Variance of Twitter Embeddings and Temporal Trends of COVID-19 cases

A Preprint

Khushbu Pahwa *1, **Ambika Sadhu** *2, **Mayank Sethi** *3, **Sargun Nagpal** 4, **Tavpritesh Sethi** +4
1. University of California, Los Angeles, California, USA.
2. National Institute of Technology, Kurukshetra, India.
3. Netaji Subhas Institute of Technology, Delhi, India.
4. Indraprastha Institute of Information Technology, Delhi, India
*Contributed Equally
+tavpriteshsethi@iiitd.ac.in

October 4, 2021

Abstract

The severity of the coronavirus pandemic necessitates the need of effective administrative decisions. Over 4 lakh people in India succumbed to COVID-19, with over 3 crore confirmed cases, and still counting. The threat of a plausible third wave continues to haunt millions. In this ever changing dynamic of the virus, predictive modeling methods can serve as an integral tool. The pandemic has further triggered an unprecedented usage of social media. This paper aims to propose a method for harnessing social media, specifically Twitter, to predict the upcoming scenarios related to COVID-19 cases. In this study, we seek to understand how the surges in COVID-19 related tweets can indicate rise in the cases. This prospective analysis can be utilised to aid administrators about timely resource allocation to lessen the severity of the damage. Using *word embeddings* to capture the semantic meaning of tweets, we identify *Significant Dimensions (SDs)*.Our methodology predicts the rise in cases with a lead time of 15 days and 30 days with R2 scores of 0.80 and 0.62 respectively. Finally, we explain the thematic utility of the SDs.

Keywords COVID-19 · Twitter · SARS-CoV-2 · Word2Vec

1 Introduction

Since late November'19, SARS-CoV-2 or Covid-19, has been spreading rapidly across all countries. On 11 March,2020, WHO declared the infectious disease as a pandemic. The pandemic has burdened the healthcare systems in almost every region, with shortage in ICU beds, healthcare workers as well as other resources as PPE, ventilators etc.[1] 2021 witnessed large-scale vaccination drives, which served as a critical tool in stabilizing the spread of virus. However, some cases of vaccine breakthrough infections have also been reported [2]. As COVID-19 continues to mutate, multiple waves of cases with varying severity each time have become a cause of concern, requiring stringent measures. With a rapid speed of transmission, the virus leaves little time for effective decision making. Though retrospective analysis plays a central role in governing administrative decisions relating to future plans of action, prospective insights can aid in controlling the extent of loss of life as well as avoid an overwhelm of medical resources.

Ever since the outbreak began, user involvement in social media applications such as Twitter, Whatsapp etc has also increased at an enormous rate. These platforms serve as a medium for people to disseminate their opinions but more importantly, for voicing concerns on the spread of the disease.The second wave in India, from April-June 2021, witnessed a surge in tweets related to fear of deaths due to the new strain, the crumbling capacity of hospitals as well as possibilities of another lockdown[3]. Given that citizens are leveraging social media platforms, a pattern exists in the latent content and the corresponding rise in COVID-19. As hotspots started to emerge in India, Brazil and the likes in 2020, a vast majority of tweets were themed around the number of cases and deaths in an affected region[4]. This shows potential in using tweets as an evidence for an upcoming surge. The semantic meaning of tweets can be exploited as a signal for early prediction of a covid wave.

In recent times, several works have carried out a predictive modelling approach using different models. Some studies deployed linear regression models to predict the number of deaths [5], the number of confirmed cases using travel history data [6] as well as the number of recoveries [7].

Our approach serves to capture the underlying relationship between these tweets and a covid surge for India. As concerns over newer waves arise, twitter serves as a platform of changing emotions, indicative of a possible rise in cases. In recent times, word embeddings have proven to be effective in various text-related tasks, as a learned representation of textual data as vectors. So, we leverage this representation of tweets to predict the number of cases in advance. With different leading periods, we train a predictive model using tweets and confirmed cases from August 2020 to June 2021. This monitoring of tweets to capture upcoming surges can be pivotal in governing administrative decisions and controlling the extent of the spread and the impact thereafter.

2 Data Collection and Pre-Processing

2.1 Datasets

For our use, the dataset consisting of tweets related to Covid-19 was downloaded from the publicly available repository maintained by Panacea Lab, Department of Computer Science, Georgia State University[8]. Apart from tweets, it also contains information about the country, its date and time of origin. For the purpose of this study, tweets pertaining to India between 1st August'20 to 31st July'21 were extracted. The number of tweets for each day varies between 100-200 and 1500 or more per day depending on the surge of Covid-19. Further, day-wise data of Covid-19 cases in India, for the period of August 2020 to July 2021, was accessed from the repository maintained by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)[9].

