Evaluate Trends in Misinformation on Twitter in the run-up to the 2020 US Presidential Elections

Abhishek Kumar Bais
Department of Computer Science
San Jose State University
San Jose, CA, USA
abhishek.bais@sjsu.edu

Jimmy Liang
Department of Computer Science
San Jose State University
San Jose, CA, USA
jimmy.liang@sjsu.edu

Rohan Kumar
Department of Computer Science
San Jose State University
San Jose, CA, USA
rohan.kumar@sjsu.edu

Samer Baslan
Department of Computer Science
San Jose State University
San Jose, CA, USA
samer.baslan@sjsu.edu

***Abstract*— Misinformation, alternative facts and falsehoods have all been used to describe consumable news both in social media and traditional news outlets. As social media, and especially Twitter, increasingly becomes the primary source of information for many Americans, having a way to determine the veracity of information reaching the people is more important now than ever. In this paper, utilizing natural language processing techniques such as Bag-of-Words (*BOW*), Bag-of-Ngrams, their TF-IDF variants and supervised learning models such as Multinomial Naïve Bayes and Logistic Regression we will analyze Twitter data referencing the two presidential candidates, Donald Trump of the Republican Party and Joe Biden of the Democrat Party to evaluate trends in misinformation on Twitter in the run-up to 2020 US presidential elections. We seek to understand the differences in velocity of spread between real and fake news.**

***Index Items* —Twitter, Natural Language Processing, Machine Learning, Bag of Words, Bag of Ngrams, TF-IDF, Supervised Learning, VADER, Sentiment Analysis, Naïve Bayes, Logistic Regression.**

I. INTRODUCTION

According to a 2019 PEW Research Survey, there are over 51 million active, adult twitter users[1] in the USA; 71% of them use twitter as their primary source of information , despite concerns of widespread misinformation, alternative facts and falsehoods on the social media platform. Every day more and more people are taking to twitter to share their opinions and engage in debate and discussion on a wide variety of subjects. As the volume of political tweets increases, so does the likelihood of misinformation and propaganda, reaching the masses. Many studies have investigated the phenomenon of misinformation on Twitter[2,3,4]. In the political space, it has been observed that twitter traffic goes up particularly in the run-up to the presidential elections. For example, in the six-month period between June 1, 2016 and November 8, 2016, a total of 171 million unique political tweets were made with the keywords "Trump" or "RealDonaldTrump" or "DonaldTrump" or "Hillary" or "Clinton" or "HillaryClinton"[5].

To gauge the extent of misinformation in circulation on social media and ascertain whether it played any role in influencing the outcome of the elections, after the 2016 US Presidential elections concluded, several investigations were undertaken, including those by the US Congress and the FBI. These investigations found that while the election results were legitimate, and that no votes were changed, there was widespread misinformation in circulation, likely enough to create echo chambers to influence the electorate's mind one way or the other.

In-order to automatically classify news items into real and fake categories, two approaches are common[10]. The first approach is a "*Fact-Checking*" approach. It is described by Conroy et al., 2015[6] as one, where the truthfulness of news content is determined after careful validation against verified known truths available in knowledge databases such as DBPedia[7] or websites such as FakeCheck.org and Snopes.com.

While "*Fact-Checking*" approaches provide high veracity, because they rely on verifying news content against known truths that have already been documented they alone cannot be the go-to method of choice for more recent news as known truths about them are yet not documented.

The second approach is a "*collective wisdom*" approach. This approach is described by Hannak et al., 2014[8] and Jin et al., 2014[9], as one that aggregates sentiments and skepticisms to news content on social media outlets such as twitter to gauge its credibility. In this approach natural language processing (*NLP*) techniques are used to extract linguistic features from news content to gain valuable insight into the nature of the text, whether mostly relaying real or fake news, from the choice of emoticons, word choices, and frequently occurring words in the news content.

In this paper, we employed a hybrid of both "*Fact-Checking*" and "*collective wisdom*" approaches to build an "*Automatic Misinformation Detection System"* and use it to evaluate trends in misinformation on twitter in the run-up to the 2020 US presidential elections.

II. METHODOLOGY

Building an "*Automatic Misinformation Detection System"* is a multi-step process. In the first step, data comprising a "*fact-*

checked" corpus of already available, human validated, labeled, 12,999 fake and 15,712 real news relaying tweets referencing "RealDonaldTrump" and "HillaryClinton" from the 2016 US presidential election cycle and an "*unlabeled*" tweet corpus of over 1.7 million tweets, 970,919 referencing "RealDonaldTrump" and 776,886 referencing "JoeBiden" from the one-month period immediately preceding the 2020 US election cycle is collected.

In the second step, both the "*fact-checked*" and the "*unlabeled*" tweet corpus is cleaned to remove hashtags, acronyms, slang words, misspellings, content-less stop words and replace emoticons with appropriate sentiment words.

