Fake and True News on Twitter During the 2016 US Election

Submitted for the Master of Science, London School of Economics, University of London

Data Science
Department of Statistics, 2019

Supervisor: Yining Chen, Department of Statistics
Partner: Pablo Barberá, Department of Methodology

Candidate Numbers:

22454,19132

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Table of Contents

List of Figures

List of Tables

I. Acknowledgement

The authors would like to convey gratitude towards the department of statistics for the opportunity to apply data science techniques to a practical real-life project, our capstone partner Pablo Barberá for his generous guidance on scoping the project, interpreting the results, and providing the dataset, our supervisor Yining Chen for his advise on sampling methods for prediction, and our program director Milan Vojnovic for his support throughout the course.

II. Executive Summary

Problem and Significance

Despite of a long history and a broad coverage of topics, fake news only came to public attention during the 2016 US election. It has been shown that links to fake news were shared more often on Twitter during the election than popular credible media outlets, such as the New York Times, Fox News, and the Washington Post, combined (Barberá, 2018). Fake news also diffused significantly farther, faster, deeper, and more broadly than traditional news (Vosoughi, Roy, and Sinan, 2018).

This paper focuses on fake news on Twitter during the 2016 US election. The subject is inherently political. However, the stakes are much bigger than politics. Our ability to vet information matters every time a mother asks Google whether her child should be vaccinated, every time a student encounters a Holocaust denial on Twitter, and every time a customer is convinced that climate change is not real. Misinformation online has the power to set social norms and create social reality. It is therefore important that we acknowledge the impact of fake news and prevent its further spread.

Objective

To prevent the spread of misinformation, we need to first detect it. Detection of fake news can be model-oriented or feature-oriented (Shu et al., 2017) (while there are some that combine both approaches (Monti et al, 2019; Volkova et al, 2017)). This paper focuses on feature-oriented detection, because we believe that it will help researchers interpret the underling characteristics that make fake news different from traditional news better than more complicated models. In particular, we will focus on both linguistic and user features because they are directly available from Twitter data. We will first conduct data analysis on linguistic and user attributes to extract important features, and then perform a classification task on labeled original tweets using these features.

Definition

The definition of fake news varies in the literature. This paper follows the definition by Allcott, Gentzkow, and Yu (2019): *news articles that are intentionally and verifiably false, and could mislead readers*. Using a list of 95 fake news domains and 116 traditional news domains subset from the list compiled by Allcott, Gentzkow, and Yu (2019), we classify tweets into two classes—"fake" or "true;" a tweet that contains an url pointing to a fake news domain is classified as "fake," and a tweet that contains an url pointing to a

traditional news domain is classified as "true" as the comparison class. We will use "fake tweets" and "true tweets" to describe original tweets that are classified as "fake" and "true" respectively in the rest of paper.

Dataset

This paper uses a Twitter dataset collected by Barberá (2018) from Oct 12th to Dec 3rd, 2016—a month before and a month after the 2016 US election day. It contains tweets that mentioned presidential candidates Hillary Clinton and Donald Trump. We processed around 220 GB of JSON files on Google's distributed computing platform using PySpark.

This paper focuses on original tweets that contained external links, retweets (reposts of original tweets), replies (comments under tweets), and quotes (reposts with comments). Original tweets are classified into two groups—fake and true, while retweets, replies, and quotes are classified based on the class of the original tweets that they were responding to. This paper aggregates these four types of tweets into two categories: posts (originals and retweets) and comments (replies and quotes). Posts are tweets that share and spread news, while comments are tweets that respond in words to the news. The rationale behind the categorization is to explore distinct characteristics between the two types of engagement with news on Twitter. Those that share and retweet a piece of news are more likely to be in support of the information, while those that comment on a piece of news are more likely to have diverse opinions about it.

In total, we will be analyzing around 7 million posts and 0.6 million comments to these posts during the two-month period.

Methods

On linguistic level analysis, a text is represented as a bag of words—a vector of words as features. Document-Feature Matrices (DFMs) are constructed from these corpora, with normalization and non-ASCII-encoded words (i.e. non-english characters), user mentions, numbers, punctuations, symbols and urls removed whenever appropriate. Dictionary methods are used to conduct sentiment and emotion analysis. We will use two-sample Kolmogorov-Smirnov tests to compare the distributions of proportions of words from different categories between classified fake and true tweets. We will calculate relative frequency to identify keywords in comments, based on which we will construct a lexicon to tag debunking and questioning signals.

On user level analysis, we build a logistic regression model for a classification task framed as whether a user shares fake posts as a function of user level features. We also build a log linear model to estimate the number of fake posts shared by a user.

Once we have extracted important features, we use five models (logistic regression, decision tree, Random Forest, Xgboost, and LSTM) to evaluate different combinations of features. We treat the task of predicting if an original tweet shares fake news or not as a classification problem, with an original tweet as an input and the class of the tweet as the target variable (fake or true).

Main Results and Implications

Our linguistic analysis shows that fake tweets that mentioned Clinton are consistently more negative than true tweets before the election date (Figure 4). Fear is the emotion that distinguishes fake and true tweets the most, with fear more prevalent in fake tweets (Table 1, Figure 5). However, this sentiment and emotion pattern disappears in comments to these tweets (Table 2, Figure 6). Keywords like “lie,” “hoax,” and “is this true” are detected in comments responding to fake tweets, while those to true tweets are more pertinent to the topics mentioned (Table 3). This is a sign that there are more debunking and questioning activities in comments to fake tweets than true tweets. Using words like “hoax,” “this is fake,” and “factcheck” and “is this true,” we tag debunking and questioning signals in comments to fake and true tweets (Table 4). About 1.38% of comments to fake tweets contain any debunking words in our lexicon and 0.45% contain any questioning words, while only 0.63% of comments to true tweets are debunking tweets and 0.12% are questioning tweets. Furthermore, after aggregating tweets on the url level, 8 out of the top 10 fake news stories contain debunking comments and 7 out of top 10 fake news stories contain questioning comments. The temporal patterns of these debunking activities and questioning activities show that questioning generally happens as soon as a post is posted and debunking happens in the first 12 hours (Figure 7). The implication of this finding is that Twitter as a community collectively fact check information on the platform. If we could utilize this crowd intelligence, we might be able to improve fake news detection. We will therefore evaluate the proportion of sentiment and emotional words in the text of a post as its linguistic features as well as the temporal debunking and questioning signals of comments to the post as debunking and questioning features for prediction.

Our user level analysis shows that fake posts are shared by a small proportion of individuals than true posts across the election period. And posts mentioning Trump is more

concentrated on a group of users than posts mentioning Clinton (Figure 8). While 0.012% (0.0056%) of users produced half of fake posts mentioning Clinton (Trump), 5.1% (4.0%) of users produced half of true posts mentioning Clinton (Trump). Furthermore, using both logistic regression and log linear regression at the user level, we examine the significance of different user attributes' effects on wether a user shares fake tweets or not and how many fake tweets the user shares (Table 5). We discover that all user attributes selected are significant at the 0.01 significance level. We will therefore use all of them as user features for prediction.

Using different combinations of features for each model, we discover that random forest with user attributes produces the highest prediction accuracy rate of 0.86 (F1 score 0.856) (Table 6, Table 7). In particular, average number of tweets per day posted by a user during the two-month period, the number of public lists that this user is a member of, and the number of tweets (including retweets) issued by the user are among the most important features ranked by random forest. The more that a user tweets per day, while the less the listed count and statuses count, the more likely that the user is going to share fake news. We also discover that if a user uses default profile on Twitter and links their profile to an url, it is more likely that the user shares fake news. This is promising because user attributes are time-invariant and directly available. Social media platforms like Twitter might be able to identify a list of users who have higher potential of spreading fake news. On the other hand, neither linguistic features nor debunking and questioning signals produce good result compared to user attributes. This might be because analyzing linguistic features using dictionary methods is highly variable for short texts like tweets and are sensitive to changes of political rhetorics. Debunking and questioning signals could potentially be promising but future study needs to better capture and represent the signals for machine learning models.

III. Introduction

i. Problem and Significance

Despite of a long history and a broad coverage of topics, fake news only came to public attention during the 2016 US election. It has been shown that links to fake news were shared more often on Twitter during the election than popular credible media outlets, such as the New York Times, Fox News, and the Washington Post, combined (Barberá, 2018). It also diffused significantly farther, faster, deeper, and more broadly than traditional news (Vosoughi, Roy, and Sinan, 2018) by studying rumors on Twitter from 2006 to 2016.

This paper focuses on fake news on Twitter during the 2016 US election. The subject is inherently political. However, the stakes are much bigger than politics. Our ability to vet information matters every time a mother asks Google whether her child should be vaccinated, every time a student encounters a Holocaust denial on Twitter, and every time a customer is convinced that climate change is not real. Misinformation online has the power to set social norms and through which it creates social reality. It is therefore important that we acknowledge the impact of fake news and prevent its further spread, just like we need to prevent the spread of epidemics.

ii. Definition

The definition of fake news varies in the literature. This paper follows the definition by Allcott, Gentzkow, and Yu (2019): news articles that are intentionally and verifiably false, and could mislead readers. It has been found that tweets about fake news often include a link to non-credible news websites (Jang et al, 2018). Therefore, we will classify tweets on the domain level; a tweet that contains a fake news domain is classified as "fake," and tweet that contains a traditional news domain is classified as "true" as the comparison class. The list of fake news and comparison domains will be discussed in the data description section. We will use "fake tweets" and "true tweets" to describe original tweets that are classified as "fake" and "true" respectively in the rest of paper.

iii. Related Work

To prevent the spread of misinformation, we need to first detect it. Detection of fake news can be model-oriented or feature-oriented (Shu et al., 2017) (while there are some that combine both approaches (Monti et al, 2019; Volkova et al, 2017)). Model-oriented approach utilizes advanced and complex models like neural networks or deep learning models to train on available features to optimize prediction performance. For

example, Ma et al (2016) trains Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU) on two fake news datasets on Twitter and Weibo and achieves high prediction accuracy of 0.881 and 0.910 respectively. Feature-oriented approach, on the other hand, focuses on determining and constructing effective combinations of features to be incorporated into less complex supervised classification models. Prediction accuracy is not the main concern. In the literature, the three most commonly studied features are network features (the spreading patterns of fake news), linguistic features (of the news contents themselves or the social media posts about them), and user features (the people that write or spread fake news on social media).

