Sentiment Analysis Methods to Mitigate Negative Effect of the COVID-19 Pandemic

Sofia A Mohamud
George Mason University
smohamu2@gmu.edu

***Abstract*—The goal of this research is to determine crucial factors that played a role in the number of confirmed COVID-19 infections within a given location. We hypothesize that political bias plays a significant role in the rise of COVID-19 cases globally and nationally; specifically, in overriding scientific reasoning for the delay or lack of deploying national policies to address the pandemic. Methods: To determine the validity of our hypothesis, we performed a literature review that identified statistical information on 1) the origins of the virus, 2) the lethality of the virus, and 3) potential parties responsible for the creation and release of the virus. In addition to the literature review, our team performed a behavioral analysis using information extracted from social media platforms to identify and determine behavior patterns associated with specific words related to the virus.**

***Index Terms*—Data mining, Sentiment, Pandemic, Machine learning**

I. Introduction

People around the globe use social media, especially Twitter, to convey, share, and gain health information. Humans share misinformation if it reinforces existing beliefs and prejudices, and this misinformation becomes viral in no time.

According to Laato et al. (2020) research model, ”a person’s trust in online information and perceived information overload are strong predictors of unverified information sharing” [1].

According to Pennycook et al. (2020) investigations of 1600 individuals, people share misinformative news without knowing the truthfulness and value of the news or information [2].

The primary threats regarding misinformation is that social media users tend to share/retweet fake COVID news more than facts-based news leading to conflicting and poor decision making which can be solved by advanced transfer learning models [3] [4].

According to the survey of 483 participants, conspiracy, political and religious misinformation beliefs about COVID-19 impairs decision-making and have a negative impact on individual responses [5]. It is hard to control the spread of misinformation on social media platforms given the popularity and reach. Nevertheless, social media platforms like Facebook and Twitter have taken significant actions ”to try to limit the proliferation of disastrous misinformation [6] regarding COVID-19 by removing fact-checked false and potentially harmful information” [5]. COVID-19 misinformation is distributed more on Twitter as compared to traditional media [7] [8].

The quick spread of COVID-19 has basically thrown the world into confusion. The initial few confirmed cases in the United States (US) were to be announced January 21, 2020 [9]–[11], [11], [12]. Fox News is the most viewed cable network in the US in comparison to CNN and MSNBC. The average age of viewers of Fox news is 68, which is higher compared to CNN and MSNBC viewers. Since both networks can reach out to people that are audiences above 65, which is a group that CDC alerts for potential higher risk from COVID-19. Fox News may apply considerable influence on the result from COVID-19. This is relatively accurate, provided that the aged spend more time watching TV in comparison to standard US citizens because they depend on television for information, facts, and news [13].

Different news platforms and experts like the Director of National Institute of Allergy and Infectious Diseases, Dr. Anthony Fauci, recommended starting from February that COVID-19 was a serious warning to the US. Observers found out that Sean Hannity of Fox News presented a particularly scornful story based on the virus. On the other hand, Tucker Carlson, was one among others who believed that coronavirus was a warning for the country starting February. Around January 28th, which was more than a month before the first COVID-19 death in the US, Tucker Carlson took to his showto talk and discuss more on the virus. His show provided more qualitative evidence [13].

II. dataset

COVID-19 Open Research Dataset Challenge is an open research dataset that was conducted by the White House [14]. It was a collaboration of leading research groups to present accurate numbers. It contains 59,000 scholarly articles that provide information on the coronavirus [5]. The research community has the leverage to access these datasets to provide more accurate results and information to the public using natural language processing and Artificial Intelligence (AI) techniques [15] [16]. Text and data mining techniques can be done through this method. The most common questions were ”What are the risk factors?”, ”What is the current information we know regarding the virus, genesis, and development?”, etc. Machine-readable articles are also provided through this dataset.