In the third step, *VADER,* a human scored, context-free text sentiment lexicon is used to perform sentiment analysis on the "*unlabeled*" tweet corpus. Based on the obtained sentiment scores, tweets are labelled as relaying mostly real or fake news. *VADER* allows for a hybrid approach to data classification be taken, enabling trends in misinformation on twitter be evaluated using more recent, "*unlabeled*", voluminous data rich in variety with older, "*fact-checked*" data. This approach also helps avoid overfit and increases confidence in the evaluations.

In the fourth step, both the "*fact-checked*" and the tweets guided by "*collective wisdom*" of aggregated sentiments and skepticisms are merged to create a labeled data corpus of 1,77,6516 tweets for misinformation trend analysis. Information in these tweets not required to evaluate misinformation trends is dropped.

In the fifth step, linguistic feature extraction using Natural Language Processing (*NLP*) techniques such as "*Bag of Words*", "*Bag of N-grams*" and their "*TF-IDF*" variants is performed on a randomly selected sample 124,353 tweets from the tweet corpus. In this paper, a sample of tweets is taken for feature extraction for lack of access to powerful compute-resources for big data mining.

In the sixth step, supervised, machine learning models such as *Multinomial Naive Bayes* and *Logistic Regression,* trained on the extracted features together with k-fold cross validation, is used to validate the accuracy of the hybrid tweet classification. Upon achieving acceptable accuracy and confidence in the classification, the now labelled 1.7 million tweets referencing the two presidential candidates Donald Trump and Joe Biden is visualized in a time-series manner, to evaluate trends in misinformation on twitter in the run-up to the 2020 US presidential elections.

An architecture framework *Automatic Misinformation Detection System* is shown in Fig 1.0 below.

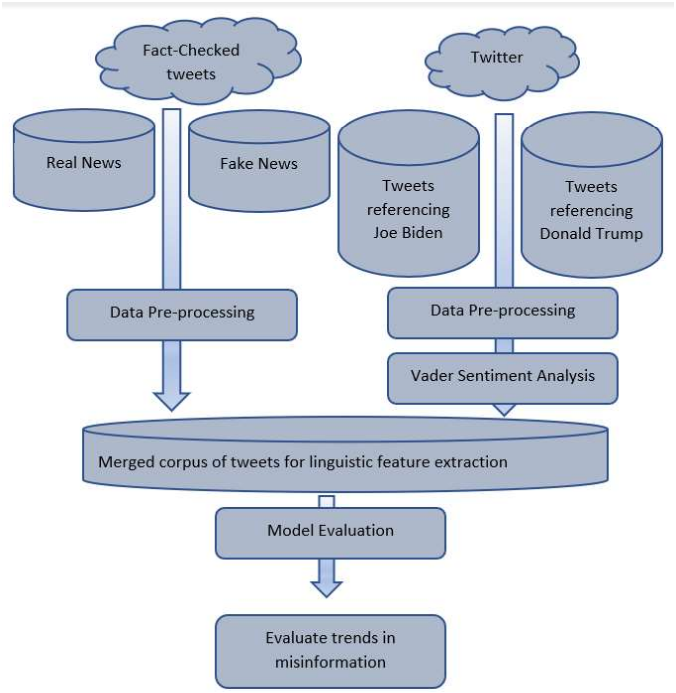

Fig 1.0. Architectural framework of Automatic Misinformation Detection System

Section III of this paper describes the data exploration performed and outlines the different data pre-processing techniques used. Section IV describes the *VADER* sentiment analysis process. Sections V describes the linguistic feature extraction techniques that *Multinomial Naïve Bayes and Logistic Regression* models leverage to validate the accuracy of classification. Section VI tabulates the results of the model evaluation using different accuracy metrics and plots time series plots of real and fake news in circulation. Finally, Section VII analyzes the time-series plots to evaluate trends in misinformation on twitter in the run-up to the 2020 US presidential elections.

III. DATA ANALYSIS

A. Data Exploration

The "*fact-checked"* corpus of training data comprises pre-labelled 12,999 fake news and 15,712 real news content in tweets referencing "RealDonaldTrump" and "HillaryClinton" from the 2016 US presidential election cycle. The 12,999 tweets in the fake news dataset relays several different types of news, dominated mostly by conspiracy theories and bias as shown in the Fig 2.0 below.

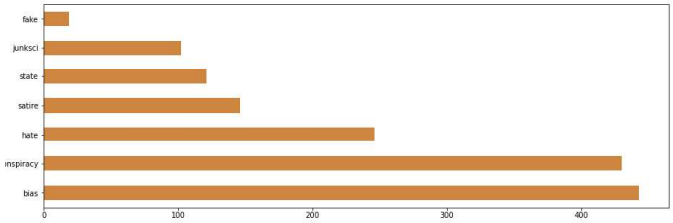

Fig 2.0. Different types of news in fake news dataset

The fake news originates from many countries around the world, the top six countries of origin being USA, Great Britain, Russia, Germany, Tuvalu, and France.