For network features, Vosoughi, Roy, and Sinan (2018) discover that the depth and breath of retweet networks can be used to distinguish fake news and true news, while a recent paper Monti et al (2019) uses a generalization of classical Convolutional Neural Networks (CNN) to graphs of fake news propagation and achieved highly accurate (92.7% ROC AUC) results. For linguistic features, Perez-Rosas et al (2018) evaluates linguistic features Ngrams, Punctuation, Psycholinguistic features (LIWC), Readability, and Syntax with a linear Support Vector Machine classifier and discover that LIWC gives the best accuracy rate (0.74). For user features, Barberá (2018), Guess and Tucker (2019), and Grinberg et al (2019) found that age and political partisanship are the factors most likely to predict the spread of misinformation. It has been shown than older republicans are more likely to spread fake news than their counterparts.

iv. Objectives

This paper focuses on feature-oriented detection, because we believe that it will help researchers interpret the underling characteristics that make fake news different from traditional news better than more complicated models. In particular, we will focus on both linguistic and user features that distinguish fake news from true news because they are directly available from Twitter data.

v. Structure of the paper

This paper will first describe the dataset used and the process of selecting domain list for classification. It will then provide an overview of methods used in both data analysis and prediction. For data analysis, we will first explore the sentiment and emotion of both posts and comments using dictionary methods as well as user attributes of posts. Posts are defined as original tweets and retweets of these tweets. Comments are defined as replies (direct comments under a post) and quotes (retweets with comments) to these

tweets. We will then identify important patterns in both linguistic and user level analysis to select features for prediction, which is framed as a supervised binary classification problem (determining if a tweet contains fake news or true news). We will use logistic regression, decision tree, random forest, extreme gradient boosting (Xgboost), and long short-term memory (LSTM) as our models. For each model, we will compare performance of different combinations of features using prediction accuracy, precision, recall, and F1 score as metrics. Lastly, we will discuss which combination of features and model gives the best performance as well as suggested future work.

vi. Brief Summary of Results

We discover significant sentiment and emotion patterns on posts of fake news and debunking and questioning activities in comments to these posts. We also discover that only a small amount of users spread fake news and that the average number of tweets per day by a user is the most important feature, with those that post more tweets per day more likely to spread fake news. However, only user attributes prove to be consistently helpful in producing high accuracy in all models. Tweets are short texts and their Linguistic features thus have high variability. Adding more features might help with strengthening linguistic cues. Debunking and questioning signals can be better captured by constructing different features or using more advanced time-series models.

IV. Data Description and Preparation

i. Domains

The lists of fake and true news domains are adopted from Allcott, Gentzkow, and Yu (2019). They compile a list of 674 fake news domains from five independent sources—Grinberg et al. (2019), PolitiFact, BuzzFeed, Guess et al. (2018), and FactCheck. This paper further subsets domain names that are listed in more than three sources, resulting in 95 fake news domains in total, including endingthefed and abcnews.com.co (Appendix A.a). Endingthefed held four out of the 10 most popular fake articles on Facebook related to the 2016 U.S. election in the prior three months before the election itself (Townsend, 2016). abcnews.com.co was a fake site created by using website spoofing—creating a website with the intention of misleading readers that the website was created by another legitimate organization. We prioritized precision over recall to achieve a precise coverage of fake news domain rather than a thorough but noisy coverage.

The list of true news domains consists of 116 major and small traditional news sites, including the New York Times and Fox News (Appendix A.b). Note that the underlying assumption that traditional news websites generate verifiably true news content could be problematic. However, the idea is to provide a comparison class to fake news, whose contents are generally different than fake news contents. Traditional news domains included in the list were selected from top sites in Alexa's News category and thus have a wide coverage and withstand the test of time and public taste.

If the domain name of an url is contained in the fake news domain list, it is classified as "fake." If it is contained in the true news domain list, it is classified as "true." If it does not belong to either of the lists, it is unclassified and we do not include them in our analysis.

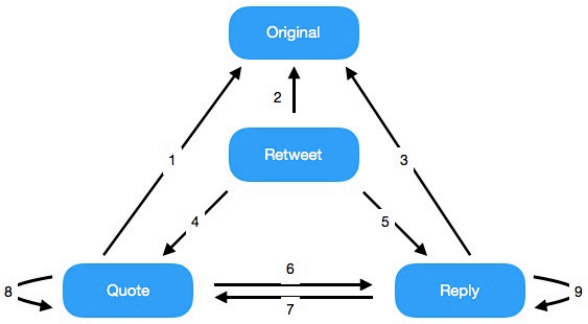


Figure 1: Type 0: original tweet; 1: quote of original tweet; 2: retweet of original tweet; 3: reply of original tweet, and so on. Note that one can also retweet a retweet seen from their feed. However, the intermediate retweet and user information is not stored in the tweet object provided by Twitter (Twitter API, 2019).

ii. Tweets

This paper uses a Twitter dataset collected by Barberá (2018) from Oct 12th to Dec 3rd, 2016—a month before and a month after the 2016 US election day (Nov 8th). It contains tweets that mentioned presidential candidates Hillary Clinton and Donald Trump. In total, this paper processed around 220 GB of josn files on Google's distributed computing platform using PySpark.

Inferring from the Twitter API, there are ten types of tweet objects (Figure 1). This paper focuses on original tweets that contained external links (type 0), retweets (type 2), replies (type 3), and quotes (type 1) to these original tweets. Original tweets are classified into two groups—fake and true—based on the domain names of urls that they contained. Retweets, replies, and quotes are classified based on the class of the original tweets that they were responding to.

This paper aggregates these four types of tweets into two categories: posts (original and retweet) and comments (replies and quotes). Posts are tweets that share and spread news, while comments are tweets that respond in words to the news. The rationale behind the categorization is to explore distinct characteristics between the two types of engagement with news on Twitter. Those that share and retweet a piece of news are more likely to be in support of the information and spread it to their followers. Those that comment on a piece of news are more likely to have diverse opinions on it.

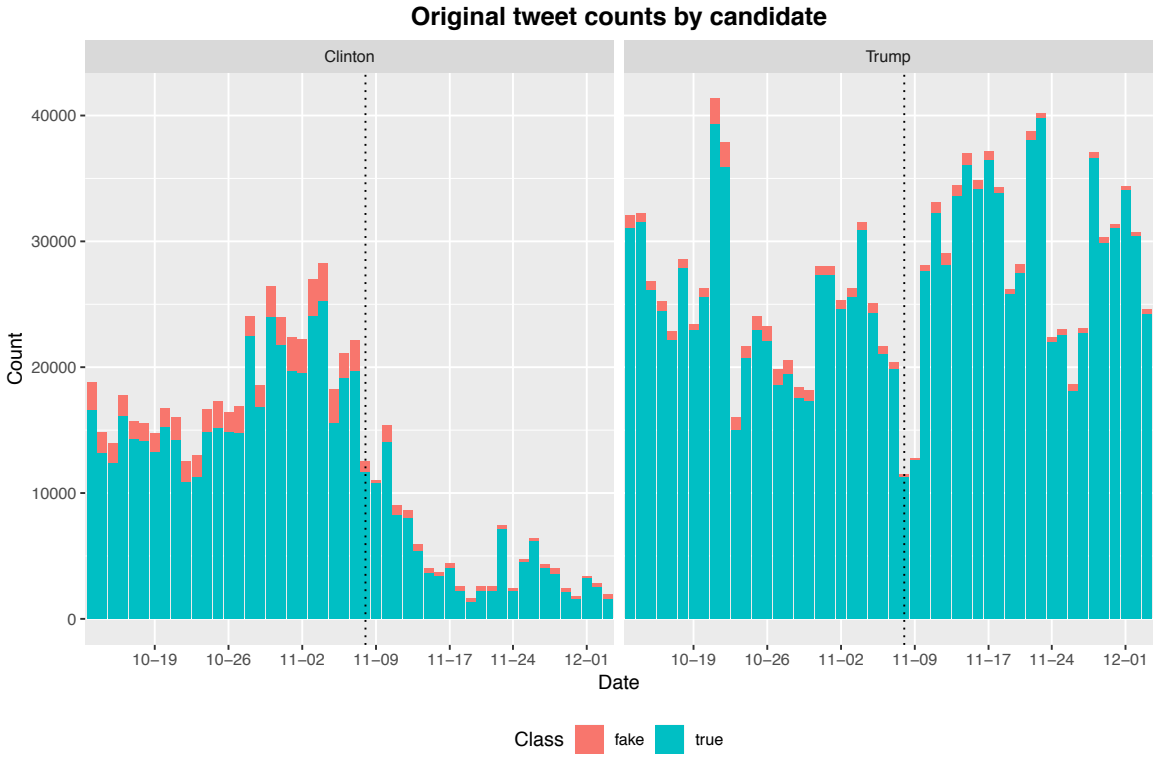


Figure 2: Counts of original tweets by class from Nov 1st to Nov 15th, 2016.

After initial processing, there are 629,085 classified (61,923 fake and 567,162 true) original tweets and 2,000,761 (114,349 fake and 1,886,412 true) retweets for Clinton and 1,382,823 classified (35,117 fake and 1,347,706 true) original tweets and 3,061,570 (60,094 fake and 3,001,476 true) retweets for Trump (Figure 2). The number of fake tweets that mentioned Clinton almost doubles that of fake tweets that mentioned Trump. The proportion of fake tweets among all classified tweets is consistently higher for Clinton over this two months period (Figure 3). Traditional news media attention overwhelmingly focused on Trump, especially after the election date.

