Analysis of Novel Corona virus Neural Sentiment Classification and Topic Discovery

Theertha Lakshmi AM
Dept. of CSE
BGS Institute of Technology
Adichunchanagiri University
BG Nagar, Karnataka, India-571448.

Dr. Ravikumar G K
Professor &Head(R&D)
Dept. of CSE
BGS Institute of Technology
Adichunchanagiri University
BG Nagar, Karnataka, India-571448

Ms.Sindhu D
Dept. of ISE
BGS Institute of Technology
Adichunchanagiri University
BG Nagar, Karnataka, India-571448

Abstract—Internet forums and open social media platforms, including virtual care boards, offer users (people/patients) who are worried regarding medical conditions a comfortable avenue to talk and exchange intelligence with one another. The COVID-19 sickness is an infection caused by a novel coronavirus that first emerged in late December 2019. As a result of this epidemic and the virus's quick global spread, the WHO announced a phase of disaster. In order to identify diverse COVID19-related concerns from general perspectives, we employed computerized retrieval of COVID-19-related conversations from social networking sites and natural language processing (NLP) techniques based on topic modeling in this study. Additionally, we look into how to classify COVID-19 comments' sentiment using LSTM recurrent neural networks. Our findings highlight the significance of using the general public's perceptions and appropriate computational tools to comprehend COVID-19-related concerns and direct pertinent decision-making. Additionally, tests showed that the study model outperformed numerous including well machine-learning techniques for COVID-19-Sentiment Identification, with an efficiency of 81.15 percent. to aid in decision-making and to comprehend COVID-19-related difficulties. Additionally, tests showed that the study method outperformed numerous existing well machine-learning techniques for COVID-19-Sentiment Classification, with an efficiency of 81.15 percent.

Keywords—COVID-19, Natural Language Processing, Topic modeling, Deep Learning.

***Abstract*—**Internet forums and open social media platforms, including virtual care boards, offer users (people/patients) who are worried regarding medical conditions a comfortable avenue to talk and exchange intelligence with one another. The COVID-19 sickness is an infection caused by a novel coronavirus that first emerged in late December 2019. As a result of this epidemic and the virus's quick global spread, the WHO announced a phase of disaster. In order to identify diverse COVID19-related concerns from general perspectives, we employed computerized retrieval of COVID-19-related conversations from social networking sites and natural language processing (NLP) techniques based on topic modeling in this study. Additionally, we look into how to classify COVID-19 comments' sentiment using LSTM recurrent neural networks. Our findings highlight the significance of using the general public's perceptions and appropriate computational tools to comprehend COVID-19-related concerns and direct pertinent decision-making. Additionally, tests showed that the study model outperformed numerous including well machine-learning techniques for COVID-19-Sentiment Identification, with an efficiency of 81.15 percent. to aid in decision-making and to comprehend COVID-19-related difficulties. Additionally, tests showed that the study method outperformed numerous existing well machine-learning techniques for COVID-19-Sentiment Classification, with an efficiency of 81.15 percent.
Keywords—COVID-19, Natural Language Processing, Topic modeling, Deep Learning.

I. INTRODUCTION:

Health care organizations can get information about people's/patients experiences by using online discussion boards like Reddit. These discussion boards are excellent resources for gathering comments from the public, which can then be analyzed for user behavior and knowledge discovery. A user can utilize keywords and information sources in a standard sub-Reddit forum to find important queries, responses, or opinions submitted by other Reddit members. An authorized user can also post new queries or topics to initiate conversations with other community leaders. In answer to each question, additional users can reflect and express their ideas and experiences. People can exchange concerns, issues, and needs to be connected to medical difficulties as well as positive and negative remarks on these discussion boards. By examining those remarks, researchers can find insightful

suggestions for enhancing health services and comprehending consumer issues. Before the end of December 2019 [1,] a new coronavirus epidemic that causes COVID-19 was discovered. A situation of emergency was declared by the WHO as a result of the virus' rapid expansion. In this paper, we employed topic modeling in natural language processing (NLP) and systematic retrieval of COVID-19-related talks from media networking sites to reveal numerous concerns linked to COVID-19 from public viewpoints. Additionally, we examine how to classify COVID-19 remarks' attitudes using LSTM recurrent neural networks. Our findings highlight the significance of utilizing the general public's perceptions and appropriate computational techniques to comprehend COVID-19-related concerns and direct pertinent decision-making.

