A Proposed Sentiment Analysis Deep Learning algorithm for analyzing COVID-19 Tweets

Harleen Kaur[1], Shafqat Ul Ahsaan[1], Bhavya Alankar[1] and Victor Chang[2*]

1. Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India

2. Artificial Intelligence and Information Systems Research Group, School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK

Email: harleen.unu@gmail.com; mailforshafqat@gmail.com; bhavya.alankar@gmail.com; ic.victor.chang@gmail.com/V.Chang@tees.ac.uk

*: corresponding author

Abstract: With the rise in cases of COVID-19, a bizarre situation of pressure was mounted on each country to make arrangements to control the population and utilize the available resources appropriately. The swiftly rising of positive cases globally created panic, anxiety and depression among people. The effect of this deadly disease was found to be directly proportional to the physical and mental health of the population. As of 28 October 2020, more than 40 million people are tested positive and more than 1 million deaths have been recorded. The most dominant tool that disturbed human life during this time is social media. The tweets regarding COVID-19, whether it was a number of positive cases or deaths, induced a wave of fear and anxiety among people living in different parts of the world. Nobody can deny the truth that social media is everywhere and everybody is connected with it directly or indirectly. This offers an opportunity for researchers and data scientists to access the data for academic and research use. The social media data contains many data that relate to real-life events like COVID-19. In this paper, an analysis of Twitter data has been done through the R programming language. We have collected the Twitter data based on hashtag keywords, including COVID-19, coronavirus, deaths, new case, recovered. In this study, we have designed an algorithm called Hybrid Heterogeneous Support Vector Machine (H-SVM) and performed the sentiment classification and classified them positive, negative and neutral sentiment scores. We have also compared the performance of the proposed algorithm on certain parameters like precision, recall, F1 score and accuracy with Recurrent Neural Network (RNN) and Support Vector Machine (SVM).

Keywords: COVID-19, Sentiment analysis, Twitter, Recurrent Neural Network (RCN), heterogeneous Euclidean overlap metric (H-EOM), Hybrid Heterogeneous Support Vector Machine (H-SVM)

1. Introduction

The world is passing through a very difficult situation due to the spread of the coronavirus. People all over the world are losing their life due to the COVID-19 pandemic. This disease has brought exceptional impacts on people both explicitly and implicitly manners. The infection and death rate due to coronavirus is increasing day-by-day. Finally, on 11 March 2020, the WHO Director-General Dr. Tedros Adhanom Ghebreyesus announced the outbreak as a pandemic (WHO, 2020). The word "pandemic" is the outspread situation of a disease that outbreaks suddenly and engulfs a geographical region such as a country or the whole world (Singhal, 2020). Some of the diseases like cholera, bubonic plague, smallpox, and influenza are declared COVID-19 as pandemic and are the major killers in history. The smallpox disease has killed over 300-500 million people worldwide in its 12,000-year of existence. The unending COVID-19 pandemic is one of the major crises of modern times.COVID-19 is merely an infectious disease transmitted via contact or through cough, sneeze, or talk to the infected person (Nishiura et al. 2020). It is now considered a new source of stress, depression, and anxiety for people due to ambiguous information circulated over social media. The false information related to COVID-19 over social media directly affects the mental health of a person. Those diagnosed as positive died because of the fear produced by the bogus information generated by social media. To control the spread of this disease, the government has framed new policies (stay home and social distancing) and imposed restrictions on the movement of people. In this situation, the internet is the primary means to get in touch with the rest of the world. The increase in using the internet during coronavirus is graphically shown in figure 1. Individuals are fully dependent on the internet, like work from home, and everybody looks at the content regarding coronavirus that circulates on social media.

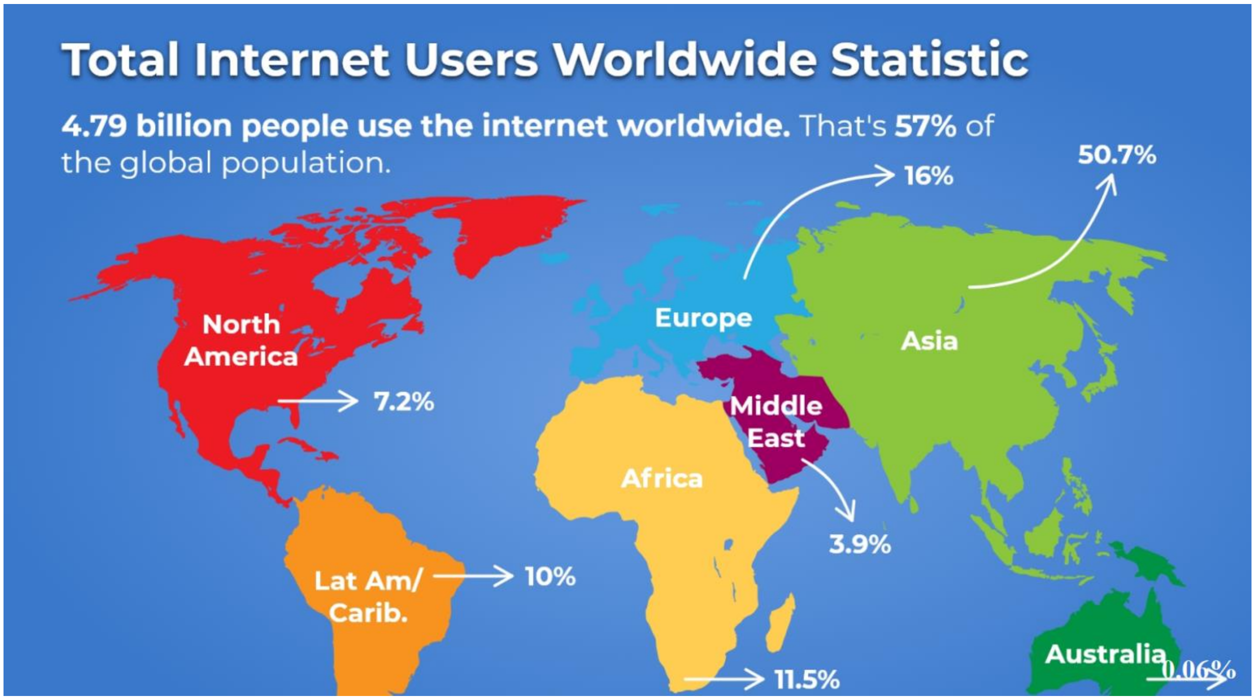

Total Internet Users Worldwide Statistic
4.79 billion people use the internet worldwide. That's 57% of the global population.
16%
50.7%
North America
7.2%
Europe
Asia
Middle East
3.9%
Africa
Lat Am/ Carib.
10%
11.5%
Australia
0.06%

Fig.1 Global Internet users in the World 2020 distribution by world regions(Source: Internet world stats)

Social media has great importance in individuals' life and connects people to the rest of the world. It is not possible for a person to work without accessing social media to cover all the updates, news like coronavirus updates, stock market updates, and some other things (Richey et al. (2018). Nowadays, people depend more on posts and tweets shared over social networking sites like Instagram, Facebook, and Twitter. It is predicted that posts shared on social media should direct people to receive authentic and foolproof information. But in most cases, the information led the people towards wrong decisions like COVID-19 information circulated over social media (Imran et al. 2020). While seeing the posts related to coronavirus, it is found that it has misled people by posting false data and figures.