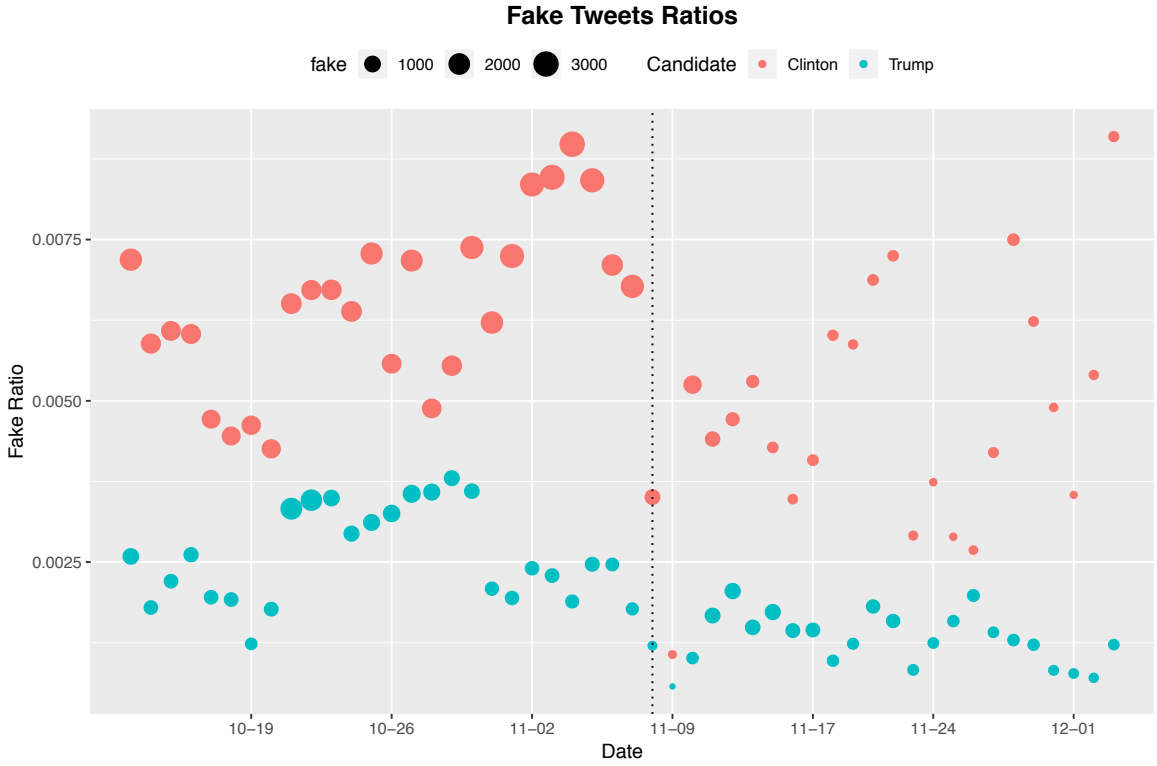


Figure 3: Fake tweet ratio over time during the two-month period for both candidates. Fake tweet ratio is calculated as the number of fake tweets divided by the sum of fake, true, and unclassified tweets.

Comments (replies and quotes) collected during this time period could be responding to original tweets before Oct 12th, 2016. This paper do not consider these comments, because the domain class of tweets before Oct 12th could not be determined. This results in 13,088 comments to fake tweets and 262,415 comments to true tweets for Clinton and 8,655 comments to fake tweets and 511,710 comments to true tweets for Trump. Note that replies contained in this dataset are those that also mentioned Hillary Clinton and Donald Trump, excluding those that did not mention their names but were responding to the original tweets that *did* mentioned their names.

In total, we will be analyzing around 7 million posts and 0.6 million comments to these posts during the two-month period. The total number of unique posts and comments is smaller than the sum of those for individual candidate because some mentioned both Clinton and Trump. Note that we have a high class imbalance in both category of tweets. To represent the reality of fake to true ratio, we will not manipulate the sample size in the data analysis part. However, we will balance the class in the prediction part for better interpretation.

V. Methods

i. Linguistic Level Analysis

A text is represented as a bag of words—a vector of words as features. Document-Feature Matrices (DFMs) are constructed from these corpora, with normalization and non-ASCII-encoded words (i.e. non-english characters), user mentions, numbers, punctuations, symbols and urls removed whenever appropriate for analysis. In the dictionary method, we will construct DFMs using one-grams, bi-grams, and trigrams because many of the dictionaries contain trigrams. We will not remove stopwords since many trigrams contain one or two stop words. We also won't remove hashtags because they are included in the NRC emotion dictionary.

On the post level, this report focuses on sentiment and emotion analysis using annotated dictionaries. Since texts in retweets are repeats of original tweets, we will only look at original tweets in linguistic analysis. Dictionaries are consisted of categories of words. We first match a text with the dictionaries, then calculate the proportions of words from different categories in the text. There are many dictionaries that categorize negative and positive words, including the Affective Norms for English Words AFINN (Nielsen, 2011), Augmented General Inquirer Positive and Negative (Stone, Dunphy, and Marshall, 1966), Loughran and McDonald Sentiment Word Lists (Loughran and McDonald, 2011), and the 2015 Lexicoder Sentiment Dictionary (Young and Soroka. 2012). We tried them all and compared results to verify if the sentiment patterns are consistent across different dictionaries. We found that except for Augmented General Inquirer Positive and Negative (Stone, Dunphy, and Marshall, 1966), the rest of the dictionaries agree with each other quiet well. We chose the 2015 Lexicoder Sentiment Dictionary (Young and Soroka. 2012) because it is the most recent one.

For emotion analysis, we will use the leading lexicon curated by the National Research Council Canada (NRC), which provides a comprehensive list of ~ 140,000 English words and ~32,000 Twitter hashtags and their weighted associations with eight basic emotions based on Plutchik (2001): anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. We will also run two-sample Kolmogorov-Smirnov tests to compare the distributions of proportions of words from different emotion categories between classified fake and true tweets.

On the comment level, we will use the same sentiment dictionary and emotion dictionary to see if patterns between fake and true tweets still persist in their comments. As

we will see later, this is not the case. To investigate the reasons behind this interesting result, we will use relative frequency analysis to identify the top keywords in the comments to both fake and true tweets with each other as reference text. Keywords are ranked by their chi-squared values, which are signed positively if the frequencies of the words exceed their expected values due to chance. The higher the chi-squared value of a word, the higher the keyness of the word in the target document relative to its reference text.

From these keywords, we discovered many debunking and questioning words, such as "hoax" and "is this true" in comments to the fake tweets. Following this insight, we hand-crafted phrases that are commonly used in fake news debunking, such as "this is fake," "hoax," and "snopes" (a reputable third party fact checking website), as well as questioning, such as "is this true," and constructed a lexicon for 43 debunking words and another for 23 questioning words (Appendix B). Using the hand-crafted dictionary, we match a text with both the "debunk" and "question" category and calculate the proportions of debunking and questioning words in the text. We tag a comment as a debunking tweet if it contains any debunking words and a questioning tweet if it contains any questioning words. Again, here we prioritize precision over recall to achieve a precise and accurate coverage of debunking and questioning signals, rather than a thorough but noisy coverage. For example, we do not include the word "fake" in the debunk category even though it is frequently used in debunking comments because it is also commonly used in non-debunking comments, such as "Hillary Clinton is so fake." Instead, we use more precise phrases that contain the word "fake," like "fake story" and "fake news."

ii. User Level Analysis

Since the sample size of comments is considerably small, we only conduct data analysis on posts. We want to investigate which user level features are important to predict both who are more likely to share fake news and the number of fake news that the users would share.

We first select and craft features following Liu and Wu (2018), such as follower count and the average number of tweets posted by a user per day. Many features change overtime, such as follower count. However, the changes tend to be small during a two-month period. We therefore select the first appearance of each feature for a user as its value throughout the period. We then clean the data. Some users' listed counts are not recorded and it makes sense to remove these user entries. After data cleaning, we get 552,077 unique

users producing posts mentioning Clinton and 884,527 unique users producing posts mentioning Trump.

For all numeric features, we log transformed them for easier interpretation. For all categorical features, we one hot encoded them. There is no need to standardize the data as we are not applying any regularization. We build a logistic regression model for a classification task framed as whether a user shares fake posts as a function of user level features. Logistic regression is a commonly used linear classification model. Here, we set the cutoff probability for the model at 50%. That is, if the output probability is greater than 50%, then we set the prediction result to be user has produced fake news. We also build a log linear model to estimate log(number of fake posts shared+1) as a function of user features. We add one to the model to avoid the case when a user shares zero fake news.

iii. Prediction

Taking an original tweet as an input and the class of the tweet as the target variable (fake or true), we treat the task of predicting an original tweet as sharing fake news domain or not as a binary classification problem. The predictive features include the following linguistic features on posts, debunking and questioning activities in comments to the posts, and user attributes of the posts:

Linguistic features

- Sentiment (2 features): proportion of negative or positive words contained in the text of the tweet. The proportion is calculated using the 2015 Lexicoder Sentiment Dictionary (Young and Soroka. 2012).

- Emotion (8 features): proportion of emotional words (eight emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) contained in the text of the tweet. The proportion is calculated using the NRC emotion lexicon.

The remaining features are only for original tweets with at least one comment.

- Accumulative Debunking Ratios (48 features): accumulative percentage of debunking tweets in comments by hour for the first two days (48 hours) since the post was published. For example, at the third hour, the debunking ratio is the accumulated

number of debunking comments divided by the accumulated number of all comments in the first three hours.

- Total Debunking Ratios (1 feature): aggregated percentage of debunking tweets in comments within the two-months period. It is calculated as the total number debunking tweets divided by the total number of comments to the post.

- Accumulative Questioning Ratios (48 features): accumulative percentage of questioning tweets in comments by hour for the first two days (48 hours) since the post was published, similar to accumulative debunking ratios above.

- Total Questioning Ratios (1 feature): aggregated percentage of questioning tweets in comments within the two-month period, similar to total debunking ratios above.

- Accumulative Comments Count (48 features): accumulative count of comments to a post by hour for the first two days (48 hours).

- Total Comments Count (1 feature): total count of comments to a post during the two-months period.