Our findings highlighted the significance of using the general public's perspectives and appropriate algorithmic tools to comprehend COVID-19-related concerns and inform pertinent decision-making. This article is organized as follows altogether. We start by giving a quick overview of internet groups for medicine. Section II provides an overview of COVID-19-related topics and a few related works. We detail the data pre-processing techniques we used in our research in section III, along with the NLP and deep learning techniques we used on the COVID-19 remarks collection. We then give the findings and conclusions. We end by discussing potential future studies focused on NLP techniques for analyzing the online platform in relation to COVID-19.

II. RELATEDWORK:

Popular techniques for evaluating textual information in online medical discussions include machine learning and DL methodologies focused on sentiment and semantic analysis. These techniques have been widely employed by researchers on internet platforms including Twitter, Reddit [2]–[7], and websites with health-related content [8], [9. For instance, Halder and colleagues [10] examined how language evolves over time to analyse a user's emotional level. They looked into user content in a sizable collection from the psychological internet discussions of healthboards.com using a recurrent neural network (RNN). Depending on user discussions in online health forums, McRoy and associates [11] looked at ways to automatically identify the information is going of breast tumor patients. By analyzing user postings from multiple subreddits from 2012 to 2018, Chakravortian and associates [12] derived topics depending on different health conditions expressed in internet discussions. A categorization method was presented by VanDam and colleagues [13] for locating articles in medical networks internet which were relevant to clinics. Within that database, the scientists gathered 9576 thread-starting posts from the healthcare reference portal WebMD.

This COVID-19-related remarks from an virtual medicare-focused community might be viewed as possibly helpful for identifying significant themes to better comprehend perspectives and emphasize user interactions and advance health policies. To the greatest of our understanding, that is the first research to employ NLP techniques to analyze COVID-19-related remarks from sub-Reddit communities, despite the fact that there are comparable efforts pertaining to a variety of health conditions in the internet forums. We recommend using topic modeling-based NLP to extract features important subjects, build a deep learning framework based on LSTM RNN for emotion categorization on COVID-19 remarks, &recognise people's positive or negative opinions as those who relate to COVID-19 concerns in order to notify responsible outcome.

Analyzing social media discussions on websites like Reddit could offer valuable information for comprehending people's ideas, which may have been challenging to do using conventional procedures like manual processes. Towards the finest of our understanding, this will be the initial study to assess comments for virtual care networks by taking into account morphological and affective aspects of COVID-related Internet posts.The text material on Reddit has been analyzed in numerous studies[43]-[45]. In this study, we examined three crucial research issues and put up a methodical framework to adequately respond to them. To respond to RQ1, we took into account an available data set that had 563,079 remarks from 10 sub-Reddits. We discovered and discovered significant underlying topic softeners regarding COVID-19 remarks connected to numerous problems.

III. RESEARCH METHODOLOGY:

This paragraph explains the research methodologies used to examine the primary components of the research, including suggesting using an unsupervised methodology along with a cooperative deep-learning model based on LSTN RNN to analyze comments from subreddits connected to COVID-19. The created framework, illustrated in Figure 2, mines and analyses remarks linked to COVID-19 using sentiment and semantic analysis.

A. Getting Ready the Model Parameters

Reddit is a social networking platform with reviews for web pages that is based in the United States. Users of this social media platform can ask and answer others' questions and comments on a variety of topics, including COVID-19. The entries are categorized into "subreddits," which are themes made up by internet consumers and include a wide range of issues like politics, research, medicine, entertainment, literature, exercise, cooking, and image-sharing. This webpage is a great resource for learning about COVID-19-related health-related issues. As the first stage in creating this classifier, the study concentrates on COVID-19-related remarks of 10 subreddits using an available database.

B. Elimination of Noisy and Stop-Words

Eliminating meaningless words and information, known as stop-words in natural language processing (NLP), from raw text is among the more crucial processes in the pre-processing of COVID19-related remarks. Furthermore, by removing stop-words, they also reduced the dimension of the characteristic region. For instance, the terms that are most frequently used in text comments—such as adjectives, punctuations, pronouns, and connecting verbs—are typically worthless and have little impact on the final product. Instances include the words "am," "is," "are," "they," "the," "these," "I," and "that."

