Data Acquisition, Pre-processing, and Exploratory Data Analysis for Tweet Turing Test: Detecting Disinformation on Twitter

John Johnson
Drexel University
jmj382@drexel.edu

Katy Matulay
Drexel University
km3868@drexel.edu

Justin Minnion
Drexel University
jm4724@drexel.edu

Jared Rubin
Drexel University
jar624@drexel.edu

College of Computing & Informatics, Drexel University, Philadelphia, PA USA

Abstract - Disinformation has infiltrated the American online social spaces such as Twitter and poses a unique and complex threat to democracy. Russian web brigades, known as trolls, associated with the Internet Research Agency (IRA)-- were highly active in their efforts to influence the 2016 U.S. Presidential election. The weapons at their disposal, over 3 million disinformation tweets, have been made public and shared online by researchers at Clemson University and the website FiveThirtyEight (538). Given the existing repository of "troll" tweets, a secondary set of "verified" tweets was obtained utilizing a Twitter Academic License with the Twitter API. The acquired 1.5 million tweets were analogous in time frame to the 538 dataset (2013-2017) and posted by accounts labeled as "verified" by Twitter. This work aims to find indicators of disinformation in Twitter accounts and tweets, using state-of-the-art Natural Language Processing (NLP) methodologies.

Keywords: NLP, disinformation, Twitter, social media, misinformation, information operations, information extraction, information retrieval

1. Introduction

1.1. What is disinformation?

As defined by Starbird et al., disinformation is a subtype of strategic information operations, used to subvert online discourse and influence public opinion by intentionally altering the information environment (1). In the age of social media and online news, online information warfare is a valuable weapon for a foreign adversary to harness and difficult to defend against.

Disinformation and strategic information operations in the context of Russian trolling have also been described as "astroturfing." According to Zelenkauskaite (2022), "the paid and orchestrated aspect of Russian trolling renders it comparable to astroturfing"; which has similarities to propaganda campaigns in that the aim is to "provide misleading information or pay people to spread misinformation by altering reality" (2). In the context of the 2016 election, Russian trolls were paid by the Russian state to create chaos in the social media space. This chaos was accomplished via distracting, deflecting attention, and sowing seeds of distrust, all in the aims of influencing public opinion.

1.2. Historical context (2016 U.S. Election)

According to the declassified 2017 Intelligence Community Assessment distributed by the Director of National Intelligence, it was assessed with a high degree of confidence that "Russian President Vladimir Putin ordered an influence campaign in 2016 aimed at the US presidential election, the consistent goals of which were to undermine public faith in the US democratic process, denigrate Secretary Clinton, and harm her electability and potential presidency" (3). By harnessing the power of social media, specifically Twitter, the IRA was able to exert a campaign of chaos, further eroding public trust in media, government, and democracy one tweet at a time.

We now know with a great degree of confidence, thanks to the joint efforts of the Federal Bureau of Investigation (FBI), the Central Intelligence Agency (CIA), and the US National Security Agency (NSA)—that "Russia's state-run propaganda machine [the IRA] contributed to the influence campaign by serving as a platform for Kremlin messaging to Russian and international audiences" (3). Several journalists also reported on this phenomenon early on and were key in exposing the operations of Russian troll farms (4) (5).

As part of the Mueller investigation, Twitter released a report that quantified just how far reaching these troll accounts had been—over a million Twitter users had been notified that they had interacted in some way with these accounts (6). Twitter then officially released a dataset of tweets and accounts from IRA associated "Russian troll" accounts, which was then further analyzed by researchers at Clemson University and finally released by FiveThirtyEight (7) in the public domain on both Kaggle and GitHub. This dataset serves as the control group of labeled "fake" accounts/tweets for our project.

1.3. Twitter platform

Twitter is a social network and real-time communication service that allows users to blog short messages of not more than 280 characters. These messages are labeled as "tweets", and can contain emojis, metadata tags, gifs, and hyperlinks. Twitter has become a service for friends, families, and coworkers to connect instantly and share information of their choice. The platform has also become an important outlet to spread news amongst the masses. It has a userbase of roughly 500 million people from all over the world. The social media site has many use cases and has been a data storage warehouse by housing information uploaded by all users.

1.4. Twitter Verification

1.4.1. Legacy Verification Policy

During the time period in which our data originates (the years 2013-2017), Twitter used a verification policy to vet certain user accounts, with the intention of providing other Twitter users a means of determining the authenticity of an account. A verified account meeting the requirements of the policy would display a blue checkmark icon adjacent to the account's display name.

The verification policy has since changed (refer to section 1.4.2 for more info on the change), the verification policy was advertised with the following meaning: "The blue Verified badge on Twitter lets people know that an account of public interest is authentic." The three core requirements for verification are that an account is 1) authentic, 2) notable, and 3) active.

To ensure *authentic* users, the legacy verification requirements called for a photo identification check and similar credentials to prove a user's identity. For a user to be consider *notable*, the account needed to represent a prominently recognized individual or organization. Finally, an *active* account meant the account must have a demonstrated history active use and of adherence to Twitter's rules (8).

1.4.2. Recent Changes

Prior to Twitter's recent acquisition and privatization by Elon Musk (on October 27, 2022), a verified account at Twitter had an extremely precise and rigorous definition (as noted in prior section). It is important to note that on November 9, 2022, Twitter stopped accepting applications for a blue verified badge under the criteria used in this research paper, and it is now referred to as "Legacy Verification policy" (8), as noted in the previous section.

We discuss the impact of these recent changes on our study in section 5.5.1.

2. Datasets

2.1. Data source #1: Russian Troll Tweets

We chose to use the 538 dataset of "Russian Troll Tweets" from the IRA accounts, publicly available on GitHub (9). The GitHub repository contains 13 pre-processed CSV files, with additional data elements added by Clemson University researchers who originally curated the dataset. The dataset contains over 2.9 million tweets from ~2800 Twitter handles, with publication dates spanning ~2013-2017. The total merged CSV is 1.2 Gb. This data source was selected because it contained labeled "fake" Twitter account data and a rich subset of pre-processed features such as: language, location, followers, following, account type, and full URLs.

2.2. *Data source #2: Twitter API*

We utilized the Twitter API with Academic Research access and Twitter's Tweet Downloader tool to export over 1 million tweets. The acquired tweets were posted during the same time frame as the 538 dataset (2013-2017) and posted by accounts labeled as "verified" by Twitter. We collected a total of, roughly, 1.5 million tweets, with each dataset containing ~57 columns.

3. Objectives

The overall goal of the project (including analysis and modeling) is to build a supervised or semi-supervised classification machine learning model that can accurately and efficiently identify potential sources of disinformation based on natural language processing (NLP) algorithms and characteristics identified as indicators of disinformation. To support and train such a model, the goal of our group for the topic of this paper will be to identify, acquire, clean, pre-process, visualize, and perform exploratory data analysis of Twitter data related to this topic.

4. Methodology

In order to carefully define our control group of verified users, we implemented a detailed methodology for curating tweets. We chose to focus on verified accounts based on Twitter's requirements for verification (refer to section 1.4) and their process to vet accounts for verification (10). The entirety of the project's code can be found in our public GitHub repository:

`github.com/disinfo-detectors/tweet-turing-test/`

4.1. Data acquisition

4.1.1. Data Source #1: FiveThirtyEight Russian Troll Tweet Dataset

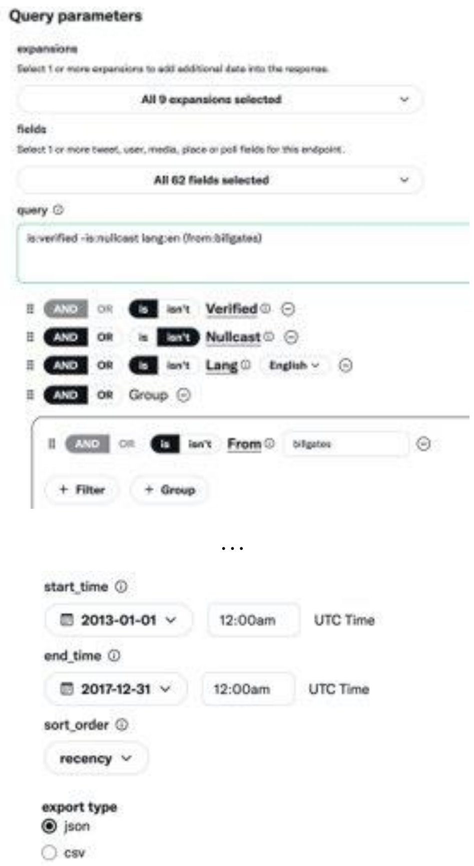
Thirteen CSV files were downloaded from GitHub. Data was integrated in a Jupyter notebook by loading all CSV files to a pandas dataframe using UTF-8 encoding, and then writing to a single, merged CSV file.

