2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and
2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)

Fake News Pattern Recognition using Linguistic Analysis

Amitabha Dey, Rafsan Zani Rafi, Shahriar Hasan Parash, Sauvik Kundu Arko and Amitabha Chakrabarty School of Engineering and Computer Science, BRAC University
Dhaka, Bangladesh
{amitabhadey.bracu, nasfar127, shahriarhasan0, sauvik.kundu240}@gmail.com, amitabha@bracu.ac.bd

Abstract— **In the wake of the 2016 US Presidential Election, the upsurge of fake news has been a subject of increased discussion and debate. In this paper, we propose a general framework that can been adopted in future elections worldwide to augment humans in making better decisions when it comes to recognizing news deception and identifying hidden bias of the author. For our study, we constructed a dataset comprising 200 tweets on "Hilary Clinton", while performing veracity assessment. We initially perform "text normalization" on tweets, explore techniques for feature extraction to classify news into categories, perform a comprehensive linguistic analysis on tweets, extract bag-of-words to find noticeable pattern, and finally apply k-nearest neighbor algorithm for classifying polarized news from credible. We later turn to some popular evaluation metrics to quantify the success rate of our framework, discuss the results of implementing knn algorithm and discuss interconnected research domains and future research directions for constructing an ideal model for fake news detection system around social media.**

Keywords-polarization, classification, fake news, linguistic analysis, knn algorithm

I. Introduction

News that we consume every day change our view of the world. The escalation of fake news has certainly put reliability in journalism and media system in jeopardy. Defining, identifying and stopping fake news from spreading has always been a top priority of governments and corporations. Yet, fake news is still prevalent in social media and is being consumed by millions of people every day. Misinformation paves way for deep-rooted problems by giving birth to hatred and reinforcing prejudices. One of the initial challenges we faced while embarking on our journey to solve the problem is that there does not yet exist a definite and unified definition of fake news and the markers needed to determine an articles legitimacy or flag a news as fake. A lot of corporations still have to resort to flagging articles manually. The task of detecting news as fake can be interpreted as classifying news written to deceive the readers on purpose. Satirical news uses irony, exaggeration and ridicule in the context of politics and otherwise, but is not intended to make readers believe in its truth value, and therefore will not considered to be fake. Moreover, news of political discourse consumed from social media plays decisive role in forming people's beliefs and opinions in distinct ways. As a result, the transparency within news is often compromised to amplify impact on society. The culture of accepting news from social media is discernible. In terms of helping to reduce the negative effects caused by fake news and also for the sake of welfare of the general public along with the news ecosystem, it's crucial that we should develop systems to detect fake news on social media and other online news portal automatically. In the latest time, diminishing trust in the mainstream media has been a pattern among people. According to Gallup polls, only 40% population of adult Americans still have trust their mass media sources to report the news fully, accurately and fairly. Derived from empirical evidence, humans are fairly ineffective at recognizing deception. Most people who are inclined towards a topic will tend to believe any information validating it to be legitimate and reliable, and vice-versa.

A. Dataset Construction

In order to test the effectiveness of the KNN algorithm and natural language processing techniques, we constructed a corpus of malicious and credible articles. The malicious articles used in our dataset were mostly articles and tweets on "Hillary Clinton" that had trended on social media and various websites that Snopes eventually investigated and labelled false. For this category, we also took into account all the tweets on "Hillary Clinton" published by President Donald Trump in an attempt to find the traits usually exhibited by deceptive news, such as ambiguity, abusive and mostly subjective and also to identify any hidden bias, totaling to 82 tweets/articles.

TABLE 1. SAMPLE OF MALICIOUS ARTICLES ON "HILLARY CLINTON" LABELLED BY SNOPES.

Date Published	Source	Fake Tweets on "Hillary Clinton"
29 May, 2017	Fresh News	BREAKING: Hillary Clinton Found Dead.

978-1-5386-5163-6/18/$31.00 ©2018 IEEE

29 Oct, 2016	TruthFeed	BREAKING : List of States Allowing you to CHANGE YOUR VOTE in Light of Hillary's Federal Investigation
10 Nov, 2016	Christian Times Newspaper	Bill Clinton just got served — by his own wife.

For credible articles, we use Twitter advanced search option to make queries and collected 20 tweets each from five most credible news sources – The Economist, Reuters, CNN, Wall Street Journal and Guardian – on the same subject in between 1st January 2016 to 1st November 2016, totaling 100 tweets.

TABLE 2. SAMPLE OF CREDIBLE ARTICLES ON "HILLARY CLINTON" WE CONSTRUCTED VIA CREDIBLE SOURCES. THESE SOURCES SCORED HIGH IN THE TRUST RATIO SCALE BASED ON TWO POPULAR SURVEYS.

