IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS WITH TWITTER DATASET FOR COVID 19

L. Sudha
Assistant professor
Department of Computer Science and Engineering
S A Engineering College, Chennai , India

I. Gowthaman
UG Student
Department of Computer Science and Engineering
S A Engineering College, Chennai, India

S. Naveen
UG Student
Department of Computer Science and Engineering
S A Engineering College, Chennai, India

M. Hemnath
UG Student
Department of Computer Science and Engineering
S A Engineering College, Chennai, India

Abstract - The main goal is to achieve a domain-specific approach to know sentiments manifested at intervals folks round the globe relating to COVID state of affairs. so on perceive this, corona-specific tweets area unit non-inheritable from twitter platform. Identification of COVID-19 sentiments from the tweets would alter hep alternatives for higher handling this pandemic state of affairs. The used dataset is extracted from twitter provided by the UCI Repository. Once gathering the tweets, the dataset is cleansed exploitation the preprocessing techniques, they're labeled and a model is developed that is effective for police investigation the particular sentiment behind a tweets associated with COVID-19. The contribution of this works is that the performance analysis of assorted machine learning classifiers exploitation planned feature set. Tweets area unit classified as positive, neutral, or negative exploitation machine learning algorithms(Multinomial Naïve mathematician, Support Vector Machine, supply Regression) and also the deep learning algorithmic rule (LSTM). Our experiments reveal that the planned model performs well in perceiving the perception of individuals regarding COVID-19 with a most accuracy. Our findings gift the high prevalence of keywords and associated terms among tweets throughout COVID-19. Further, this work clarifies the method on pandemics and lead public health authorities for a higher society.

I. INTRODUCTION

Sentiment analysis has attracted a great deal of attention from researchers over time and has seen forceful changes within the manner the analysis is completed to assemble and extract the feeling that has been pictured directly or indirectly by humans once expressing themselves on social media platforms. the advance within the handiness of web across the planet over the course of your time has created it doable to urge most of the planet population on-line that successively has accumulated the quantity of data that's being generated by the social media handles, e.g. Twitter, Facebook, etc. Now, with this walloping quantity of data being generated each single day, the {information} scientists have began to mine the precious information from the chunk of knowledge which will eventually enable them to review the various opinions of citizens on social media platforms. Sentiment analysis of tweets could be a well-liked topic for classifying unseen tweets into varied classes. Twitter has become a go-to choice for virtually each research worker out there for sentiment analysis. However, regardless of however well liked the platform is, the info that twitter generates within the style of tweets has its own shortcomings. Tweets is a string of most one hundred forty characters that have currently been accumulated up to 280 characters as of 2020. The new limit is applicable to the majority the languages supported by twitter excluding Korean, Chinese and Japanese as they'll incorporate a great deal a lot of data in a very lesser variety of words. Tweets area unit ordinarily informal knowledge which can contain slang words, emoticons, abbreviations, incomplete expressions etc. to form the foremost out of this type of knowledge, specialists have studied totally different tweets and have additionally analysed their sentiments extensively. COVID-19 could be a extremely communicable disease originated by a fresh uncovered coronavirus. The symptoms of COVID-19 vary from being delicate to moderate respiratory disease. those that area unit recent or area unit experiencing any in progress medical problems area unit a lot of at risk of develop of import malady. Corona viruses area unit a gathering of connected polymer infections that cause sicknesses in homeothermic animals and feathered creatures. In people, these infections cause metastasis plot contamination that may go from mellow to deadly. Mellow diseases incorporate some instances of the essential cold (which is to boot caused by totally different infections,