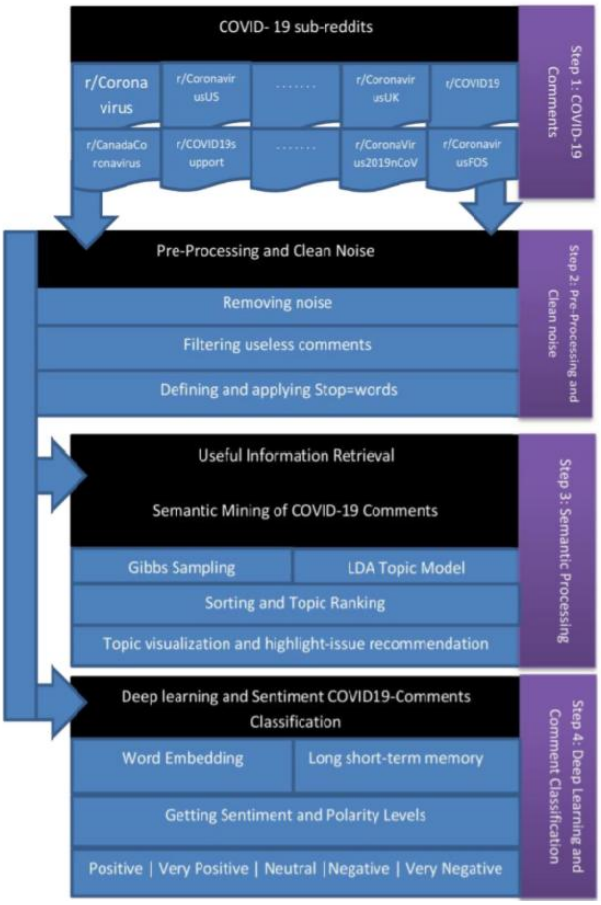

Fig1. A summary of the survey methodology used to interpret remarks made about COVID-19 in a relevant way.

C. The mining of COVID-19 comments and semantic retrieval

A useful methodology used in NLP is text-document modeling, which models particular publications and collections of documents using terms found in the texts. One form of document modeling method for meaningful retrieval in natural language processing is topic modeling. Latent Dirichlet Allocation (LDA) and Semantic Similarity Segmentation are popular topic modeling techniques (PLSA). The LDA has a deep fundamental architecture and allows for conditional network training, which is one of its major benefits.LDA has the potential to reduce complexity, making it appropriate for large-scale corpora. Each paragraph in a collection is characterized by a randomized combination of latent subjects in the probability framework known as LDA.A dispersion over terms is used to characterize each concealed subject. The most significant benefit of LDA above LSI is that it takes into account the fact that the text articles in a large corpus have a number of hidden subjects, which in turn cause redistributions over keywords formed in the publications. A further advantage of LDA is that simple inference methods can be offered on previously undiscovered formats. Employing the LDA framework here has as its goal the extraction of semantic characteristics.

Algorithm 1: Pre-Processing and Removing the Noise to Prepare the Input Data.

Input: A group of COVID-19–related comments as main document context
Output: Text in a string.

1: d_i = Get data(); getting COVID-19 comments as pure data.
2: **For** d_i.row (all record) ! = last record **do**
3: $d_{i2} = d_i$.cleanData(d_i); removing stop-words, clean noise
4: $d_{i2} = d_{i2}$.arranged(); processing to arrange dataset.
5: **end for**
6: **return** d_{i2} as a string

In the final phase, we used subject modeling relying on an LDA Prediction structure and Gibbs selection [20] to recover semantic information from COVID-19-related statements and uncover latent topics. However, COVID-19 remarks may depend on a variety of topics that users on Reddit have addressed. This process allows us to identify and learn about these

important issues or topics. Given that the homogeneity Autoregressive probabilities from which the continuous subject probabilities are generated, we regarded a gathering of documentation, likes that COVID-19-related remarks and phrases, as topics(K) based on the LDA model. The preceding calculation was used to calculate and acquire the percentage of observed value D from each COVID-19-related statement in a collection.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}(p\ \theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d \quad (1)$$

determining specifications for the subject of Dirichlet prior and also taking into account specifications for the term Dirichlet before. N is the amount of the vocabulary, and M is the quantity of textual information. The possibilities of topics at the manuscript were calculated using a composite of Dirichletmultinomials. Using two Dirichletmultinomials, the topic-word probabilities were likewise determined to be (,). Additionally, the variables at the textual level were designated as dd, which can sampling each article. Every text statement's word-level parameters (zdn and wdn) were examined.