The coronavirus had already disturbed people mentally; now, the opinions and tweets on COVID-19 prove alarming and a root of apprehension that needs to be highlighted to deal with disingenuous information from multiple resources. The main focus of this paper is that people should refrain from posting data over social media as they may generally cripple the impact at some point in emergencies. People must take the responsibility of sharing those data, which proves fruitful for the general public. Agencies should place fact-checkers to block false information from circulation over the internet (Alamoodi et al. 2020, Sanders et al., 2020).

In this paper, the study of sentiment analysis has been carried out through the tweets generated within the time period of COVID-19through the R programming language. We have collected the Twitter data based on hashtag keywords, including COVID-19, coronavirus, deaths, new case, recovered. We performed the sentiment classification using Recurrent Neural Network (RNN) and Support Vector Machine (SVM) and classified them as positive, negative and neutral sentiment scores.

2. Related Works

Balahur (2013) analyzes Twitter datasets using supervised machine learning methods like support vector machines, unigram, and bigram. The results generated by deploying these techniques over Twitter data clearly specified that unigram and bigram methods outshine support vector machines. The results include unique tags, modifiers and emotive words that are used to enhance the performance rating of emotions. Jianqiang and Xiaolin (2018) presented a word embedding approach that used unsupervised learning as a base. In this proposed method, hidden contextual semantic relationships and characterization between words and tweets are used. The embedded words are combined with the characteristics of n-gram and mood polarity score to form a set of emotional features and incorporated into a deep convolutional neural network.

Ortis et al. (2018) introduced a multimodal embedding space method that performed an analysis of multiple sets of images and extorted text from them. The model used a support vector machine on the text properties and examined the emotions from the images. By applying multiple-output support vector regression and multiple-input multiple-output approach, Han et al. (2020) have proposed a new multi-step time series prediction model. Also, they have conducted a comparative analysis of three primary prediction models in their work. The model was validated with both the simulated and real-world datasets. The quantitative and thorough appraisals were conducted based on expectation accuracy and computational expense. For the prediction of time series data, different data analysis techniques were applied in different studies in the literature, Baboota and Kaur (2019).

An optimized monthly streamflow time series prediction model was designed by considering different data analysis techniques by Yu et al. (2020). In the first stage of the proposed model, phase space reconstruction was conducted by applying the Correlation integral and False Nearest Neighbours (FNN) method. For the comparison purpose, the result is compared with four types of models. KNN model was performing better than other models and in case of superiority, the ARMA model was giving a better result. Authors have also used the moving average of streamflow time series data as input to the ANN model. Leskovec (2011) in his study investigated the techniques for analyzing, modeling and optimizing social media. In this paper, a series of steps, how to gather social media data, analyze the data and build prediction models for data analysis. The methods for monitoring the flow of emotions over the network and polarization development are discussed within this paper.

Mikolov et al. (2010) deployed Recurrent Neural Networks with statistical language modeling to forecast the upcoming word within the textual data. The experimental study showed a decent fall in the word error rate. Mikolov et al. (2011) presented a recurrent neural network language model (RNN LM) and observed that it outshines other language processing modeling methods. The training and testing phases of the proposed RNN model are 15-fold faster and accurate concerning other language modeling techniques. Nallapati et al. (2017) proposed an interpretable neural sequence model based on Recurrent Neural Networks, used to extract words from documents. The experimental results demonstrated that the proposed model outshines the other state-of-the-art deep learning models. Liu et al. (2016) introduced a multi-tasking learning approach that works on the principle of recurrent neural networks. The authors have proposed three different techniques for information shearing to represent model text used to perform a specific task and constitute a number of layers that work on shearing the same set of protocols. The presented model can enhance the text classification while using additional interrelated responsibilities.

Aras et al. (2016) introduced an effective approach to extend the concept of the Layer-wise Relevance Propagation (LRP) process with respect to recurring frameworks like Long Short Term Memory (LSTM) by using multiplicative interactions via the application of an extended version of LRP. The emotional sentiments within a sentence were predicted by using the LSTM model to check the relevance of the output in reference to the number of sentiments within a sentence. The experimental results also demonstrated whether the decision of the classifier is relevant to a particular class or against it and how these classifiers execute well against the gradient-based decomposition. A lexicon-based classification system has been proposed by Muhammad et al. (2016), where contextual grasp approaches are incorporated by means of the global and local context. The authors' have also introduced a hybridized approach for general-purpose lexicon, sentiwordnet, having genre-specific vocabulary.

The social media platform like Facebook, Instagram, and Twitter plays a key role in generating data and circulating content within no time. In the circulation of content related to a particular subject, there is a decent increase in the number of hate speeches. To filter these types of speeches Schmidt and Wiegand (2017) proposed a filtering tool for natural language processing. The output predicted that character-level methods are better than token-level approaches. The authors' in the presented methodology showed that a lexical list of resources could be helpful in ranking in combination with others. Pandey et al. (2017) proposed a novel metaheuristic approach on the basis of cuckoo and K- means searching strategies. In this method, the best possible cluster heads are found based on the content of the sentimental subject taken from the Twitter dataset.

Wang and Li (2015) modified the text classification methods to predict motions for sentiment analysis for the image data. Their strategy confirmed that textual and visual features for tagging emotions within an image are unsatisfactory for the prediction. The authors' have carried out the experimental results over two datasets and demonstrated that the proposed approach enhances the accuracy with respect to state-of-the-art methods. Xu et al. (2019) presented a novel Hierarchical Deep Fusion (HDF) method for emotional analysis. In the proposed model, the relationship among the features of images, text and sentimental content has been analyzed. The authors' have used three-level Hierarchal Long Short Term Memory (H-LSTM) to combine visual content with textual content in order to explore the inter-modal association of text and image at different levels.