overpoweringly rhinoviruses), whereas increasingly deadly assortments will cause severe acute respiratory syndrome, MERS, and COVID-19. facet effects in numerous species change: in chickens, they cause Associate in Nursing higher metastasis plot malady, whereas in dairy farm animals and pigs they cause loose bowels. Antibodies or antiviral medications to forestall or treat human coronavirus diseases aren't found nonetheless. Coronavirus spreads from humans to humans in many manners. It causes metastasis issues and causes respiration issues. it's a coffee death rate compared to severe acute respiratory syndrome or MERS. However, as there's no immunizing agent obtainable and thanks to the character of transmission and unfold of virus, countries round the world taken imprisonment and isolation because the solely preventive steps. As a result people's movement being restricted they spent an honest quantity of your time in home or place of keep. This creates a perfect state of affairs for folks to precise their views on social networking sites together with Twitter. Like in different elements of the planet folks of Republic of India additionally specific their views concerning COVID-19 on twitter. during this paper, those tweets area unit wont to analyze the behavior and sentiment of individuals from Dec 2019 to could 2020. folks were terribly negative concerning COVID within the period of time. several weren't assured to win the fight against COVID. however because the imprisonment happened, folks became optimistic concerning it. Positive tweets area unit enormously high within the quantity within the month of April and should. This work provides new insights on COVID-19 and people's thoughts concerning it. Social media is today's thanks to browse folks. there's an honest variety of papers appeared within the literature since the eruption causes countries to travel in imprisonment state. Sentiment analysis and knowledge image of Worldwide Twitter knowledge on Covid nineteen has been conferred.

II. LITERATURE SURVEY

Assigning sentiment labels to documents is, initially sight, a customary multi-label classification task. several approaches are used for this task, however the present progressive solutions use deep neural networks (DNNs). As such, it looks doubtless that normal machine learning algorithms, like these, can offer an efficient approach. we have a tendency to describe another approach, involving the utilization of possibilities to construct a weighted lexicon of sentiment terms, then modifying the lexicon and scheming optimum thresholds for every category. we have a tendency to show that this approach outperforms the utilization of DNNs and alternative normal algorithms. we have a tendency to believe that DNNs don't seem to be a universal cure-all which listening to the character of the information that you just making an attempt|try|are attempting to find out from is a lot of vital than trying out ever a lot of powerful general purpose machine learning algorithms[1]. The usage of social media is apace increasing day by day. The impact of social group changes is bending towards the peoples' opinions shared on social media. Twitter has received a lot of attention thanks to its time period nature. we have a tendency to investigate recent social changes in MeToo movement by developing Socio-Analyzer. we have a tendency to used our four-phase approach to implement Socio-Analyzer. a complete of 393,869 static and stream knowledge is collected from the information world web site and analyzed employing a classifier. The classifiers determine and categorise the information into 3 classes (positive, neutral, and negative). Our results showed that the utmost peoples' opinion is neutral. consecutive higher variety of peoples' opinion is contrary and compared the results with TextBlob. we have a tendency to validate the 765 tweets of weather knowledge and generalize the results to MeToo knowledge. The exactitude values of Socio-Analyzer and TextBlob area unit seventy.74% and 72.92%, severally, once thought of neutral tweets as positive[2]. Deep neural networks have introduced novel and helpful tools to the machine learning community. altogether completely different types of classifiers will most likely turn out use of those tools what is more to bolster their performance and generality. This paper reviews this state of the art for deep learning classifier technologies that ar getting used outside of deep neural networks. Non-network classifiers will use several parts found in deep neural network architectures. throughout this paper, we've got an inclination to tend to review the feature learning, improvement, and regularization ways in which people core of deep network technologies. we've got an inclination to tend to then survey non-neural network learning algorithms that create innovative use of those ways in which to bolster classification. as a results of the many opportunities and challenges still exist, we've got an inclination to tend to discussion directions which can be pursued to expand the world of deep learning for a range of classification algorithms[3]. The happening of Corona Virus sickness 2019 (COVID-19) could be a grave international public health emergency. Nowadays, social media has become the most channel through that the general public will acquire data and categorical their opinions and feelings. This study explored belief within the early stages of COVID-19 in China by analyzing Sina-Weibo (a Twitter-like microblogging system in China) texts in terms of area, time, and content. Temporal changes at intervals one-hour intervals and also the spatial distribution of COVID-19-related Weibo texts were analyzed. supported the latent Dirichlet allocation model and also the random forest algorithmic rule, a subject extraction and classification model was developed to hierarchically determine seven COVID-19-relevant topics and thirteen sub-topics from Weibo texts[4]. To provide AN union literature on the detection of Abusive language on Twitter mistreatment tongue process (NLP). during this study, the survey has been conducted on completely different strategies and analysis conducted on the categories of Abusive language employed in social media, why it's important? however it's been detected in real time social media platforms and therefore the performance metrics that area unit employed by researchers in evaluating the performance of the detection of abusive language on Twitter by the users. Giving AN union review of past methodologies, together with strategies, vital options and core algorithms, this study arranges and depicts the current condition regarding this space[5]. Word illustration has invariably been a very important analysis space within the history of language process (NLP). Understanding such complicated text knowledge is imperative, providing it's wealthy in data and may be used wide