Date Published	Source	Credible Tweets on "Hillary Clinton"
28 Oct, 2016	Reuters	Head of FBI says agency to review more emails related to Hillary Clinton's private email
25 Oct, 2016	The Economist	A back-of-the-envelope calculation shows that Hillary Clinton has a 96% chance of winning the White House
25 Oct, 2016	CNN	Hillary Clinton is leading Donald Trump by 7 points in North Carolina, a new poll shows

B. *Data Exploration*

To determine which features may be the most effective for classification, we compared the article sentiment magnitude vs. length of article, the sentiment of credible and malicious article and the relationship between average length of each word and the number of words in the article to understand the relationship between select features of the malicious and credible articles.

II. PREPROCESSING

Before performing our analysis, we cleaned our data by making the data structured. The process is widely known as "text normalization". We discuss the steps of our process in greater lengths below.

A. *Parts-of-Speech tagging*

The NLTK framework is used for processing natural languages and providing comprehensive support for various NLP related tasks. We used NLTK library to assign parts-of-speech to each token in an attempt to recognize words of importance. The output generated by module is shown in Fig.1.

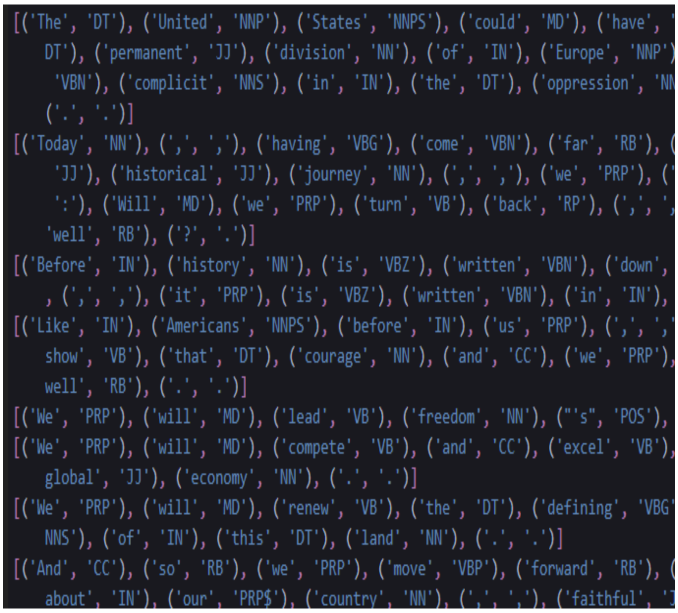

Figure 1. POS Tagging Each Token After Tokenizing Tweets

Named Entity Recognition

This is the first step after POS tagging towards information extraction from unstructured data. Named Entity Recognition is part of the information extraction. It is used to prevent splitting up tokens, like proper nouns (e.g. White House) which can lose their inherent meaning when broken up.

Figure 2. POS Tagging Each Token After Tokenizing Tweets

B. *Searching Global News*

We used Event Registry API to search for world news events. The system is capable to identify a bulk of articles that has coherence with the same event. It is able to identify groups of articles in different languages that describe the same event and represent them as a unique event. From articles in each event it can then extract events significant information, such as event related place, time and date, personalities who are related and what is it talking about. Extracted information is kept in a database. A dedicated interface for the users is available that allows users to search for events using advanced search options, to visualize and aggregate the search results, to analyze individual events and to identify correlated events. We used named entity recognition mentioned above to extract searchable and meaningful token from the headline of the article and simultaneously performed queries in Event Registry based on those extracted tokens to fetch news published by various news outlets or portals around the globe on that particular topic. The output of our module is shown in Fig 3.

```
q = QueryArticlesIter(keywords=QueryItems.OR(["Hillary", "Russia"]))
```


```
{'id': '11779704', 'uri': '833562650', 'lang': 'eng', 'isDuplicate': True, 'date': '2018-03-17', 'time': '19:53:00', '
    dateTime': '2018-03-17T19:53:00Z', 'sim': 0, 'url': 'http://www.vvdailypress.com/zz/news/20180317/
    latest-ex-cia-director-accuses-trump-of-moral-turpitude?rssfeed=true', 'title': "Trump lauds firing of ex-top FBI
    official as 'great day'", 'body': 'Attorney General Jeff Sessions fired Andrew McCabe, a former FBI deputy
    director long scorned by Trump\n\nWASHINGTON -- In what President Donald Trump called "a great day for Democracy,"
    Attorney General Jeff Sessions fired Andrew McCabe, a former FBI deputy director long scorned by Trump, two days
    before McCabe\'s scheduled retirement date, acting on the recommendation of bureau disciplinary officials.\n\n
    McCabe suggested the move was part of the Trump administration\'s "war on the FBI." Trump tweeted in praise of
```


Figure 3. Returns Related News Articles Based On Query

C. *Performing Image Search*

We designed this module to collect all the noun phrases extracted from the headline of the article and concatenates to construct an URL to display all the Google Images based on the query. This will act as a visual aid to what sort of content the user is consuming. The output of our module is shown in Fig 4.