3. Methodology

As the coronavirus cases are increasing exponentially, researchers and medical experts are in a race to develop novel rapid point-of-care diagnostics to manage the spreading of this disease. The situation became so worse that individuals wondered and deluged with fear and anxiety about what will come next. It gave rise to a wave of panic and uneasiness, and people found helpless making the situation even worse. Unpredictability and restlessness go hand-in-hand. Many real-life facts that people want to know when this deadly disease will be over, when the vaccine will be available, or when schools will reopen, or when it will be safe to visit nears and dears. While COVID-19 may create anxiety and depression, the internet can be the only source available for the public to release the anxiety to some level. However, circulating false information over the internet makes them more anxious and distressed.

3.1 Social Media vs. Misleading Information

Even if COVID-19 may occur once in a lifetime, the practice of dealing with such situations is still essential. However, some countries have successfully managed the outbreak, while some other countries have broken down badly to handle the given situation. The era we live in makes it unacceptable that social media has an important role in our lives. Social media is everywhere and everybody is connected with it directly or indirectly. In this pandemic situation, the government has framed new policies (stay home and social distancing) and imposed restrictions on the movement of people. It would be better if social media platforms would have guided us properly in this grave-like situation. Contradictory to assumptions, it has been seen that people were busy circulating spurious substances or false information over social media (Goa et al. 2020). Due to this lockdown, millions of people came in touch with social media for the first time to remain updated. It would be better if valid information could be shared and people could stay updated regarding this deadly disease that has gripped the whole world. While circulating the inappropriate content regarding COVID-19, it has created an alarming situation among people, leading to mental disturbances. People believe that using social media is very bad (Tasnim et al., 2020). The facts about coronavirus are that it is air born, and it stays over surfaces for hours. It attacks senior citizens easily; it causes breathlessness; it causes death in few days; it is uncured etc., doing multiple rounds on social media at an unexpected pace(Rajkumar 2020, Ni et al. 2020).

3.2 Deep Learning

Deep learning is a sub-domain of machine learning that consists of algorithms called neural networks, which are proposed to represent a high-level generalization of data processing through multiple layers that are piled up among each other alternating linear and nonlinear transformations(Litjens et al. 2017). Another class of neural networks known as deep neural networks, which include tens or even hundreds of layers in the form of a heap with one layer on top of another, proved to be a major advancement in speech, image, and text processing. These methods are now considered the most recent stage in the development of neural networks to make predictions from imaging data. One of the most attractive and efficient types of deep neural networks is called Recurrent Neural Networks (RNN) (Lecun et al. 1998, Krizhevskyet al. 2012) that have the capability to learn automatically to discover and combine local image characteristics in rising levels of generalization to enable prediction of result finally. RNN's are most effective only when applied to large data sets, particularly processing medical imaging databases of sufficient size are now usual(Zhang et al. 2018). One of the most striking advantages of RNN's is their ability to deal with complex associations among the input image and the complicated outputs to capture by manual measurements. Deep learning algorithms are best suited when applied to massive image-based data sets to determine and test new imaging attributes (Do et al., 2019).

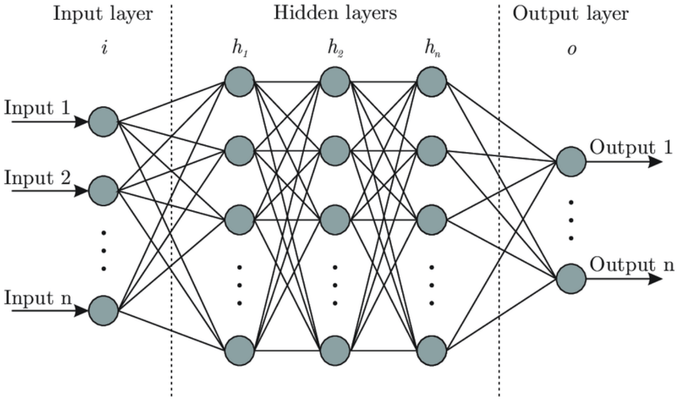


Fig.2 Structure of deep learning scheme

The deployment of neural networks for data processing has increased to a great extent for practical applications. The neural networks are trained using back-propagation plus stochastic gradient descent to adjust the node biases and weights of the edges. In back-propagation, the output produced by the neural network is compared with the expected result. After that, the difference between the final output and the expected output is calculated and the result is propagated backward through all the nodes from the output layer to the input layer (Tang et al., 2015). The stochastic gradient descent method is used to calculate the error at each node to update the weights and biases in the network to lessen the training error. The repetition of this process continues until the final output matches the expected result, and the training error minimizes to an adequate level.

3.3 Recurrent Neural Networks

Recurrent Neural Networks (RNN)belong to supervised machine learning, consisting of artificial neurons having one or more feedback loops. To train an RNN, a training dataset that constitutes input and target data is fed at the input layer of the neural network using back-propagation plus stochastic gradient descent to adjust the node biases and weights of the edges(Wang et al. 2016). The output produced by the neural network is compared with the expected result. After that, the difference between the final output and the expected output is calculated and the result is propagated backward through all the nodes from the output layer to the input layer. The stochastic gradient descent method is used to calculate the error at each node to update the weights and biases in the network to lessen the training error (Ouyang et al., 2015). The repetition of this process continues until the final output matches the expected result, and the training error minimizes to an adequate level.

A simple RNN consists of three layers- the input layer, hidden layer and output layer, as shown in Fig. 2. The input layer has N inputs, and the input is provided in the form of vectors through time t such that $\{\ldots..\{x_{t-1}, x_t, x_{t+1} \ldots\ldots\ldots\}\}$where$x_t = (x_1, x_2, x_3 \ldots\ldots\ldots x_n)$.The input layer units are strongly connected to the hidden layer units (Dos et al., 2014). Within the hidden layer, the units are characterized by a weight matrixW_{IH}. The hidden layer consists of M hidden units represented as$h_t = (h_1, h_2, h_3, \ldots\ldots.h_M)$, that are linked with each other. The state of the system by hidden layer is characterized as

$$h_t = f_H(O_t) \qquad (1)$$

Where

$$O_t = W_{IH}x_t + W_{HH}h_{t-1} + b_h \qquad (2)$$

$f_H(.)$ represents the activation function of the hidden layer, b_h depicts the bias vector of hidden units. The output layer is connected with the units in the hidden layer through weighted connections W_{HO}. The output layer consists of P elements $y_t = (y_1, y_2, y_3, \ldots..y_p)$ and are evaluated as

$$y_t = fo(W_{HO}h_t + b_o) \qquad (3)$$

Where $fo(.)$ depicts the activation function and b_o the bias vector of the output layer.