User features

- Number of days after account creation (1 feature): the number of days since a user opened his account until 3rd Dec. 2016 (end of the two-month period).

- Average number of tweets per day (1 feature): the average number of posts and comments an user produces per day between 12th Oct 2016 and 3rd Dec 2016. It is a measure of the activeness of a user during the two-month period.

- Default Profile (1 feature): true if the user uses the default profile setting on Twitter, false otherwise.

- Default profile image (1 feature): true if the user uses the default profile image, false otherwise.

- Profile url (1 feature): true if the user links an url to his/her profile, false otherwise.

- Length of user description (1 feature): the number of characters (including space and punctuations) in the user's profile description.

- Length of username (1 feature): the number of characters in the user's username.

- Listed count (1 feature): the number of public lists the user is a member of. A list can be compiled by any user based on the user's interest or focus on Twitter.

- Statuses count (1 feature): the total count of tweets (both posts and comments) by the user since account creation.

- Follower count (1 feature): the count of followers of the user. The user does not necessarily follow his/her followers.

- Friend count (1 feature): the count of friends of an user. A friend of a user follows the user and is followed by the user.

- Verified (1 feature): true if the user is verified as a public figure by Twitter, false otherwise.

We will use the following five models to evaluate different combinations of features:

Models

- Logistic Regression: A commonly used linear model for classification.

- Decision Tree: A tree-based supervised learning model.

- Random Forest: A method that randomly selects features to build multiple decision trees and average the results.

- Extreme Gradient Boosting (Xgboost): a model in the form of an ensemble of decision trees that utilizes boosting (converting a set of weak learners to a strong one) and a gradient descent optimization process at each iteration (Chen and Guestrin, 2016).

- Long Short-Term Memory: A RNN model that deals with sequential data. It was proposed as a solution to the vanishing gradient problem that standard RNN faces with feedback gates that keep or forget information to learn long-term dependencies of inputs (Hochreiter and Schmidhuber, 1997). In our model, at each time point, the input is presented as a vector of three elements: accumulative debunking ratio, questioning ratio, and comment count before that hour. We use an Adam optimizer with a learning rate of 0.01, four epochs, and a batch size of 128.

Since the focus of this paper is to identify important features for fake-news detections, instead of prediction accuracy, we will use the default parameters in all models. The features that contain temporal patterns—time dependent, that is—will only be used in LSTM. These features are trained on one layer of LSTM and the outputs are concatenated with other time-invariant features as inputs to a dense layer with sigmoid activation function. All other features that are time-invariant will be used in Logistic Regression, Xgboost, Decision Tree, and Random Forest.

Data is separated into two groups for prediction: 1) original tweets with comments and 2) original tweets with or without comments. The reason for separating the data this way is because only a small portion (less than 5%) of original tweets have any comments. When we are evaluating debunking and questioning features, including all original tweets will introduce a significant amount of missing data, impacting the performance of our models.

For better interpretation, we will also make the two classes balanced by randomly selecting the same number of tweets from the fake and true class. Since there are significantly more true tweets than fake tweets, we will keep all fake tweets and randomly sample from the true class. After sampling, we have 4,669 original tweets with comments and 97,029 original tweets with or without comments for each class for prediction. Note

that this might introduce some bias to our sample, because there is more randomness in the true class than in the fake class.

For each combination of features and models, we use a 0.75/0.25 train/test split. The performance metrics we will use to evaluate the models are accuracy rate, precision, recall, and F1 score.

iv. Feature Importance

We used random forests algorithms to train multiple trees on linguistic, user, and debunking and questioning features. We recorded the decrease in classification error due to split of a feature during the training of each tree. We averaged the decrease in error across all the trees, which can be seen as a measure of the importance of a feature in predicting fake news.

VI. Data Analysis

i. Linguistic level analysis

Posts

We first construct a DFM with tweets grouped by the class for each day. Values of the DFMs are normalized in proportions. We then calculate the proportions of positive and negative words using the 2015 Lexicoder Sentiment Dictionary (Young and Soroka. 2012). Finally, we compare the ratio of positive words between fake tweets and true tweets. Figure 4 shows the trends of positive ratio by class for both Clinton and Trump over the two-month period.

A key pattern is that before the election date, fake tweets that mentioned Clinton are consistently more negative than true tweets, while sentiment for Clinton and Trump is quite similar in true tweets (Figure 4). This agrees with previous finding that fake news are more pro-Trump than pro-Clinton and that the goal of fake news during the 2016 US election period is to provoke negative opinions towards Clinton on social media (Silverman, 2016; Allcott and Gentzkow, 2017; Guess and Tucker, 2019). As soon as the election date passed, there is no obvious pattern between the two classes nor the two candidates. The surge of positive word ratio right after the election date, especially for

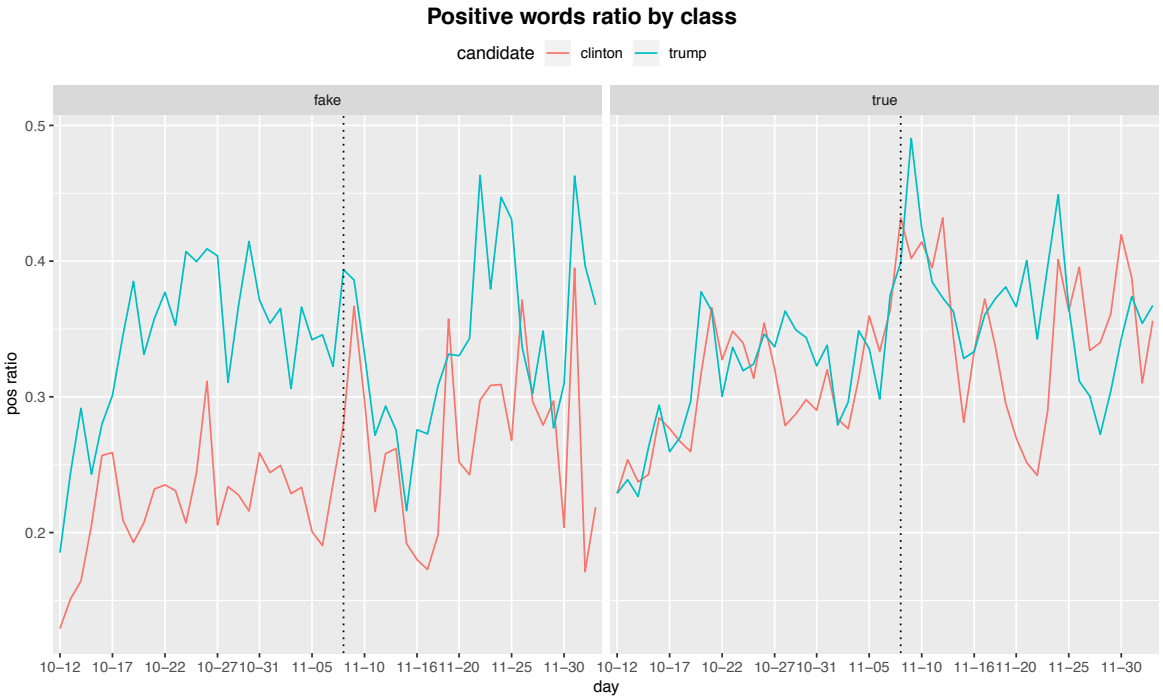


Figure 4: Trends of positive words ratio in **original tweets** over the two-month period. Positive words ratio is calculated as the proportion of positive words divided by the sum of proportion of positive words and proportion of negative words.

Trump, might be because of the positive election result for Trump. We can also see that the positive word ratios are below 0.5 for all days for both classes and both candidates. This suggests that tweets that contain news that discuss the two presidential candidates, either fake or true, use more negative words than positive words, reflecting the generally negative tone in political reporting on social media.

Sentiment analysis gives a general sense of emotion in these tweets. To get a more granular view of emotions included in these tests, we turn to the NRC emotion dictionary. However, the two dictionaries do not contain each other in any way. The NRC emotion dictionary categorizes words that generate a certain basic emotion while the 2015 Lexicoder sentiment dictionary categorizes words that are themselves negative or positive.

Tweets from each day are grouped into one single DFM without being grouped by the classes they belong to. For each emotion, we calculate the fraction of words in the tweets related to the emotion, resulting in a vector of emotion weights for each tweet that summed to 100% across the emotions. Since some tweets only contain one word, or a string of words without spaces, we get some rare vectors where the emotion score for one or more particular emotions is 100%. This scenario is rare (10 to 50 out of thousands of tweets) but it will drastically increase the standard deviation of distribution for each emotion. Therefore, we will treat these tweets as outliers and remove them from the emotion analysis. However, even after removing these extreme outliers, we still have exceptionally high standard deviation relative to the mean. This might be explained by other extreme large values that generate large standard deviations or the large number of zeros in the vector (a tweet might not contain one or multiple of the eight emotions) that generate small means.

For both Clinton and Trump, the distributions of proportions of words from each emotion are significantly different for fake and true tweets (p values ≈ 0.000) (Table 1). Fear appears to be the emotion that distinguishes fake tweets from true tweets the best, with fear more prevalent in fake tweets. In fact, most emotions are more prevalent in fake tweets for both candidates. This suggests that fake tweets contain more emotional words than true tweets on average (Figure 5). The second most distinguishable emotion for tweets that mentioned Clinton is anger, a relatively negative emotion, while that for Trump is surprise, a relatively positive emotion. This agrees with our previous finding that fake tweets that mentioned Clinton are more negative than those mentioned Trump.

In addition, both fake and true tweets that mentioned Trump have average higher scores for surprise. This suggests that tweets that mentioned Trump during the election period tend to contain more surprising words.