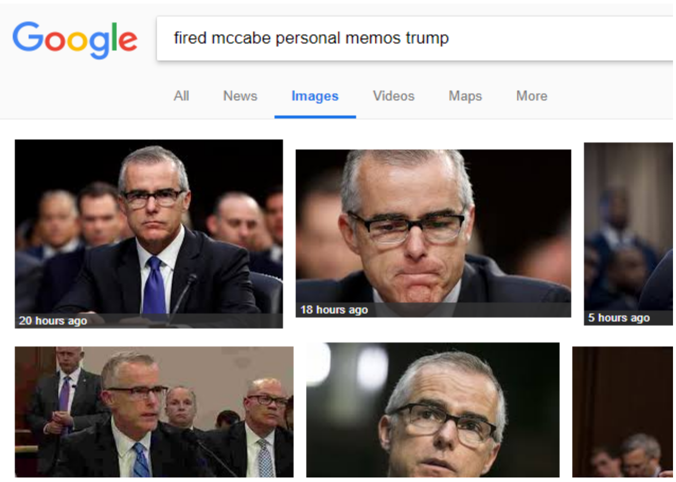

Figure 4. Based On The Headline, It Is Able to Fetch Relevant Images from Google

III. Sentiment Analysis

Stance detection for tweets involves detecting if the tweet is in FAVOR or AGAINST a particular target which can be a person, a trending topic, etc. For this paper, we decided on our target as "Hillary Clinton". The idea was to collect two sets of dataset, one should be labeled as "Legitimate" and the other as "Fake". We performed sentiment analysis on 100 credible tweets on "Hillary Clinton", measuring Polarity (-1 to 1). The result is shown in Fig 5.

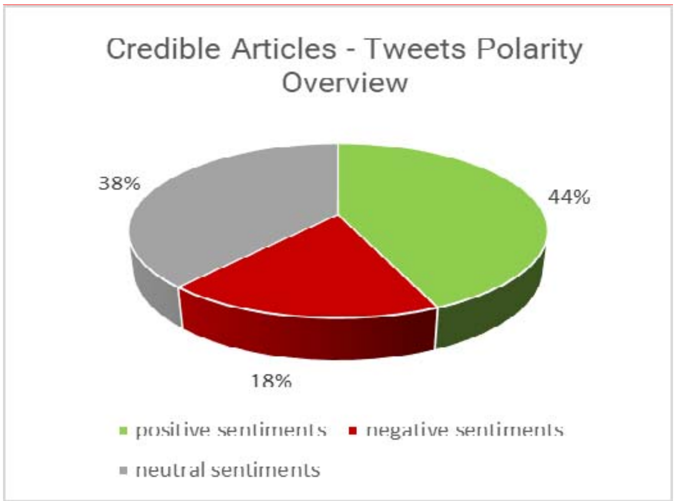

Figure 5. Percentage of Credible Tweets With Positive, Negative and Neutral Sentiments

The pie chart in Fig.5 tells us that the credible tweets on "Hillary" were mostly Positive or Neutral in terms of polarity during that certain time period. We can therefore hold it as one our markers for constructing predictive model.

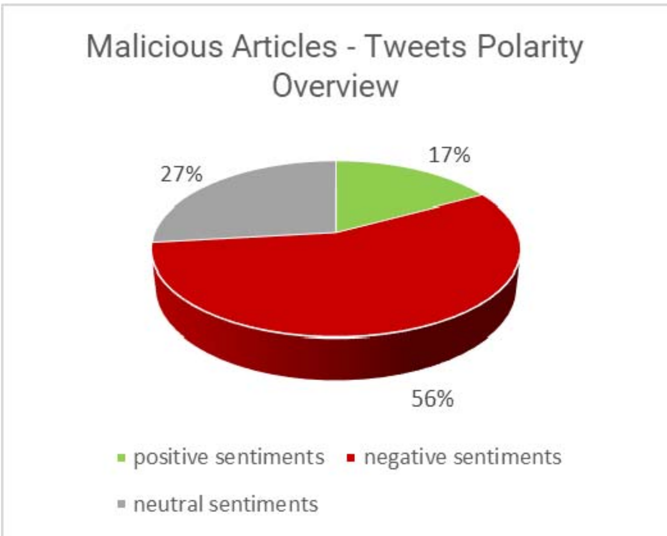

Figure 6. Percentage of Malicious Tweets with Positive, Negative and Neutral Sentiments

We later performed Sentiment Analysis on the malicious tweets, labelled by Snopes, and Donald Trump's tweets on "Hillary Clinton" published within a certain period. The result, shown in Fig.6, exposes the hidden bias towards the subject by the authors.