Algorithm 2: General Process for Semantic-Comment-Mining via Topic Model.

Input: A group of COVID-19–related comments as main document context
Output: A set of topics from the documents as integer values;

1: Pre-process and remove noise and clean data by Algorithm 1.
2: **for** each topic $k \in \{1, 2, \ldots, \mathrm{k}\}$ **do**
3: word-probability under the topic of sampling —— or the word distribution for topic *k* among COVID-19–related comments
4: - ϕ ~Dirichlet(β)
5: **end for**
6: **for** each COVID-19-related comments-document $d \in \{1, 2, \ldots, D\}$ **do**
7: The topic distribution for document *m*
8: *dθ~Dirichlet*(α)
9: **for** per word in COVID-19–related content-document *d* **do**
10: sampling the distribution of topics in the COVID-19-related comments-documents to obtain the topic of the word: *.Zd ~ Mul*(ϕ)
11: word-sampling under the topic, $W_d\tilde{M}ul(\phi)$
12: **end for**
13: **end for**

As part of our methodology for identifying hidden patterns and semantically analyzing, Algorithm2 outlines a general procedure. The quantity of COVID-19-related responses is the primary input for this document's environment. Based on Algorithm 1, The raw info is processed in Line 1 to remove distortion and punctuation marks. The likelihood of the gradually shifting from Topic K[i] is computed in Lines 2–5. The topic dispersion likelihood from the COVID-19-Content-Document is computed in lines 6 through 11. The variables were calculated for the manuscript and phrase of the architecture, as shown in Equation 1. More specifically, the LDA treats phrases as a statistical model combination of a pre-defined count of underlying features and topics as multivariate regression probabilities in publications. Algorithm 3's lines 1-3 illustrate the semantic mining process used to identify latent concepts.

D. Identification of sentiment using COVID-19 and supervised learning

DNN used with success for a variety of ML applications, including NLP-based approaches that use sentiment elements for deep categorization.DNN may simulate high-level representations and reduce dimensionality by combining non-linear manipulations with numerous processing elements based on complicated structural models. RNNs are common models that have been shown to be important and strong in most NLP works. RNNs are designed to employ sequential data, and the result is improved by keeping track of earlier computations.

Algorithm 3: COVID-19–Related Comments Mining and Topic Recommendation.

Input: Importing latent topics from Algorithm 2
Output: Recommended top highlight topics of various aspects of COVID-19 comments

1: Extract semantic contents, training the LDA Topic Model
2: Determining the top topics recommended based on the value of the topic probability of all data.
3: Ranking and sorting the most meaningful topics recommended of COVID-19 comments
4: **return** A list of recommended highlight topics

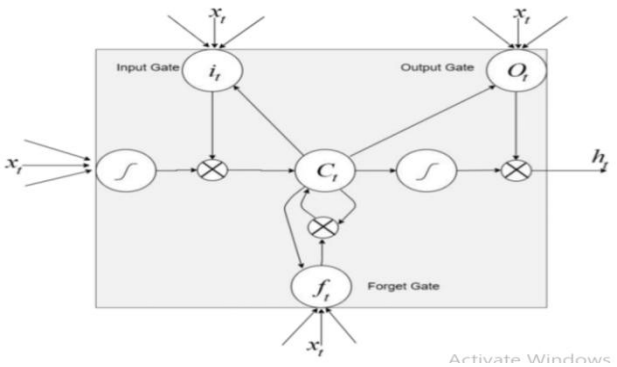

Fig2:The basic design of an LSTM memory module. This construction has three gates (ft,it,ot) and a storage unit, as indicated (ct).In fact, RNNs have a storage feature that stores data that has already been computed. However, due to curvature disappearing or bursting, basic RNNs face considerable difficulties but are incapable to establish long-term relationships. The advantage of LSTM [30, [31] units is that they may overcome this difficulty by modifying the knowledge in a current block utilizing three separate gates. Every LSTM cell's equation formally expressed as