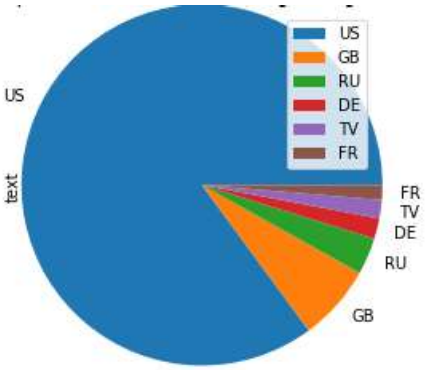

Fig 3.0. Top 6 countries of origin of fake news

Fake news also came from multiple sources. The now banned "*Iranian presstv*", a Russian propaganda outlet "*activistpost*", a white supremacist linked website "*truthfeed*" were the top spreaders.

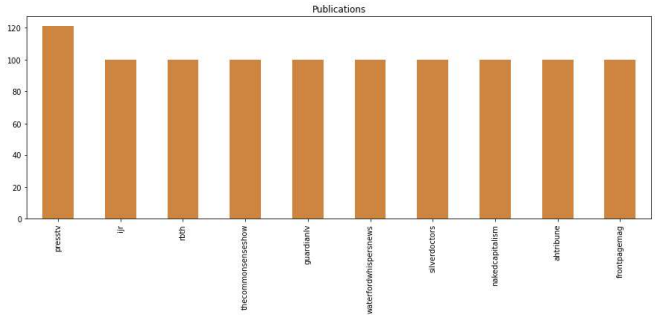

Fig 4.0. Top sources of fake news

The 15,712 tweets in real news dataset also comes from a variety of different sources with Reuters being the largest source.

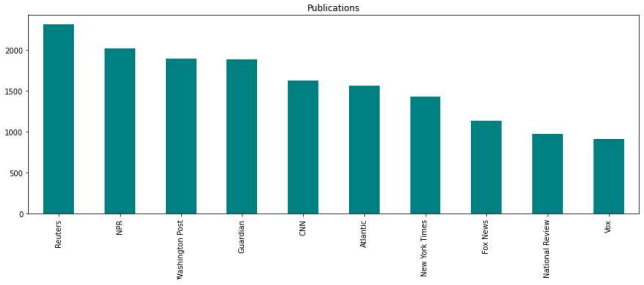

Fig 5.0. Top sources of real news

Certain unique characteristics such as longer titles and more widespread use of uppercase is observed in tweets labelled fake than real. These are shown in Fig 6.0, and Fig 7.0 below

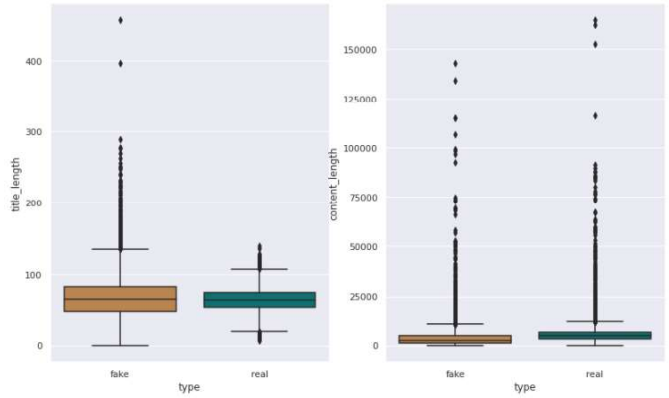

Fig 6.0. Title, Content length comparison of tweets labelled fake, real

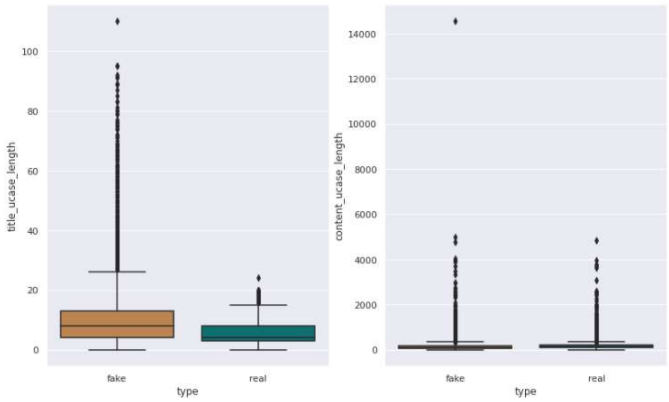

Fig 7.0. Use of uppercase in tweets labelled fake, real

The 1.7 million tweets referencing "RealDonaldTrump" or "JoeBiden" collected over the one-month period immediately preceding the 2020 US presidential elections, included 970,919 tweets referencing Donald Trump and 776,886 tweets referencing Joe Biden. The tweets were also made from all over the world including tweets with a geo-tag in Antarctica!