across varied applications. during this survey, we have a tendency to explore totally different word illustration models and its power of expression, from the classical to modern state-of-the- art word illustration language models (LMS). we have a tendency to describe a spread of text illustration strategies, and model styles have blossomed within the context of information science, as well as SOTA LMs. These models will remodel giant volumes of text into effective vector representations capturing an equivalent linguistics data. Further, such representations are often used by varied machine learning (ML) algorithms for a spread of information science connected tasks. In the end, this survey in short discusses the normally used mil and metric capacity unit primarily based classifiers, analysis metrics and also the applications of those word embeddings in numerous information science tasks[6]. Along with the emergence of the net, the speedy development of hand-held devices has democratized content creation because of the intensive use of social media associated has resulted in an explosion of short informal texts. though a sentiment analysis of those texts is efficacious for several reasons, this task usually|is usually|is commonly} perceived as a challenge only if these texts area unit often short, informal, noisy, and made in language ambiguities, like equivocalness. Moreover, most of the prevailing sentiment analysis ways area unit supported clean knowledge. during this paper, we tend to gift DICET, a electrical device-based technique for sentiment analysis that encodes illustration from a transformer and applies deep intelligent discourse embedding to reinforce the standard of tweets by removing noise whereas taking word sentiments, polysemy, syntax, and linguistics information under consideration[7].

