contained tweets related to the claim discussed in the Snopes article. Similar to this work, in constructing our data set, we scraped information from PolitiFact to establish ground truths. These ground truths were later used in an attempt to label data in an unsupervised fashion. Previously referenced data sets are rather dated (from 2011 [2] and 2013 [4]), though still applicable.

To accomplish our unsupervised automated labeling, we found Kusner et al. [3] and their proposed Word Mover's Distance function to be the most valuable method we could adopt for our work. Leveraging word2vec word embeddings, the application of Word Mover's Distance, in testing, outperformed seven existing state of the art approaches for document distances. It also outperformed 6 of 8 real-world classification tasks. Given the success and ease of implementation, we found this method to be the most accessible and beneficial for our application.

Given the growing age of the data sets used in most previous work and the prime opportunity for collecting information during an election cycle, on top of a worldwide pandemic, we adopted an approach similar to Ma et al. [5]. In addition, our literature review aided in identifying embedding methods that would allow for the unsupervised labeling of our data set to improve its value for future research.

3 METHODOLOGY

In this section, we discuss our data collection and labeling procedures. Analysis and properties of our data sets are discussed in the next section.

3.1 Ground Truth

In later sections, we discuss the exploration that we conducted into automatically labeling misinformation. To do this exploration, we wanted to investigate if we could successfully use previous ground truth data to label data effectively. PolitiFact provided this ground truth data in a convenient format. We chose to collect data from political figures on Twitter that had been mentioned in PolitiFact articles multiple times since there was already an available ground truth for statements that might arise in Twitter discourse

We collected all PolitiFact articles that were archived up to early October, which is when we first started gathering tweets for our data set. PolitiFact posts provide the subject that is responsible for the claim in question, a summary of the claim, and a rating for the claim. The rating is on a 6 point scale (True, Mostly True, Half True, Barely True, False, Pants on Fire), which is particularly useful as the nuances of the truth value of information are not always captured within a binary classification. Forcing a binary classification ignores a lot of the intricacies in language. Within PolitiFact articles, the specification of who made a claim and a summary of it is helpful as it allows for easier matching within a given tweet, due to both having a short length.

The PolitiFact articles that we used were only those that existed in the first few weeks of October. Notably, all articles predate our actual data collection. In hindsight, we should have used updated articles for our final unsupervised labeling attempts. As we discuss later, we had some hindrances in our development in which time was a primary factor. However, we do not think that the negative results that we found were directly a result of using the older articles.

In total, we collected 5,085 PolitiFact articles. We chose Twitter accounts to monitor based on the authors mentioned in PolitiFact. The subset that we chose were authors that appeared multiple times in claims made on PolitiFact. We chose them with the reasoning that if they had appeared multiple times in PolitiFact, and therefore had made multiple claims that had been investigated, there may be chance they either make a claim worth investigation in the future, or others may discuss their previous claims on social media. In the instances where a non-named, non-person entity (Facebook, Bloggers, Twitter, etc.) was the maker of a claim, we omitted them from the list.

3.2 Data Collection

To collect our data, we utilized the Twitter API. Twitter was chosen due to the presence of many prominent political figures on the platform that are active and the ability to query and interact with posts easily. The Twitter API enables the retrieval of tweets through the construction of rules that they use within their search engine. The queries resemble boolean operators in traditional search engines. Our rules took the form of "-has:media @NAME OR from:NAME". This collects tweets in which the user is tagged or if the user themselves makes the tweet. Our analysis for misinformation is contained within the text of each post, and any attached media is ignored. This rule was utilized for all 83 users that we collected tweets from. These users were selected based on the most common subject of the PolitiFact posts (those that made the claims being reviewed by PolitiFact).

With Twitter's streaming API, the tweets are delivered to our client at the moment they are made. This means that we have a record of tweets before they are possibly deleted, or the posting accounts are deleted or suspended. We collected two sets of tweets for our data set, which we refer to as our Control set and our Filtered set. The Twitter API offers a feature in which a random 1% of all tweets made in a day can be captured in-stream, and this composed our Control data set. The filtered stream is all tweets that adhere to the rule we described above. We ran our stream client 24/7 starting on 10/12 for the filtered stream, and 10/21 for the control stream.