”Uncover COVID-19 challenge” is another dataset by Roche Data Science Coalition [5]. Numerous datasets can be found here that have been collected globally through 20 different sources. A few of the sources are Johns Hopkins,

```
In [49]: df=pd.read_csv("all-states-history.csv")
         print (df.shape)

         (15409, 42)
```

Figure 1

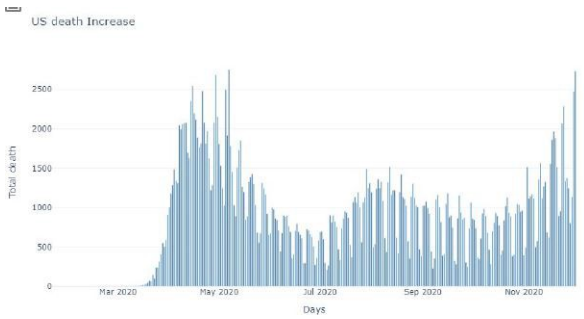

Figure 2

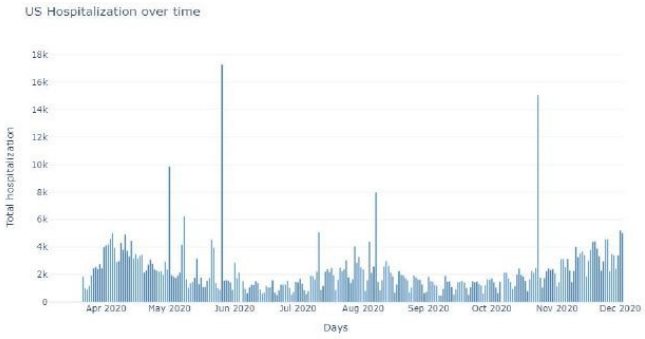

Figure 3

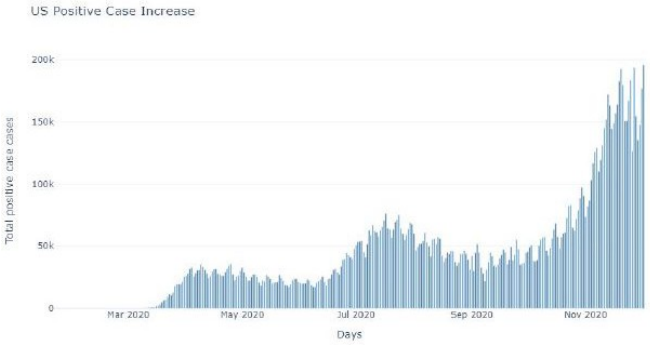

Figure 4

the WHO, etc., The goal of this dataset is to provide information on a global scale like "What is the number of cases globally?", "Which country is the most affected by this pandemic?", "What age group is the most affected?", and others.

In addition to these datasets, there has been an increasing number of efforts using Twitter data analysis. Various tweet analytics tools are available that help in determining features of social interactions of Twitter users and their behaviors. For this research, we extracted Twitter data related to COVID-19 misin formation using Twitter API.

This paper is based on COVID-19 data [14] [17]. The dataset was acquired from the COVID-19 Tracking project and New York Times. COVID-19 data, which is used to draw visualization on different questions like the number of confirmed cases, total number of deaths, number of recoveries and, so on for worldwide cases. This data set has 156292 records and eight fields.

Another data set used is the all-states-history.csv extracted from COVIDtracking.com. This data has 15409 records (rows) and42 columns. Figures 1, 2, 3 show the code used to extract the data, the increase in deaths, hospitalizations, and positive for the US.

The fourth visualization displays the number of positive COVID-19 cases inthe US. The graphs used in figures 2, 3, and 4 show the cumulative daily count of deaths, the cumulative daily count of hospitalizations, and the cumulative daily count of new confirmed cases in the US. The X-axis indicatesthe date of observation.

III. METHODS

For desired results, COVID-19 related dataset from Kaggle and COVIDtracking.com was processed and analyzed. Cases, deaths, and hospitalization numbers based on these datasets reflect cumulative totals since January 22, 2020 until December 2, 2020. Based on the frequency of how Kaggle and COVID-19 tracking website updates this data, they may not indicate the exact numbers of daily, hospitalization, and death cases as reported by State and Local government organizations or the news media. This paper uses daily new cases, hospitalization, and COVID-19 related death datasets for figures 5,6, 7.

For misinformation, controversy, and hoax, Twitter data was collected using Twitter API. This data set had 17560 records and seven fields.