2.2 Data-Cleaning and Pre-Processing

Before using the tweets, some amount of cleaning is performed on the text to convert it to a format suitable for modelling. Data cleaning step in the proposed pipeline consists of removal of hashtags, mentions, URLs, and emojis. The cleaned data is passed through a Tokenizer that "splits" the phrase into "tokens". This step is followed by lemmatization (an evolution of stemming, that reduces the word to its *lemma* or core representation). Lastly, we carry out removal of stopwords excluding negation words, and punctuation removal, since these add only little meaning.

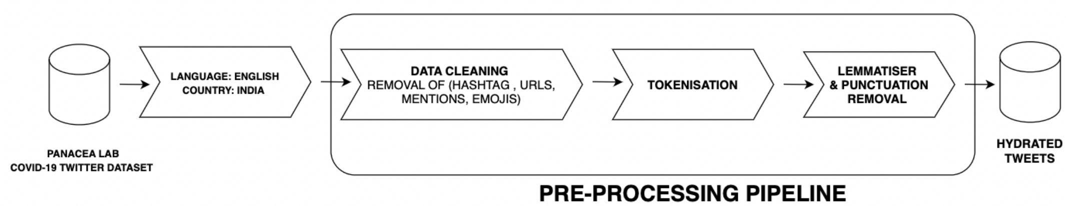

Figure 1: Pre-Processing of the Tweets

2.3 Word Embeddings

Every "token" or "word" obtained after cleaning needs to be converted into a numerical representation. In this study, we leverage an unsupervised low-dimensional representation of words using the Word2Vec model [10]. The gensim [11] implementation of Word2Vec, with default parameters was utilised. Each vector is learned from the tweet corpus with a window size of 5. Thus, for every word in a tweet, we get a 100-Dimensional vector representation. To obtain the individual tweet vectors, we average out the vectors corresponding to each word present in it. Further, for attaining a single representation of each day, again an average is taken over the vectors of all the tweets of that day. Thus, each day is represented as a 100-dimensional vector.

3 Feature Selection

Out of 100 dimensions of the day-wise vectors, *"Significant Dimensions"* or *SDs* are selected as the ones that show a "leading" relationship with the new case count. "Leading" means that these dimensions show a rise/fall before a certain period, post which the case count shows a change. Thus, temporal trend analysis of these SDs can facilitate early intervention for covid-spread control [1].

Two mechanisms were utilised for SD selection-

3.1 Cross Correlation Analysis

Cross correlation is a technique used to model the relationship between two time series. This helps to capture the underlying trend between the time series of two variables. To determine the significant features having a leading trend with respect to the new cases, CCF (Cross Correlation Function) has been used. Since, the raw new-case series (response variable) is non-stationary. In order to perform the CCF analysis, it had to be first made stationary which entails that there should not be any change in its statistical properties over time. Stationarity of the new-case series was obtained on transformation employing second order differentiation on the raw series. In order to make all the 100 dimensions of the tweet vectors stationary, rolling variance over 3 days has been used as the transformation. The raw and stationary new case series are as shown in Fig. 2 and Fig. 3 below.

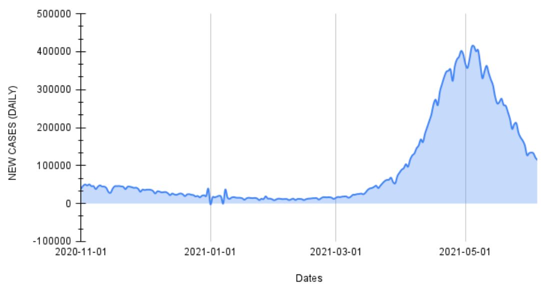

Figure 2: New-Case Series

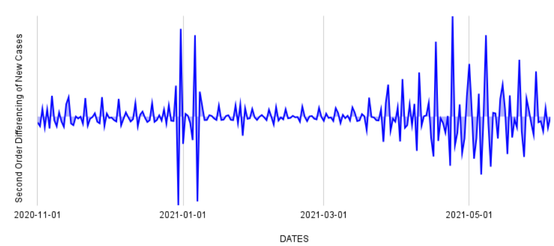

Figure 3: Stationary New-Case Series

The CCF plots obtained for the cross-correlation between the dimensions and new-cases serves as a means to identify the lag of the dimensions (x-variable) that might be useful predictors for the new-cases (y-variable). 40 dimensions were identified as significant i.e. having a leading trend with respect to the new-cases. The CCF plots for some of the dimensions are shown in Fig.