It is important to note that for the filtered stream, we initially only collected tweets from Donald Trump (@realDonaldTrump) as he is very active on Twitter, and many users engage with his comments. This was a result of attempting to find a successful way to interact with the API. We thought this would be satisfactory for preliminary data. We started running our revised rules (including all 83 authors) on November 3rd. In hindsight, we wished that we were able to deploy the revised rules prior to the election occurring. However, as we discuss later, a small proportion of tweets we collected did not mention or involve Donald Trump. This only accounts for around 10% of the total size of the collected tweets. That is, by capturing tweets involving Donald Trump, we captured a far larger volume of discourse than with the remaining 82 authors.

While we have sufficient data in terms of quantity from the days prior to the election, we might have not captured some trends of misinformation. We feel that for the purposes of misinformation detection though, this will not have many noticeable effects, but it is worth noting this abnormality.

3.3 Automated Labeling

Due to the volume of data that we expected to collect, and the general utility that it might provide if successful, we explored trying to label our tweets using an unsupervised approach. We wanted to leverage existing fact checks that are updated periodically, like PolitiFact, in order to develop a method that could, given some tweet, suggest the most relevant PolitiFact article. To do this, we utilized embeddings from FastText along with WordMover's Distance to attempt attaining a notable top-N accuracy.

FastText [1] is a model that represents words as a list of n-grams and can be used to generate word embeddings from them. The use of n-grams allows the model to operate on words not found in the original vocabulary. This is particularly applicable to our use case since they can be used to handle hashtags and the natural variation in writing styles. These embeddings are vector representations of each word contained in a tweet. The embeddings can be used in conjunction with other algorithms, like Word Mover's Distance, to perform natural language processing tasks, like document similarity.

We use Word Mover's Distance [3] in order to determine the similarity between each tweet and PolitiFact article. Word Mover's Distance is a suitable algorithm for this application because it can find distances between two texts that share no similar words, ignoring stop words. The algorithm relies on the assumption that similar words should have similar word vectors. It measures the distance between two documents as the minimum distance between the embedded words in one document to the embedded words in the other document, in manner similar to Earth Mover's Distance. Unfortunately, Word Mover's distance is slow to use in practice, and even relaxed approaches like Relaxed Word Moving Distance [8] run in $O(n^2)$ time, where n is the number of unique words.

Our evaluation of this approach largely comes from our own observations as we were manually labeling data. We don't have an explicit way to measure the accuracy in an efficient manner for each tweet other than our observations of effectiveness. With this in mind, we made sure to pay careful attention to the recommendations that were made through this methodology as we were labeling. From all observations, there were 0 instances in which the article recommendations contained any articles that were relevant, or even close to the subject of the tweet. We noticed the primary failure point was that any time that Donald Trump was mentioned or tagged in a tweet, there was an automatic association with 3 or 4 articles that would always appear, leaving 2 wildcard articles, which were still never relevant in the tweets we labeled. It may be the case we didn't label any tweets that had information related to a published PolitiFact article. Still, of the tweets we labeled, no relevant articles were ever suggested.

We attribute these poor results to a few different factors. The first is that due to the relatively small size of the PolitiFact article collection. We didn't feel that 5,085 articles were enough of a training basis for FastText to create meaningful embeddings. Instead, we utilized FastText's provided Wikipedia corpus for training. This likely created a more generalized setting for FastText, and it wasn't particularly tuned for the task at hand, which was likely detrimental. The other factor is that we didn't have the bandwidth this semester to properly tune or work more on this. Calculating the distances for our data set was a very time-intensive procedure (multiple days), and this made it intimidating to tune and continue working on while we still needed to label tweets. We would also like to note that we attempted our same approach with doc2vec over FastText in order to more closely emulate the original methodology of Kusner et. al. [3], and it produced similarly poor results.