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (2)$$

$$i_t = \sigma\left(W_f x_t + U_i h_{t-1} + b_i\right) \quad (3)$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \quad (4)$$

W, U, and b stand for the variables in the valves and cell phases, respectively. These three equations, eqs. 2-4, individually establish the forget (ft), input (it), and production (ot) valves for one LSTM cell. According to Figure 3, the remember gate in an LSTM layer chooses which initial knowledge from of the current block is remembered. The new knowledge that is stored in the memory module is controlled or determined by the input gate. The amount of data to be disclosed from the memory storage cell is controlled or decided by the output nodes. The calculations for the cell-memory/input unit are:

$$\tilde{C}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \quad (5)$$

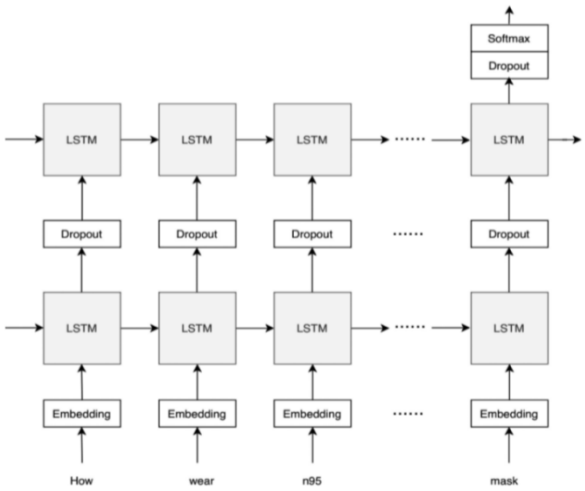

Fig3. The LSTM's structural components for COVID-19 sentiment categorization.

IV. PROPOSED APPROACH

We give a thorough explanation of the data collection process and experimental findings in this section, followed by a detailed analysis of the findings. We evaluated 563,079 Reddit comments about COVID-19. The whole information is accessible on Kaggle and was gathered between January 20, 2020, and March 19, 2020. 1 To execute reasoning and preserve the LDA prediction model to recover hidden patterns, we used MALLET2. Our deep-learning model was implemented using the Python tool Keras3.

A. Sentiment and Polarity Results

Text and comments can be classified depending on phrase polarity using sentiment classification, a useful NLP tool for information extraction [33]–[35]. This method has a wide range of applicability in many fields, including information extraction in internet medical groups. Identifying the general tone of this COVID-19-related remarks was the major goal of this investigation. We determined if each comment was on average neutral, favorable, or negative. Figure 10 displays the general tone of the database's comments combined with the typical tone of remarks outlining the words COVID-19. We used SentiStrength to provide negative and positive ratings for each of the polar remarks in our dataset consisting, and we then used the different values as direct guidelines for drawing conclusions about the polarity/sentiment of the COVID-19 remarks.

B. Deep Classification and Feature Analysis

We built up this information to autonomously categorize the sentiment of the COVID-19 comments for all of these data; depending on the Sentistrength sentiment score, we classified each remark as very positive, positive, extremely negative, negative, or neutral. There were 338,666 COVID-19-related remarks in the training dataset and 112,888 in the test dataset. With the use of the Support Vector Machine (SVM), Naive Bayes, Logistic Regression, and K Nearest Neighbors (KNN) methods, we tested the suggested LSTM-model in this research together with supervised machine learning approaches.