3.2 Boruta Analysis

Boruta Algorithm is a wrapper feature selection method that has been used for feature selection of the dimensions.Most of the traditional feature selection algorithms rely on a small subset of features which yields a minimal error on a chosen classifier.While fitting a random forest model on a data set, one can recursively get rid of features in each iteration which didn't perform well in the process. This will eventually lead to a minimal optimal subset of features as the method minimizes the error of random forest model. This happens by selecting an over-pruned version of the input data set, which in turn, throws away some relevant features.Boruta, on the other hand, finds all features which are either strongly or weakly relevant to the decision variable. This makes it well suited for biomedical applications.

4 Results

4.1 Model Development and Evaluation

Random Forest Regressor has been employed for modelling purpose.In our analysis, the response variable considered is the new case series, obtained by first order differencing on the confirmed case series. The stationarity of new case series

[1]Dimension numbering is 0-indexed

Algorithm 1: Boruta Algorithm

1 Extend the information system by adding copies of all variables (the information system is always extended by at least 5 shadow attributes, even if the number of attributes in the original set is lower than 5).
2 Shuffle the added attributes to remove their correlations with the response.
3 Run a random forest classifier on the extended information system and gather the Z scores computed.
4 Find the maximum Z score among shadow attributes (MZSA), and then assign a hit to every attribute that scored better than MZSA.
5 For each attribute with undetermined importance perform a two-sided test of equality with the MZSA.
6 Deem the attributes which have importance significantly lower than MZSA as 'unimportant' and permanently remove them from the information system.
7 Remove all shadow attributes.
8 Repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

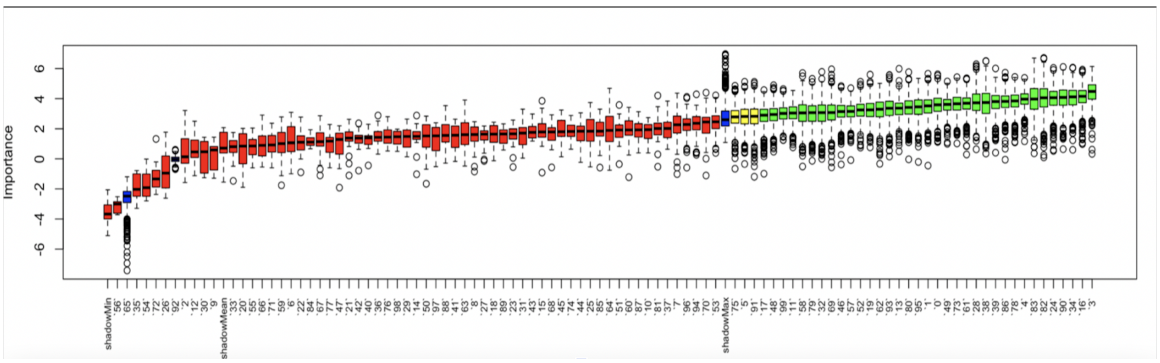

Figure 4: Importance of Dimensions through Boruta Analysis

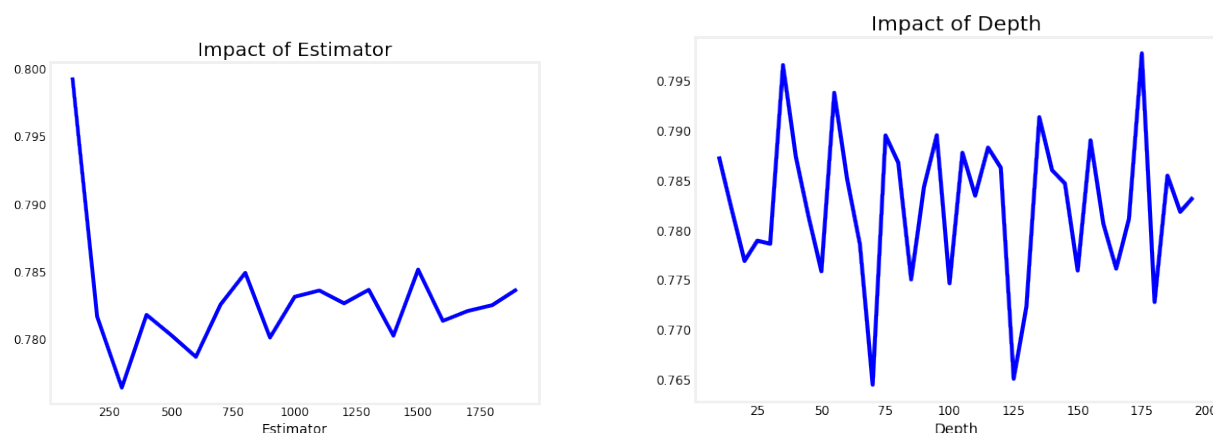

Figure 5: R-Square scores with varying parameters for 15-day lead from CCF Plots

is attained on second-order differencing with a P-value of $0.01 < 0.05$ .The predictors are embedding dimension values, which attain stationarity on first-order differencing.