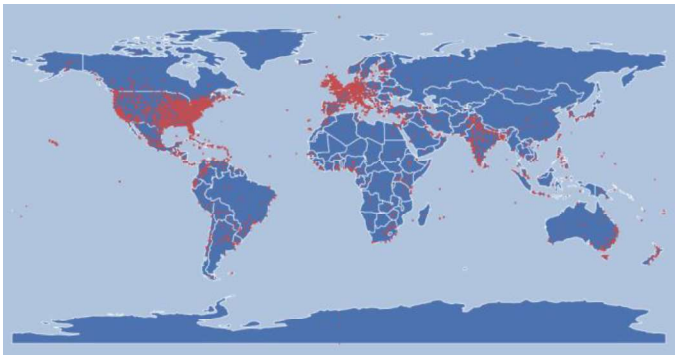

Fig 8.0. Origin of tweets referencing Donald Trump, Joe Biden

The top five countries of origin of most tweets for which geographic location was available is shown in Fig 9.0. These countries are USA, UK, Canada, and Germany and France. 337,079 tweets originated from the USA. The tweets came from multiple mediums such as Twitter web, Iphone, Android, Ipad,

Tweet deck and Hootsuite. The vast majority of tweets, over 400,000 came from Twitter web.

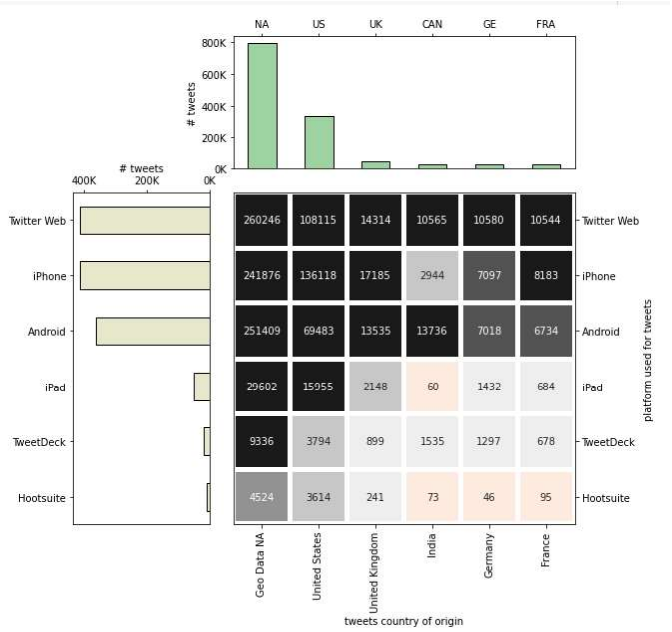

Fig 9.0. Tweets quantified by medium and country of origin

B. Data Preprocessing

Although the dataset contained both English and non-English tweets, as shown in Fig 10.0 below, over 60% of the tweets were made in the English language.

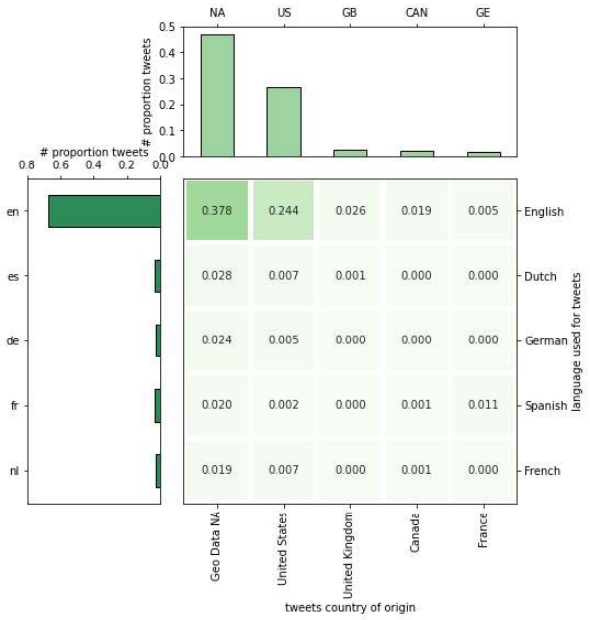

Fig 10.0. Tweets quantified by source language

The tweets contained emoticons, whitespace both trailing and leading, numbers, links, user information, single characters, repeat characters, erroneous punctuation, spelling mistakes, etc. These features increase dimensionality and require careful data processing to achieve accurate sentiment analysis results. Pre-processing of twitter data involves several techniques, many of them are outlined by Angiani et al[12]. In our paper we employed the data processing techniques outlined in the Table 1.0 below.

High dimension data construct	Preprocessing Techniques
Emoticons	Replace with 'positive', 'negative' keywords
Trailing or leading white spaces	Remove
Numbers	Remove
Links/ URLs	Remove
User Information	Remove
Single character words	Remove
Repeat characters in words	Replace with single character
Erroneous punctuations	Remove
Word Contractions	Replace with expanded form
Unnecessary capitalizations	Replace with lowercase
Contentless words	Stop words removal
Inconsistent word form	Lemmatize

Table 1.0. Preprocessing techniques used

Emoticons like *"sad", "weep", "love", wink", "laugh"* and *"smiley"* were recognized and replaced by *"positive"* or *"negative"* keywords. Over 100-word contractions such as "aren't", "couldn't" were replaced with their expanded form words such as "are not", "could not". Several tweets contained words with repeat characters such as *"baaaaaaad"*. From a sentiment analysis perspective, *"baaaaaaad"* and *"bad"* relay the same emotion. Replacing repeat characters by a single character allowed differently spelled words, conveying the same meaning to be treated identically. In this way, using various pre-processing techniques we transformed raw twitter data to a more feature extraction friendly form, suitable for linguistic feature extraction.

