Machine learning Model to Detect Emergency in the Global Pandemic

Rini Raju
George Mason University
rriniraj@gmu.edu
Accepted as a working paper in data mining Conference 2020

***Abstract*—It is crucial to use advanced machine learning models to improve disaster and emergency response in critical events around the world. In this paper we introduce a new model, which can highlight the essential help that people need in the time of emergency. Based on the user comments, we choose the emergency response that can use the optimal resources to address the maximum needs. The new features in the model helps to analyse each person response from political, social and health perspective. This approach helps to recognize different type of users to improve emergency response in the time of global pandemic. Also, collecting pandemic data from different online resources, makes this research more powerful in feature extraction to improve the model accuracy based on emergency data. This model can help health applications to improve disaster response time and services.**

***Index Terms*—Information extraction, Social network, Data mining, Machine learning**

I. Introduction

An attempt to try and slow the spread of new cases of COVID-19 is becoming the main goal to keep the size of our health-care systems and meet the requirements of the patients. While social distancing and practicing hygiene will help to an extent, it is possible for the number of patients to grow our system's capacity in different locations in the country. This unlucky reality is the reason why many hospitals are delaying elective surgeries in order to make sure the resources like intensive care unit (ICU), gowns, beds and gloves are made available for the patients affected with COVID-19. [1]

The campaign like public health that are uplifting social distancing must provide additional tips from staying safe any harm that may be physical. The public must take into terms of two results based on today's climate; as the virus continues to spread, the risk is higher for putting themselves at risk as front liners continue to provide care to those with the disease and the total time and resources that are taken into consideration for taking care of the ones affected will create a tension in our hospital systems [1].

The above graph depicts the importance of Social distancing. It displays the daily exposure of cases without any stern restrictions and daily exposure with social distancing in place. In the month of April, as shown above social distancing was not taken into practice, but moving to may, there was a 50% reduction in daily contacts and whereas, in June, there was about 75% reduction in daily contacts. Therefore, it is important to understand and study the importance of social distancing and other factors.

COVID-19 has affected and is still affecting day to day life, business, world trade, movement, and the global economy [2] [3] [4] [5]. Various sectors and industries such as healthcare, social service, finance, education, and tourism are also affected by this disease. With the flu season approaching stress on the healthcare system, especially hospitals will be greatly impacted with the overlap of COVID-19 and flu infections. [6] [7] [8]Also, the opening of universities might increase the spread (and number of cases) of COVID-19 as students from different parts of the country will come together. The students from the hardest hit areas of the COVID-19 may be asymptomatic but might transfer the virus to other students. [9] [10] [11] [12]

Therefore, it is necessary for the public to educate themselves on the critical situation of COVID-19 and take necessary precautions to protect themselves, family, and people around them. However, with the rise and persistent use of social media misinformation regarding the seriousness and dangers regarding COVID-19 have risen exponentially. [13]

Individuals by and large recognize the integral part social media plays in daily life. Misinformation can be mitigated by advanced by detection methods [14] [15]. Transfer learning models can be used to analyse data [16] [17] [18] [19].People are more inclined to express their thoughts and feelings on specific topics through social media platforms like Twitter. An analysis of twitter users has shown that individuals either seek information from social media or express their opinion regarding the COVID-19 [20]. Data gathered from social media platforms provide a great deal of benefit and added information to state, local, [21] and federal government; as well as the health industry in support of making informed and focused decisions on how to educate people on COVID-19 related topics.

II. Dataset

This paper is based on COVID-19 data [22]. This dataset was acquired from COVID-19 Tracking project and NY times. COVID-19 adat which is used to draw a visualization on different questions like the number of confirmed cases, total number of deaths, number of recovery and so on for worldwide cases. This data set has 156292 records and 8 fields.