EXAMPLES OF COVID-19 COMMENTS FROM THE REDDIT CORPUS

Polarity	People's Comment	Score of the words
Positive	I hope loved ones remain safe healthy.	I[0] hope[2] loved[3] ones[0] remain[0] safe[1] healthy[0]
	Ah yes manbaby magnificent immune system. better luck next time covid-19	Ah[0] yes[0] manbaby[0] magnificent[3] immune[0] system[0] better[0] luck[2] next[0] time[0] covid[0] 19[0]
Negative	Greed prejudice racism hate kill faster covid-19 .	Greed[-2] prejudice racism[-1] hate[-3] kill[-1] faster[0] covid[0]
	So much bullshit one thread alone. scary times	So[0] much[0] bullshit[-2] one[0] thread[0] alone[0] scary[-3] times[0]
Neutral	I heard radio likely official guidance next 10 - 14 days	I[0] heard[0] radio[0] likely[0] official[0] guidance[0] next[0] 10[0] 14[0] days[0]
	Everyone wear mask case unintentionally spreading everyone	Everyone[0] wear[0] mask[0] case[0] unintentionally[0] spreading[0] everyone[0]

Nevertheless, theyutilised the Scikit-learn tool, a Python toolkit that supports numerous machine-learning techniques, to implement the Learning models to sentiment categorization. We selected these strategies for the COVID-19 [42] due to their high precision and efficacy for sentiment categorization in natural language processing. Figure3 depicts the most precise model for classifying a COVID-19 statement as having a very positive, positive, extremely negative, negative, or neutral feeling. Our method, which used the LSTM model to classify all COVID-19 comments into the given dataset, had an accuracy rate of 81.15 percent, exceeding that of conventional machine-learning methods. The sentiment and semantic methodologies, in our opinion, can produce insightful findings with a summary of how users' and people's feelings about the incident.

C. Discussion and Useful Results

Analyzing social media remarks on websites like Reddit could offer valuable information for comprehending individualideas, that also could be challenging to do using conventional approaches like manual procedures. Towards the authors ' ability, this is the first objective was to analyze comments for virtual care networks while taking into account semantic and sentiment aspects of COVID-related remarks from Reddit. The text material on Reddit has been analyzed in numerous studies[43]-[45]. In this study, we looked into three significant research topics and put forth a methodical framework that adequately answers them. To respond to RQ1, we took into account an available data set that had 563,079 remarks from 10 sub-Reddits. We discovered and discovered significant active themes of phrases regarding COVID-19 remarks relating to several difficulties.

V. RESULTS:

The resulting LDA results were interpreted using several visuals. Since a topic is described as a categorized probability across words, When applied to manuscripts, the deterministic LDA framework postulates that every item in a group was composed of a variety of latent (unobserved) subjects. It is feasible to identify numerous words that are likely connected to needs and emphasize interactions of the individuals or members on Reddit in relation to the top-ranked subjects for the COVID-19 remarks. For the purpose of identifying RQ2 and RQ3, we created a two-layer LSTM to detect relevant latent features and sentiment categorization on COVID-19-related worries using medical discussion boards on Reddit.We did this in order to obtain the directional phrases for each remark and to identify RQ2 and RQ3. We showed that our deep learning strategy for sentiment categorization based on LSTM outperforms various other popular machine learning techniques.

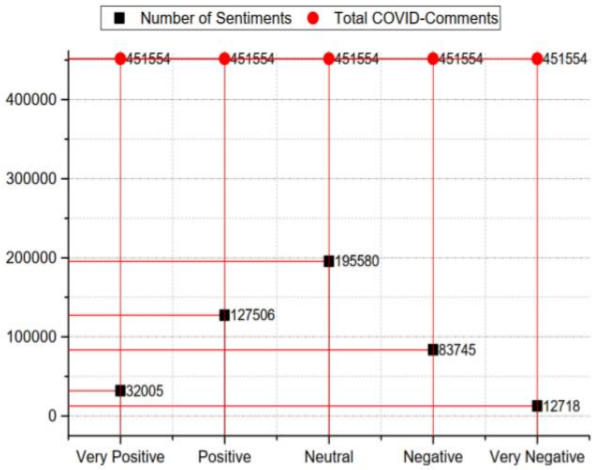

Fig4. Reddit data on the proportion of COVID-19 responses with positive, negative, and impartial sentiments.

English-language material was the only type of text that was considered for this study, which was a selecting criterion. As a result, the outcomes do not take into account comments submitted in other countries. Additionally, the comments that were obtained for this study were just from January 20 and March 19, 2020. As a result, the time spaceamong these completion of the investigation and the time frame of our analysis may have had some influence on the timelines and four outcomes.