Figures 8, 9, 10 are based on data analyzed using Python, Natural Language Processing (NLP), and Tableau.

```
# pass these authorization details to tweepy
api = tw.API(auth, wait_on_rate_limit=True)
# test authentication
try:
    api.verify_credentials()
    print("Authentication OK")
except:
    print("Error during authentication")
# Extracting Specific Tweets from Twitter
search_words = 'plandemic'
new_search = search_words +" -filter:retweets"
```

Figure 5

```
#Add time your file was created to discr
filename = 'Covid19_tweet_Worldwide.csv'
```

Figure 6

```
# Open/Create a file to append data
with open (filename, 'a', newline='') as csvFile:
    csvWriter = csv.writer(csvFile)
    for tweet in tw.Cursor(api.search,q=new_search, count
                           lang="en",
                           tweet_mode= 'extended',
                           since='2020-11-01',
                           until ='2020-12-03').items():
        #tweets_encoded = tweet.text.encode('utf-8')
        #tweets_decoded = tweets_encoded.decode('utf-8')
        tweetCountTest += 1
        print(tweetCountTest)
        print (tweet.created_at, tweet.full_text)
       # if tweet.coordinates or tweet.geo:
        csvWriter.writerow([tweet.created_at, tweet.full
```

Figure 7

Implementation of these analysis processes included cleaning/standardizing andpre-processing of the dataset to get rid of any duplicate values or outliers.

The place field of Twitter data was in JSON format. Therefore, this field was split into type Coordinate, City, State, Country Code, and Country.

This paper then studied and analyzed the cleaned data for better representation of the result. With the transformed data, we performed an exploratory and predictive analysis.

Based on the major factors identified in the analysis the following questions were derived:

1. How did preemptive measures impact overall transmission/infection rates?

2. How did controversy impact preemptive measures and transmission infection rates?

3. Compare the number of retweeted COVID-19 misinformation/conspiracies in relation to hotspots (i.e., States with the highest number of infections, hospitalizations, and or deaths).

This paper recognizes that the nature of this topic is ideolog- ical in some regard, research on ideological assumptions must also be taken into consideration when extrapolating answers to these questions. Our initial approach, was to gather numerical-based data for the purposes of providing a quantitative analysis. After we completed the analysis, we took the data and developed visualizations using Tableau and Python.

Using various libraries, we then imported the related information and developed visualizations of the results for a better understanding. We then used a Forecast and Time Series model to forecast future COVID-19 case and death trends. Our team also incorporated a secondary approach to this research. We performed a social media Sentiment Analysis to explore COVID-19 tweets that potentially impacted how individuals interpreted the seriousness of COVID-19. This methodology (or Behavioral Analysis) allowed us to identify a specific user

```
data.head()
```

	Date_Time	text	username	user_location	retweet_count	favourite_count	Place
0	12/2/2020 23:51	b'@realDonaldTrump Wait! Didn\xe2\x80\x99t you say the Samething about COVID19? Hoax/FakeNews#'	b'DavidTashawn'	b'Mexico, MO'	0	0	NaN
1	12/2/2020 23:48	b'Dr. Henry on people who think #COVID19 is a hoax: "Might I suggest that you don\'t follow Twit...	b'katsleplan'	b'Surrey, British Columbia'	1	7	NaN
2	12/2/2020 23:32	b'#COVIDIOTS\xf0\x9f\x92\xa5You will soon understand your ignorance protected a HOAX that was me...	b'AndreeaSaint'	b'Canada'	0	2	NaN
3	12/2/2020 23:27	b'@stephanieinnes @azcentral I can't believe there are still people who think #COVID\xe3\x83\xbc...	b'roanneaz63'	b'Arizona'	0	9	NaN
4	12/2/2020 23:21	b'Many Republicans predicted that as soon as Democrats won in November the Covid19 virus would d...	b'mkevane'	b'Tweets are personal.'	1	6	NaN

```
print('Dataset size:',data.shape)
print('Columns are:',data.columns)
```

```
Dataset size: (17560, 7)
Columns are: Index(['Date_Time', 'text', 'username', 'user_location', 'retweet_count',
       'favourite_count', 'Place'],
```