We considered varying parameters in terms of depth of the random forest as well as the number of estimators. Using these different values, two leading periods are tested- 30-day and 15 day. In case of 30 days, we use the dimension values to predict "new cases" after 30 days, i.e. a "shift" of 30 days. Similarly, 15-day lead time is used to predict new cases after a fortnight. Different models are trained using SDs from both CCF plots and Boruta, using the dimension values from 1 November to 15 July. R2 scores are calculated using the data from 1-15 August for the 15-day lag model.

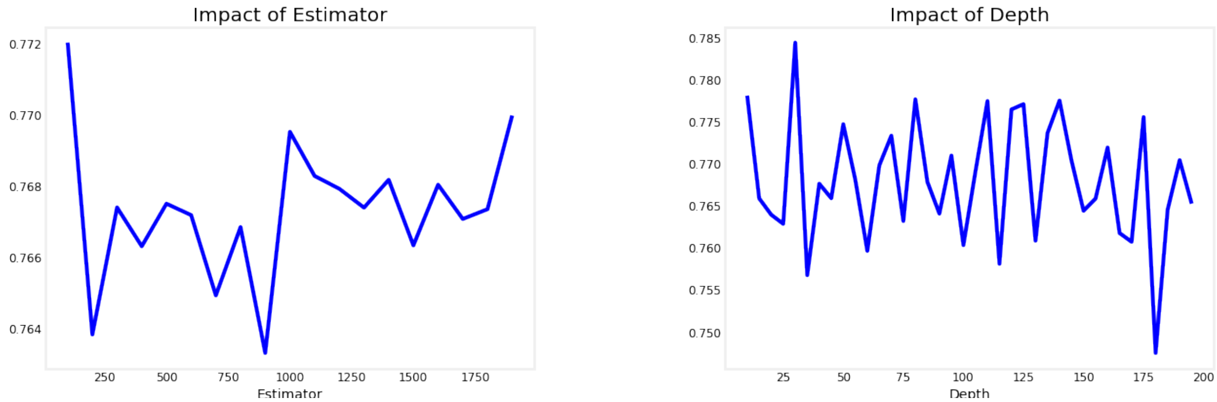

Figure 6: R-Square scores with varying parameters for 15-day lead from Boruta Analysis