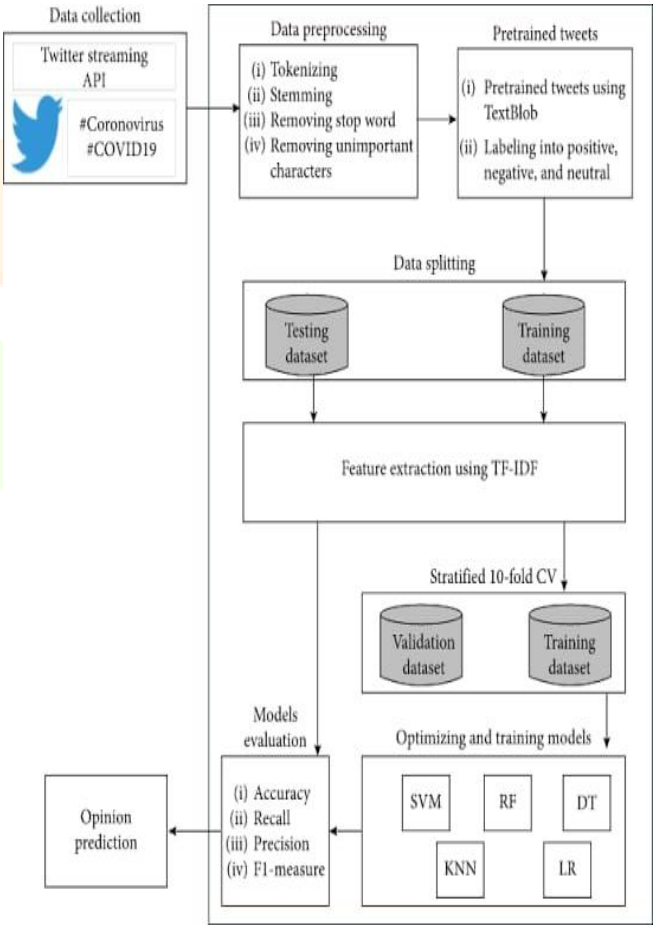
III. EXISTING SYSTEM

Sentiment analysis and textual analytics is presented, as well as that on ML methods, twitter and NLP. Significant data challenges are evolving and need to be addressed, and strategic information characteristics restructuring data, as well as the ML techniques. The analysis of past epidemics, crisis situational analysis and tracking, has also involved tracking Twitter data. A better understanding of the US's geographical spread concerning the valances of both healthy and unhealthy sentiment. The people in rural areas tweet less than those in urban areas and suburbs, using the spatial distribution of analyzed tweets. This work also notes that food tweets per capita were less in small urban areas than in larger towns and cities. It was revealed using linear regression that in low-income areas, tweets related more to unhealthy sentiment. sentiment analytics has also had avenues for the use of Twitter data.

IV. PROPOSED SYSTEM

Introduced to beat all the disadvantages that arise within the existing system. this method can increase the accuracy of the classification results by classifying the info supported the tweets.As raw tweets ar typically short, unstructured, informal, and noisy, the primary step of sentiment analysis is to preprocess the info. The Sentiment timeline helps to know trends in positive and negative sentiment over time. Multinomial Naïve Bayes algorithmic program predicts the tag of a text like a chunk of email or article. It calculates the likelihood of every tag for a given sample then provides the tag with the best likelihood as output. The multinomial distribution commonly needs whole number feature counts.In the SVM algorithmic program, we tend to plot every knowledge item as some extent in n-dimensional area (where n is range of options you have) with the worth of every feature being the worth of a specific coordinate. Support Vectors ar merely the co-ordinates of individual observation. The SVM classifier may be a frontier that best segregates the 2 categories (hyper-plane/ line). Logisticregression is estimating the parameters of a provision model (a sort of binary regression). Mathematically, a binary provision model includes a variable with 2 doable values, like pass/fail that is diagrammatical by associate degree indicator variable, wherever the 2 values or labelled "0" and "1.

V. SYSTEM ARCHITECTURE



VI. MODULES

A. Data Selection and Loading - The data choice is that the method of choosing the information for police work the negative tweets from the twitter dataset. The dataset that contains the data regarding the UserName, ScreenName, Location, OriginalTweet,,Sentiment.COVIDSENTI contains 2 months' value of tweets. Our search was restricted to tweets in English. The keywords used ensured a matter corpus targeted on the COVID-19 and associated phenomena.

knowledge assortment was performed to gather all tweets from a given region, however this point separated by individual dates. This section was separated from the month-wise assortment as a result of month-wise analysis of tweets needed quick knowledge assortment to modify straightforward and speedy verification of concepts, whereas day-wise analysis needed a lot of knowledge points and computation-intensive graphical calculations to be done on massive knowledge points.

B. Data Preprocessing - Data pre-processing is that the method of removing the unwanted information from the dataset. virtually each social media platform uses hashtags to represent topics, i.e., #COVID-19, #StayHome, #StaySafe, and #Coronavirus. In most cases, hashtags area unit inessential to sentiment and may have an effect on the performance. Thus, in our opening move, we tend to performed basic improvement of the text by removing inessential hashtags, simply the hashtag character not the hashtag text. The next process is to avoid recognizing a similar word as a special word thanks to capitalization, we tend to fold all capitalized letters to character.

C. Exploratory Data Analysis - Exploratory information associatealysis is an approach of analyzing information sets to summarize their main characteristics, usually mistreatment applied math graphics and different information mental image strategies. A applied math model are often used or not, however primarily EDA is for seeing what the info will tell North American nation on the far side the formal modeling or hypothesis testing task.