Figure 8

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17560 entries, 0 to 17559
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date_Time       17560 non-null  object
 1   text            17560 non-null  object
 2   username        17560 non-null  object
 3   user_location   17560 non-null  object
 4   retweet_count   17560 non-null  int64
 5   favourite_count 17560 non-null  int64
 6   Place           277 non-null    object
```

Figure 9

group and keywords used when commenting on COVID-19 information. It helps to identify the number of people who believe on the controversy related to COVID such as:

COVID-19 is man made in the laboratory?

It is not real and is a hoax?

5G technology is responsible for the global pandemic?

As with our initial approach in figures 13, 14, and 15, the Sentiment Analysis leveraged both Python and R/R Studio as programming languages and visualization tools. We developed a custom code using Python that used an application programming interface (API) to interact with Twitter. The Python code contains keys and tokens created within the Twitter

```
def clean(sentence):
    #make everything lowecase
    sentence = sentence.lower()
    #remove urls
    sentence = re.sub(r'http\S+', " ", sentence)
    # remove mentions
    sentence = re.sub(r'@\w+',' ',sentence)

    # remove hastags
    sentence = re.sub(r'#\w+', ' ', sentence)

    # remove digits
    sentence = re.sub(r'\d+', ' ', sentence)

    # remove html tags
    sentence = re.sub('r<.*?>',' ', sentence)

    #remove stop words
    sentence = sentence.split()
    sentence = " ".join([word for word in sentence if not word in stopwords])
    # remove punctuation
    sentence  = "".join([char for char in sentence if char not in string.punctuation])
    sentence = re.sub('[0-9]+', '', sentence)
     #remove
    sentence = re.sub(r'\bx\d*\b', '', sentence)

    return sentence
```

Figure 10

```
#tokenization
def tokenization(sentence):
    sentence = re.split('\W+', sentence)
    return sentence

data['Tweet_tokenized'] = data['text'].apply(lambda x: tokenization(x.lower()))
data.head()
```

Figure 11

Field Name	Table	Remote Field Name
retweet_count	Covid19_tweet_Worldwide.csv	F7
favourite_count	Covid19_tweet_Worldwide.csv	F8
place	Covid19_tweet_Worldwide.csv	F9
ID		F9 - Split 3
place - Split 7		F9 - Split 7
placesplit		F9 - Split 1
type		F9 - Split 1 - Split 1
coordinate		F9 - Split 2
city		F9 - Split 7 - Split 1
State		F9 - Split 7 - Split 2
place - Split 8		F9 - Split 8
country_codde		F9 - Split 8 - Split 1
place - Split 9		F9 - Split 9
country		F9 - Split 9 - Split 1

Figure 12

Development Console to allow for requests, formatting, and extraction of Twitter data as it relates to COVID-19.

In addition to the API logic, our custom code also contains logic for error handling, data cleansing (i.e., removal of special characters and links), classification and parsing of Tweets, and calculating positive/negative/neutral Sentiment Analysis percentage ratings.

IV. RESULT ANALYSIS

Our team performed several iterations of the Sentiment Analysis using several COVID-19 keywords, such as:

- COVID-19
- COVID-19 hoax
- COVID-19 conspiracy
- COVID-19 fake news
- COVID-19 spike

```
# keys and tokens from the Twitter Dev Console
consumer_key = 'xxxxxxxxxx'
consumer_secret = 'xxxxxxxxxx'
access_token = 'xxxxxxxxxx'
access_token_secret = 'xxxxxxxxxx'
```

Figure 13

```
def clean_tweet(self, tweet):
    '''
    Utility function to clean tweet text by removing links, special characters
    using simple regex statements.
    '''
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())
```

Figure 14

```
def main():
    # creating object of TwitterClient Class
    api = TwitterClient()
    # calling function to get tweets
    tweets = api.get_tweets(query = 'COVID-19', count = 200)

    # picking positive tweets from tweets
    ptweets = [tweet for tweet in tweets if tweet['sentiment'] == 'positive']
    # percentage of positive tweets
    print("Positive tweets percentage: {} %".format(100*len(ptweets)/len(tweets)))
    # picking negative tweets from tweets
    ntweets = [tweet for tweet in tweets if tweet['sentiment'] == 'negative']
    # percentage of negative tweets
    print("Negative tweets percentage: {} %".format(100*len(ntweets)/len(tweets)))
    # percentage of neutral tweets
    print("Neutral tweets percentage: {} %".format(100*(len(tweets) -(len( ntweets )+len( ptweets)))/len(tweets)))