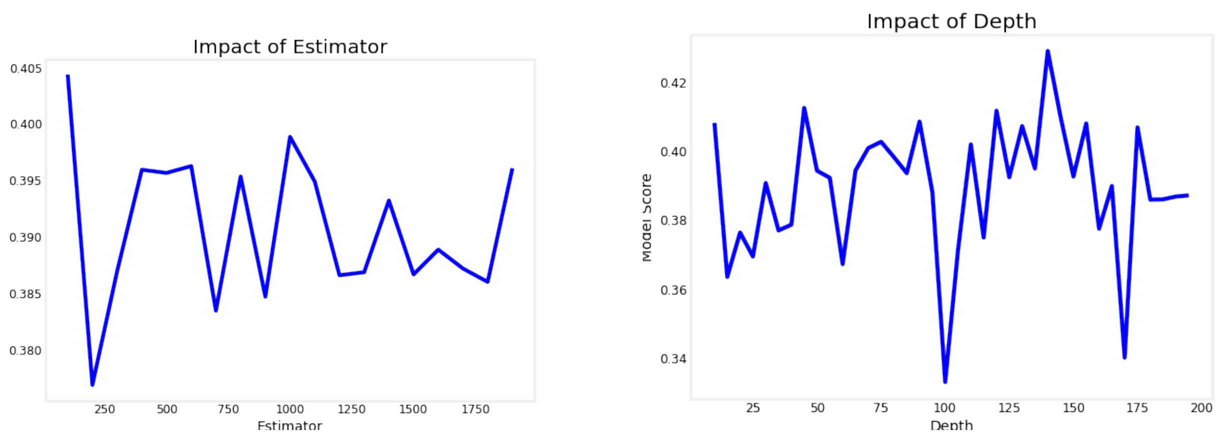

Figure 7: R-Square scores with varying parameters for 30-day lead from CCF Plots

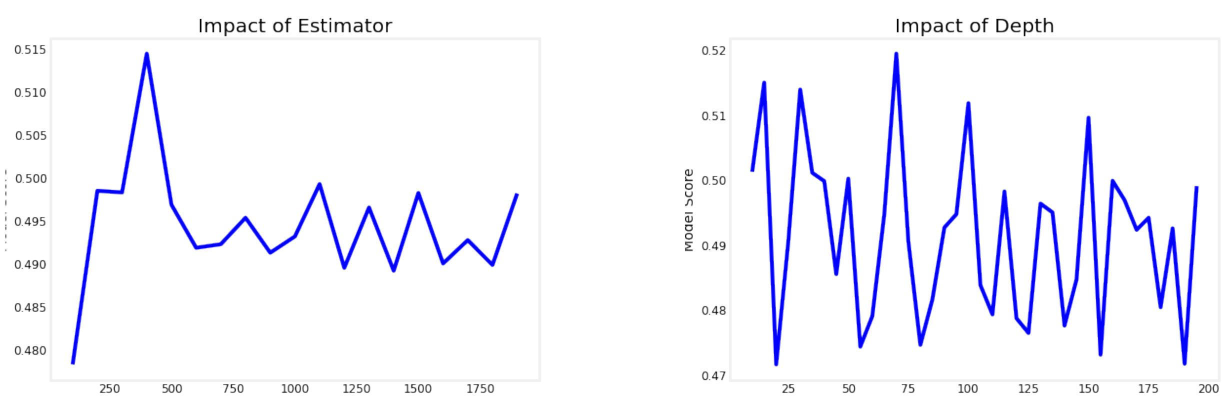

Figure 8: R-Square scores with varying parameters for 30-day lead from Boruta Analysis

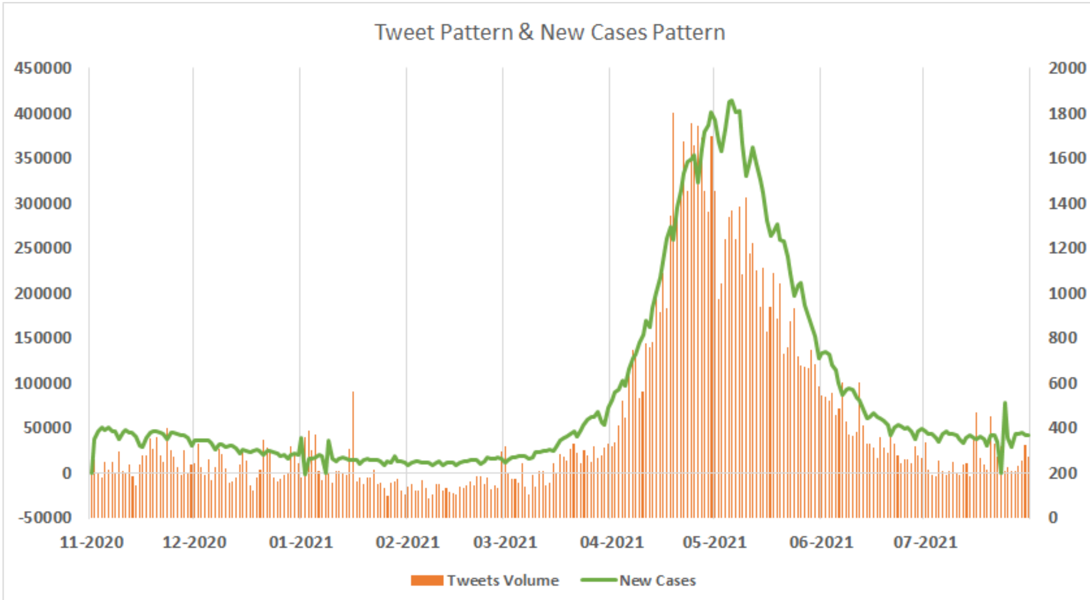

Figure 9: Tweet and New Cases trend

4.2 Social Media trends and Tweet-based language modeling capture rapidly changing COVID-19 environment

As shown in the figure, the period of April-May 2021 saw a peak in tweets about covid,owing to the second wave.Likewise, as the cases started to reduce, the number of corresponding tweets also showed a similar decreasing trend.This confirms that the trends in usage of Twitter can capture the changing COVID scenario and their analysis can be useful in deriving insights.