The objectives of EDA area unit to:

• Suggest hypotheses regarding the causes of discovered phenomena

• Assess assumptions on that applied math logical thinking are going to be based mostly

• Support the choice of applicable applied math tools and techniques

• Provide a basis for more information assortment through surveys or experiments.

D. Splitting DataSet into Train and Test Data - Data cacophonous is that the act of partitioning on the market information into 2 parts, typically for cross-validator functions. One portion of the info is employed to develop a prophetic model. and also the alternative to guage the model's performance. Separating information into coaching and testing sets is a very important a part of evaluating data processing models. Typically, after you separate a knowledge set into a coaching set and testing set, most of the info is employed for coaching, and a smaller portion of the info is employed for testing. The train-test split procedure is employed to estimate the performance of machine learning algorithms after they area unit accustomed build predictions on information not accustomed train the model.

E. Future Extraction - Vectorization techniques and word embeddings area unit used for feature extraction. Similarly, for word embeddings, pretrained Word2Vec, GloVe, and fastText embeddings trained on Common Crawl and Wikipedia area unit used and have 300-D vectors. additionally, we have a tendency to used hybrid models, like hybrid ranking (HyRank) and ImprovedWord Vector (IWV) that incorporate sentiment and context of tweets for Twitter sentiment analysis.

F. Classification - Classification could be a method associated with categorization, the method within which concepts and objects ar recognized, differentiated, and understood. during this project, the multinomial NB, SVM, LR and LSTM classification algorithmic rule is employed for classifying the info. Multinomial Naïve Bayes algorithmic rule predicts the tag of a text like a chunk of email or article. It calculates the chance of every tag for a given sample so provides the tag with the very best chance as output. The multinomial distribution usually needs whole number feature counts.

G. Result Generation - The Final Result can get generated supported the general classification and prediction. The performance of this planned approach is evaluated victimization some measures like,

Accuracy - Accuracy of classifier refers to the power of classifier. It predicts the category label properly and also the accuracy of the predictor refers to however well a given predictor will guess the worth of foretold attribute for a replacement information.

AC= (TP+TN)/(TP+TN+FP+FN)

Precision -preciseness is outlined because the variety of true positives divided by the quantity of true positives and the quantity of false positives.

Precision=TP/ (TP+FP)

Recall- Recall is that the variety of correct results divided by the quantity of results that ought to are came back. In binary classification, recall is named sensitivity. It is viewed because the likelihood that a relevant document is retrieved by the question.

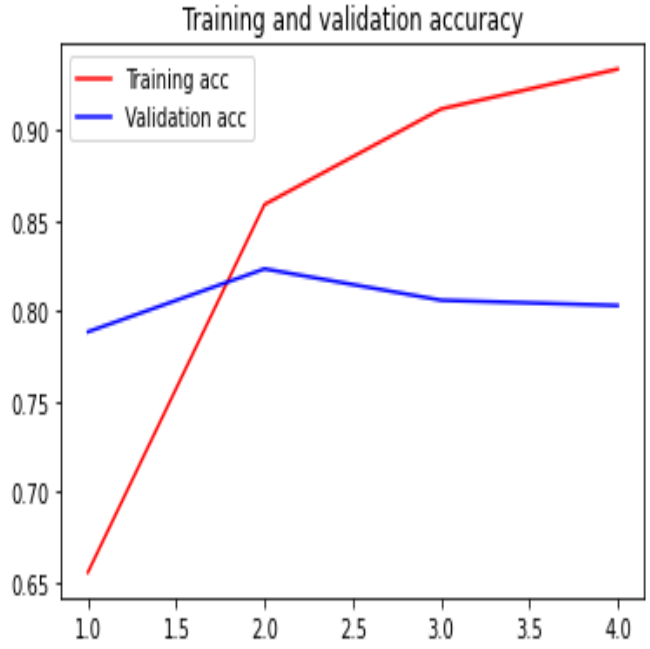
Recall=TP/(TP+FN)