3.4 Manual Labeling

Due to the negative results we achieved with our attempts at automated labeling, and the time we had left in the semester, we manually labeled as many tweets as we could in the time after Fall break. We hoped to label enough to try a few different classification models on our data set; however, we ran out of time in the semester. As is detailed in the Discussion section, labeling the tweets ended up being a more challenging process than we had envisioned at the start.

We created an interface to help with our labeling, which can be seen in figure 1. We utilized the automated labeling results that we had in order to recommend PolitiFact articles that were semantically close to the given tweet. We did this in order to evaluate the results of the calculations. While our results weren't satisfactory, it still gave us some assistance in labeling, though not much. Any assistance is preferred when attempting to label such an extensive data set manually.

These recommend articles can be seen highlighted in yellow in figure 1. The original tweet is highlighted in green, and a summarized version is highlighted in blue. This interface helps us label faster when it comes to substanceless tweets, but made little difference when labeling tweets that had substance, as we often needed to verify the information presented in a tweet with outside sources.

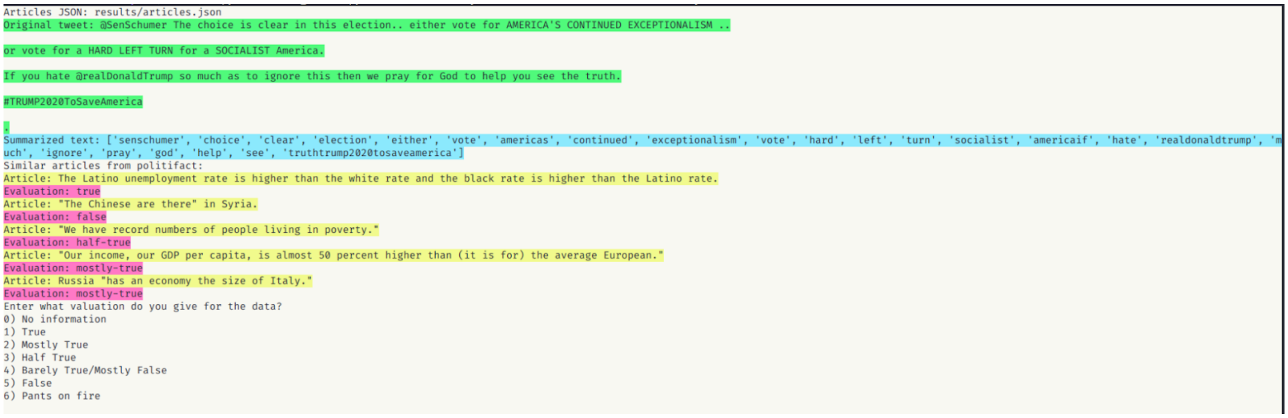
We utilized the same rating scale that PolitiFact uses. We used this scale to match any suggested PolitiFact articles that were correct and actually matched. We also liked this scale as we thought it captured the essence and nuances of misinformation nicely. That is to say, misinformation is rarely accurately summarized through a binary classification.

When it comes actually labeling tweets on a scale, it is difficult to standardize how we label tweets among the three of us. This is particularly an issue when we introduce a scale, as we might each have weightings for what constitutes something as 'Half True' compared to 'Mostly True'. We don't foresee this being an issue until we have a significant number of tweets of substance labeled.

Many of the tweets we labeled were baseless Ad Hominem attacks or general nonsensical tweets that had no informational value. Tweets that had actual informational value were very sparse. From this, we each labelled disjoint sets of tweets in order to cover as many as possible within the time we had available. Once we identify the tweets that had informational value to them (regardless of if the information was true or false), we plan to finalize labeling of tweets based on a consensus of our three ratings.