4.1.2. Data Source #2: Twitter Academic Research API and Downloader API Tool

Using the academic research bearer token, we were able to access the Twitter Tweet Downloader tool (11) and create custom queries to retrieve bulk exports of tweets as json files. We curated two distinct datasets using the tool: Dataset (2A) Tweets by a small number of manually chosen subset of verified accounts spanning the entire 2013-2017 time period of interest, and (2B) Tweets by a large number of randomly chosen verified accounts during randomized samples of time within the 2013-2017 time period of interest. Dataset 2A contained ~500k tweets, whereas Dataset 2B contained ~1 million. JSON files were then loaded, normalized, converted to pandas dataframes, merged, and further pre-processed. Further illustration of the data acquisition and pre-processing pipeline can be found in the Appendix as **Figure 23** and **Figure 24**.

4.2. Pre-processing

Pre-processing was split into multiple parts, as the datasets had unique features that required independent pre-processing in some cases. Many of these cases were predominantly focused on text and natural language processing due to the context contained in the dataset. Tools such as NLTK (Natural Language Toolkit) were used to identify the linguistics commonly used on Twitter. Other methods were also used to extract certain characters for further analysis.



Query parameters

expansions

Select 1 or more expansions to add additional data into the response.

All 9 expansions selected

fields

Select 1 or more tweet, user, media, place or poll fields for this endpoint.

All 62 fields selected

query

is:verified -is:nullcast lang:en (from:billgates)

AND OR is isn't Verified

AND OR is isn't Nullcast

AND OR is isn't Lang English

AND OR Group

AND OR is isn't From billgates

+ Filter + Group

…

start_time

2013-01-01 12:00am UTC Time

end_time

2017-12-31 12:00am UTC Time

sort_order

recency

export type

● json

○ csv

Figure 1 – Image of tweet downloader query parameters

A "tweet" is a main component of Twitter, and an important feature that will be fed into the model. A tweet can contain up to 280 characters[i]. These characters can be emojis, symbols, letters, numbers, and spaces. Additionally, a tweet can also include up to 4 photos, a GIF, or a video. To use this feature, it must be structured in a specific way. The python package, NLTK, supplies a module called "TweetTokenizer" that parses words into tokens. The tool does recognize retweets, hyperlinks, metadata tags, and gifs and can keep them together for easier recognition. With the help of this module, 3 million tweets were tokenized. Although hyperlinks, metadata tags, and gifs are a part of a tweet, they were separated into other features due to best practices.

After tokenization, symbols and numerical values were evaluated within a tweet. Python has a great

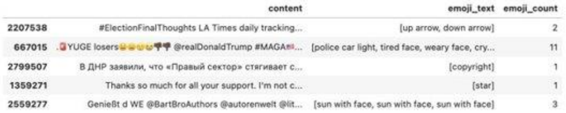
[i] Note on maximum tweet length: prior to November 2017, the maximum tweet length for most Twitter users was 140 characters (19).

module called regex that can strip certain symbols. This process was used to clean the tokens to ensure only words were captured. Additionally, any tokens that contained numerical values were excluded and can be evaluated in future work. The next step in the process was to verify if any of the tokens were a stop word. If it ended up being marked as one, it was removed. The word bank of stop words was included in the NLTK toolkit and have deemed certain words that are ‘fluff’ or are plain useless. With stop words already being identified by the toolkit, attention was allocated to focus on filtering additional words or on feature engineering to conduct a better model.

As shown in **Figure 23** and **Figure 24** in the Appendix, the datasets required some filtering, feature extraction, column name alignment, and addition of new label columns to differentiate data sources.

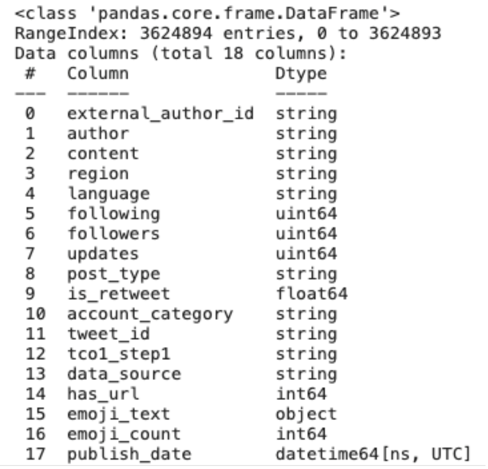
4.2.1. *Emoji Pre-processing*

In order to extract the text tags for emojis, a function was built using the *demoji* package. Text strings for each emoji were captured in a list and then a count of the emojis/tweet was derived. An example of this is shown below in **Figure 2**



	content	emoji_text	emoji_count
2207538	#ElectionFinalThoughts LA Times daily tracking...	[up arrow, down arrow]	2
667015	.🚨YUGE losers😩😩😩😩👎👎 @realDonaldTrump #MAGA🇺🇸...	[police car light, tired face, weary face, cry...	11
2799507	В ДНР заявили, что «Правый сектор» стягивает с...	[copyright]	1
1359271	Thanks so much for all your support. I'm not c...	[star]	1
2559277	Genießt d WE @BartBroAuthors @autorenwelt @lit...	[sun with face, sun with face, sun with face]	3