3.3.1 Activation function

In neural networks, each node has an activation function associated with it and the activation function determines the output of that node while providing input or set of inputs. There are a number of activation functions related to neural networks; however, "sigmoid" and "tanh" are commonly used. The activation function is basically used in the output layer combined with the loss function to train a classification model (Chen et al., 2018). The “sigmoid” and “tanh” activation functions are described as

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \qquad (4)$$

and

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad (5)$$

The “tanh” activation function can be defined as a scaled “sigmoid” and is represented as

$$\sigma(x) = \frac{tanh(x/2) + 1}{2} \qquad (6)$$

3.3.2 Loss function

The efficiency of the network is determined by the loss function while comparing the output $y_t$with the target output z_t and is described as

$$\mathrm{L(y,z)} = \sum_{\mathrm{t=1}}^{\mathrm{T}} \mathrm{L_t(y_t, z_t)} \qquad (7)$$

3.4 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm based on vector theory. In order to classify the data using this approach, the data is plotted in the form of vectors on the space. Hyper-planes are used to make the decisions and classify the data points by keeping the different categories of data as far as possible from one another. The labeled data points are used to train the machine and generate the hyper-planes. When new data are entered, the machine easily categorizes it as one of the available classes (Zainuddin and Selamat 2014). The SVMs are practically applied using a kernel. The ability to learn the hyper-plane is made using linear algebra, in which the observations are not directly used, rather their inner product is. The inner product is calculated by finding the sum of the product of each pair of values of input. For instance, inner product of input vectors [a, b] and [c, d] would be (a*c) + (b*d). The prediction of the inputs is made using the dot product of input x and support vector x_i that is calculated by using the following equation:

$$f(x) = B_0 + \ \mathrm{sum}\big(\mathrm{a_i} * (\mathrm{x}, \mathrm{x_i})\big) \qquad (8)$$

In (8), the inner product of input (x) would be calculated with all the support vectors in the data and the coefficients of $\mathrm{B_0}$ and $\mathrm{a_i}$ (for input) should be estimated using a learning algorithm while training.

3.4.1 Kernel trick

The data that is not linearly separable requires the transformation of the input space into a feature space by means of transformation functionϕ, based on dot products:

$$\begin{cases} Max_a \left\{ \sum_{i,j=1}^{m} a_i - \frac{1}{2} \sum_{i,j=1}^{m} \sum_{i,j=1}^{m} a_i a_j u_i u_j \langle \phi(x_i).\phi(x_j) \rangle \right\} \\ a_i \geq 0, i = 1, \dots \dots m \\ \sum_{i=1}^{m} a_i u_i = 0 \end{cases} \qquad (9)$$

where

$\emptyset(\mathrm{x_i})$ is the input x that has been transformed into the $\mathrm{i^{th}}$ element.

The kernel trick will replace the dot product, as it is impossible to compute the scalar product (Cornuejols and Miclet 2011). It is defined in (10) by Hilbert Schmidt Theory:

$$\langle \emptyset(x).\emptyset(x_i) \rangle = \sum_{i=1}^{\infty} \lambda_i \, \emptyset_i(x) \emptyset_i(y) = \ K(x, y) \qquad (10)$$

Where:

λ_i is the weighting coefficient of the $\mathrm{i^{th}}$ element.

The basic SVM requires much time while training the model in the context of data that is not linearly separable. Additionally, the standard SVM classifier is not optimal for handling big data as it does not provide the correct results. Therefore, a modified version of SVM that is parallel support vector machines has been introduced. The different types of kernels used with the Support Vector Machine in sections 3.4.2, 3.4.3 and 3.4.4.

3.4.2 Linear Kernel SVM

The kernel is a dot product that can be calculated as:

$$(x, x_i) = sum(x * x_i) \qquad (11)$$

The kernel is used to define the similarity or the measure of the distance between the support vectors and the incoming data. Some other types of kernels are also available, namely, polynomial kernel and radial kernel, which deal with more complex data in a higher dimensional plane and allows the lines to separate the classes not linearly separable easily.

3.4.3 Polynomial Kernel

A polynomial kernel is used instead of dot-product and allows the curved lines in the input space using the following equation:

$$K(x, x_i) = 1 + \mathrm{sum}(\mathrm{x} * \mathrm{x_i})^{\mathrm{d}} \qquad (12)$$

The polynomial degree (d) must be defined beforehand to the learning algorithm; it becomes a linear kernel if the degree is one.

3.4.4 Radial Kernel

This is the most complex type of kernel that can generate complex regions like polygons in space. The equation used is:

$$k(x, x_i) = exp(-\gamma * sum\left(\left(x - x_i^2\right)\right) \qquad (13)$$

Whereγ should be specified to the algorithm, in the range of 0-1. A good γ would be 0.1.

3.5 Hybrid Heterogeneous Support Vector Machine (H-SVM)

3.5.1 Algorithm Overview

Distance learning is one of the most important types of learning techniques used to classify heterogeneous data. The proposed multi-kernel-based heterogeneous support vector machine has been used two types of distance learning metrics, such as Euclidean distance and heterogeneous Euclidean overlap metric (H-EOM). The H-EOM measures the distance between nominal features by exploiting the label information of attributes. Hybrid heterogeneous support vector machine maps nominal features into real space by minimizing the generalization error. The proposed algorithm uses H-EOM to calculate the space concerning the ith nominal features because the space between numerical features is a normalized Euclidean distance. In contrast, the distance between unknown features is the maximum space.

Algorithm 1: Hybrid Heterogeneous Support Vector Machine (H-SVM) algorithm

Input:

Let X=[X_1, X_x,..., X_n] be Heterogeneous Datasets.

Output:

A support vector machine model with a mapping information table for each nominal attribute.

```
Iteration:
1: iteration i← 0
2: Initialize each nominal attribute $\mathscr{b}_1^{\mathscr{k}}$ using HEOM$\{\mathbf{p_i q_i}\} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{d_i} \|\mathbf{p_i}^{\mathbf{i}}, \mathbf{q_i}^{\mathbf{i}}\|^{\mathbf{2}}}$
3: while
    Stop condition not fulfill
            Do
4:Compute: margin and kernel matrix$\sum_{t=1}^{n} \beta_t \mathcal{K}_t \langle x_i, x_j \rangle$ by **optimization problem** of equation
5: Compute: radius □ by solving a quadratic equation or estimated □ with a variance of the $\mathcal{K}_{n\times n}$
6: Calculate: $\frac{\partial \mathcal{T}}{\partial \mathscr{b}_1^{\mathscr{k}}}$ for each nominal attribute through Generalization error bound
7: update mapping cost
        $\mathscr{b}_1^{\mathscr{k}\,\mathrm{i+1}} \leftarrow \mathscr{b}_1^{\mathscr{k}\,\mathrm{i}} + \gamma \frac{\partial \mathcal{T}}{\partial \mathscr{b}_1^{\mathscr{k}}}$
      and calculate the error $\mathcal{T}$ with a γ step
8: i = i + 1
9: End
```

3.6 Sentiment Analysis

Sentiment analysis is the process of analyzing text with the help of machine learning to identify the polarity of text. Researchers and data scientists use different sources of text to perform sentiment analysis. However, Twitter is one of the dominant sources used for sentiment analysis. Twitter is considered to be a key pillar for social networks. It acts as a podium for celebrities, sportspersons, politicians, field experts, and scientists, etc., to declare their opinions on a topic. Twitter can be used without any limitations. Users are allowed to generate substance varying from day-to-day events to far-reaching incidents (Agarwal et al., 2011). The impact of social media networks on our lives has become so large that first-hand information regarding national to international events is collected through social media, Massari, L. (2010).