4 RESULTS

This section discusses the results, both positive and negative, that we found during our investigation of this project. We discuss the limitations we faced when finding negative results and what we would like to do in future work on this project to overcome those limitations in order to find success.



```
Articles JSON: results/articles.json
Original tweet: @SenSchumer The choice is clear in this election.. either vote for AMERICA'S CONTINUED EXCEPTIONALISM ..

or vote for a HARD LEFT TURN for a SOCIALIST America.

If you hate @realDonaldTrump so much as to ignore this then we pray for God to help you see the truth.

#TRUMP2020ToSaveAmerica

.
Summarized text: ['senschumer', 'choice', 'clear', 'election', 'either', 'vote', 'americas', 'continued', 'exceptionalism', 'vote', 'hard', 'left', 'turn', 'socialist', 'americaif', 'hate', 'realdonaldtrump', 'm
uch', 'ignore', 'pray', 'god', 'help', 'see', 'truthtrump2020tosaveamerica']
Similar articles from politifact:
Article: The Latino unemployment rate is higher than the white rate and the black rate is higher than the Latino rate.
Evaluation: true
Article: "The Chinese are there" in Syria.
Evaluation: false
Article: "We have record numbers of people living in poverty."
Evaluation: half-true
Article: "Our income, our GDP per capita, is almost 50 percent higher than (it is for) the average European."
Evaluation: mostly-true
Article: Russia "has an economy the size of Italy."
Evaluation: mostly-true
Enter what valuation do you give for the data?
0) No information
1) True
2) Mostly True
3) Half True
4) Barely True/Mostly False
5) False
6) Pants on fire
```

Figure 1: Our tweet labeling interface.

4.1 Data Set

When we discuss our data sets in this section, we make distinctions between the data set in which all tweets are included and the data set that we collected after 11/3, in which we deployed our revised rules to include all 83 political figures. These distinctions are made due to variations in the volume of tweets that occur on any given day and variation of some qualities of the data set. We first discuss our control data set in order to establish a baseline to compare to when later discussing different qualities of the filtered data set, which contains tweets relating to politicians and political commentators. We discuss the results of our manually labeled data in the next section. This section focuses on the overall set of tweets that we collected throughout the semester, which are not necessarily labeled. This is done to give a sense of the general nature of the data included in our collection.

We present some measurements of the tweets that we collected in table 1. Control-Election and Filtered-Election are the subsets of tweets from Control-All and Filtered-All that were collected starting on 11/3. We stopped collecting tweets on 11/18 and 11/22, respectively. These dates coincided with Fall break, and we felt that we had captured enough data after the election, in which many trends of misinformation and disputes of election results had occurred.

Overall, we collected 524,112 tweets in our filtered set and 357,086 tweets in our control set. We didn't focus efforts on labeling the control set, as the misinformation that we may find might span multiple topics. We had pulled a ground truth set from PolitiFact, so we had hoped to only work with political misinformation this semester. However, we collected the control data set, as we feel our collection of tweets can be used for more than just misinformation detection, and it can serve as a focused snapshot of an impactful year. Having our filtered data set of political tweets, along with the corresponding control set on the collected days, can be useful across many other applications and domains.

Our collection of tweets is too large to be able to provide a detailed description of its contents, in terms of what it might be useful for and what the average tweet looks like. We can, however, discuss the appearance of the data set at a high level. We provide an anecdotal discussion of the data set later in our Discussion section (Manual Labeling subsection).

As shown in table 1, the number of unique authors in proportion to the total number of tweets is closer to 1 in the Control sets and is notably less than 1 in the Filtered data sets. We feel that this is an important distinction to note, as it indicates on some level that some authors are actively posting and engaging with tweets. It may also indicate on some level that information is shared by multiple entities. The Control data set ratio is less relevant as we expect it to be close to 1 since the control set is randomly selected across all tweets made in a day.
Another metric that we think is important to note is the number of retweets in a given set. Again, this is less impactful within the control set since they are randomly sampled tweets, but it is notable in our filtered sets. Since we are capturing all tweets from a user or when they are mentioned, and due to internal mechanics of the Twitter API, these retweets appear as duplicates within the set. This effectively means that the Filtered data set has 268,000 unique tweets. While this is a notable reduction, there is some value in being able to track who retweeted a tweet, or how often it was retweeted. We can also see that the proportion of retweets for Filtered is significantly above that of Control. It is hard to assess implications and causes of this, but it generally can show that information, tweets, and possibly misinformation might spread at a higher volume.