Further, visualization of word embeddings learned from tweets during this period show a similar pattern. t-SNE is a variation of the Stochastic Neighbor Embedding used primarily for visualizing high-dimensional data [12]. Since word embeddings are a geometrical representation of words in vector space, similar words are closer to each other. To visualize and interpret this similarity amongst the tweets, we used the TensorFlow Embedding Projector [13].Figure shows the t-SNE plot obtained for the day-wise averaged tweets. Here, each point in space represents one day.The cluster on left mostly contains dates from April to May, again corroborating the existence of a relationship between tweets and cases. Moreover, since word embeddings capture the semantic meaning of words in vector form, this clustering pattern indicates that there's a pattern associated with the type of tweets posted during this period. These visualizations serve as the basis for our hypotheses to use tweets for case count prediction.

4.3 Temporal trends in daily tweets mimic Covid-19 surge

A deeper analysis of the specific tweets mentioned over the different time periods provides insights into the current public concerns. As is clear from the frequency plots of different terms in three separate phases- December-January-February(before the second wave), April-May (during the second wave) and June-July(post second wave). From Fig. 11, it can be seen that there is a clear variation in the occurence of some key words pertaining to the pandemic and its management across the 3 periods chosen for analysis.The daily new covid cases plot as in Fig. 2, can be bifurcated into three periods: 1) Dec, Jan, Feb (2020) where there is a plateau in the series, 2) April,May (where there is a sudden rise), and then 3) June, July(2021) where there is again a decline. It can be validated that the words like "government", "cases", "lockdown", "reports", "mask", "help", "need", "pandemic" are higher in percentage for the April-May period when India witnessed the highest caseloads.National movements such as the farmers' movement, election in several states during the period led to an increase the risk of transmission of COVID-19.

4.4 Rolling Variance in the latent space dimensions captures variability in the COVID-related tweets

In this study, first analysis was carried out on the line plots of the embedding dimensions with the daily new cases. Upon visual inspection, rolling variance over 3 days seemed to be a reasonable approximating feature. So for the purpose of

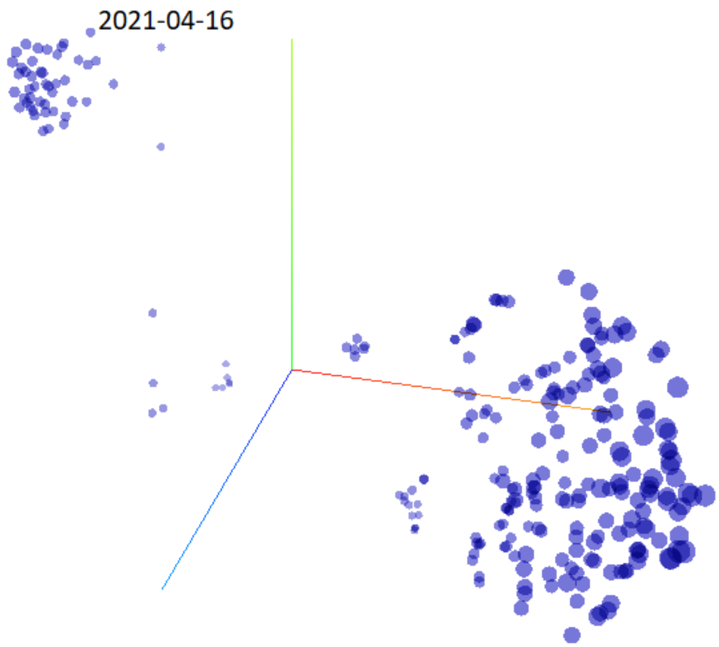

Figure 10: t-SNE plot for word2vec vectors of tweets

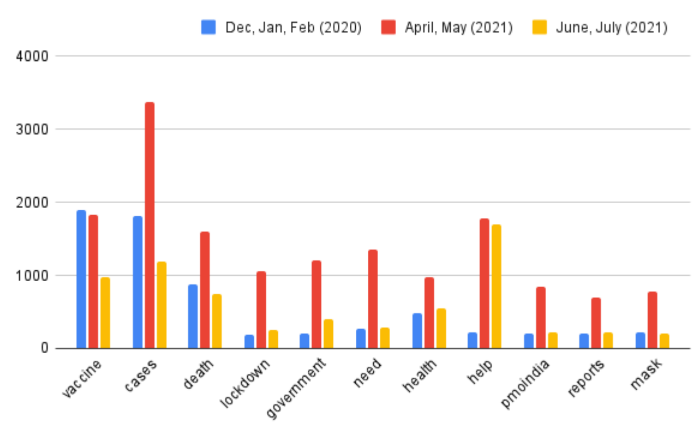

Figure 11: Word Frequency variation of common words across DEC-JAN-FEB(2020), APRIL-MAY(2021), and JUNE-JULY(2021)