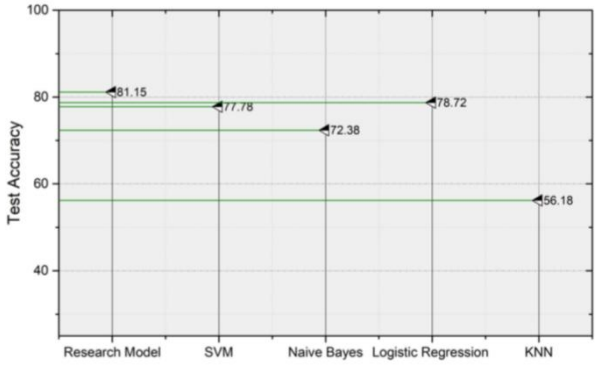

Fig5. Reliability of the COVID-19 sentiment-classification techniques employing various characteristics.

Finally, the research showed that we were able to derive some useful knowledge from remarks connected to COVID-19 thanks to a systematic framework that combined NLP and deep learning techniques related on subject modeling and an LSTM design. These quantitative findings can be helpful for assessing the pro and con tendencies of an internet community as well as for gathering user comments to aid researchers and physicians in better understanding the behavior of individuals in a life-or-death crisis. Regarding upcoming research, they plan to evaluate various social networking sites, including Facebook pages, using composite fuzzy deep-learning algorithms which may be used in the coming enabling sentiment characterisation as an unique method of collecting relevant residual subjects from opening remarks.

VI. CONCLUSION AND FUTURE WORK:

For the understanding, it's the only research that examines the relationship among the sentiment of COVID-19 remarks and semantic Reddit themes. The main goal of this study was to illustrate an original application of natural language processing (NLP) based upon the LSTM network to identify relevant salient issues with segmentation techniques on COVID-19-related issues through medical websites, which include a sub-Reddit. The findings of this paper, in our opinion, will help in evaluating people's problems and priorities with regard to COVID-19-related problems. Additionally, our study finds can help in refining realistic COVID-19 public health services and intervention programs.

REFERENCES:

1. M. Malta, A. W. Rimoin, and S. A. Strathdee, "The coronavirus 2019nCoVepidemic:Ishindsight20/20?,"EClinicalMedicine,vol.20,pp.1–2, 2020.
2. J. Thomas, A. V. Prabhu, D. E. Heron, and S. Beriwal, "Reddit and radiation therapy: A descriptive analysis of posts and comments over 7 yearsbypatientsandhealthcareprofessionals,"Adv.Radiat.Oncol.,vol.4, no. 2, pp. 345–353, 2019.
3. G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitterdataduringcriticaleventsthrough Bayesiannetworksclassifiers," Future Gener. Comput.Syst., vol. 106, pp. 92–104, 2020.
4. J.M.Barros,P.Buitelaar,J.Duggan,andD.Rebholz-Schuhmann,"Unsupervised classification of health content on reddit," in Proc. 9th Int. Conf. Digital Public Health, 2019, pp. 85–89.

5. M. Roy, N. Moreau, C. Rousseau, A. Mercier, A. Wilson, and L. Atlani-Duault, "Ebola and localized blame on social media: Analysis of Twitter and Facebook conversations during the 2014–2015 Ebola epidemic," Culture, Medicine, Psychiatry, vol. 44, no. 1, pp. 56–79, 2020.
6. B. Alshemali, "Improving the reliability of deep neural networks in NLP: A review,"Knowl.-Based Syst., vol. 191, 2020, Art. no. 105210..
7. A. Park, and M. Conway, "Tracking health related discussions on Reddit for public health applications," in Proc Amer. Med. Inform. Assoc., 2017, Art.no. 1362, vol. 2017.
8. G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," Decis. Support Syst., vol. 57, pp. 77–93, 2014.
9. O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment analysis using deep learning approaches: an overview," Sci. China Inform. Sci., vol. 63, no. 1, pp. 1–36, 2020
10. W. Liu, "Research on Cloud Computing Security Problems and Strategy", IEEEE conference on Consumer Electronics, Communications and Networks, April-2012, pp. 1216 – 1219.