Figure 2 – Emoji text and count pre-processing



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3624894 entries, 0 to 3624893
Data columns (total 18 columns):
 #   Column               Dtype
---  ------               -----
 0   external_author_id   string
 1   author               string
 2   content              string
 3   region               string
 4   language             string
 5   following            uint64
 6   followers            uint64
 7   updates              uint64
 8   post_type            string
 9   is_retweet           float64
 10  account_category     string
 11  tweet_id             string
 12  tco1_step1           string
 13  data_source          string
 14  has_url              int64
 15  emoji_text           object
 16  emoji_count          int64
 17  publish_date         datetime64[ns, UTC]
```

Figure 3 – Final dataset columns and data types

After all the datasets were merged, cleaned, and pre-processed a final dataset was exported as parquet file. The data dictionary for the final merged dataset can be found as **Table 1** in the Appendix.

5. Results

5.1. Exploratory data analysis

Exploratory data analysis (EDA) was conducted on the pre-processed data. The EDA process was iterative, with lessons learned during EDA fed back as improvements and modifications to pre-processing. The sections that follow explain the EDA results.

5.2. Multivariate Feature Analysis

Given the high dimensionality of the datasets and importance of contextualizing based on positive/negative class (troll/verified), only multivariate analysis will be highlighted herein.

5.2.1. *Numeric Feature Correlation*

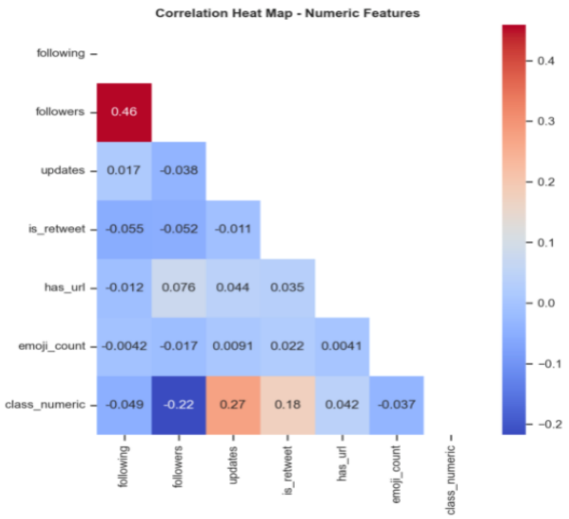


Figure 4 – Correlation Heat Map

The only features identified as moderately correlated (refer to **Figure 4**) were followers and following at 0.46. Class was slightly negatively correlated to followers and weakly positive correlated to updates.

5.2.2. *Accounts per data source*

The data sources were labeled in three categories. The number of accounts within each data source category are visualized in **Figure 5**.

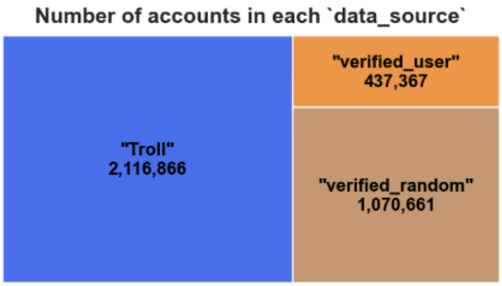


Figure 5 – Accounts per data source

5.2.3. *Following/Followers vs. Account Category*

The features "following" and "followers" were used to examine differences between account categories. This was further expanded to derive a new feature for the ratio of following to followers (discussed in the next section).

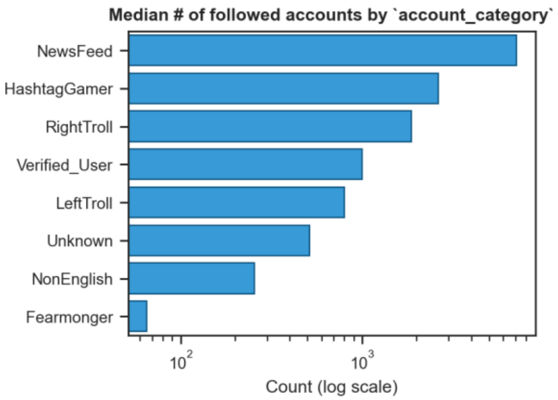


Figure 6 – Median # followed accounts by account category

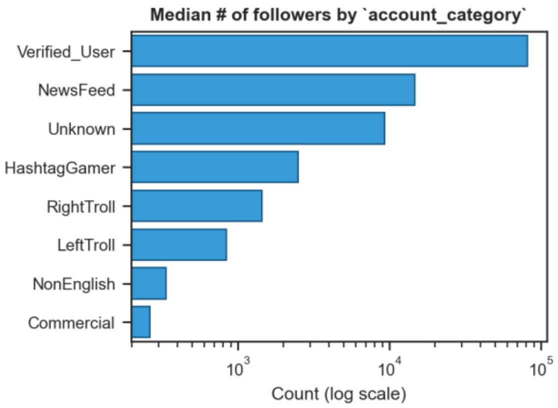


Figure 7 – Median # followers by account category

5.3. Feature Engineering

New features were derived for use in the follow-on ML project to identify key attributes in numeric form.

5.3.1. *Following Ratio*

Using the columns "following" and "followers", the following ratio was derived to show if there was an imbalance specific to troll account. As the graph in **Figure 8** depicts, when compared against the account category, it becomes apparent that verified users tend to have the smallest median following ratio, meaning that they have far more followers than accounts they follow. Whereas some trolls tend to follow more accounts than they have followers.

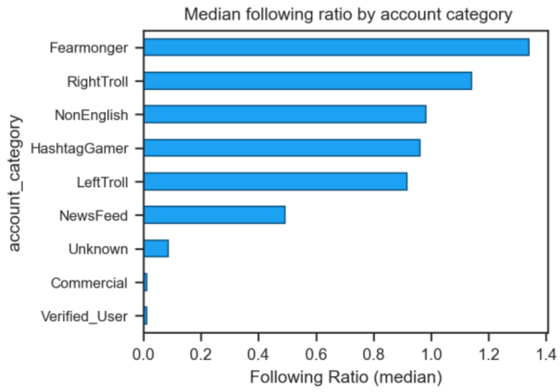


Figure 8 – Median following ratio by account category

5.3.2. *Emoji Count / Emoji Text*

Using the emoji count and emoji text columns derived in pre-processing, it was possible to determine the most frequently used emojis for the entire dataset, and for subsets based on data source.

For verified users, the most frequently used emoji was "face with tears of joy", whereas for trolls it was "flag: United States". Because the trolls used emojis more frequently than verified users, this same emoji was the most frequently used across the entire dataset. There are many overlapping emojis used by both groups, but it is evident that the American Flag played a larger role in the context of tweets made by trolls, which makes sense given the agenda of influencing American politics.

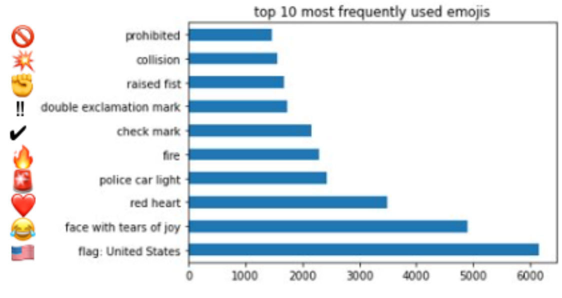