Overall, we feel the collection of tweets that we gathered this semester serves as a good slice of political discourse around the election and can be impactful in multiple domains. As we discuss later, more cleaning and labeling of this data set would be required before we would want to share the data set publicly, but we feel it holds a lot of potential value.

4.2 Manual Labeling

As mentioned in prior sections, most of the value in the data set at this point lies within our ability to manually label it, due to the

Data Set	Total Count	Unique Authors	Total Retweets	Non-Donald Trump Tweets	Tweet/Author	Retweet Proportion
Control-All	357,086	346,570	140,091	349,019	1.03	39.2%
Filtered-All	524,112	359,362	256,527	72,172	1.46	48.9%
Control-Election	181,450	178,200	71,175	176,348	1.02	39.2%
Filtered-Election	312,012	244,850	145,906	53,936	1.27	46.8%

Table 1: Relevant metrics regarding all tweets collected. The suffix of '-Election' indicates only tweets collected after 11/3 in which we included filtered rules.

negative results that we found with our exploration into automated labeling. With time constraints this semester, we started labeling upon returning from Fall break and attempted to label as many as we found time to do. In total, we have 21,476 labeled tweets on the scale of 0-6, in which 1-5 refer to ratings given by PolitiFact of the truthfulness of a claim, and 0 denotes that there is no informational value included in the tweet.

The ratings and their meanings can be seen in table 2. As mentioned previously, we derived this scale to match with PoltiFact articles in the event that our unsupervised labeling was accurate. Our labeling to this point was somewhat subjective, as the tweets that each of us labeled were disjoint. We did this primarily as we wanted to make a first pass at the data.

The distinctions between most measurements in table 2 are self-explanatory. Still, we wanted to clarify each point in how we interpreted it to set a loose standard that we could adhere to when we were labeling. The rating of 0 was utilized when we found a tweet that had no substantial or information or made no claim. These tweets generally consisted of targeted political rhetoric that didn't make claims (name-calling, opinionated descriptions, etc.) but rather were based on opinions or views of how things ought to be.

We labeled all other tweets using ratings 1-6. The lines for standards blur within these ratings. Generally, things were only marked as True if all information in the tweet was verifiably correct. Mostly True and Half True were used if approximately 75% and 50% of the information in the claim were true, respectively. Barely True corresponds to approximately 25% true, while False and Pants on Fire blur a line. We reserved Pants on Fire for the most egregious tweets that were intentionally and maliciously being deceptive. Unfortunately, our labeling is still subjective even though we tried to define some standards for it, and it is difficult to measure the weighting each person gave to a claim being 75%, 50% or 25% true.

The tweets we labeled consisted of 1,973 unique tweets. Since retweets were common within our data set and were essentially duplicates, we propagated the rating to all occurrences after we came to this realization. Upon propagation, we have labeled 21,476 tweets within our collection. Future labeling will progress faster with the propagation of retweets. Per table 1, around half of the tweets in our Filtered set are retweets. However, even with this propagation, there are many unique tweets that we still need to label.

Our labeling only included a limited number of days that we able to start, but unable to finish fully. Each day in our data set contains approximately between 3,000 and 5,000 tweets. We tried to label from a variety of days in order to gather some variation in our tweets before and after the election. Our currently labeled tweets span from 10/12/20 to 10/17/20 and 11/11/20 to 11/14/20.

A breakdown of ratings in our manual labeling can be seen in table 3. We focus on the results from the breakdown of unique tweets only, as metrics become distorted based on variations in the number of retweeted tweets and how often they were retweeted. In our labeled tweets, we observed that close to 86.6% of tweets that were devoid of information that we gave a rating of 0. The remaining 13.4% were given a rating of some truth value. Interestingly, the labels of 1 (True) and 5 (False) had similar proportions, and labels between seemed to be close in proportion as well. We reserved a label of 6 (Pants on Fire) for especially egregious cases.