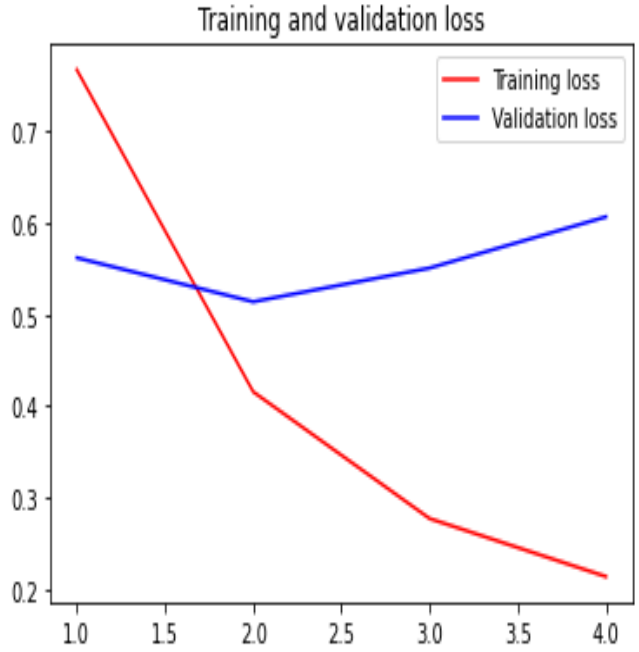
F-Measure- F live (F1 score or F score) may be a live of a check's accuracy and is outlined because the weighted mean of the preciseness and recall of the test.

F- measure=2TP/(2TP+FP+FN)

```
------Multinomial NB------
              precision    recall  f1-score   support

    Negative       0.63      0.61      0.62     12626
     Neutral       0.61      0.20      0.30      6074
    Positive       0.58      0.77      0.67     14225

    accuracy                           0.60     32925
   macro avg       0.61      0.53      0.53     32925
weighted avg       0.61      0.60      0.58     32925


------Support Vector Machine------
              precision    recall  f1-score   support

    Negative       0.72      0.73      0.72     12626
     Neutral       0.60      0.49      0.54      6074
    Positive       0.73      0.79      0.76     14225

    accuracy                           0.71     32925
   macro avg       0.69      0.67      0.67     32925
weighted avg       0.70      0.71      0.71     32925

------Logistic Regression------
              precision    recall  f1-score   support

    Negative       0.78      0.73      0.75     12626
     Neutral       0.60      0.71      0.65      6074
    Positive       0.80      0.78      0.79     14225

    accuracy                           0.75     32925
   macro avg       0.73      0.74      0.73     32925
weighted avg       0.75      0.75      0.75     32925
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 54, 32)            1155744
_________________________________________________________________
bidirectional (Bidirectional (None, 54, 512)           591872
_________________________________________________________________
flatten (Flatten)            (None, 27648)             0
_________________________________________________________________
dense (Dense)                (None, 3)                 82947
=================================================================
Total params: 1,830,563
Trainable params: 1,830,563
Non-trainable params: 0
_________________________________________________________________
Train on 32925 samples, validate on 8232 samples
Epoch 1/4
32925/32925 [==============================] - 356s 11ms/sample - loss: 0.7671 - ac
0.6555 - val_loss: 0.5619 - val_acc: 0.7886
Epoch 2/4
32925/32925 [==============================] - 352s 11ms/sample - loss: 0.4156 - ac
0.8592 - val_loss: 0.5139 - val_acc: 0.8235
Epoch 3/4
32925/32925 [==============================] - 352s 11ms/sample - loss: 0.2773 - ac
0.9121 - val_loss: 0.5507 - val_acc: 0.8061
Epoch 4/4
32925/32925 [==============================] - 349s 11ms/sample - loss: 0.2140 - ac
0.9342 - val_loss: 0.6066 - val_acc: 0.8032
```





VII. FUTURE ENHANCEMENT

In the future, it holds the potential to discover the mindset of people. In this work, we have only used tweets in English language. Tweets in other Indian languages if collected will have a better representation of people's sentiments. In near future systems are moving towards Analytics and Data to find the Prediction through machine learning or deep learning.