Figure 9 – Top 10 emojis (entire dataset)

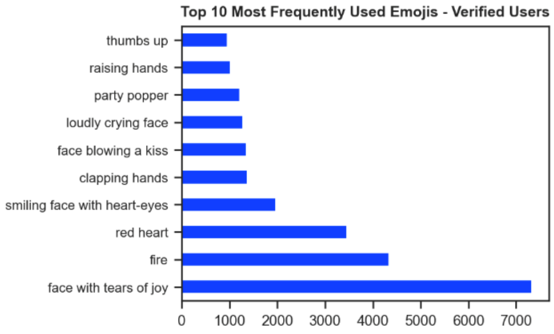


Figure 10 – Top 10 emojis (verified users)

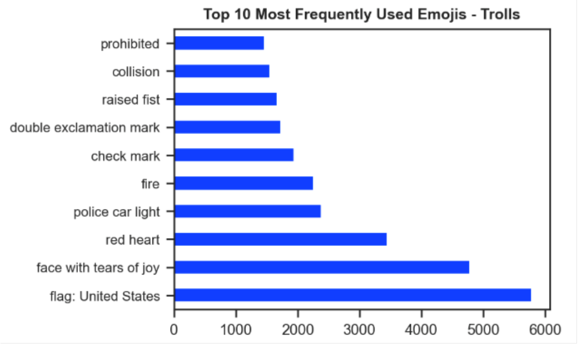


Figure 11 – Top 10 Emojis (trolls)

5.3.3. *Region*

The region feature contained both a significant number of null values as well as a significant number of unique values. It's not uncommon for a tongue-in-cheek value to be entered by a user, e.g., NASA's account likes to list "Pale Blue Dot" in reference to the view of Earth from a 1990 photograph taken from space by the Voyager 1 space probe (12).

Among the troll data, the region value was determined by a commercial tool from Salesforce called Social Studio (13).

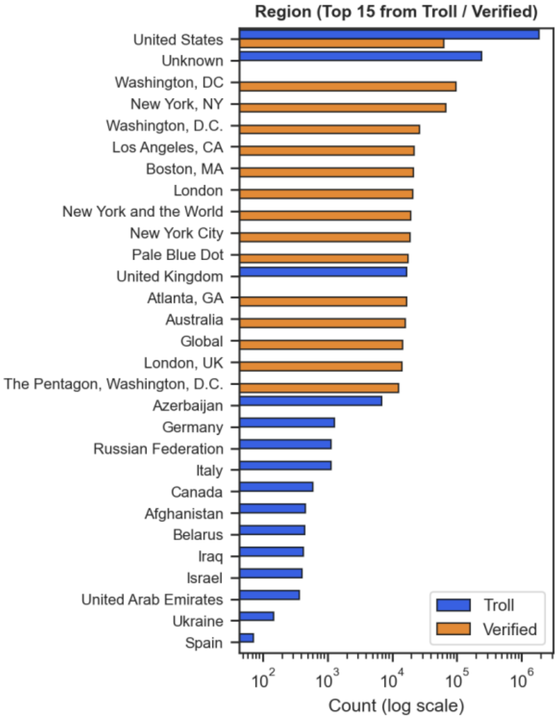


Figure 12 – Region vs. Class

Looking at account handles vs. region and data source (**Figure 13**) we primarily see troll handles, as these users had the highest count of tweets/author. The only verified account to rank high enough to be in the top 10 is @AskTSA, the U.S. Transportation Security Administration's account for asking travel questions.

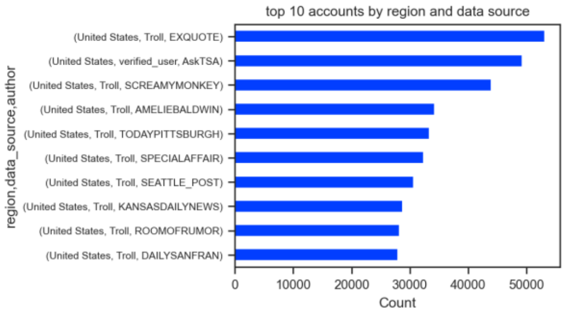


Figure 13 – Top 10 accounts by region and data source

5.3.4. *Generic Tweet vs. Retweet*

The binary column "is_retweet" was built from existing labels (for troll tweets from troll dataset) and tweet object metadata (for verified tweets from Twitter

API). The troll dataset had a higher percentage of retweets **Figure 14**, but both datasets are more heavily comprised of original tweets.

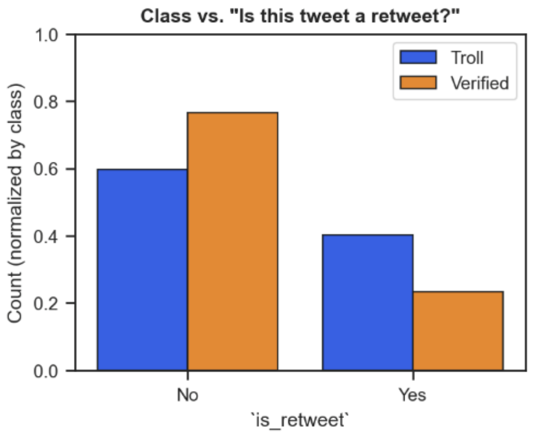


Figure 14 – Retweet vs. Class

5.3.5. *Russian Alphabet Letters*

A function was created to clean a tweet (hashtags, links, retweets) into just pure text so that it could then be analyzed at the character level for Russian alphabet characters. The regex “findall” function was used with the character set `[\u0400-\u04FF]`, which represents the Russian alphabet start and finish characters.

After getting a count of Russian alphabet characters, a percentage of total characters was calculated. Total percentage was bracketed at 25% to further analyze. What was found that several tweets labeled as English language, were from account category = ‘NonEnglish’, so additional cleaning was done. Accounts with > 25% Russian characters were plotted against account category and region in **Figure 15**.

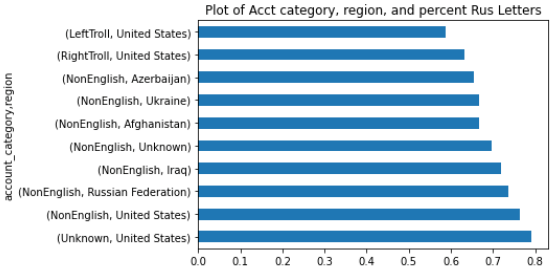


Figure 15 – Account category, region, and percent Russian letters/tweet

5.4. NLP

With tokenizing the tweets, deeper analysis was done to compare the verbiage of the different twitter users. This included comparing word lengths, tweet lengths, frequent words, and n-grams.

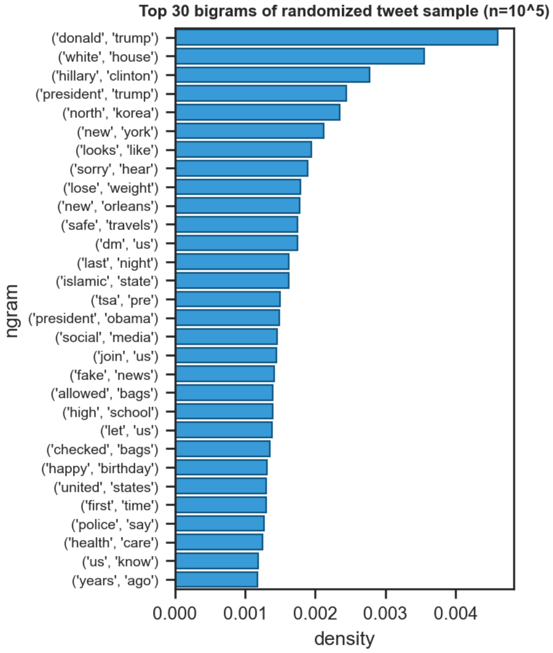


Figure 16 – Top 30 bigrams observed in a randomized sample of $n=10^5$ tweets, ranked by density

5.4.1. *Tweet Length*

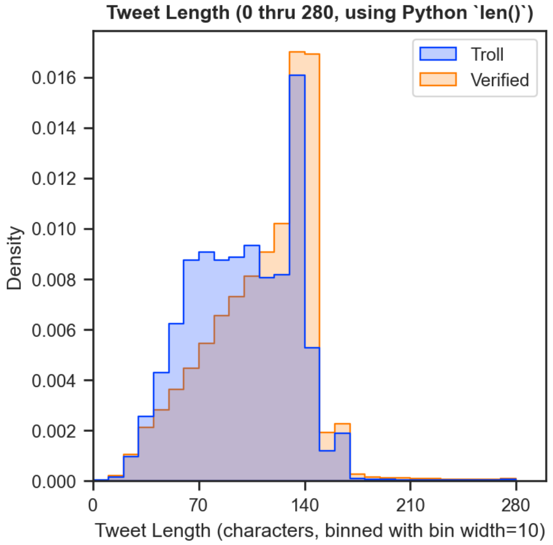


Figure 17 – Tweet length by class

The graph in **Figure 17** shows the distribution of tweet length for the entire dataset among each of our two classes. There appears to be some difference in the range of (40, 110) characters where troll tweets tend to be more prevalent.

An additional caveat exists for tweet length: Twitter uses a distinct method for counting the characters of a tweet to determine its length. Twitter's length method gives allowances for URLs, emoji characters, and non-Latin characters. The graph in **Figure 17** uses Python's built-in *len* function to calculate tweet length, so it is not completely representative of how Twitter would measure the length of a tweet.

Twitter's character counting method was simulated using an external library ("tweet_counter") and used to visualize a new distribution of tweet lengths. This distribution is compared against the same tweets measured by Python *len* in **Figure 21** in the Appendix.

5.4.2. Word Cloud

Comparing the words users frequently chose was challenging to display in a digestible way on conventional graphs. Word clouds were implemented to have a nice visualization of the verbiage being used.

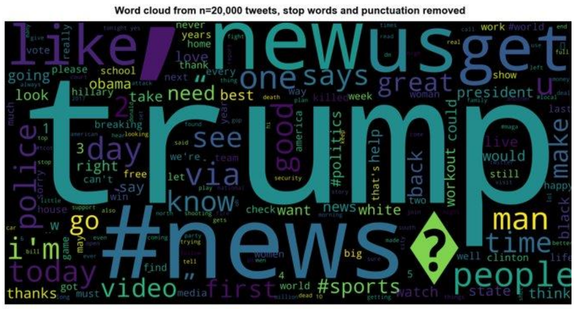