IV. VADER SENTIMENT ANALYSIS

Valence Aware Dictionary for Sentiment Reasoning *(VADER)* is a sentiment lexicon developed by Hutto and Gilbert[11] that enhances well-established sentiment word-banks such as (LIWC, ANEW, and GI) by adding emoticons, acronyms and initialisms to extract over 9,000 linguistic features from text data. Hutto and Gilbert then used crowdsourcing to obtain score texts on a scale ranging from "[–4] Extremely Negative" to "[4] Extremely Positive", with allowance for "[0] Neutral (or Neither, N/A)".

V. LINGUSTIC FEATURE EXTRACTION

To evaluate the tweet classification, machine learning models such as *Multinomial Naive Bayes* and *Logistic Regression* from the scikit-learn python library are trained and classifications validated using a k-fold cross validation process. In other to facilitate robust and accurate model evaluation, multiple Natural Language Processing (*NLP*) techniques for linguistic feature extraction such Bag-of-Words *(BoW)*, Bag-of-Ngrams, and their Term Frequency – Inverse Document Frequency *(TF-IDF)* variants are used.

A. Bag of Words (BOW)

In this approach, the twitter data tweet corpus is represented in terms of a *BoW* vocabulary discarding grammar but keeping multiplicity. Based on the frequency of occurrence of each word a numeric vector state model of fixed length is created. Linguistic features are then extracted from the vector state model representation.

B. Bag of N-grams

In this approach, the twitter data corpus is represented as an unordered set of *N-grams*. Although the bag of *N-grams* representation closely resembles the *BoW* representation, by associating weights to group of words that occur together, more accuracate feature extraction can be performed. For N = 1 the *BoW* model can be viewed as a special case of the N-gram model. In our paper we used N=2 and N=3 grams. Fig 11.0 shows the top ten, 2-gram, and 3-grams of the tweets referencing the two presidential candidates.

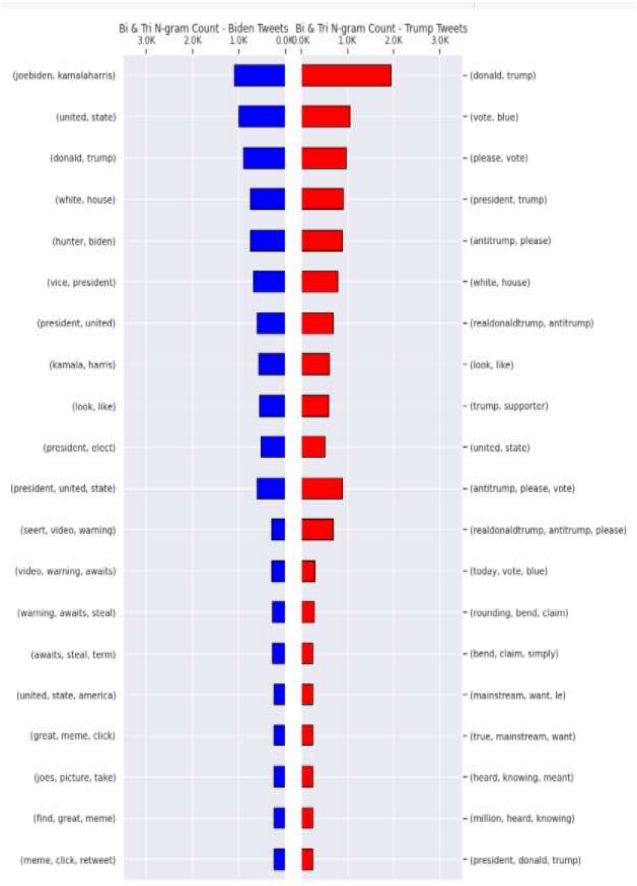

Fig 11.0. Top ten words in 2-gram, 3-gram of tweet corpus

C. Term Frequency-Inverse Document Frequency (TF-IDF)

In this approach, the relevance of words in determining the nature of the tweet, whether relaying mostly real or fake news is additionally accounted for. Mathematically, Term Frequency, *TF(d, t)* is defined as

$$TF(d,t) = \begin{cases} 0 & if\ freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & otherwise. \end{cases}$$

while, Inverse Document frequency is defined as,

$$IDF(t) = \log \frac{1+|d|}{|d_t|},$$

where d, is the entire tweet corpus while d_t is the set of tweets containing term t. If $|d_t| << |d|$, the term t will have a large IDF scaling factor and vice-versa.

Words found uniquely in real or fake news are given higher importance than words found in both tweet categories. The Term Frequency-Inverse Document Frequency is defined as

$$TF\text{-}IDF(d,t) = TF(d,t) \times IDF(t).$$

Once the linguistic features are extracted, the *Multinomial Naïve Bayes* and *Logistic Regression* models are trained and validated using the accuracy and precision and recall metrics, defined underneath

$$Accuracy = (TP + TN)\_ (TP + FP + TN + FN) \quad -1$$

$$Precision = TP\ (TP+FP) \quad -2$$

$$Recall = TP\ (TP+FN) \quad -3$$

where, TP is true positive (the occurrences of correctly identified tweets relaying real or fake news) and FP is false positive (the occurrences of incorrectly identified tweets relaying real or fake news).