VIII. CONCLUSION

From the analysis, it is observed that people were mostly expressing their thoughts with positive sentiments. Since the explosion of COVID-19 conspiracy theories, social media has been widely used both for and against misinformation and misconceptions. We address the issue of Twitter sentiment on COVID-19-related Twitter posts. We benchmark sentiment analysis methods in the analysis of COVID-19-related sentiment. This analysis was unique in its way as Lockdown shows peaks and were positive. Most people were criticizing Lockdown on

the face but twitter was full of positive tweets. This analysis can be further taken to new possibilities of Emotion analysis. Rather than having Positive, Negative, and Neutral tweets, we can analyze based on emotions. Text can also represent the emotions of a person writing. Tweets can have emotions like Hate, Respect, Agreement, Anger, Happiness. Each tweet can have multiple emotions and we can have the emotion having a maximum score.In these, people can have multiple emotions and several amazing insights can be generated.

IX. REFERENCES

[1]T. Ahmad, A. Ramsay, and H. Ahmed, "Detecting emotions in English and Arabic tweets," Information, vol. 10, no. 3, p. 98, Mar. 2019.

[2] A. Bandi and A. Fellah, "Socio-analyzer: A sentiment analysis using social media data," in Proc. 28th Int. Conf. Softw. Eng. Data Eng., in EPiC Series in Computing, vol. 64, F. Harris, S. Dascalu, S. Sharma, and R. Wu, Eds. Amsterdam, The Netherlands: EasyChair, 2019, pp. 61–67.

[3] R. Moradi, R. Berangi and B. Minaei, "A survey of regularization strategies for deep models", *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 3947-3986, Aug. 2020.

[4] X. Han, J. Wang, M. Zhang and X. Wang, "Using social media to mine and analyze public opinion related to COVID-19 in China", *Int. J. Environ. Res. Public Health*, vol. 17, no. 8, pp. 2788, Apr. 2020.

[5] U. Naseem, S. K. Khan, M. Farasat and F. Ali, "Abusive language detection: A comprehensive review", *Indian J. Sci. Technol.*, vol. 12, no. 45, pp. 1-13, 2019.

[6] U. Naseem, I. Razzak, S. K. Khan and M. Prasad, "A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models", *arXiv:2010.15036*, 2020, [online] Available: http://arxiv.org/abs/2010.15036.

[7] U. Naseem, I. Razzak, K. Musial and M. Imran, "Transformer based deep intelligent contextual embedding for Twitter sentiment analysis", *Future Gener. Comput. Syst.*, vol. 113, pp. 58-69, Dec. 2020.

[8] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," Information, vol. 9, no. 5, p. 127, May 2018.

[9] X. Carreras and L. Màrquez, "Boosting trees for anti-spam email filtering," 2001, arXiv:cs/0109015. [Online]. Available: https://arxiv.org/ abs/cs/0109015 .

[10] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson, "The pandemic of social media panic travels faster than the COVID-19 outbreak," J. Travel Med., vol. 27, no. 3, Apr. 2020, Art. no. taaa031.

[11] Z. Jianqiang, G. Xiaolin and Z. Xuejun, "Deep convolution neural networks for Twitter sentiment analysis", *IEEE Access*, vol. 6, pp. 23253-23260, 2018.

[12] H. Wang et al., "Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan China", *Cell Discovery*, vol. 6, no. 1, pp. 1-8, Dec. 2020.

[13] I. Fung et al., "Pedagogical demonstration of Twitter data analysis: A case study of world AIDS day 2014", *Data*, vol. 4, no. 2, pp. 84, Jun. 2019.