Sentiment analysis is the method used to evaluate a sentence or a word on the basis of sentiment. There are mainly two approaches used for sentiment analysis. One method is to use the dictionary where each word is represented by a numerical value as polarity (Kumar and Vadlamani, 2015). The next method is machine learning, where statistical methods are employed to find out the vectorized value of a word via word embedding (Yadav et al. 2020). After that, the machine learning algorithm is trained using the digitized value of a word or a sentence, as shown in Fig 3. There are multiple machine learning algorithms used for sentiment analysis like Support Vector Machine (SVM), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Random Forest, Naïve Bayes, and Long Short-Term Memory (LSTM), Kuko, M., & Pourhomayoun, M. (2020). In this study, we have performed the sentiment classification using Recurrent Neural Network (RNN) and Support Vector Machine (SVM) and classified them as positive, negative and neutral sentiment scores.

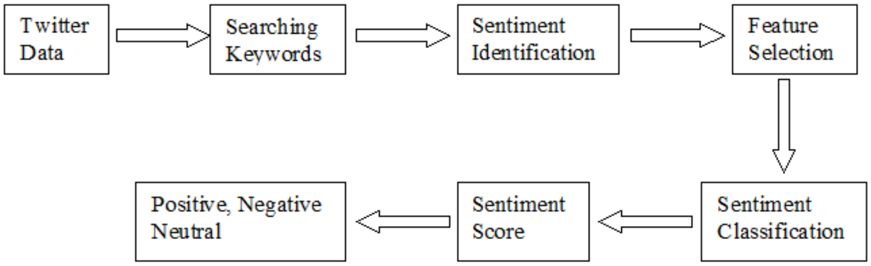


Fig. 3. Flowchart of the sentiment analysis process

4. Results and Discussion

In this paper, the dataset is taken from the IEEE data port where Twitter data about COVID-19 is easily available, Lamsal (2020). This dataset contains IDs and sentiment scores of the tweets related to the COVID-19 pandemic on Twitter. In the experiment, we have performed the sentiment analysis of this data using deep machine learning strategies like Recurrent Neural Network and Support Vector Machine through the R platform. In this study, the Twitter data has been analyzed based on hashtag keywords like (COVID-19, coronavirus, deaths, new case, recovered) and classified the data into positive (strongly positive, weakly positive), negative (strongly negative, weakly negative) and neutral sentiment scores as shown in Figs 8a, 8b and Fig 9a, 9b. The sentiment analysis is done over a different number of tweets and tried to define the sentiment scores linked with this data. Here, the analysis is carried out first on 20 tweets. After that, over 50 tweets have taken 250 tweets and, finally, over 500 tweets and classified them into different sentiment values categories.

We performed H-SVM, RNN and SVM-based sentiment analysis, which classifies the tweets based on their sentiment values. Tweets for a sample of 20, 50 and 250 tweets are considered during the experiment. The obtained results are represented using bar graphs. It has been observed, positive and neutral tweets are plotted in fig 10a and 10b with a sample of 250 tweets. The third experiment has been carried out with a sample of 600 tweets with SVM and RNN is shown in 11a and 11b. Here RNN has a maximum positive to total count ratio and SVM has neutral tweets.

The total number of sentiments for tweets is classified and categorized for the most frequently occurred keyword (COVID-19, coronavirus, deaths, new case, recovered) in the Twitter data during the COVID-19 pandemic shown below in fig 12 and 13. In the future, a similar classification experiment can be built on extracted tweets in different languages, Facebook data and Instagram data.