Figure 18 – Word cloud of a randomized sample of n=20,000 tweets, with stop words and punctuation removed. A higher resolution rendition of this word cloud is in the Appendix as **Figure 22**.

5.5. Discussion

Originally obtained was a dataset with 2.9 million rows. After implementing a data acquisition once more tweets were pulled from Twitter's API, the dataset has expanded to 3.6 million tweets. With the magnitude of the dataset, efficient methods have been utilized to store our data in the best memory cost effective way. Using Parquet files has increased the speed of our processes, which will help optimize the different ML models when added later.

Tokenization and pre-processing have also transformed our data into an acceptable structure when the implementation of ML models is applied. From EDA, we were able to gain more insight into what demonstrates a fake account and tweet and derive additional features that will hopefully be useful in implementing machine learning (ML) models to classify accounts/tweets as real/fake. Overall, the project has moved into the right direction to create fully optimized ML models.

5.5.1. Account Verification Validity

As noted in section 1.4.2, Twitter's policy for account verification has recently changed. As has been demonstrated, a foundational assumption of our study is that a tweet obtained from a verified account is authentic and able to be labeled as such. This begs the question of whether this assumption is still valid.

Our assessment on this assumption is this: even though portions of our data were obtained recently (immediately prior to the policy change), the period during which our data was generated does not overlap with the revised policy. We can verify (using archived versions of the verification policy pages from Twitter's website) that the policy was consistent with how we have presented it both during the time period we acquired data (Sept/Oct 2022) (14) and during the time -period we are studying (years 2013-2017) (15).

Based on this, we conclude our original assumption is valid. Further, despite the uncertain future of the Twitter verification program or its policies, we believe there are still valuable insights to be gained by undertaking this study on the acquired and pre-processed data we have herein described.

6. Conclusion

The aim of the project thus far has been to acquire, pre-process, and analyze the target dataset(s) in preparation for applying additional NLP methodologies and ML models. Given the results thus far, we plan to expand upon the existing NLP analysis to implement additional quantitative methods that will enable us to derive additional text features for an applicable ML model. We successfully acquired, pre-processed, and analyzed over 3.6 million tweets from 2013-2017 utilizing the Twitter API and academic license. Utilizing our methodology for identifying verified accounts from the same time period of the 538 dataset, we were able to capture a baseline control dataset of equivalent size and context.

7. Future works

Planned topics of expansion include further analysis of hashtags, gifs, and hyperlinks in the context of NLP; quantifiable features such as sentiment, TF-IDF, and tweet length; and implementation of ML/deep learning models such as word embeddings to find word similarity with embeddings like BERT, GloVe, and word2vec. Since we found that our data can deal with multi-lingual expressions, one BERT alternative coined “TwHIN-BERT” is a BERT pre-trained word embedding on 7 billion Tweets in 100 different languages, which can be useful in sourcing multi-lingual expressions amongst real and fake Twitter users (16) In addition, we aim to expand our analysis to include probabilistic topic modeling methods in comparing the most frequent topics amongst different users that are classified as either real or fake. Older topic modeling methods like LDA, LSA, and LSI can be implemented, with metrics like perplexity and coherence calculations. An example of newer topic modeling methodologies that can be explored are the family of neural topic models, in which the AI community recently aimed to bridge deep learning with existing topic modeling methods (17) .