A

Emotion	Mean		Standard Deviation		KS test	
	fake	true	fake	true	D	p
fear	**1.45**	0.73	1.66	1.21	0.21	0.00*
anger	**1.09**	0.82	1.42	1.33	0.11	0.00*
sadness	**0.91**	0.63	1.33	1.14	0.10	0.00*
joy	**0.65**	0.40	1.18	0.86	0.08	0.00*
surprise	0.82	**0.99**	1.26	1.43	0.07	0.00*
anticipation	**0.99**	0.84	1.42	1.31	0.06	0.00*
disgust	**0.52**	0.50	1.00	1.21	0.04	0.00*
trust	**1.56**	1.47	1.85	1.77	0.04	0.00*

B

Emotion	Mean		Standard Deviation		KS test	
	fake	true	fake	true	D	p
fear	**1.21**	0.69	1.62	1.21	0.14	0.00*
surprise	**2.32**	2.20	1.47	1.88	0.11	0.00*
anticipation	**1.04**	0.79	1.46	1.31	0.08	0.00*
anger	**0.86**	0.68	1.32	1.21	0.07	0.00*
sadness	**0.76**	0.71	1.25	1.42	0.06	0.00*
joy	0.52	**0.68**	1.02	1.44	0.05	0.00*
disgust	**0.55**	0.49	1.21	1.20	0.03	0.00*
trust	**1.29**	1.27	1.63	1.65	0.03	0.00*

Table 1: Mean and standard deviation of weights for each emotion in **original tweets**. Emotions are ranked by D statistics. **(A)** Clinton. **(B)** Trump.

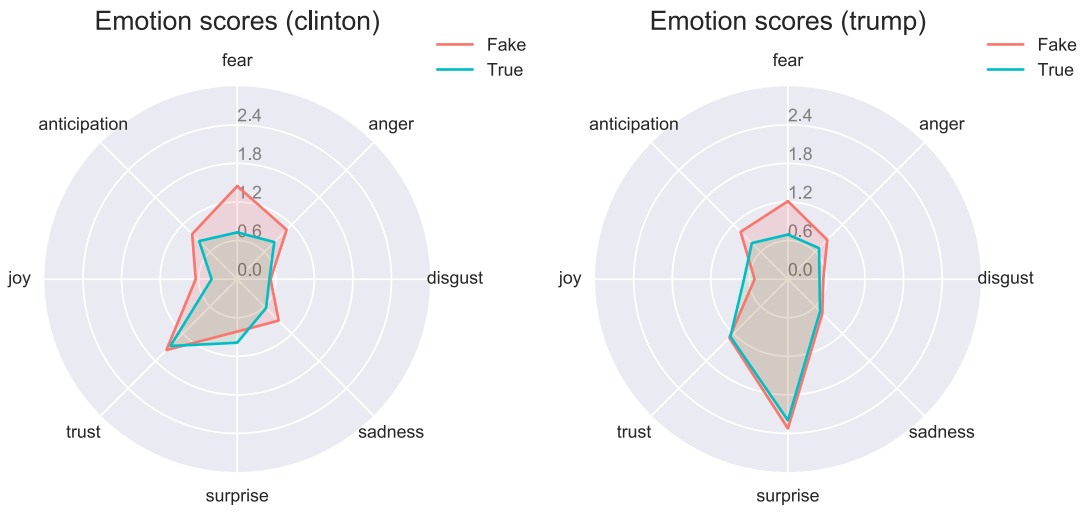


Figure 5: Radar charts of mean weights of each eight basic emotion in **original tweets** by candidate.

Comments

We conduct the same sentiment and emotion analysis on comments to these posts. It appears that surprise and disgust are the two emotions that distinguish comments to fake tweets from comments to true tweets the most (Table 2). Surprise is more prevalent in comments to true tweets that mentioned Clinton and disgust is more prevalent in comments to fake tweets that mentioned Trump. This is somewhat inconsistent with previous results by Vosoughi, Roy, and Sinan (2018), who concluded that fake rumors inspired responses expressing both greater surprise *and* disgust. We can observe that the emotion differences are quite variable in comments, with much higher standard deviation and lower D statistics (Table 2). This suggests that emotion pattern varies more in comments than in posts.

Similar high variability can also be observed in sentiment pattern for comments to fake tweets (Figure 6). In contrast, there is more stability for comments to true tweets.

A

Emotion	Mean		Standard Deviation		KS test	
	fake	true	fake	true	D	p
surprise	0.66	**1.02**	1.46	1.83	0.08	0.00*
fear	**1.40**	1.10	2.31	1.95	0.04	0.00*
trust	**2.32**	2.01	3.11	2.75	0.04	0.00*
joy	**1.03**	0.83	1.94	1.65	0.03	0.00*
disgust	**0.91**	0.77	1.80	1.57	0.03	0.00*
anticipation	**1.36**	1.27	2.27	2.11	0.02	0.00*
anger	1.35	**1.37**	2.25	2.22	0.02	0.08
sadness	1.08	**1.10**	1.97	1.93	0.01	0.18

B

Emotion	Mean		Standard Deviation		KS test	
	fake	true	fake	true	D	p
disgust	**1.56**	0.54	4.19	1.24	0.11	0.00*
trust	**2.44**	1.81	3.24	2.53	0.08	0.00*
anticipation	**1.56**	1.13	2.43	1.93	0.07	0.00*
joy	**1.15**	0.77	2.06	1.55	0.06	0.00*
surprise	1.71	**1.88**	2.48	2.49	0.05	0.00*
anger	**1.23**	1.03	2.09	1.83	0.04	0.00*
fear	**1.13**	0.96	2.03	1.77	0.03	0.00*
sadness	**0.98**	0.82	1.80	1.60	0.03	0.00*

Table 2: Mean and standard deviation of weights for each emotion in **comments**. **(A)** Clinton. (**B**) Trump.

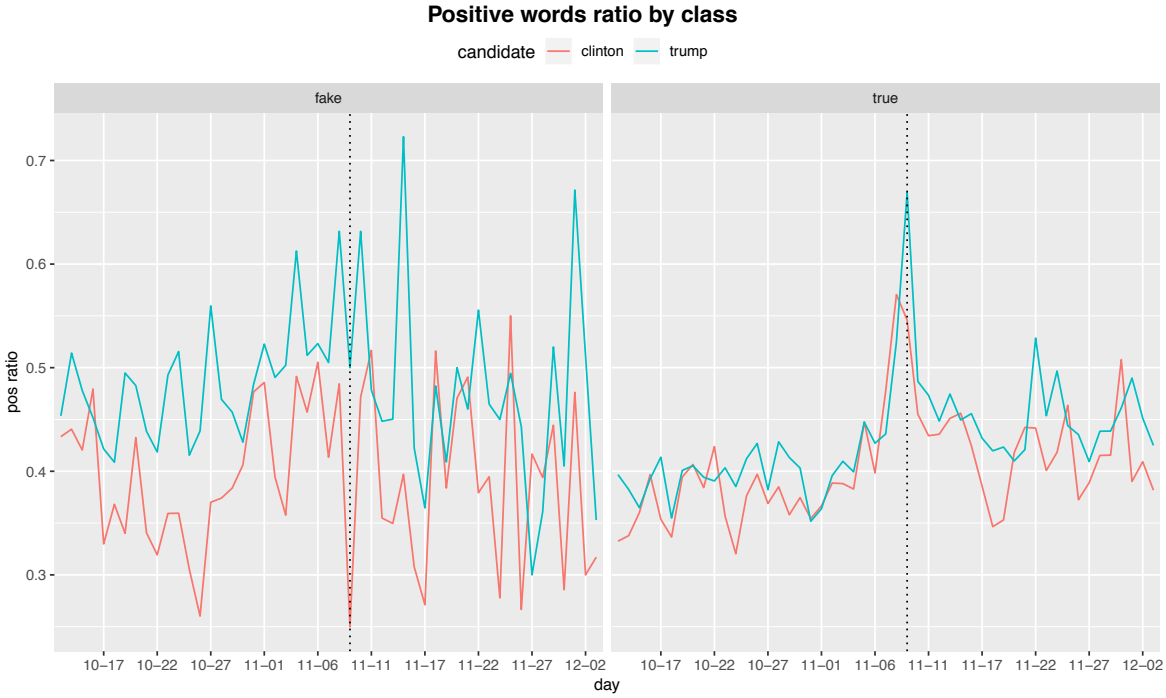


Figure 6: Trends of positive words ratio in **comments** over the two-month period. Positive words ratio is calculated as the proportion of positive words divided by the sum of proportion of positive words and proportion of negative words.

The high variability of sentiment and emotion suggests that there might be more mixed opinions in comments, especially for those responding to fake tweets. To investigate this supposition, we extract the keywords from comments responding to both classes using relative frequency analysis (Table 3). Within comments to fake tweets, phrases like "deplorable lie," "hoax," and "snopes" (reputable third-party fact-checking website) are clear signals of fact-checking efforts, while phrases like "is this true" are signals of questioning the validity of the news. Within comments to true tweets, keywords are more

pertinent to the subjects of the news, such as the two presidential candidates and the election. It is therefore not unreasonable to suspect more debunking and questioning signals within comments to fake tweets than those to true tweets. In fact, some older papers, including Mendoza et al (2010), Zhao et al (2015), and Starbird et al (2015), suggest that debunking and questioning activities are prevalent in public responses to fake news and can potentially be used to detect rumors.

	Fake		True	
order	**replies**	**quotes**	**replies**	**quotes**
1	deplorable_lie	true	trump	trump
2	assange_as	is_this_true	you	gop
3	as_a_pedo	this_true	and	zach
4	assange	this_is_true	emails	vote
5	frame_assange_as	@seanhannity	#imwithher	comey
6	assange_as_a	is_true	vote	#nevertrump
7	hoax	divorce	for_clinton	fbi
8	frame_assange	snopes	recount	emails
9	it's_a_deplorable	this_for_real	clinton_campaign	election
10	a_hoax	is_this_for	trump_is	trump's
11	…	…	…	…

Table 3: Top 10 keywords in comments (replies and quotes) from both classes. Keywords are ranked by chi-squared values of the words.

To verify this hypothesis, we hand-crafted a list of common debunking and questioning words selected and extended from the top keywords above and tagged comments that contain at least one of these words (Table 4). There are 43 debunking words and 23 questioning words (appendix B). After tagging, about 1.38% of comments to fake tweets are debunking tweets and 0.45% are questioning tweets, while only 0.63% of comments to true tweets are debunking tweets and 0.12% are questioning tweets.