    # printing first 5 positive tweets
    print("\n\nPositive tweets:")
    for tweet in ptweets[:10]:
        print(tweet['text'])

    # printing first 5 negative tweets
    print("\n\nNegative tweets:")
    for tweet in ntweets[:10]:
        print(tweet['text'])
```

Figure 15

- COVID-19 hospitalizations
- COVID-19 death

Within each iteration, the Sentiment Analysis provided three quantifiable metrics for analysis that are Positive, Negative, and Neutral Percentages.

Our analysis shows that the percentages for the search terms of COVID-19 hoax and COVID-19 conspiracy vary between one and three percent. Whereas the COVID-19 fake news search showed significant variation between the previous two search terms.

V. DISCUSSION

Recognizing that the words hoax, conspiracy, fake news, death, and hospitalizations all have negative connotations associated with them, ideally, the expectation for a Sentiment Analysis associated with these words would reflect a relatively low positive percentage. However, the analysis shows that is not the case with the COVID-19 hoax and COVID-19 conspiracy results. Both iterations of these analyses show that there is a positive percentage rating of over thirty percent. When compared to the COVID-19 fake news and COVID-19 hospitalization searches, the results fell in-line with expectations, reflecting less than fifteen percent for the Positive Percentage rating.

VI. CONCLUSION

The results and interpretation of results for our research have been limited due to three specific reasons; time, sample size, and availability of geo location data. However, our research used the bootstrapping method for each iteration performed.; this method allowed for a thorough review of the underlying data as a means of error handling and validation. Sample sizes greater than 20,000 tweets would require a significant amount of additional time for review and validation. As well as an increase in errors, misclassification of sentiment, and an invalid interpretation of results and we addressed this problem in this research. Also, sentiment analysis on online users helps us to find the real source of wrong behaviors to improve public health.

REFERENCES

[1] S. Laato, A. K. M. N. Islam, M. N. Islam, and E. Whelan, "What drives unverified information sharing and cyberchondria during the covid-19 pandemic?," *European Journal of Information Systems*, vol. 29, no. 3, pp. 288–305, 2020.

[2] C. P. Rodr´ıguez, B. V. Carballido, G. Redondo-Sama, M. Guo, M. Ramis, and R. Flecha, "False news around covid-19 circulated less on sina weibo than on twitter. how to overcome false information?," *International and Multidisciplinary Journal of Social Sciences*, vol. 9, no. 2, pp. 107–128, 2020.

[3] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[4] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," *Psychological Science*, vol. 31, no. 7, pp. 770–780, 2020. PMID: 32603243.

[5] Z. Barua, S. Barua, S. Aktar, N. Kabir, and M. Li, "Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation," *Progress in Disaster Science*, vol. 8, p. 100119, 2020.

[6] M. Heidari and J. H. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pp. 0542–0547, 2020.

[7] M. Heidari, J. H. J. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[8] A. Bridgman, E. Merkley, P. J. Loewen, T. Owen, D. Ruths, L. Teichmann, and O. Zhilin, "The causes and consequences of covid-19 misperceptions: understanding the role of news and social media," *The Harvard Kennedy School (HKS)*, 2020.

[9] A. Ghenai and Y. Mejova, "Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 518–518, 2017.

[10] T. Tran, P. Rad, R. Valecha, and H. R. Rao, "Misinformation harms during crises: When the human and machine loops interact," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 4644–4646, 2019.

[11] M. Angeline, Y. Safitri, and A. Luthfia, "Can the damage be undone? analyzing misinformation during covid-19 outbreak in indonesia," in *2020 International Conference on Information Management and Technology (ICIMTech)*, pp. 360–364, 2020.

[12] H. X. L. Ng and J. Y. Loke, "Analysing public opinion and misinformation in a covid-19 telegram group chat," *IEEE Internet Computing*, pp. 1–1, 2020.

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020.

[14] "Covid-19 open research dataset challenge (cord-19)." https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge, 2020. Online; Retrived 21 September 2020.

[15] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[16] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[17] "Covid-19 twitter chatter dataset for scientific use." http://www.panacealab.org/covid19/, 2020.