5 DISCUSSION

In this section, we discuss the results we presented in the previous section in more detail, along with reflection and discussion of the limitations we faced during our project, with how we plan to continue the project and overcome these limitations.

5.1 Data Set

We feel that our data set collection was successful and will provide a lot of utility for different research areas after we finish cleaning and labeling it. We offer a good balance of tweets regarding Donald Trump before and after the election, as well as a subset of political figures after the election. We also captured a month's worth of tweets during the COVID-19 pandemic. This data set will serve as a useful snapshot of political discourse and behavior on social media within the pandemic and election.

We had a few points that we had wished we could have improved upon but were overall happy with our work in this area. The main issue that we wished we would have executed differently was to have the filtered set of politicians' tweets span the entire time we were collecting tweets. This would have provided a more comprehensive data set and would have let there be continuity between all subsets.

Another point that we wish we could have improved upon was to expand the number of politicians we collected from. We collected 83 politicians' tweets, many of which are well known and active on Twitter. However, the ones that we had recorded, had multiple occurrences within our set of PolitiFact articles. As shown in table 1, the number of non-Donald Trump tweets is meager compared to the overall data set's size, so having more non-Donald Trump tweet references would have potentially given us more data to work with. Since misinformation was seemingly sparse in our data set, more collected tweets would have been essential for making our data set effective.

5.2 Labeling

Our approach to automated and manual labeling was not of the quality and quantity, respectively, that we had envisioned when we started this project. We attribute this to a lack of a main project focus on our part. Our focus became divided as we collected the data set and realized the sheer magnitude of the data set we were accumulating. Our initial plans for this project were to collect a data set, label it, evaluate the data set using a novel detection methodology, and compare our methodology to other existing methods.

In hindsight, we metaphorically bit off a bit more than we could chew with the experience and time that we had as a group. After realizing the magnitude of our data set, our inclination was to change gears and look for unsupervised methods for data set classification so that it might be feasible to have our data set labeled by the end of the semester. As mentioned in the previous section, our results were not what we expected, and the time needed to fine-tune and further pursue this approach was not realistic. From here, we felt that the most significant contribution that we could make, while still making the project successful in some regard, was to label the data set manually.

Even with three people, this turned out to be a more intensive task than we anticipated, in both time and effort. The primary factors for this being the case were due to the nature of the tweets and the nature of misinformation. Going into labeling, we expected tweets would be a bit more dense with misinformation than they actually were. We were surprised that it was relatively sparse within our data. This limitation increased the amount of time we had to spend labeling before finding a tweet with any relevant substance.

The other confounding factor for manually labeling data was our judgment of the information in a tweet when we did find a tweet of substance. We expected that the variety of misinformation that we would find would mostly be glaringly wrong to some degree. What we ended up finding, anecdotally, was that many of the claims weren't glaringly wrong but were rooted in some confounded or partial truth. This often meant that when we found a tweet of substance, we would have to manually verify each part of the information in order to accurately label it with the scale that we used from PolitiFact.

These factors, coupled with the time we had left in the semester to label tweets, made it challenging to produce a number of tweets that were significant or relevant. As previously shown in our results section, we classified most tweets as 'No Information' because they usually corresponded with Ad Hominem attacks or other logical fallacies regarding current (at the time of tweet publication) events, without making some some factual statement, whether it be true or untrue.

5.3 Future Work

Given the difficulties encountered throughout the process, we have identified a number of areas of future work that could significantly improve our results. As identified previously in this work, additional labeling could improve our data set, and serves as a practical next step. When completed, our data set could be made public to aid in future research and to provide an updated snapshot of the misinformation landscape that exists on Twitter.

Another area of future work that could be improved upon is in regards to available ground truth articles relative to misinformation trends. Our articles were collected prior to collecting our data to identify relevant authors during data collection. While beneficial in this regard, this approach does not allow for newer misinformation trends, which may be present in the data collected. For example, claims of election fraud were captured in our post-election data by various accounts. At the time of writing, investigations are being performed on these claims with no general consensus on these accusations' legitimacy. Leveraging our existing data set in combination with an updated collection of Snopes/Politifact articles would improve our ability to label accurately.