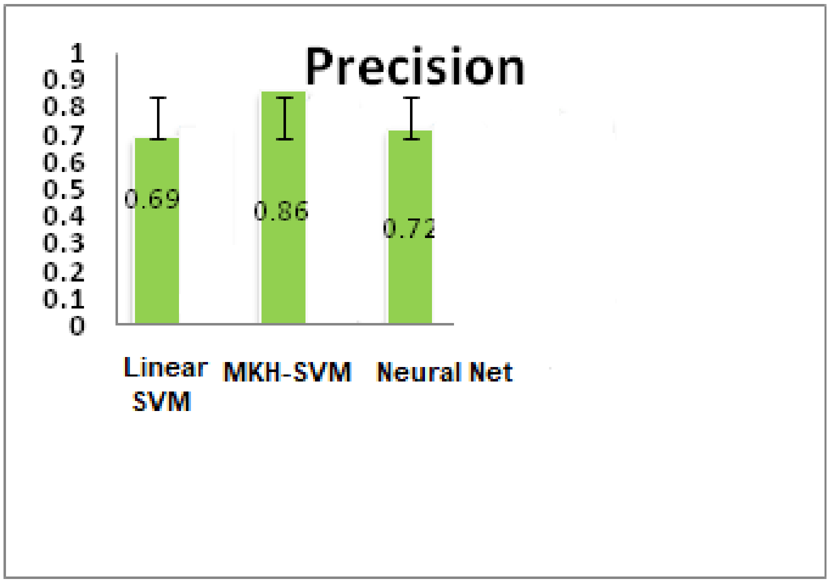


Fig.4 Precision

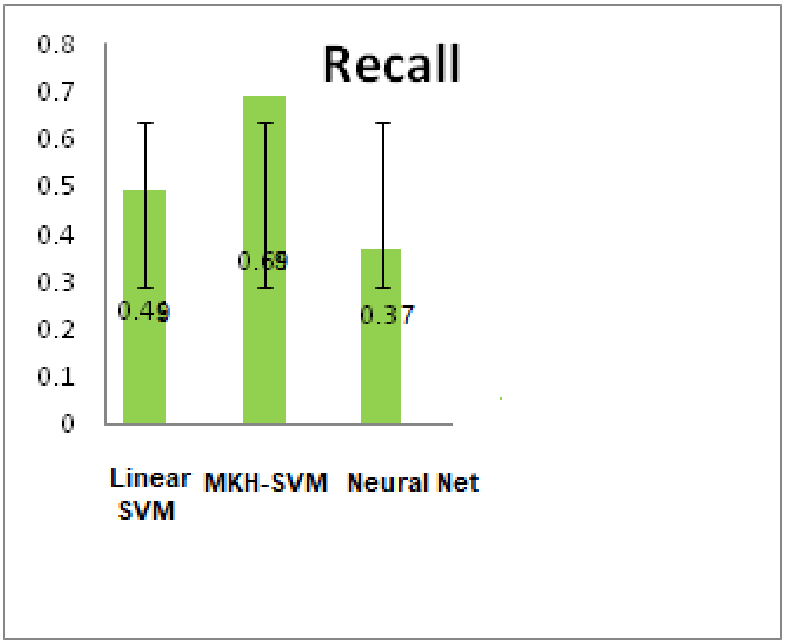


Fig.5 Recall

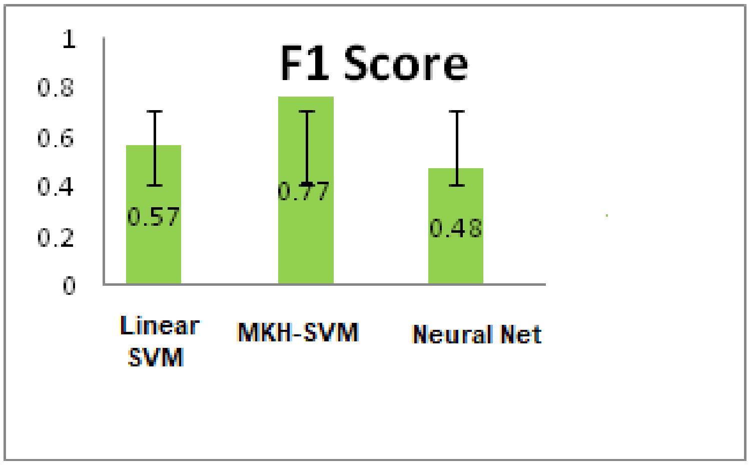


Fig. 6 F1 Score

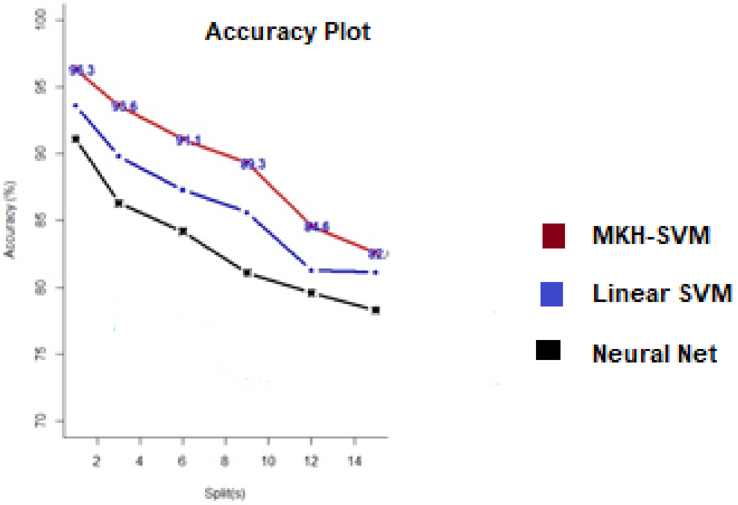
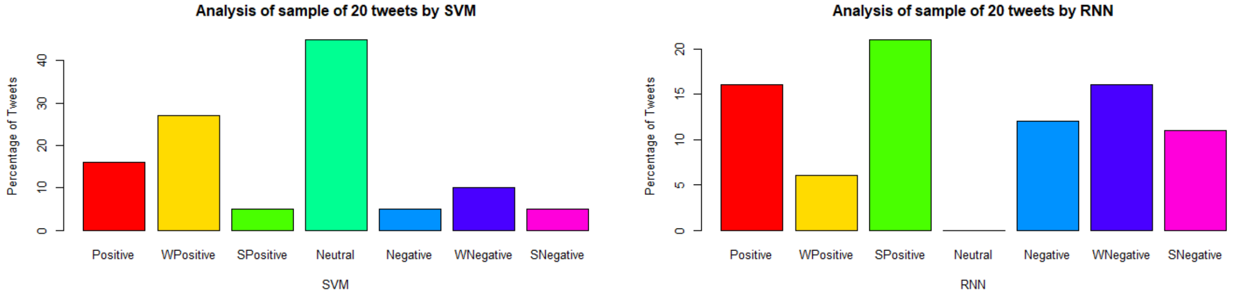
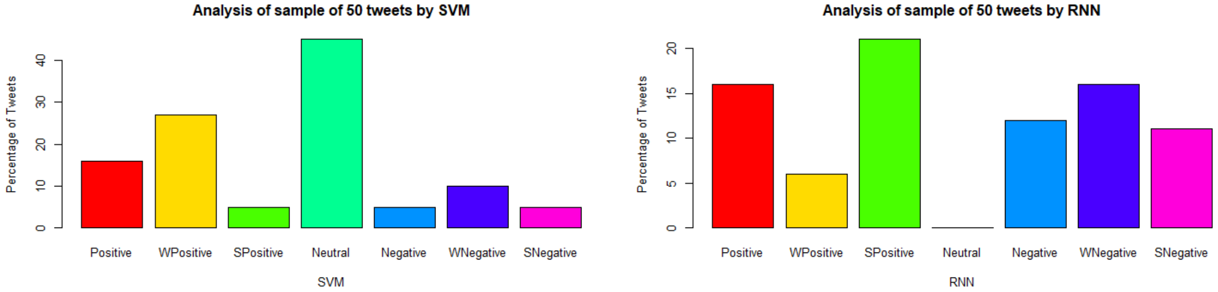