We also aim to implement Named-Entity Recognition (NER) to find the most prevalent named entities within the set of Tweets that we have obtained using open-source Python libraries like spaCy. NER could potentially be used to identify key elements in the text after we do a preliminary topic modeling analysis, which could glean more insight as to what accounts are specifically talking about. Before implementing NER models, we will require part-of-speech (POS) tagging to classify part-of-speech for each word. One limitation is that this could take time considering the size of our dataset, so that will be considered.

The second installation of this project will involve applied machine learning models to classify tweets/accounts as real or fake. This can be implemented using a Bag-of-Words model and then training a Naïve Bayes, Logistic Regression, Decision Tree, and/or Support Vector Machine (SVM) models, depending on our use-case.

8. References

1. *Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations.* **Starbird, Kate, Arif, Ahmer and Wilson, Tom.** Nov 2019, 2019, J Proc. ACM Hum.-Comput. Interact., Vol. 3, pp. 1-26.

2. **Zelenkauskaite, Asta.** *Creating Chaos Online: Disinformation and Subverted Post-Publics.* [prod.] University of Michigan Press. Ann Arbor, MI : University of Michigan Press, 2022. Vol. https://doi.org/10.3998/mpub.12237294. EPUB.

3. **Office of the Director of National Intelligence.** *Assessing Russian Activities and Intentions in Recent US Elections.* 06 January 2017.

4. **Chen, A.** The Agency. [Online] 06 07, 2015. https://www.nytimes.com/2015/06/07/magazine/the-agency.html?_r=0.

5. **Elliott, Chris.** The readers' editor on… pro-Russia trolling below the line on Ukraine stories. [Online] May 04, 2014. https://www.theguardian.com/commentisfree/2014/may/04/pro-russia-trolls-ukraine-guardian-online.

6. **Policy, Twitter Public.** Update on Twitter's review of the 2016 US election. [Online] January 19, 2018. https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html.

7. **Roeder, Oliver.** Why We're Sharing 3 Million Russian Troll Tweets. [Online] FiveThirtyEight, JUL 31, 2018. [Cited: DEC 03, 2022.] https://fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/.

8. **Twitter.** Legacy verification policy. [Online] November 09, 2022. https://help.twitter.com/en/managing-your-account/legacy-verification-policy#:~:text=This%20page%20documents%20the%20criteria,badged%20prior%20to%20this%20date..

9. **fivethirtyeight/russian-troll-tweets.** Github. [Online] Aug 27, 2018. https://github.com/fivethirtyeight/russian-troll-tweets/.

10. **Twitter.** About Twitter Verified Accounts. [Online] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts#requirements.

11. —. Tweet Downloader. *Twitter API Tools.* [Online] https://developer.twitter.com/apitools/downloader.

12. **Pale Blue Dot. *Wikipedia.* [Online] [Cited: December 8, 2022.] https://en.wikipedia.org/wiki/Pale_Blue_Dot.**

13. **Salesforce. Region classification for sources and posts in Social Studio. *Salesforce.com Help.* [Online] [Cited: December 8, 2022.] https://help.salesforce.com/s/articleView?id=000386921&type=1.**

14. **Twitter. About Verified Accounts (Web Archived Version). *Twitter Help Center.* [Online] September 01, 2022. https://web.archive.org/web/20220901050421/https:/help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.**

15. **—. About verified accounts (Web Archived Version). *Twitter Help Center.* [Online] December 31, 2017. https://web.archive.org/web/20180101035730/https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.**

16. *TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations.* Zhang, Xinyang, et al. s.l. : arXiv, 2022.

17. *Topic Modelling Meets Deep Neural Networks: A Survey.* Zhao, He, et al. s.l. : arXiv, 2021.

18. Linvill, D.L and Warren, P.L. Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building. [Online] http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf.