	Examples of debunking comments to fake tweets
1	This went around a while ago. It is a hoax.
2	My apologies. ...this story is fake. We will no longer follow this "verified" News Agency websites on Internet.
3	Sorry, folks, I support Trump, but that article isn't true.
4	It would appear that this, too, is one of those fake news reports, if snopes is to be believed.
5	I AM A TRUMP SUPPORTER AND THIS STORY IS FAKE SO QUIT SPREADING IT. The site mentioned is a malware site.
6	…

Table 4: Examples of debunking comments to fake tweets.

After aggregating tweets at the url level and counting their numbers of post share, we identify top links to fake news. The top 10 fake news include "Donald Trump Protester Speaks Out I was Paid to Protest" by abcnew.com.co and "Michelle Obama Deletes Hillary Clinton Twitter" by yournewswire.com. Out of the top 10 fake news, 8 inspire at least one debunking comment and 7 inspire at least one questioning comment. This suggests that our lexicon has high precision.

Like Starbird et al (2015) has suggested, the temporal pattern of enquiring signals can help improve real time prediction. We will next investigate the temporal pattern of both debunking and questioning comments.

Figure 7 shows the accumulative debunking and questioning ratio by hour since a post was posted. The aggregation is at the class level. Questioning comments occurred as soon as the post was first posted then there would be an increase of debunking activities which then stabilized after 12 hours. There also tends to be a second bump of both questioning and debunking on the second day after a fake post was posted. This suggests that people tend to question a piece of information before anyone starts fact checking it.

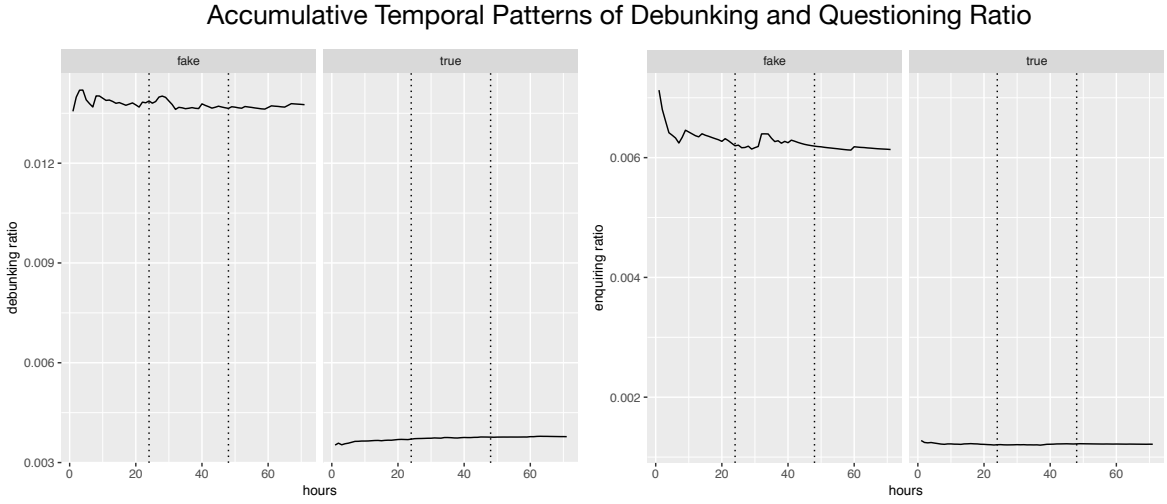


Figure 7: Accumulative temporal patterns of debunking (left) and enquiring (right) ratio at the class level.

The implications of the above linguistic analysis on both the post and comment level are two-fold. First, sentiment and emotion can serve as important features in the post level for fake news detection. Second, comments to fake tweets contain higher proportion of debunking and questioning tweets than comments to true tweets. We could potentially utilize crowd intelligence of the Twitter community together with sentiment and emotion signal in original posts to improve fake news detection.

ii. user level analysis

Fake news seems to be mostly produced by a small proportion individuals as the number of users decrease exponentially with the number of news (Figure 8). Indeed, 50 percent of fake news were produced by 0.012% users for tweets mentioning Clinton and by 0.0056% users for tweets mentioning Trump. True news on the other hand is far less concentrated. The number of users also decrease exponentially with the number of news but less quickly. 5.1% users and 4.0% users produced 50% of true news for tweets mentioning Hillary and Trump. This finding is consistent with Guess et al (2019) on Facebook.

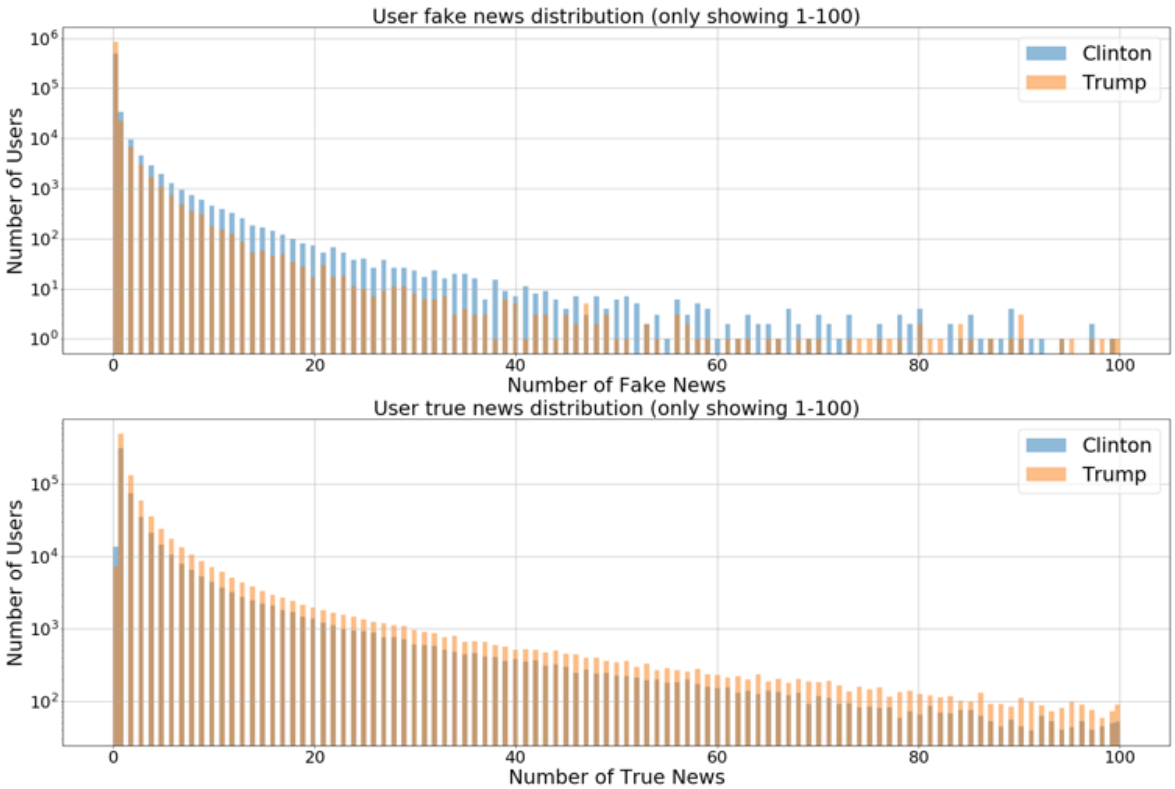


Figure 8: Distribution of the number of news shared by users.

In the logistic regression model, across the two cases, all user attributes except log user friends count (only statistically significant in Trump case) are statistically significant at 1 percent significant level (Table 5). Effect dimensions of attributes are consistent across two cases. Probability of sharing fake news is positively correlated with user default profile and user default image and negatively correlated with user profile url and user description log. Therefore, the more effort user puts in their profile, the less likely they have spread any fake news. Probability of sharing fake news is slightly negatively correlated with the number of days after account creation.

In the log linear model, statistically significant effect sizes (at 1 percent significance level) are identified for all user attributes except user default profile in Clinton case and except for number of days after account creation log for Trump and combined case (Table 5). The effect directions are consistent across the Trump and Clinton responses.

	Logit Model		Log Linear Model	
Features	**Coefficient**	**p**	**Coefficient**	**p**
const	-3.94	0.00*	-0.03	0.00*
verified	-3.02	0.00*	-0.06	0.00*
default_profile	0.22	0.00*	0.01	0.00*
default_profile_image	0.40	0.00*	0.02	0.00*
url	-0.43	0.00*	-0.01	0.00*
followers_count_log	0.18	0.00*	0.01	0.00*
friends_count_log	0.01	0.00*	0.00	0.00*
favourites_count_log	-0.02	0.00*	-0.00	0.00*
statuses_count_log	-0.07	0.00*	-0.01	0.00*
listed_count_log	-0.09	0.00*	-0.01	0.00*
number_of_days_after_account_creation_log	-0.02	0.00*	0.00	0.25
average_number_of_tweets_per_day_log	1.00	0.00*	0.15	0.00*
description_length_log	-0.02	0.00*	-0.00	0.00*

Table 5: Coefficients and p values of user attributes of logistic regression and log linear model.

VII. Prediction

i. Classification Performance

For original tweets with comments, random forest consistently produces the best accuracy rate and Xgboost produces the best F1 score (Table 6). The best combination of features is user attributes with linguistic features. Surprisingly, debunking and questioning ratios, both accumulative and aggregated, do not produce a satisfactory prediction accuracy. Different combinations of debunking and questioning ratios with other features also produce low performance.

There are many potential reasons for this low performance. First, the lexicon is small and has low recall, with only 2,772 tagged debunking tweets and 908 tagged questioning tweets out of all 620,468 comments. The debunking and questioning signals are thus very small compared to the sample size and do not have a huge impact on prediction accuracy. Second, even though the lexicon might have high precision in picking up actual debunking and questioning comments to the fake class, it also identifies many false signals in the true class. Even though these tweets used the common debunking or questioning words, they might not be debunking nor questioning the post. For example, "breaking: founder of fake news site still publishing fake news," and "Google News introduces fact check feature _ just in time. Fact checks are essential today. FoxNoise is the worst org" contain the word "fake news" and "fake check," which are both common debunking words. However, the comments are responding to the topic of fake news but not to debunk a piece of fake news.