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

          (15409, 42)
```

Fig. 1. Figure 1

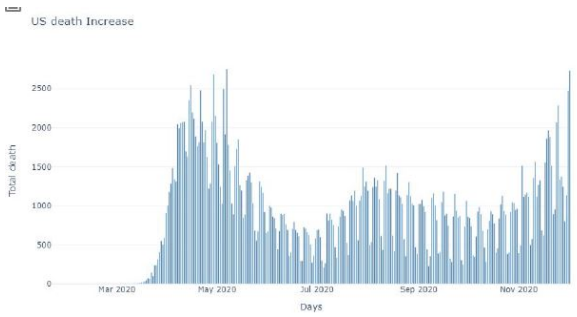

Fig. 2.

Another data set used is all-states-history.csv extracted from COVIDtracking.com. This data contains the daily positive, hospitalized, deaths and recovered cases for US States. This data has 15409 records (rows) and 42 columns. 1 3 2

The above visualization displays the COVID-19 cases in the USA. The graph shows the daily cumulative count of confirmed cases, daily cumulative count of deaths and daily count of new confirmed cases in the USA. The X-axis indicates the date of observation.

We also extracted tweet data related to COVID-19 misinformation using Twitter API.

III. Methods

For desired results, COVID-19 related dataset from Kaggle and COVIDtracking.com is 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 these data, they may not indicate the exact numbers of daily, hospitalization, and death cases as reported by the state and local government organizations or the

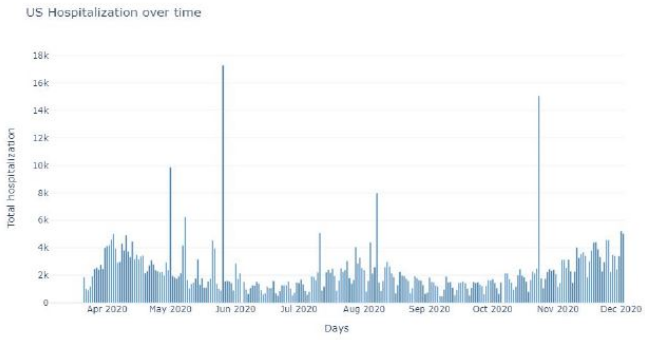

Fig. 3.

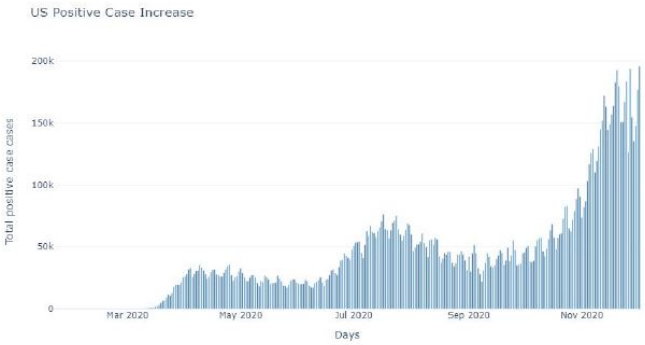

Fig. 4. Figure 1

```
# 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"
```

Fig. 5.

news media. This paper uses daily new cases, hospitalization, and COVID-19 related death data sets. 5 6 7

For misinformation, controversy and hoax, Twitter data is collected using Twitter API. This data set has 17560 records and 7 fields.

Data is analyzed using Python, Natural Language Processing (NLP) and Tableau. Implementation of these analysis processes includes cleaning/standardizing and pre-processing of the dataset to get rid of any duplicate values or outliers.8 9 10

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 studies and analyzes 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 outlined in Section I the following questions were derived:

1. How did pre-emptive measures impact overall transmission/infection rates?

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

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

Fig. 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
```

Fig. 7.

```
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'katsieplain'	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 #COVIDhxe3\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'],
```

Fig. 8.

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 ideological in some regard, research on ideological assumptions must also be taken into consideration when extrapolating answers to these questions. Our initial approach as previously explained is 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.1112

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 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?

```
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
```

Fig. 9.

```
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
```

Fig. 10.

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

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

Fig. 11.

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

Fig. 12.

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

Fig. 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())
```

Fig. 14.

As with our initial approach, 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 Development Console to allow for requests, formatting, and extraction of Twitter as it relates to COVID-19. 131415

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 key words, such as:

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

Within each of these iterations, the Sentiment Analysis provided three quantifiable metrics for analysis that are Positive, Negative, and Neutral Percentages. The following images

```
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'])
```

Fig. 15.

reflect the amounts for each of the three metrics for each of the iterations previously identified.

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. Limitations

In addition, geo location services are available on Twitter. End users are not required to enter in location information. Without this requirement, we were unable to confidently relate sentiment to a given geographical location within the United States, preventing us from being able to assess political impacts on perception to a given state.

VII. conclusion

the results and interpretation of results for our research has been limited due specific reasons; But also improves the several aspect for emergency response. As a Future work we will develop the automated application which has a potential to detect the important online information in social media to find the best emergency response.

References

[1] S. Lockey, "What's important: What is our role in the covid-19 pandemic?," *Journal of Bone and Joint Surgery*, vol. 102, pp. 931–932, 2020.

[2] Abid Haleem, Mohd Javaid, "Effects of covid-19 pandemic in daily life." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7147210/, 2020. Online; accessed 3 April 2020.

[3] L. Bursztyn, A. Rao, C. Roth, and D. Yanagizawa-Drott, "Misinformation during a pandemic," *Institute for Economics*, 2020.

[4] 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.

[5] A. Gokaslan and V. Cohen, "Openwebtext corpus." https://skylion007.github.io/OpenWebTextCorpus/, 2016.

[6] 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.

[7] "Coronavirus disease (covid-19)." https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19, 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] 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.

[11] P. Cihan, "Fuzzy rule-based system for predicting daily case in covid-19 outbreak," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–4, 2020.

[12] 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.

[13] S. Song, Y. Zhao, X. Song, and Q. Zhu, "The role of health literacy on credibility judgment of online health misinformation," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–3, 2019.

[14] 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.

[15] 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.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[17] 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.

[18] 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.

[19] 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.

[20] H. H. Drias and Y. Drias, "Mining twitter data on covid-19 for sentiment analysis and frequent patterns discovery," *medRxiv*, 2020.

[21] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 19–27, IEEE Computer Society, 2015.

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