19. Tsukayama, Hayley. Twitter is officially doubling the character limit to 280. *The Washington Post.* November 7, 2017.

Appendix

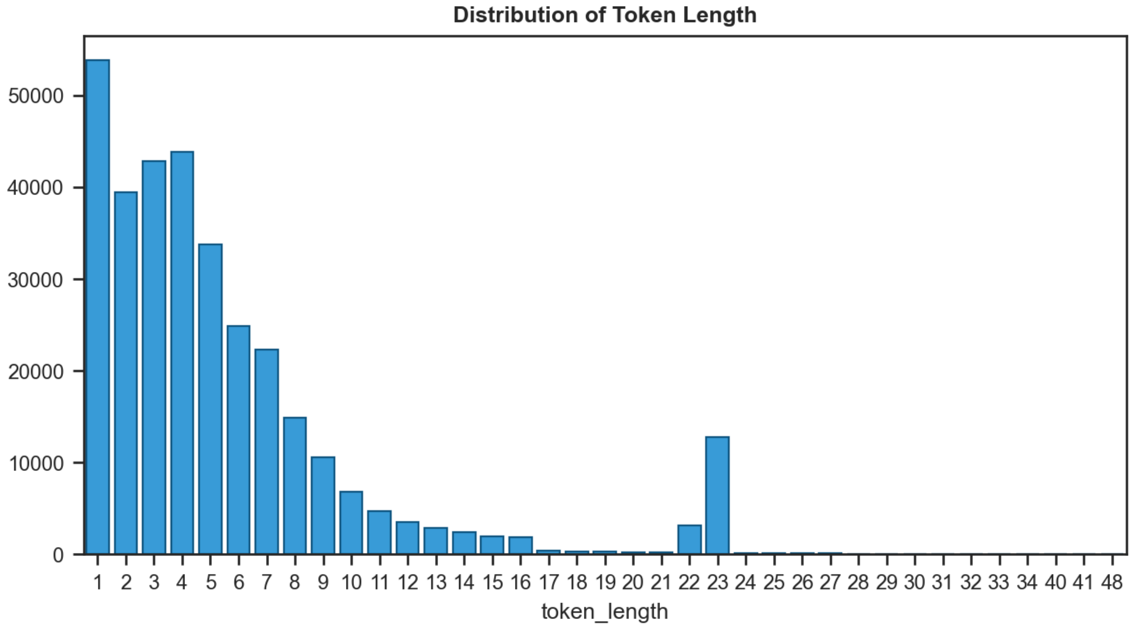


Figure 19 – Distribution of token length (characters) observed in a randomized sample of *n=20,000* tweets.

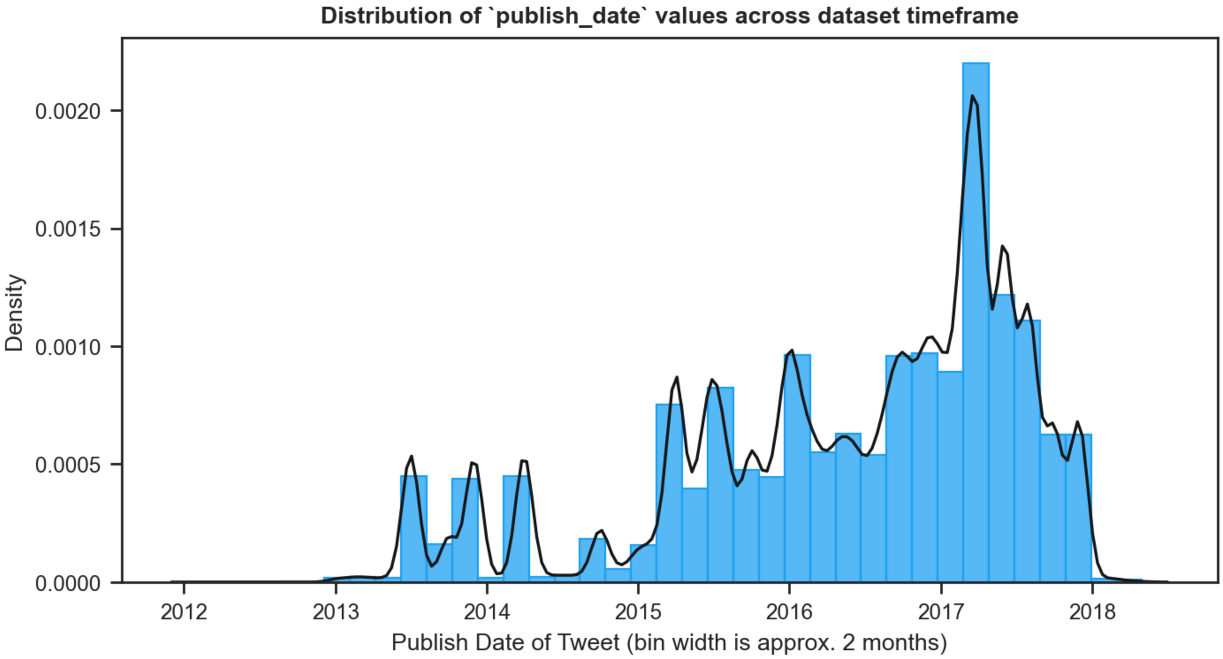


Figure 20 – Distribution of publish date (bin width = ~2 month) observed in a randomized sample of *n=20,000* tweets.

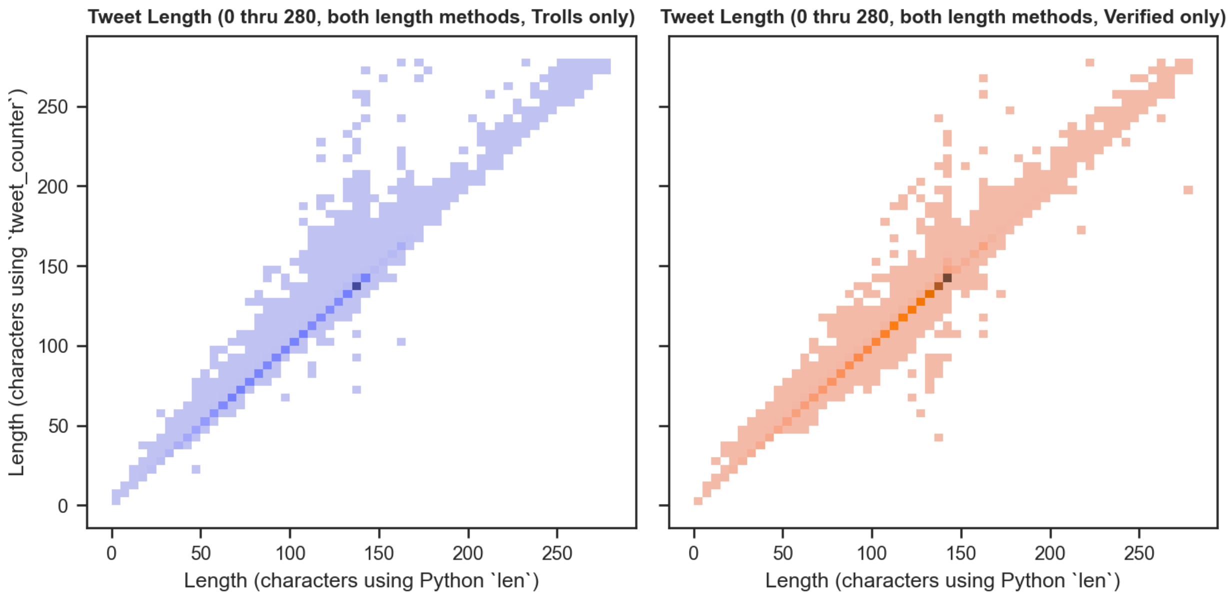


Figure 21 – Comparison of distributions of tweet length when calculated by either the Python *len* function or an approximation of Twitter's distinct character counting method. The distribution for "Verified only" tweets appears to stay closer to the diagonal (indicating the two methods generally agree more). By contrast, the "Trolls only" distribution appears to skew upward from the diagonal, suggesting the Twitter method considers tweets to be longer than their simple Python *len* character count.

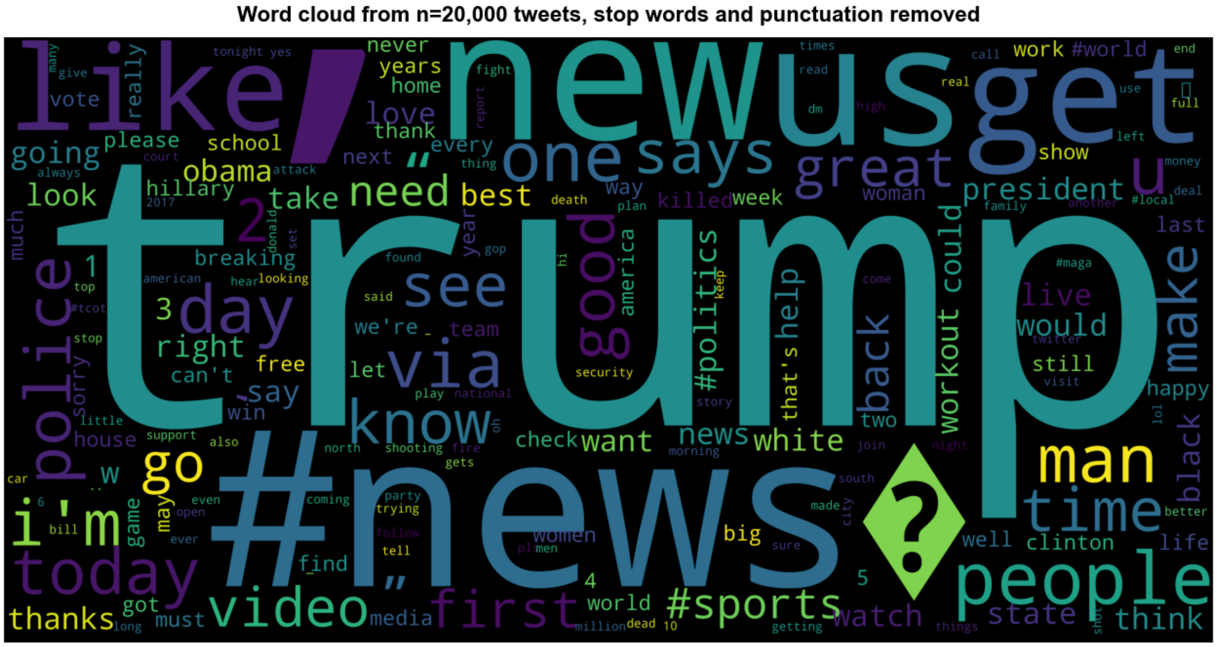


Figure 22 – A higher resolution rendition of the word cloud previously shown in **Figure 18**.

Table 1 – Data Dictionary

Feature	Description	Data Type	Notes
external_author_id	Author ID	Nominal	Unique identifier / primary key for twitter accounts
author	Author Twitter Handle	Categorical	The Twitter handle (i.e. username) of this tweet's author
content	Tweet Content	String Object	The text of the tweet itself, including hashtags, hyperlinks, mentions, etc.
region	Region	Categorical	A region classification. For troll tweets, determined by Social Studio
language	Language	Categorical	The language of the tweet
following	Number of Other Accounts Followed	Continuous Numerical	The number of other accounts that this tweet's account is following.
followers	Number of Other Accounts Following	Continuous Numerical	The number of other accounts that follow this tweet's account.
updates	Number of Update Actions	Continuous Numerical	A composite measure of public metrics for interaction by other users with a tweet.
is_retweet	Retweet	Binary	Binary indicator of whether or not the tweet is a retweet.
account_category	Account Category	Nominal	A categorization, primarily for differentiating types of troll accounts.
tweet_id	Tweet ID	Nominal	Unique identifier for each tweet, assigned by Twitter.
tco1_step1	First URL from within Tweet Content	Nominal	The first URL redirected to after http://t.co/...
data_source	Data Source	Categorical	Either "Troll", "verified_user", or "verified_random"
has_url	Tweet Has URL	Binary	Does this tweet contain a URL?
emoji_text	Emoji Text	String / List of Strings	A list of natural language descriptions of each unique emoji used.
emoji_count	Number of Emoji	Continuous Numerical	The quantity of emoji used within the `content` of this tweet.
publish_date	Publish Date of Tweet	Interval	The original publish date of this tweet.

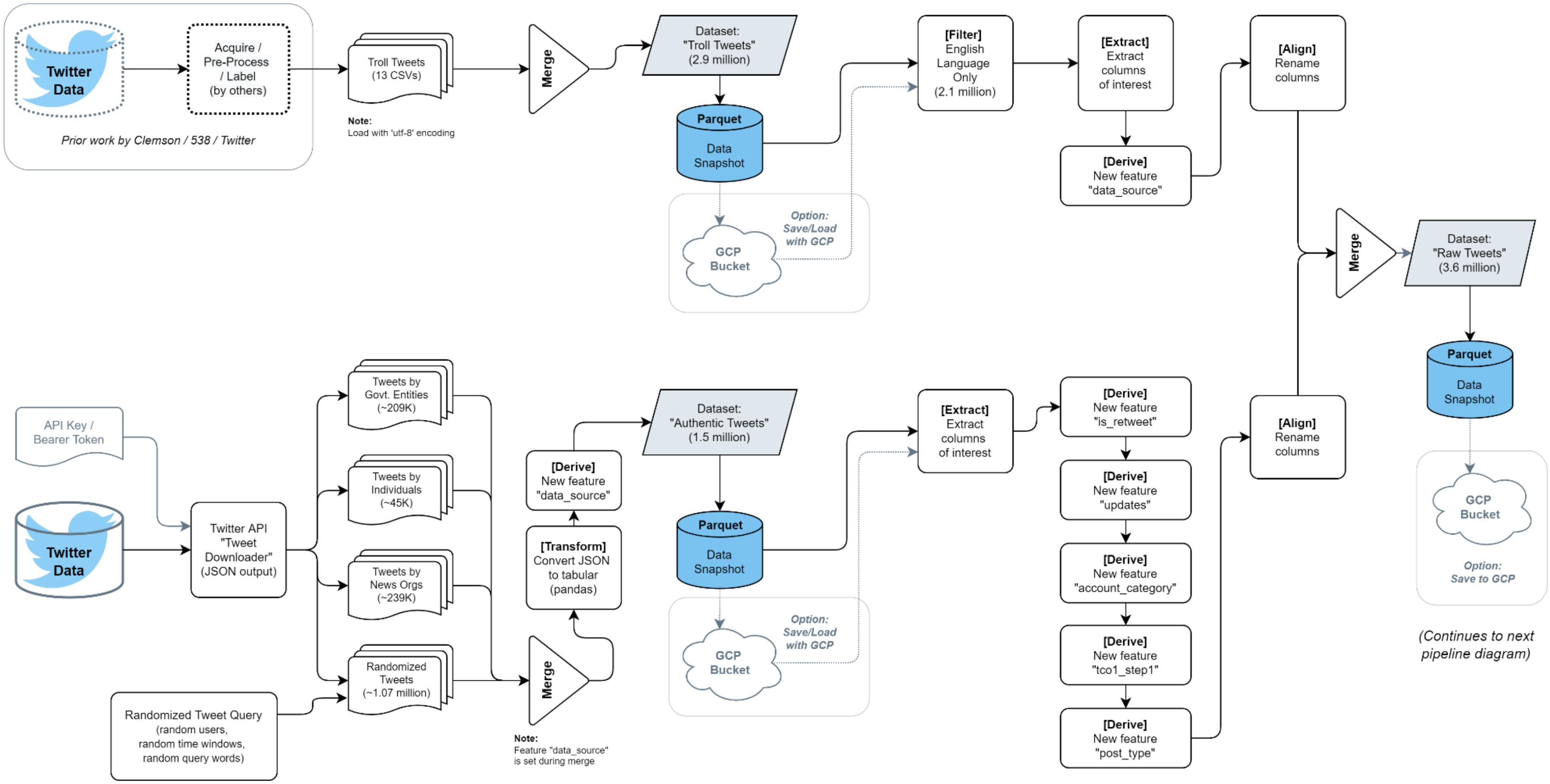

Twitter Data
Acquire / Pre-Process / Label (by others)
Prior work by Clemson / 538 / Twitter
Troll Tweets (13 CSVs)
Note:
Load with 'utf-8' encoding
Merge
Dataset: "Troll Tweets" (2.9 million)
Parquet
Data Snapshot
GCP Bucket
Option: Save/Load with GCP
[Filter] English Language Only (2.1 million)
[Extract] Extract columns of interest
[Derive] New feature "data_source"
[Align] Rename columns
Merge
Dataset: "Raw Tweets" (3.6 million)
Parquet
Data Snapshot
GCP Bucket
Option: Save to GCP
(Continues to next pipeline diagram)
API Key / Bearer Token
Twitter Data
Twitter API "Tweet Downloader" (JSON output)
Randomized Tweet Query (random users, random time windows, random query words)
Tweets by Govt. Entities (~209K)
Tweets by Individuals (~45K)
Tweets by News Orgs (~239K)
Randomized Tweets (~1.07 million)
Merge
Note:
Feature "data_source" is set during merge
[Transform] Convert JSON to tabular (pandas)
[Derive] New feature "data_source"
Dataset: "Authentic Tweets" (1.5 million)
Parquet
Data Snapshot
GCP Bucket
Option: Save/Load with GCP
[Extract] Extract columns of interest
[Derive] New feature "is_retweet"
[Derive] New feature "updates"
[Derive] New feature "account_category"
[Derive] New feature "tco1_step1"