VI. RESULTS AND DISCUSSION

An evaluation of the tweet classification using *Multinomial Naïve Bayes* and the *Logistic Regression* supervised learning models built using different linguistic features is captured in the Table 2.0 below.

Feature Extraction Technique	**Classifier**	**Accuracy**	**Precision**	**Recall**
Bag of words	Logistic Regression	84.49%	92.6%	62%
Bag of words + Tf-IDF		88%	93.5%	73%
Bag of n-grams		89.37%	94.8%	75%
Bag of n-grams + Tf-IDF		83.6%	94.8%	55%
Bag of words	Naïve Bayes	74.1%	76.3%	52.7%
Bag of words + Tf-IDF		80.66%	82.7%	72%

Bag of n-grams		76.6%	76.4%	78%
Bag of n-grams + TF-IDF		74.2%	98.39%	19%

Table 2.0. Model Evaluation using different linguistic feature extraction techniques

The ROC curves for the model evaluation are shown below.

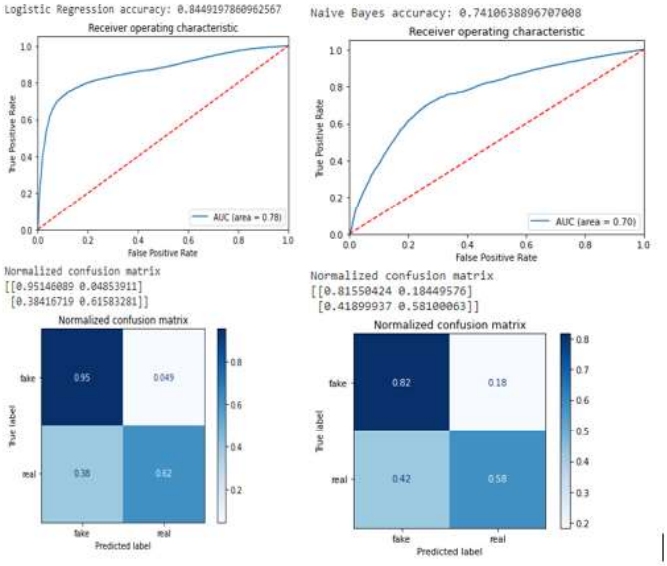

Fig 12.0 ROC Curve for BoW

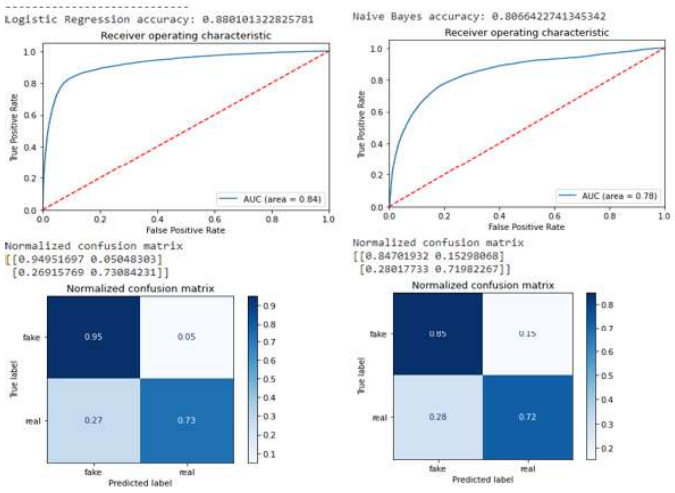

Fig 13.0 ROC Curve for BoW with Tf-IDF

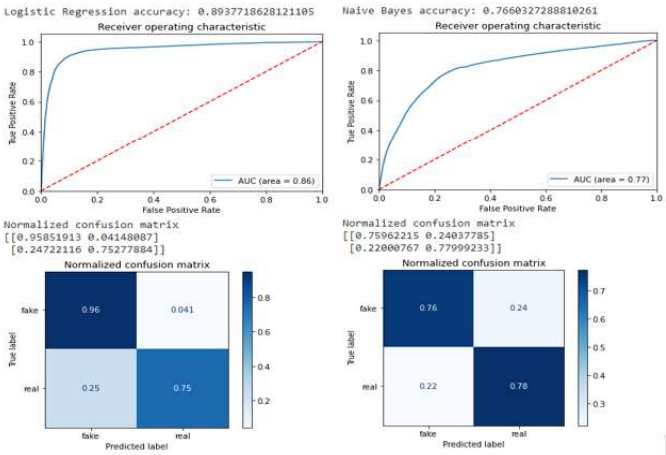

Fig 14.0 ROC Curve for Bag of Ngrams

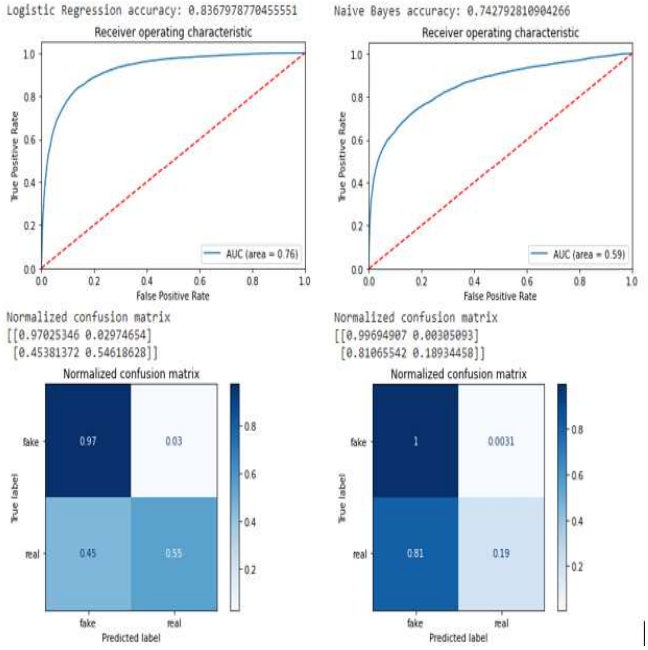

Fig 15.0 ROC Curve for Bag of Ngrams with Tf-IDF

In-order to understand the difference in velocities of spread of real and fake news, "*Reach-Score*" of a tweet "*t*" is defined as under

*Reach-Score(t) = L * R*
where,
t = Tweet labelled real or fake