It is also worth noting that the term "fake news" has become more and more common in political rhetorics, especially on social media, where the public and politicians often use the term to show disagreement with a point of view that is different than their own, even they are not verifiably fake. For example, "Dumb bitch! Donald doesn't talk to women his age. Let alone 5 years older. Fake story to hide Hillary the Cunt" and "Throw her and Comey in jail for corruption! This is not true America! It's time to act and stand together against the dictatorship is IN WH" both use debunking words "fake story" and "not true." However, the comments are intended to express the users' opinions, rather than objectively comment on the veracity of the news. In this case, the static lexicon is sensitive to the use of political rhetorics, which is constantly changing.

Original Tweets With Comments

Features	Model	Accuracy	Precision	Recall	F1
User Alone	Logistic	0.699	0.673	0.773	0.720
	Xgboost	0.744	0.710	**0.826**	**0.764**
	Decision Tree	0.739	0.727	0.767	0.746
	Random Forest	**0.759**	**0.764**	0.750	0.757
Linguistic Alone	Logistic	0.550	0.567	0.421	0.483
	Xgboost	0.597	0.586	0.662	0.621
	Decision Tree	0.626	0.610	0.701	0.652
	Random Forest	**0.628**	**0.611**	**0.708**	**0.656**
Debunking & Quesitoning Alone	**Logistic**	**0.527**	**0.520**	0.700	0.597
	Xgboost	0.527	0.516	0.882	0.651
	Decision Tree	0.522	0.513	0.893	0.652
	Random Forest	0.523	0.513	0.893	0.652
*Temporal	LSTM	0.523	0.514	**0.897**	**0.654**
User and Linguistic	Logistic	0.704	0.685	0.756	0.719
	Xgboost	0.752	0.725	**0.813**	**0.766**
	Decision Tree	0.728	0.711	0.767	0.738
	Random Forest	**0.769**	**0.788**	0.737	0.762
User and Debunking & Quesitoning	Logistic	0.701	0.675	0.774	0.721
	Xgboost	0.746	0.713	**0.824**	**0.764**
	Decision Tree	0.737	0.729	0.755	0.742
	Random Forest	**0.763**	**0.769**	0.751	0.760
*Temporal	LSTM	0.630	0.627	0.646	0.636
Linguistic and Debunking & Quesitoning	Logistic	0.560	0.580	0.438	0.499
	Xgboost	0.615	0.601	0.681	0.639
	Decision Tree	0.616	0.602	0.686	0.641
	Random Forest	**0.639**	**0.627**	**0.686**	**0.655**
*Temporal	LSTM	0.544	0.537	0.647	0.587
All	Logistic	0.702	0.683	0.753	0.716
	Xgboost	0.750	0.721	**0.817**	**0.766**
	Decision Tree	0.721	0.707	0.755	0.730
	Random Forest	**0.757**	**0.767**	0.740	0.753
*Temporal	LSTM	0.576	0.580	0.556	0.568

Table 6: Performance of models using different combinations of features, only with original tweets with comments as inputs. The class is balanced as we have randomly sampled the same number of fake and true tweets. The number of records from each class is 4669. Train-test split ratio is 0.75.
* debunking and questioning ratios are included as time-series features in this model. Features are recorded for the first 48 hours since a post was published. Features are accumulative. For example, at the third hour, the debunking ratio is the accumulated number of debunking comments divided by the accumulated number of all comments in the first three hours.

Additionally, temporal signals of debunking and questioning activities seem not to be useful neither. The LSTM model only achieves a low prediction accuracy of 0.523 with temporal features alone, 0.63 with user attributes, 0.544 with linguistic features, and 0.576 with all other features (Table 6). This might be because there is a higher variability at the tweet level than at the class level where the temporal patterns are more distinctive between the fake and true class. The accumulative values for individual tweets can vary a lot at each time point (hr) and at each time point there are different numbers of tweets that contain any comments. The representation of debunking and questioning signals as ratios might be problematic as well. For less popular tweets, which are the majority of tweets, they only attract a few number of comments. In many cases, tweets only have one comment at a later stage after it is posted. If this one comment is tagged as debunking or questioning, the debunking or questioning ratio would be 1 for all later hours and 0 for all earlier hours, while for more popular tweets, even though they might attract more debunking and questioning comments, their debunking and questioning comments would be very small throughout the two-days period. Representing the signals as different features or using additional features like the aggregated scores of emotional words of the tagged debunking and questioning comments, and popularity of the post might add more distinctive characteristics of posts that invite debunking or questioning comments.

Original Tweets With or Without Comments

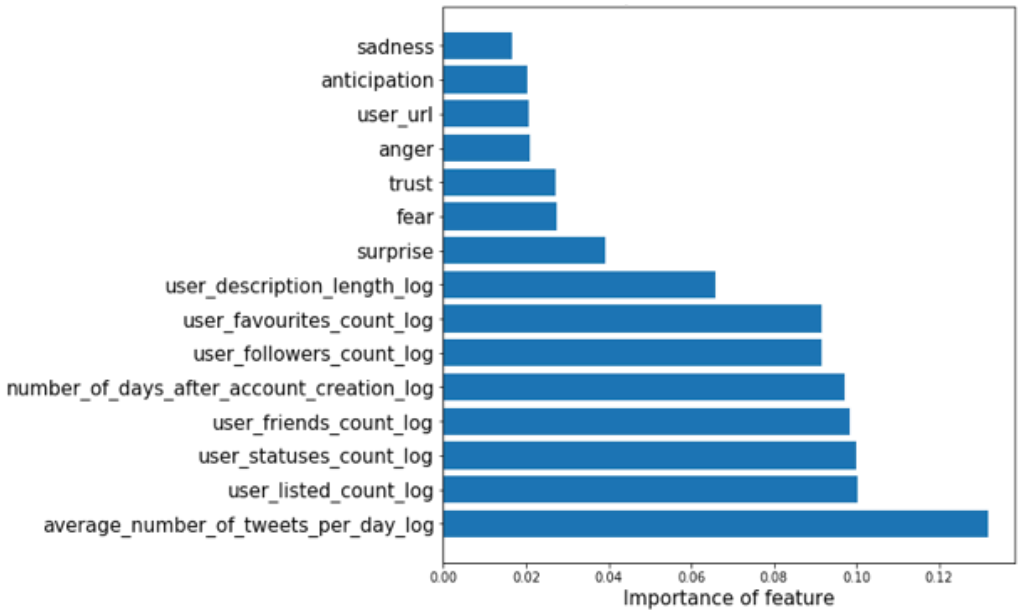
Features	Model	Accuracy	Precision	Recal	F1
User Alone	Logistic	0.672	0.663	0.703	0.682
	Xgboost	0.706	0.682	0.774	0.725
	Decision Tree	0.822	0.807	**0.846**	0.826
	Random Forest	**0.860**	**0.881**	0.832	**0.856**
Linguistic Alone	Logistic	0.567	0.552	0.715	0.623
	Xgboost	0.606	0.587	0.715	0.645
	Decision Tree	0.676	0.652	0.756	0.700
	Random Forest	**0.677**	**0.653**	**0.758**	**0.701**
User and Linguistic	Logistic	0.683	0.675	0.706	0.690
	Xgboost	0.723	0.702	0.777	0.738
	Decision Tree	0.778	0.765	**0.802**	0.783
	Random Forest	**0.829**	**0.850**	0.800	**0.824**

Table 7: Performance of models using different combinations of features, with original tweets with or without comments as inputs. The class is balanced as we have randomly sampled the same number of fake and true tweets. The number of records from each class is 97,029. Train-test split ratio is 0.75.

For original tweets with or without comments, random forest produces the best performance in both prediction accuracy rate (0.86) and F1 score (0.856) (Table 7). The best combination of features is user attributes alone. Adding linguistic features weakens model performance, instead of strengthening it. This might be because there is a lot more variability at the tweet level than the class level. In the descriptive data analysis part, we aggregated sentiment and emotion features of individual tweets at the class level by taking the means. The distinctive patterns that we discover that separate fake tweets and true tweets are at the class level. We also randomly sample data for prediction while we keep all data during linguistic level analysis. This might have caused many patterns we discover as insignificant.

ii. Feature Importance

Since Random Forest gives the best performance, we use the model to evaluate the importance of features. We perform feature importance test for all random forest models with different combinations of features. All models produce similar top important features. The top 8 features are all features from the user attribute list. Average number of tweets per day, user listed count, user statuses count, number of days after account creation, and user follower count are among the most import features overall. Among linguistic features, fear and surprise are the two most important features that distinguish fake tweets from true tweets.



Feature 9: Feature importance of all features on all features.

VIII. Future Work

The weakness of our model lies largely on our linguistic features extraction. To better utilize linguistic cues on posts, it might be important to include more features than just sentiment and emotion. For example, including LIWC features on posts might increase the effectiveness of linguistic features.

To better utilize debunking and questioning signals, there are a few directions that researchers can explore. First, inspired by Tschiatschek et al (2018), which concludes that any approach that is not learning about users' correcting accuracy is prone to failure in the presence of adversarial/spam users (who want to promote fake news), we suspect that analyzing users who debunk and question fake news might help with improving performance. However, we still need a lot stronger debunking and questioning signals (more data) to produce significant results. Second, improving the LSTM model, for example, by adding on more layers, might help the model pick up the temporal patterns of debunking and questioning ratios more effectively. Third, it might be interesting to include aggregated linguistic features of comments as well, since comments with more emotional words are less likely to be debunking or questioning even if they contain some debunking or questioning words. Last but not least, we can avoid the limitation of dictionary methods by training a classifier specific to our question to filter debunking and questioning words.