While we focused on FastText calculations of WordMover's Distance for matching of embeddings, several alternative methods for automated labeling could be investigated, potentially yielding better results. Examples such as Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) have both been leveraged in prior work to establish a latent and low-dimensional representation of documents. Compared to bag of words (BOW) approaches, these methods have been noted to provide a more coherent document representation [3] .

Lastly, social media platforms such as Twitter have started to implement their own notification system alerting users to potential inaccuracies of some information being posted. Investigating methods of incorporating these flags of misinformation into a model could result in further improved accuracy and performance from what our PolitiFact article matching is able to provide.

6 CONCLUSION

In this work, we presented a brief exploration towards unsupervised labeling and contributed a data set that captures close to a month's worth of Twitter activity that is centralized around the 2020 Presidential election, and during the COVID-19 pandemic. We collected 524,112 tweets in total from various political figures in a time frame from 10/12 to 11/22. We labeled 1,973 unique tweets and 21,476 tweets when accounting for retweets. We discussed our observations from our time spent manually labeling, as well as discussed a high-level view of our data set.

We would also like to share the list of political figures we recorded along with data that we labeled in addition to this report.
https://github.com/hadyer2/598GW-Paper/

REFERENCES

[1] Piotr Bojanowski et al. “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[2] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. “Information Credibility on Twitter”. In: *Proceedings of the 20th International Conference on World Wide Web*. WWW ’11. Hyderabad, India: Association for Computing Machinery, 2011, pp. 675–684. ISBN: 9781450306324. DOI: 10.1145/1963405.1963500. URL: https://doi.org/10.1145/1963405.1963500.

[3] Matt Kusner et al. “From word embeddings to document distances”. In: *International conference on machine learning*. 2015, pp. 957–966.

[4] S. Kwon et al. “Prominent Features of Rumor Propagation in Online Social Media”. In: *2013 IEEE 13th International Conference on Data Mining*. 2013, pp. 1103–1108.

[5] Jing Ma et al. “Detecting Rumors from Microblogs with Recurrent Neural Networks”. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI’16. New York, New York, USA: AAAI Press, 2016, pp. 3818–3824. ISBN: 9781577357704.

[6] Gordon Pennycook. “Fighting misinformation on social media using crowdsourced judgments of news source quality”. In: *Proceedings of the National Academy of Sciences* 116 (Jan. 2019), p. 201806781. DOI: 10.1073/pnas.1806781116.

[7] Gordon Pennycook et al. *Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention.* Mar. 2020. DOI: 10.31234/osf.io/uhbk9. URL: psyarxiv.com/uhbk9.

[8] Matheus Werner and Eduardo Laber. “Speeding up Word Mover’s Distance and its variants via properties of distances between embeddings”. In: *arXiv preprint arXiv:1912.00509* (2019).

[9] Liang Wu et al. “Misinformation in Social Media: Definition, Manipulation, and Detection”. In: *SIGKDD Explor. Newsl.* 21.2 (Nov. 2019), pp. 80–90. ISSN: 1931-0145. DOI: 10.1145/3373464.3373475. URL: https://doi.org/10.1145/3373464.3373475.

Rating	Meaning
0	No Information
1	True
2	Mostly True
3	Half True
4	Barely True
5	False
6	Pants on Fire

Table 2: Ratings for Manual Labeling

Rating	Quantity (Retweets Included)	Proportion (Retweets Included)	Quantity (Unique Only)	Proportion (Unique Only)
0	13,916	64.8%	1,708	86.6%
1	616	2.8%	82	4.2%
2	54	0.25%	27	1.4%
3	193	0.90%	35	1.8%
4	1,707	7.9%	29	1.5%
5	4,652	21.7%	86	4.4%
6	338	1.6%	6	0.30%
Total	21,476		1,973	

Table 3: Rating Distribution for Manual Labeling