A. *Bag-of-Words (BoW) model*

Perhaps the simplest method of representing texts is the bag of words approach, which regards each word as a single, but significant unit. We created a BoW model for the two datasets that we prepared in an attempt to find which word frequencies in each categorized tweets. We filtered through the 50"Hillary" tweets made by Donald Trump and appended only the noun-phrases in our corpus. We then created the BoW model based on the corpus. Our results are shown in figures below.

```
[[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  Tweet 1
 [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  Tweet 2
```

Figure 7. Every Index Represents a Unique Word from The Entire Corpus of Tweets. 0 Means There Are No Occurrence of The Word, 1 Means The Word Occurred Once, and So On.

Our findings from this analysis are that Trump's polarized tweets were very subjective in nature. He often used superlative words and the word "big" frequented in his tweets which is not specific in nature, to describe a situation and "phony" and "crooked", which is abusive in nature, to describe Hillary Clinton. We performed the same comprehensive analysis on the credible "Hillary" tweets as well to perform a comparative study.


```
{'crooked': 17, 'hillary': 36, 'clinton': 10, 'was': 88, 'john': 40, 'podesta':
 62, 'big': 6, 'money': 51, 'russia': 71, 'dollars': 26, 'wife': 90, '
political': 63, 'run': 70, 'drain': 29, 'swamp': 80, 'clintons': 11, 'dems'
: 21, 'ads': 0, 'facebook': 32, 'media': 50, 'they': 83, 'wow': 91, 'james'
: 39, 'comey': 13, 'terrible': 81, 'candidate': 9, 'hits': 37, 'dnc': 25, '
fbi': 34, 'director': 24, 'free': 35, 'pass': 57, 'bad': 3, 'deeds': 19, '
phony': 60, 'new': 53, 'polls': 64, 'fake': 33, 'intelligence': 38, 'bill':
 7, 'uranium': 87, 'russian': 72, 'speech': 76, 'korea': 43, 'nukes': 54, '
election': 30, 'loss': 46, 'steve': 78, 'bannon': 4, 'thanks': 82, 'donald'
: 27, 'authorities': 2, 'whereas': 89, 'special': 75, 'council': 15, '
pocahontas': 61, 'primaries': 66, 'lets': 45, 'dept': 23, 'confirms': 14, '
many': 49, 'real': 69, 'story': 79, 'collusion': 12, 'donna': 28, 'book': 8
, 'primary': 67, 'crazy': 16, 'bernie': 5, 'justice': 41, 'department': 22,
 'sorry': 74, 'paul': 59, 'manafort': 48, 'trump': 85, 'presidential': 65,
'movie': 52, 'star': 77, 'and': 1, 'kasich': 42, 'passion': 58, 'democratic
': 20, 'party': 56, 'sanders': 73, 'unfair': 86, 'obama': 55, 'lynch': 47,
'law': 44, 'enforcement': 31, 'decisions': 18, 'purposes': 68, 'totally':
84}
```


Figure 8. Vocabulary Corpus for The Trump's "Hillary" Tweets. It Represents All The Unique Noun-Phrases Present In These Tweets With Their Associated Unique Id Number.

Based on our findings, we can assert that legitimate tweets in general are more specific and more objective in nature. They are rather critical than abusive. They refrain from using strong words if not absolutely necessary and would only do so to quote somebody. Fake news, on the other hand, is rather ambiguous, and even if not necessarily abusive, it is highly polarized. Their hidden bias can be reflected on their sentiment analysis.

```
word 6 occurs 3 times
word 7 occurs 1 times
word 8 occurs 1 times
word 9 occurs 1 times
word 10 occurs 4 times
word 11 occurs 1 times
word 12 occurs 1 times
```

Figure 9. This Tells Us The Occurrence of Each Word In All of The 50 Tweets

K-nearest neighbors (KNN) is a classification technique used for both classification and regression predictive problems. At its most basic level, it is essentially classification by finding the most similar data points in the training data, and making an educated guess based on their classifications. The steps of applying KNN algorithm were - 1. Computing a distance value between the item to be classified and every item in the training data-set 2. Picking the k closest data points (the items with the k lowest distances) 3. Conducting a majority vote among those data points the dominating classification in that pool is decided as the final classification. There are many different ways to compute distance, as it is a fairly ambiguous notion, and the proper metric to use is always going to be determined by the data-set and the classification task. Two popular ones, however, are Euclidean distance and Cosine similarity. We went for Euclidean distance, the formula of which is shown in Fig.10.