L = Number of Likes of Tweet -1

R = Number of Re-tweets of Tweet -2

To give due importance to tweets not liked or re-tweeted even once, a correction of 1 unit is applied to L, R before computing *Reach-Score(t)* values.

A time-series analysis of all 1.7 million tweets referencing the two presidential candidates Donald Trump of the republican party and Joe Biden of the democrat party from the one-month immediately preceding the 2020 US presidential elections is captured in the Fig 16.0 and Fig 17.0 below.

Fig 16.0 captures the reach score of fake and real news circulation in tweets referencing both the presidential candidates as a percentage of tweets in their respective corpus.

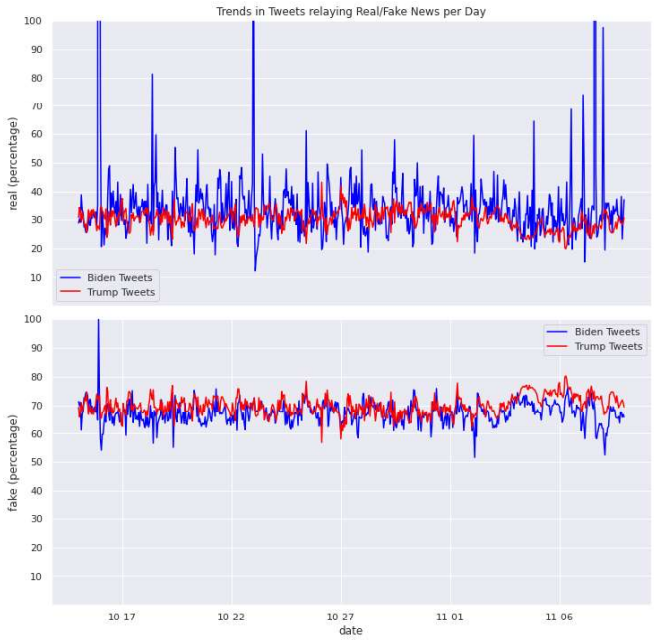

Fig 16.0. Reach-Score as percentage of tweets

Fig 13.0 captures mean reach score of fake and real news in circulation in tweets referencing both the presidential candidates as a percentage of tweets in their respective corpus.

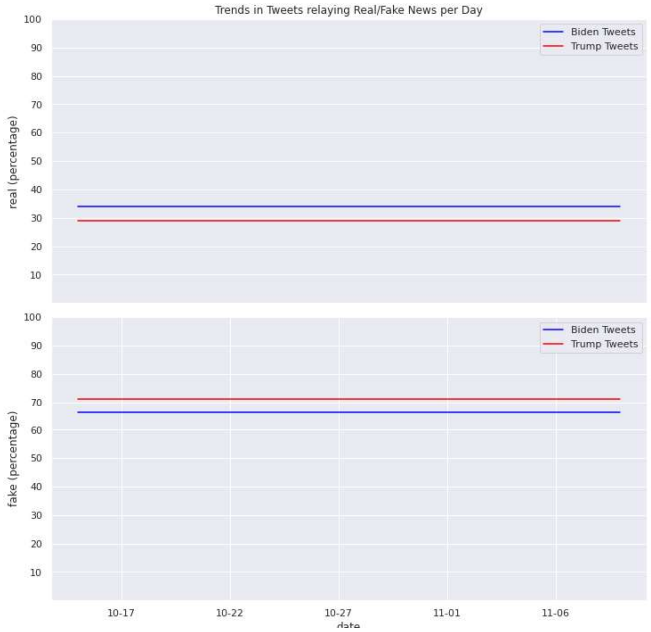

Fig 17.0. Mean Reach-Score as percentage of tweets

VII. CONCLUSION

This paper presents a hybrid approach combining "*fact-checking*" and "*collective wisdom*" of aggregated sentiments and skepticisms to create an "*Automatic Misinformation Detection System*" that classifies tweets as relaying mostly real or fake news.