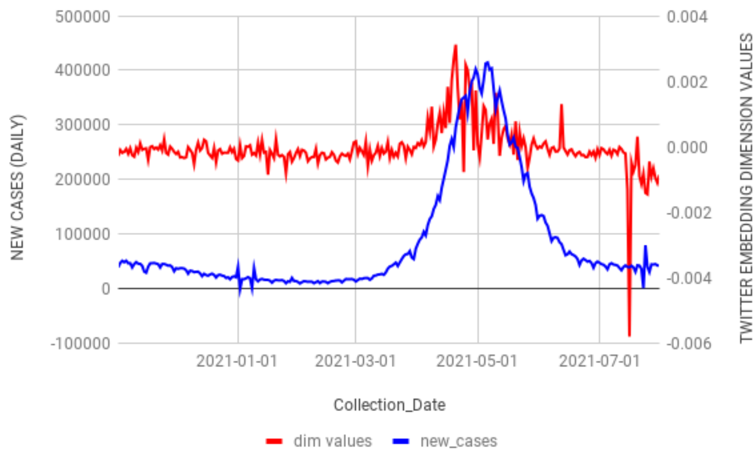

Figure 12: Variation of Twitter Embedding Dimension values with New Cases (Daily)

feature selection, the embedding dimensions were transformed using rolling variance. This was implemented using the pandas.core.window.rolling.Rolling.var package.

4.5 Dimensions formed from Word Embedding are predictive of new COVID-19 caseloads

In this section, we capture the trend of our predicted caseloads for COVID-19 and the actual data reported by JHU. In Fig. 13., the 3-day moving average of the predicted case series using the overlapping dimensions between Cross-Correlation Analysis and Boruta; and the actual case series are plotted. The solid lines represent the trend lines while the dotted lines represent the series in the same colour for which the trend is plotted.The trend lines for predicted and actual caseloads are similar, and thus point out towards the effectiveness of the chosen twitter embedding dimensions for predcition of caseloads.

Figure 13: 3-day Moving-Average of Predicted and Actual Case loads

Dimension	Keywords	Inferences
3	"government", "leader", "political", "provided", "vaccination", "exam", "decision", "control"	1.People in administration and government have neglected the situation of Coronavirus in India concurrently, 2. Effects and responses of the government leaders ensuring about taking full efforts to control the spread
24	"hope", "calm", "together", "fight", "stay at home"	1. General public, as well as political leaders, tweeted about the need to stay hopeful and calm in order to f
69	"Enforcement", "Protocol", "High Courts", "Shocking", "Vaccine"	1. Messages relating law enforcement bodies like High courts and Supreme courts, and decisions ordered b
37	"Decision", "Oxygen", "Admission", "Lockdown", "Election", "Adani", "Ambani"	1. While some tweets appreciate the efforts of some local governments, others criticise the government for 2. With the dearth in oxygen on the rise, tweets revolved around the need for setting up of oxygen plants, w
14,26	"ventilator", "ICU", "beds", "tocilizumab", "oxygen"	1. With the shortage of oxygen, ICU beds and medicines in every region, there was a hike in demand for re

Table 1: Themes represented by SDs

5 Discussion

Throughout the pandemic, researchers from across the world developed novel research methodologies for the spatio-temporal predictions of COVID-19 cases using both statistical and deep learning techniques.Some studies relied on clinical data [14] relied on data from clinical records of patients, to predict the risk of COVID-19 progression.However, gathering clinical data is a difficult task.On the other hand, some works focused on symptoms mentioned in tweets and google searches [16], thus establishing the utility of social media as a signalling mechanism.

By using word embeddings to represent tweets, in this work we rely on the ability of word vectors to capture varying semantic meaning, with changing context of the words in the individual tweets. Clustering patterns in the tSNE visualizations confirm the presence of temporal trends in the tweets.

In this pipeline, vectors for each day have been obtained by averaging the individual tweet vectors. As indicated by cross correlation analysis, values of the vector dimensions depict a relationship with COVID cases. Given that vector values change rapidly, 3-day rolling variance appeared to be a good feature.To carry forth a "prospective" analysis, we select dimensions(SDs) that show a leading relationship with the new cases, both based on cross correlation as well as boruta. The corresponding R2 scores from the Random Forest Regressor highlight the utility of leveraging tweet embeddings as a surveillance mechanism and validate the selection of SDs. Although exact values may not match, merely a rising/falling trend in an upcoming time window can help govern policy making.