[Derive] New feature "post_type"
[Align] Rename columns

Figure 23 – Data Acquisition and Transformation Pipeline (1 of 2)

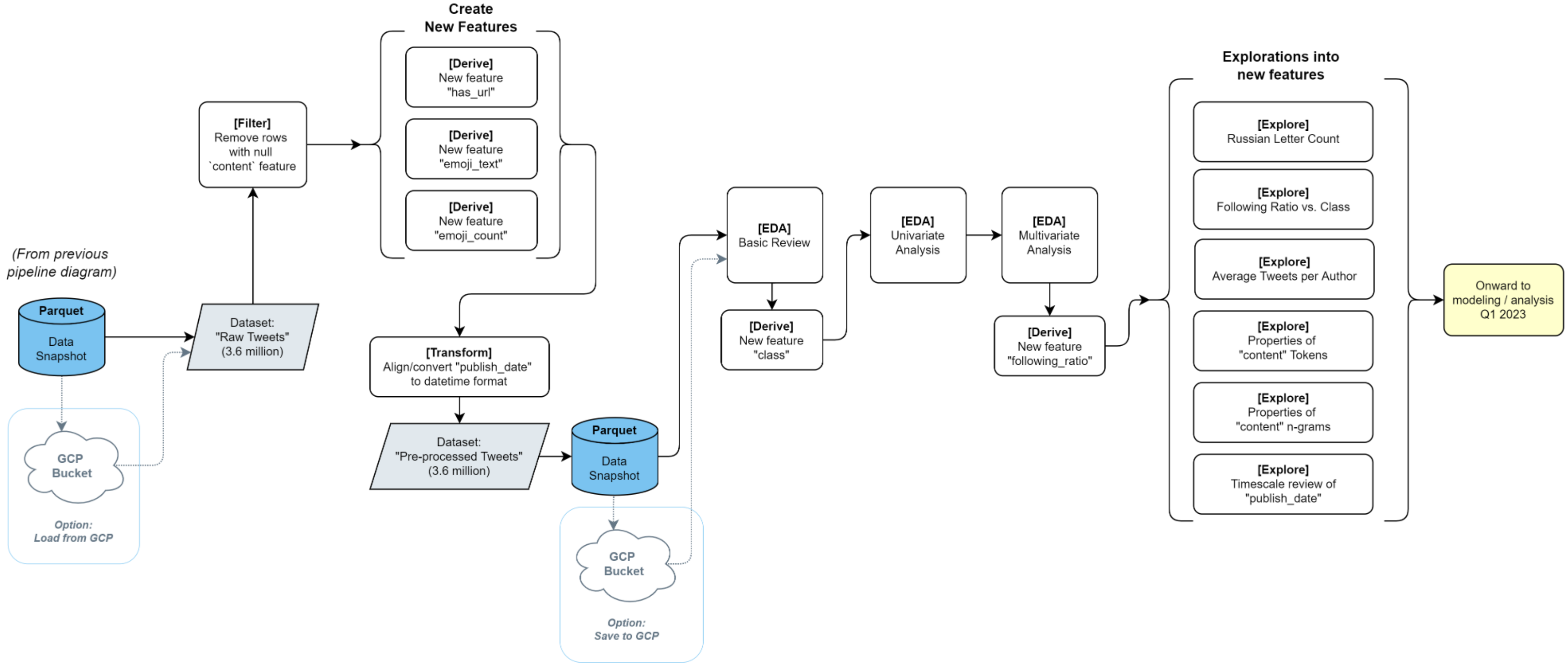

(From previous pipeline diagram)
Parquet
Data Snapshot
GCP Bucket
Option: Load from GCP
Dataset: "Raw Tweets" (3.6 million)
[Filter] Remove rows with null `content` feature
Create New Features
[Derive] New feature "has_url"
[Derive] New feature "emoji_text"
[Derive] New feature "emoji_count"
[Transform] Align/convert "publish_date" to datetime format
Dataset: "Pre-processed Tweets" (3.6 million)
Parquet
Data Snapshot
GCP Bucket
Option: Save to GCP
[EDA] Basic Review
[Derive] New feature "class"
[EDA] Univariate Analysis
[EDA] Multivariate Analysis
[Derive] New feature "following_ratio"
Explorations into new features
[Explore] Russian Letter Count
[Explore] Following Ratio vs. Class
[Explore] Average Tweets per Author
[Explore] Properties of "content" Tokens
[Explore] Properties of "content" n-grams
[Explore] Timescale review of "publish_date"
Onward to modeling / analysis Q1 2023

Figure 24 – Data Acquisition and Transformation Pipeline

Table 2 – List of accounts used for “verified” Twitter dataset (1 of 2)

Account Type	Account Handle	Account Name	# Tweets Pulled	Timeframe
Verified News Organizations	abcnews	ABC News	239,165 *(Total across all news orgs)*	October 2013 October 2014 October 2015 October 2016 October 2017 (Applies to all news orgs listed)
	AP	Associated Press		
	TheAtlantic	The Atlantic		
	business	Bloomberg		
	nytimes	The New York Times (News)		
	NPR	NPR		
	propublica	ProPublica		
	TIME	TIME		
	USATODAY	USA Today		
	axios	Axios		
	BBCWorld	BBC		
	csmonitor	The Christian Science Monitor		
	Forbes	Forbes		
	Newsweek	Newsweek		
	Reuters	Reuters		
	RealClearNews	Real Clear Politics		
	thehill	The Hill		
	WSJ	The Wall Street Journal (News)		
	FoxBusiness	Fox Business		
	TheIJR	Independent Journal Review		
Account Type	**Account Handle**	**Account Name**	**# Tweets Pulled**	**Timeframe**
Verified Prominent Individuals	MittRomney	Mitt Romney	289	October 1, 2013 – Dec 31, 2017
	MichelleObama	Michelle Obama	70	
	BarackObama	Barack Obama	7,027	
	HillaryClinton	Hillary Clinton	9,825	
	SarahPalinUSA	Sarah Palin	8,566	
	Mike_Pence	Mike Pence	3,936	
	elonmusk	Elon Musk	3,115	
	billgates	Bill Gates	1,800	
	CondoleezzaRice	Condoleezza Rice	308	
	DonaldJTrumpJr	Donald J Trump Jr	10,362	

*(Table continues onto next page in **Table 3**)*

Table 3 – List of accounts used for "verified" Twitter dataset (2 of 2)

Account Type	Account Handle	Account Name	# Tweets Pulled	Timeframe
Verified Government Organizations	LibraryCongress	Library of Congress	208,771	October 1, 2013 – Dec 31, 2017
	CDCgov	Centers for Disease Control		
	HHSgov	Dept of Health and Human Services		
	StateDept	Dept of State		
	TheJusticeDept	Dept of Justice		
	US_FDA	Food and Drug Administration		
	FBI	Federal Bureau of Investigation		
	NIH	National Institute of Health		
	askTSA	Transportation Safety Administration		
	NSF	National Science Foundation		
	NASA	National Aeronautics and Space Administration		
	smithsonian	Smithsonian Institute		
	USDA	Dept of Agriculture		
	FCC	Federal Communications Commission		
	PeaceCorps	Peace Corps		
	EPAgov	Environmental Protection Agency		
	USGS	US Geological Survey		
	FEMA	Federal Emergency Management Agency		
	USARMY	US Army		
	USNWSgov	National Weather Service		
	DeptofDefense	Dept. of Defense		
Account Type	**Account Handle**	**Account Name**	**# Tweets Pulled**	**Timeframe**
Randomized Tweets from Verified Accounts	*(various)*	*(various)*	95,432	2014-04-23 08:00 AM to 2014-04-23 12:00 PM
			99,835	2017-12-03 12:00 AM to 2017-12-03 04:00 AM
			130,685	2015-04-26 08:00 PM to 2015-04-27 12:00 AM
			114,964	2017-03-20 12:00 AM to 2017-03-20 04:00 AM
			93,315	2013-12-13 08:00 AM to 2013-12-13 12:00 PM
			180,535	2017-06-29 08:00 PM to 2017-06-30 12:00 AM
			183,010	2017-04-14 12:00 PM to 2017-04-14 04:00 PM
			76,445	2016-01-19 08:00 AM to 2016-01-19 12:00 PM
			96,440	2013-07-06 12:00 AM to 2013-07-06 04:00 AM
		Total Tweets (Table 2 and Table 3)	**1,563,587**	
		Total Accounts (Table 2 and Table 3)	**~140k**	