Fig.7 Accuracy



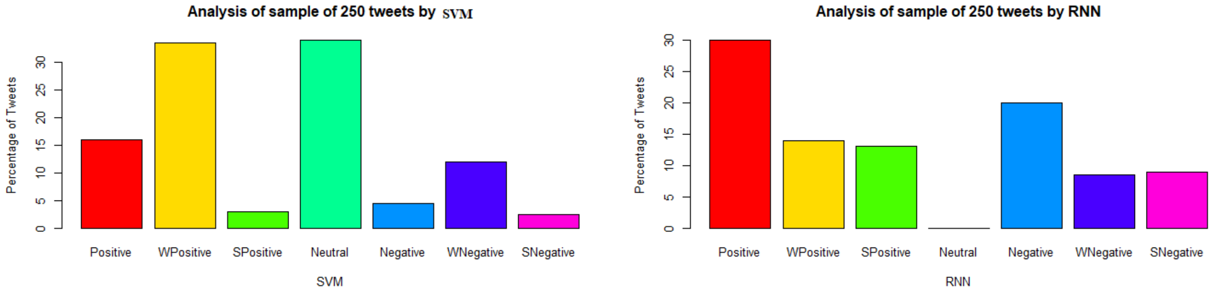
(a) **(b)**

Fig.8 Analysis of a sample of 20 tweets (a) SVM result (b) RNN result



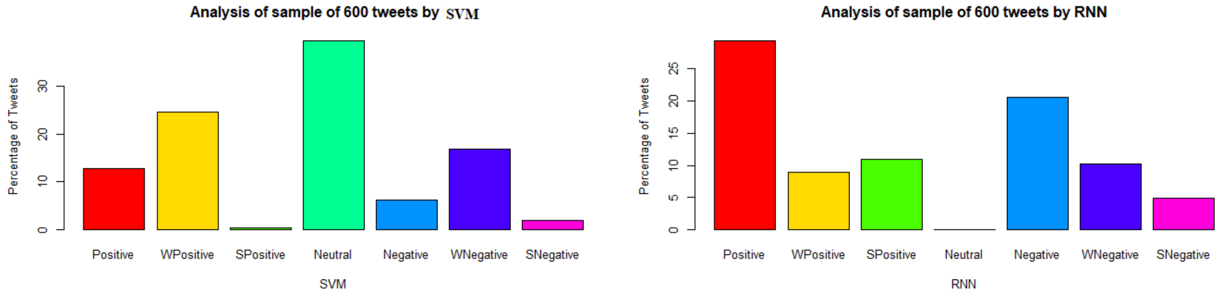
(a) **(b)**

Fig.9 Analysis of a sample of 50 tweets using (a) SVM (b) RNN



(a) **(b)**

Fig.10 Analysis of a sample of 250 tweets (a) SVM result (b) RNN result



(a) **(b)**

Fig.11 Analysis of a sample of 600 tweets (a) SVM (b) RNN

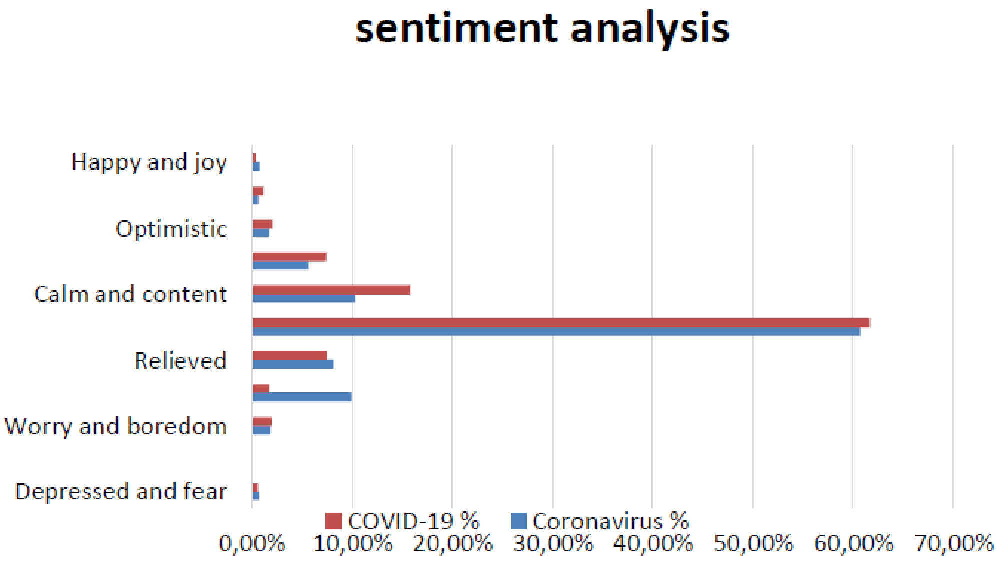


Fig.12 Sentiment analysis

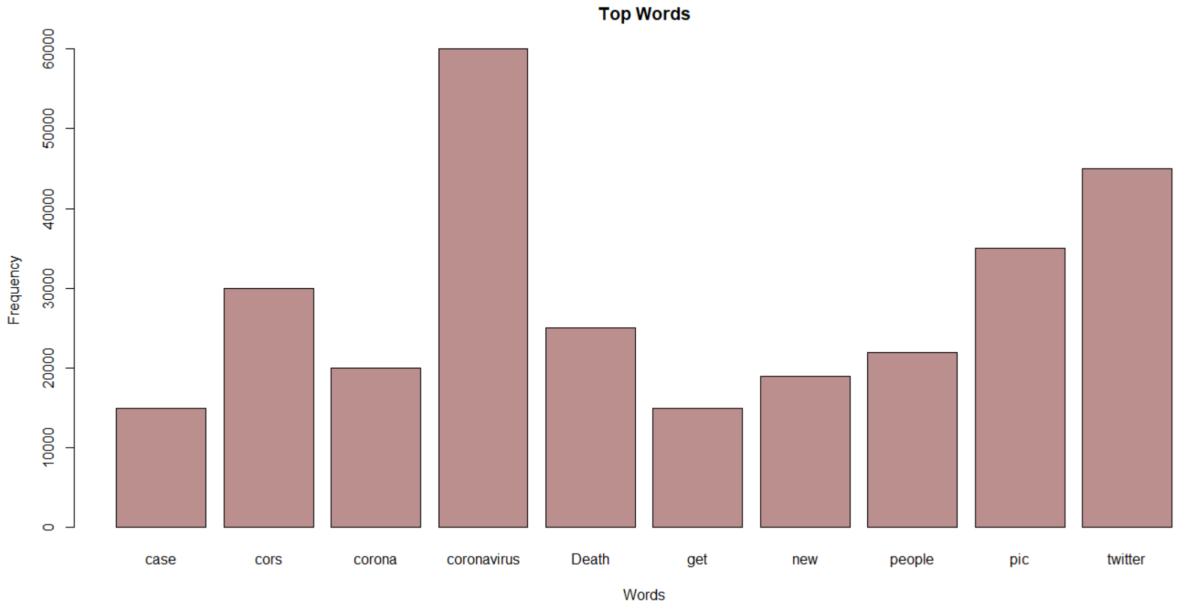


Fig. 13 Most frequently used words over Twitter handle during COVID-19.

Conclusion

With the rise in the cases of COVID-19, a wave of fear and anxiety originated throughout the world. The government imposed restrictions on the movement of people. Hence, a large number of the population rely on social media to update themselves, particularly on social media network platforms like Facebook, Twitter, Instagram, etc. The content circulated over social media regarding coronavirus has a direct impact on the lives of people. Sometimes it was handled positively by people and sometimes, it posed a negative impact on the daily routine. We collected the Twitter data based on hashtag keywords, including COVID-19, coronavirus, deaths, new case, recovered. In this paper, sentiment analysis was conducted to determine the impact of Twitter data analysis on the mental health status of the people. We performed the sentiment classification by using Recurrent Neural Network (RNN) and Support Vector Machine (SVM) and classified them as positive, negative and neutral sentiment scores.