The "*fact-checked*" corpus contains already available, human labelled, 15,712 tweets relaying real news from reputed news sources such as "*Reuters*", "*New York Times*", "*Washington Post*", "*NPR"* and "*Guardian*", and 12,999 tweets relaying fake news from sources known to peddle hate, conspiracy theories, satire and misinformation such as the banned "*Iranian presstv*", a Russian propaganda outlet "*activistpost*", a white-supremacist linked website "*truthfeed*", the "*conservativetribune*", "*onion*" and "*thespoof*" from the 2016 US presidential election cycle referencing Donald Trump and Hillary Clinton.

The corpus for "*collective wisdom*", sentiment and skepticism aggregation contain over 1.7 million tweets from the one-month immediately preceding the 2020 US presidential elections referencing Donald Trump and Joe Biden.

The system uses supervised learning models such as "*Logistic Regression*" and "*Multinomial Naïve Bayes*"; trained using multiple linguistic feature extraction techniques such as "*BoW*", "*Bag of N-grams*" and their Term Frequency Inverse Document Frequency (*Tf-IDF)* variants to validate the accuracy of classification over various metrics including ROC/AUC curve, precision and recall. An average accuracy of classification of 86.35% using the "*Logistic Regression*" model and 76.39% using the "*Multinomial Naïve Bayes*" model was achieved. For the same dataset, the "*Multinomial Naïve Bayes*" algorithm converged for all linguistic feature extraction techniques while the "*Logistic Regression Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm (lbfgs)*" algorithm failed to converge even after 100 iterations. Also, when the "*BoW*" technique was used, the "*Multinomial Naïve Bayes*" algorithm evaluated 5652 tweets/sec while 3454 tweets/sec using its Tf-IDF variant, 1.6X slowdown. When the "*Bag of N-grams*" technique was used, the "*Multinomial Naïve Bayes*" algorithm evaluated about the same 1,222 tweets/sec, both without and with its Tf-IDF variant.

Accepting the evaluated classification as reasonably accurate, time-series plots of trends in real and fake tweets were drawn and analyzed. From them, it was observed that misinformation has a high 65%-71% "*Reach-Score*" while real news has a low 30%-35% "*Reach-Score*" among the masses, in the tweets referencing both Joe Biden and Donald Trump. There is also lower circulation of misinformation in tweets referencing Joe Biden than in tweets referencing Donald Trump. When compared to a 2018 PEW Research Survey of how much misinformation and real news Americans believe normally circulates on twitter[13], it is observed that 8-14% higher

misinformation and 7-12% higher real news circulates on twitter in run-up to the elections than otherwise.

The trends show, much like the 2016 US presidential election cycle, the magnitude and reach of misinformation on twitter continues to be a serious challenge in the political space, particularly in the run-up to the elections, and has the potential to create echo chambers, big enough to influence the electorate's mind one way or the other.
e completion of this work.

IX. REFERENCES

[1] PEW Research, "10 facts about Americans and Twitter", Aug, 2019.

[2] Fang Jin, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, Chang-Tien Lu, Naren Ramakrishnan, ""Misinformation propagation in the age of twitter", IEEE, Volume:47, Issue 12, 2014

[3] Suchita Jain, Vanya Sharma, Rishabh Kaushal, "Towards automated real-time detection of misinformation on Twitter", 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

[4] Mabrook S. Al-Rakhami ; Atif M. Al-Amri, "Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter", IEEE, Volume 8, 2020

[5] Alexandre Bovet , Flaviano Morone & Hernán A. MakseJ, "Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump ", Scientific Reports, April 2017

[6] Nadia K. Conroy, Victoria L.Rubin, Yimin Chen, "Automatic deception detection: Methods for finding fake news", Proceedings of the Association for Information Science and Technology, 2016.

[7] DBPedia, https://wiki.dbpedia.org/

[8] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber, "Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations", Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14), Ann Arbor, MI, 2014

[9] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang, "News credibility evaluation on microblog with a hierarchical propagation model", Data Mining (ICDM), 2014 IEEE International Conference on, pages 230–239. IEEE.

[10] Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea1, "Automatic Detection of Fake News", Proceedings of the International Conference on Computational Linguistics (COLING 2018), New Mexico, NM, August 2018.

[11] C.J Hutto, Eric. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", 2015 Conference: Proceedings of the Eighth International AAAI Conference on Weblogs and Social MediaAt: Ann Arbor, MI.

[12] Giulio Angiani, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi "A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter", 2016 KDWeb.

[13] Pew Research Center, "News Use Across Social Media Platforms", Sept 10, 2018