IX. Conclusion

This paper takes the feature-oriented approach to detect fake news on Twitter during the 2016 US election, using a dataset collected one month before and one month after the election date. The tweets are those that mentioned the two presidential candidates Hillary Clinton and Donald Trump. We first perform linguistic and user features extraction and then use several supervised machine learning models to predict if an original tweet shares a fake news domain or not. We discover significant sentiment and emotion patterns on posts of fake news and debunking and questioning activities in comments to these posts. We also discover that only a small amount of users spread fake news and the average number of tweets per day by a user is the most important feature, with those that post more tweets per day more likely to spread fake news. However, only user attributes prove to be consistently helpful in producing high accuracy in all models. Tweets are short texts and their Linguistic features thus have high variability. Adding more features might help with stabling linguistic cues. Debunking and questioning signals can be better captured by constructing different features or using more advanced time-series models.

X. Limitations

One of the limitations is that the comments included in the dataset are limited to those that mentioned the two candidates. Our analysis and prediction power on comments are therefore limited and might have excluded many important signals. To have a more effective representation of comments to posts that share fake news, it is important to include all comments to the posts. Another major limitation is that the dictionary method is easily affected by the changes of usage of words and need to be constantly updated.

XI. References

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the Diffusion of Misinformation on Social Media. *The National Bureau of Economic Research*. Working Paper No. 25500.

Allcott. H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*. 31(2): 211–236.

Barberá, P. (2018). Explaining the Spread of Misinformation on Social Media: Evidence from the 2016 U.S. Presidential Election. Note prepared for *the APSA Comparative Politics Newsletter*. Available at http://pablobarbera.com/static/barbera-CP-note.pdf. Accessed April 20th, 2019.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785-794, San Francisco, California, USA — August 13 - 17, 2016.

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*. 363(6425):374-378.

Guess A, Nyhan B, Reifler J (2018). Selective exposure to misinformation: evidence from the consumption of fake news during the 2016 US presidential campaign. Working Paper. European Research Council. Available at https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf. Accessed April 20th, 2019.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science*. Vol. 5, no. 1, eaau4586.

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*. Volume 9 Issue 8, Pages 1735-1780.

Jang, S.M., Geng, T., Li, J.Q., Xia, R., Huang, C. Kim, H., and Tang, J. (2018). A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*. Volume 84, Pages 103-113.

Liu, Y. and Wu, B.Y.F. (2018). Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *AAAI Publications, Thirty-Second AAAI Conference on Artificial Intelligence,* pp 354-361.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., and Cha, M. (2016). Detecting Rumors from Microblogs with Recurrent Neural Networks. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16),* pages: 3818-3824, New York, New York, US, 9–15 July 2016.

Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? *Proceedings of 1st Workshop on Social Media Analytics (SOMA '10)*, Washington, DC, USA.

Monti, F., Frasca, F. Eynard, F., Eynard, D., Mannion, D., and Bronstein, M.M. (2019). Fake News Detection on Social Media using Geometric Deep Learning. *Cornell University Press, Computer Science, Social and Information Networks*. arXiv:1902.06673 [cs.SI]

Nielsen, F.Å. (2011). A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages*: 93-98.

Perez-Rosas, V.P., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic Detection of Fake News. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401 Santa Fe, New Mexico, USA, August 20-26, 2018.

Plutchik, R. (2001). The Nature of Emotions. *American Scientist*. Vol. 89, Issue 4, p. 344-350.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. Cornell University Press. arXiv:1708.01967 [cs.SI].

Silverman, C. (2016). This Analysis Shows how Fake Election News Stories Outperformed Real News on Facebook. *BuzzFeed News*, November 16. Available at https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook. Accessed April 20th, 2019.

Starbird, K., Spiro, E.S., Arif, A., Chou, F.J., Narasimhan, S., Maddock, J., Shanahan, K., and Robinson, J. (2015). Expressed Uncertainty and Denials as Signals of Online Rumoring. *University of Washington*. Available at http://faculty.washington.edu/kstarbi/ExpressedUncertainty-Final.pdf.

Stone, P.J., Dunphy, D.C., and Smith, M.S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Townsend, T. (2016). Meet the Romanian Trump Fan Behind a Major Fake News Site. *Inc. magazine*, ISSN 0162-8968. Accessed April 21st, 2019.

Twitter API Developer. Tweet Object. Avalilable at https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html. Accessed March, 2019.

Tschiatschek, S., Singla, A., Rodriguez, M.G, Merchant, A., and Krause, A. (2018). Fake News Detection in Social Networks via Crowd Signals. Proceedings of *WWW '18 Companion*, April 23–27, 2018, Lyon, France.

Vosoughi, S., Roy, D., and Sinan, A. (2018). The spread of true and false news online. *Science*. Vol. 359, Issue 6380, pp. 1146-1151.

Young, L. and Stuart S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication* 29(2): 205-231.

Zhao, Z., Resnick, P., and Mei, Q. (2015) Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. *Proceedings to International World Wide Web Conference Committee (IW3C2),* Pages 1395-1405, Florence, Italy — May 18 - 22, 2015.

XII. Appendices

A. list of fake news domains and traditional news domains

a. List of 95 fake news domains

yournewswire.com	newslo.com	bluevisionpost.com
angrypatriotmovement.com	usa-television.com	stgeorgegazette.com
conservativedailypost.com	conservativearmy88.com	dailynews5.com
worldnewsdailyreport.com	stuppid.com	newsfeedhunter.com
usanewsflash.com	mainerepublicemailalert.com	flashnewscorner.com
en-volve.com	wetheproudpatriots.com	floridasunpost.com
tmzworldnews.com	thenewyorkevening.com	breakingtop.world
freedomsfinalstand.com	satiratribune.com	persecutes.com
awarenessact.com	christiantimesnewspaper.com	fedsalert.com
huzlers.com	newsexaminer.net	theusaconservative.com
usasupreme.com	politicot.com	weconservative.com
now8news.com	uspoln.com	freedomcrossroads.us
teddystick.com	theseattletribune.com	mississippiherald.com
neonnettle.com	usadailypost.us	alabamaobserver.com
americanjournalreview.com	houstonchronicle-tv.com	wazanews.tk
usherald.com	londonwebnews.com	worldpoliticsnow.com
react365.com	whatdoesitmean.com	bostonleader.com
consnation.com	therightists.com	washingtonevening.com
the-postillon.com	civictribune.com	empiresports.co
washingtonfeed.com	theexaminer.site	americanflavor.news
empirenews.net	channel18news.com	prntly.com
politicops.com	ourlandofthefree.com	usapoliticstoday.com
theracketreport.com	usadailytime.com	usapoliticsnow.com
abcnews.com.co	ushealthyadvisor.com	empireherald.com
notallowedto.com	guerilla.news	associatedmediacoverage.com
dailyusaupdate.com	ky12news.com	denverinquirer.com
realnewsrightnow.com	snoopack.com	federalisttribune.com
viralactions.com	news4ktla.com	kmt11.com
president45donaldtrump.com	redinfo.us	redcountry.us

channel23news.com	asamericanasapplepie.org	thelastlineofdefense.org
nationalreport.net	undergroundnewsreport.com	usadailyinfo.com
superstation95.com	buzzfeedusa.com	

b. List of 116 traditional news domains

cnn.com	bakersfield.com	macon.com
nytimes.com	bendbulletin.com	myrtlebeachonline.com
theguardian.com	bnd.com	naplesnews.com
washingtonpost.com	broadcastingcable.com	nashvillescene.com
foxnews.com	charlestoncitypaper.com	news.cornell.edu
huffingtonpost.com	chicagomaroon.com	news.usc.edu
usatoday.com	collegian.psu.edu	newseum.org
wsj.com	columbian.com	news-journalonline.com
cnbc.com	dailynebraskan.com	news-leader.com
reuters.com	dailynexus.com	newstimes.com
time.com	dailynorthwestern.com	nwfdailynews.com
nypost.com	dailypress.com	pjstar.com
usnews.com	dailyprogress.com	presstelegram.com
cbsnews.com	dailytexanonline.com	rapidcityjournal.com
chron.com	dailytrojan.com	readingeagle.com
thehill.com	dcourier.com	redandblack.com
nbcnews.com	delcotimes.com	rgj.com
theatlantic.com	durangoherald.com	sacurrent.com
latimes.com	fair.org	santacruzsentinel.com
abcnews.go.com	fredericksburg.com	santafenewmexican.com
thedailybeast.com	globegazette.com	sgvtribune.com
sfgate.com	greenvilleonline.com	signalscv.com
newsweek.com	greenwichtime.com	siouxcityjournal.com
chicagotribune.com	havasunews.com	standard.net
economist.com	hcn.org	stanforddaily.com
theroot.com	heraldnet.com	steynonline.com
voanews.com	heraldsun.com	studlife.com
nj.com	heraldtimesonline.com	tallahassee.com

miamiherald.com	ibj.com	theday.com
mercurynews.com	independent.com	theeagle.com
bostonglobe.com	islandpacket.com	theledger.com
seattletimes.com	jou.ufl.edu	timesleader.com
oregonlive.com	journalism.org	ubm.com
washingtontimes.com	journalismjobs.com	vcstar.com
azcentral.com	journaltimes.com	wacotrib.com
ajc.com	kitv.com	wcfcourier.com
philly.com	knoxnews.com	wvgazettemail.com
sacbee.com	lacrossetribune.com	yakimaherald.com
aspentimes.com	leadertelegram.com	

B. list of debunking and questioning words

a. List of 43 debunking words

hoax	aint true	fact checking	hoax site
satire	not fact	factcheck	hoax website
not true	not facts	factchecked	misinformation
isn't true	not truth	factchecking	mis information
wasn't true	debunk	fake news	disinformation
aren't true	debunked	fake story	dis information
weren't true	snopes	fake stories	mis info
isnt true	snoops	fake site	misinfo
wasnt true	politifact	fake website	dis info
werent true	fact check	satire site	disinfo
ain't true	fact checked	satire website	

b. List of 25 questioning words

is this true	is this fake	is this satire	is this hoax
is that true	is that fake	is that satire	is that hoax
is it true	is it fake	is it satire	is it hoax
was this true	was this fake	was this satire	was this hoax
was that true	was that fake	was that satire	was that hoax
was it true	was it fake	was it satire	was it hoax
not sure			