Acknowledgment

This research work is catalyzed and supported by the National Council for Science and Technology Communication (NCSTC), Department of Science and Technology, Ministry of Science and Technology (Govt. of India), New Delhi, India (Grant Recipient: Dr. Harleen Kaur). This work is partly supported by VC Research (VCR 000116) for Prof Chang.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* (pp. 30-38).

Alamoodi, A., Zaidan, B., Zaidan, A., Albahri, O., Mohammed, K., Malik, R., & Hameed, H. (2020). Sentiment Analysis and Its Applications in Fighting COVID-19 and Infectious Diseases: A Systematic Review. *Expert Systems with Applications*, 114155.

Ardalani-Farsa, M., & Zolfaghari, S. (2010). Chaotic time series prediction with residual analysis method using hybrid Elman–NARX neural networks. *Neurocomputing*, *73*(13-15), 2540-2553.

Arras, L., Montavon, G., Müller, K. R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. arXiv preprint arXiv:1706.07206.

Baboota, R. and Kaur, H. (2019) Predictive Analysis and Modelling Football Results using Machine Learning approach for English Premier League, vol. 35, Issue 2, International Journal of Forecasting, Elsevier

Balahur, A. (2013, June). Sentiment analysis in social media texts. In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 120-128).

Chen, S., Peng, C., Cai, L., & Guo, L. (2018, July). A deep neural network model for target-based sentiment analysis. In *2018 IEEE International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7).

Cornuéjols, A., & Miclet, L. (2011). *Apprentissage artificiel: concepts et algorithmes*. Editions Eyrolles.

Do, H. H., Prasad, P. W. C., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, *118*, 272-299.

Dos Santos, C., & Gatti, M. (2014, August). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69-78).

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., ... & Dai, J. (2020). Mental health problems and social media exposure during COVID-19 outbreak. *Plos one*, *15*(4), e0231924.

Han, J., Qian, K., Song, M., Yang, Z., Ren, Z., Liu, S., & Li, X. (2020). An Early Study on Intelligent Analysis of Speech under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety. *arXiv preprint arXiv:2005.00096*.

Imran, A. S., Doudpota, S. M., Kastrati, Z., & Bhatra, R. (2020). Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning--a Case Study on COVID-19. *arXiv preprint arXiv:2008.10031*.

Internet Users Worldwide Statistic, Available at: https://www.broadbandsearch.net/blog/internet-statistics

Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. IEEE Access, 6, 23253-23260.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Kumar Ravi, Vadlamani Ravi (2015) A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems, vol. 89, 14-46.

Lamsal R. (2020). Coronavirus (COVID-19) Tweets Dataset. IEEE Dataport.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Leskovec, J. (2011, March). Social media analytics: tracking, modeling and predicting the flow of information through networks. In Proceedings of the 20th international conference companion on World wide web (pp. 277-278).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis, 42,* 60-88.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In Eleventh annual conference of the international speech communication association.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011, May). Extensions of recurrent neural network language model. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5528-5531). IEEE.

Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. Knowledge-based systems, 108, 92-101.

Nallapati, R., Zhai, F., & Zhou, B. (2017, February). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

Ni, M. Y., Yang, L., Leung, C. M., Li, N., Yao, X. I., Wang, Y.,& Liao, Q. (2020). Mental health, risk factors, and social media use during the COVID-19 epidemic and cordon sanitaire among the community and health professionals in Wuhan, China: Cross-sectional survey. *JMIR mental health*, *7*(5), e19009.

Nishiura, H., Oshitani, H., Kobayashi, T., Saito, T., Sunagawa, T., Matsui, T., & Suzuki, M. (2020). Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19).

Ortis, A., Farinella, G. M., Torrisi, G., & Battiato, S. (2018, September). Visual sentiment analysis based on on objective text description of images. In 2018 International conference on content-based multimedia indexing (CBMI) (pp. 1-6). IEEE.

Ouyang, X., Zhou, P., Li, C. H., & Liu, L. (2015, October). Sentiment analysis using convolutional neural network. In *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing* (pp. 2359-2364). IEEE.

Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management, 53(4), 764-779.

Rajkumar, R. P. (2020). COVID-19 and mental health: A review of the existing literature. *Asian journal of psychiatry*, 102066.

Sanders, A., White, R., Severson, L., Ma, R., McQueen, R., Paulo, H. C. A.& Bennett, K. P. (2020). Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse.

Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media (pp. 1-10).

Singhal, T. (2020). A review of coronavirus disease-2019 (COVID-19). *The Indian Journal of Pediatrics*, 1-6.

Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(6), 292-303.

Tasnim, S., Hossain, M. M., & Mazumder, H. (2020). Impact of rumors and misinformation on COVID-19 in social media. *Journal of preventive medicine and public health*, *53*(3), 171-174.

Wang, X., Jiang, W., & Luo, Z. (2016, December). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2428-2437).

Wang, Y., & Li, B. (2015, November). Sentiment analysis for social media images. In 2015 IEEE international conference on data mining workshop (ICDMW) (pp. 1584-1591). IEEE.

World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): situation report, 30.

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, *53*(6), 4335-4385.

Yu, C. S., Lin, Y. J., Lin, C. H., Lin, S. Y., Wu, J. L., & Chang, S. S. (2020). Development of an Online Health Care Assessment for Preventive Medicine: A Machine Learning Approach. *Journal of Medical Internet Research*, *22*(6), e18585.

Zainuddin, N., & Selamat, A. (2014, September). Sentiment analysis using support vector machine. In *2014*IEEE I*nternational Conference on Computer, Communications, and Control Technology (I4CT)* (pp. 333-337).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

Xu, J., Huang, F., Zhang, X., Wang, S., Li, C., Li, Z., & He, Y. (2019). Sentiment analysis of social images via hierarchical deep fusion of content and links. Applied Soft Computing, 80, 387-399.

Richey, M., Gonibeed, A. & Ravishankar, M.N. The Perils and Promises of Self-Disclosure on Social Media. *Inf Syst Front* **20,** 425–437 (2018). https://doi.org/10.1007/s10796-017-9806-7

Massari, L. (2010). Analysis of MySpace user profiles. *Information Systems Frontiers, 12*(4), 361–367. doi:10.1007/s10796-009-9206-8

Kuko, M., & Pourhomayoun, M. (2020). Single and clustered cervical cell classification with ensemble and deep learning methods. *Information Systems Frontiers, 22*(5), 1039–1051.