Corresponding to every dimension, there are tweets that contribute the highest to it. This is identified using the absolute value of tweet vector along that dimension. Thus, for every dimension, tweet vectors are sorted based on the absolute value of the vector, and a subset of top 30 tweets was selected. Table 1 lists the key messages highlighted by 6 SDs, shortlisted from dimensions common to both CCF and Boruta analysis. The ability of vector dimensions to capture specific aspects is demonstrated by the tokens discovered in SDs. The selected SDs are able to portray public emotions over the pandemic, government actions, demands for resources etc. These insights coupled with spatial data can be harnessed for tracking the situation in different areas. Some topics of discussion, that were trending around the second wave, were not discovered in the SDs.These include "Kumbh Mela", the religious festival that led to a huge number of devotees overcrowding the venue. This Mela became a point of controversy for becoming a leading cause for the second wave. However, none of the top tweets mentioned specifically the mela. This can be attributed to the averaging being done to obtain the vectors. Similarly, the SDs do not indicate any mention of *mucormycosis* or black fungus. But, since the incidence of these cases was also very low, the number of people who tweeted about it is also significantly less.

Thus,our contributions in this work are twofold. First, it paves way for research towards leveraging social media as a 15 day lead-tracker for Covid-19, which can enable timely interventions for better disease control. Though the exact number of new cases may not match, a rising/falling trend can provide a prospective analysis as to how the situation is going to change.Secondly, it introduces *Significant Dimensions (SDs)* and associated themes.Analysing the messages associated with these SDs can aid in selecting the domains where administrative decisions could be channeled. These themes can be coupled with locational data, to analyse spatial trends, similar to [15].

However,there are several limitations to this study. First, we have leveraged Word2Vec with default parameters for generating word embeddings. Given that the vector for a day is an average of the tweets for that day, there is an associated loss of context. A comparison with other language models such as BERT would enable selecting the one that gives the best results. Further, instead of averaging the individual tweets, learning day-wise embeddings could capture context better, and provide a deeper understanding of thematic distribution across dimensions. Altering the parameters for Word2Vec is also a possible extension of this study.Currently, the study is limited to India. Covering more countries across the world can help in identifying "global" dimensions, whose tracking can lead to a more generic analysis. Moreover, the results are limited to English language only. Including vernacular languages can support a "local" analysis as well. The results obtained are also limited by the number of active twitter users, as well as users who actually tweet about the pandemic. Lastly, current study uses the rolling variance of raw values of the dimensions for prediction. A future extension could be to explore a different feature, such as entropy, or combining a set of features.

References

[1] McMahon, Devon E., et al. "Global resource shortages during COVID-19: Bad news for low-income countries." PLoS neglected tropical diseases 14.7 (2020): e0008412.

[2] COVID, CDC, et al. "COVID-19 Vaccine Breakthrough Infections Reported to CDC—United States, January 1–April 30, 2021." Morbidity and Mortality Weekly Report 70.21 (2021): 792.

[3] Sv, Praveen, Remya Lathabhavan, and Rajesh Ittamalla. "What concerns Indian general public on second wave of COVID-19? A report on social media opinions." Diabetes & metabolic syndrome 15.3 (2021): 829.

[4] Osakwe, Zainab Toteh, et al. "Identifying public concerns and reactions during the COVID-19 pandemic on Twitter: A text-mining analysis." Public Health Nursing 38.2 (2021): 145-151.

[5] Ghosal, Samit, et al. "Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases-March 14th 2020)." Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14.4 (2020): 311-315.

[6] Ogundokun, Roseline O., et al. "Predictive modelling of COVID-19 confirmed cases in Nigeria." Infectious Disease Modelling 5 (2020): 543-548.

[7] Rustam, Furqan, et al. "COVID-19 future forecasting using supervised machine learning models." IEEE access 8 (2020): 101489-101499.

[8] Banda, Juan M., et al. "A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration." Epidemiologia 2.3 (2021): 315-324.

[9] https://github.com/GCGImdea/coronasurveys

[10] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[11] Rehurek, Radim, and Petr Sojka. "Software framework for topic modelling with large corpora." In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. 2010.

[12] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

[13] Smilkov, Daniel, et al. "Embedding projector: Interactive visualization and interpretation of embeddings." arXiv preprint arXiv:1611.05469 (2016).

[14] Meng, Zirui, et al. "Development and validation of a predictive model for severe COVID-19: A case-control study in China." Frontiers in Medicine 8 (2021): 619.

[15] Bisanzio, Donal, et al. "Use of Twitter social media activity as a proxy for human mobility to predict the spatiotemporal spread of COVID-19 at global scale." Geospatial health 15.1 (2020).

[16] Yousefinaghani, Samira, et al. "Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada." Frontiers in public health 9 (2021): 359.