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

Figure 10. General Formula for Euclidean Distance. Euclidean Distance Is Essentially The Magnitude of The Vector Obtained By Subtracting The Training Data Point From The Point to Be Classified.

The polarity of our randomly selected tweet on Hillary was -0.3. The application of KNN algorithm required us to first find the k nearest neighbors of our test data. We decided on k=3, i.e. the three nearest neighbors. By applying the Euclidian equation, we calculated the distances and located its three nearest neighbors. For this particular sample, out of the 3 nearest neighbors, 2 was labeled fake and 1 was labeled legitimate. The test sample tweet was fake. Our system also predicted the tweet to be fake with an accuracy of 66.66%. The accuracy can be fine-tuned by increasing the number of tweets in our dataset.

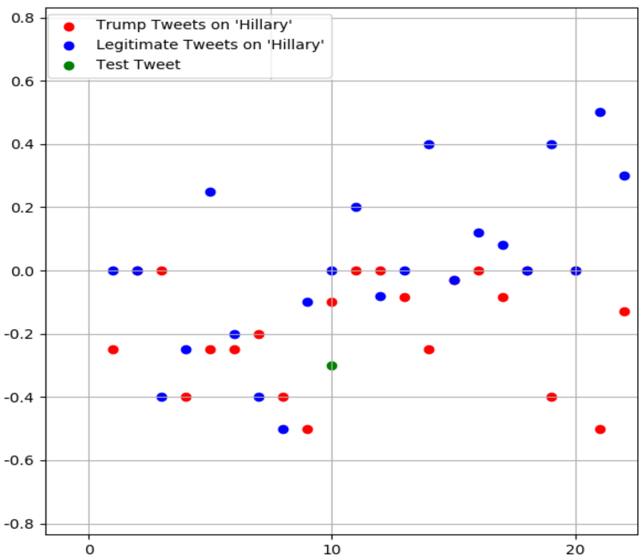

Figure 11. Polarity vs. Tweets. Red plots represent Trumps tweets and Blue plots represent credible tweets. The single green plot represents our test data.

IV. RELATED WORKS

Although current approaches taken to validate news and facts and filtering out deceptive news is limited, there is a good foundation on which more sophisticated systems can be build. Organizations like Snopes and PolitiFact verify claims manually and flag deceptive stories in an attempt to prevent the widespread of viral fake news story. However, manual fact checking and source validation is very time consuming and is not feasible when taken into account the speed at which viral news travels. Miraj Patel in his paper [1] uses a set of articles flagged as false by Snopes and implements various machine learning algorithms (k-nearest neighbors, support vector machines, and long short-term memory networks) to see how accurately the training model is able to predict whether a news is fake based on sentiment-related features. Hadeer Ahmed in his paper [2] developed an n-gram to automatically detect fake contents with a focus on fake reviews and news. He also discusses two extraction and six machine learning classification techniques and also investigates the impact of keystroke features on the accuracy of n-gram models.

V. CONCLUSION

In this paper, we discussed the computational linguistics implementations we have used to perform linguistic analysis on tweets to observe patterns exhibited by legitimate and fake or ambiguous news. We have designed separate independent modules that are aimed to assist humans in making better decisions of detecting deception in news. We deconstructed the grammar of the tweets for in-depth analysis and constructed a comprehensive BoW model based on the categorized labeled tweets. We have compared how polarity and subjectivity varies between legitimate and polarized tweets, considering the topic of the tweets to be same. Our future research on this is going to be about performing in-depth stance detection analysis on top of the BoW models that we constructed in this research. We have designed separate independent modules that are aimed to assist humans in making better decisions of detecting deception in news. We deconstructed the grammar of the tweets for in-depth analysis and constructed a comprehensive BoW model based on the categorized labeled tweets. We have compared how polarity and subjectivity varies between legitimate and polarized tweets, considering the topic of the tweets to be same. Our future research on this is going to be about performing in-depth stance detection analysis on top of the BoW models that we constructed in this research.

REFERENCES

[1] Miraj Patel, “Detection of Maliciously Authored News Articles,” December 11, 2017.

[2] Hadeer Ahmed, “Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity”, 2017