Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach

Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang

***Abstract*—Internet forums and public social media, such as online healthcare forums, provide a convenient channel for users (people/patients) concerned about health issues to discuss and share information with each other. In late December 2019, an outbreak of a novel coronavirus (infection from which results in the disease named COVID-19) was reported, and, due to the rapid spread of the virus in other parts of the world, the World Health Organization declared a state of emergency. In this paper, we used automated extraction of COVID-19–related discussions from social media and a natural language process (NLP) method based on topic modeling to uncover various issues related to COVID-19 from public opinions. Moreover, we also investigate how to use LSTM recurrent neural network for sentiment classification of COVID-19 comments. Our findings shed light on the importance of using public opinions and suitable computational techniques to understand issues surrounding COVID-19 and to guide related decision-making. In addition, experiments demonstrated that the research model achieved an accuracy of 81.15% – a higher accuracy than that of several other well-known machine-learning algorithms for COVID-19–Sentiment Classification.**

***Index Terms*—Coronavirus, COVID-19, Natural Language Processing, Topic modeling, Deep Learning.**

Manuscript received April 21, 2020; revised May 25, 2020 and June 6, 2020; accepted June 7, 2020. Date of publication June 9, 2020; date of current version October 5, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants 61941113, 81674099, 61502233, in part by the Fundamental Research Fund for the Central Universities under Grants 30918015103, 30918012204, in part by Nanjing Science and Technology Development Plan under Project 201805036, and "13th Five-Year" equipment field fund under Grant 61403120501, in part by the China Academy of Engineering Consulting Research under Project 2019-ZD-1-02-02, and in part by National Social Science Foundation under Grant 18BTQ073. *(Corresponding authors: Yongli Wang; Hamed Jelodar.)*

Hamed Jelodar and Yongli Wang are with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jelodarh@gmail.com; yongliwang@njust.edu.cn).

Rita Orji is with the Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada (e-mail: rita.orji@dal.ca).

Shucheng Huang is with the School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China (e-mail: schuang6@126.com).

Digital Object Identifier 10.1109/JBHI.2020.3001216

I. Introduction

ONLINE discussion forums, such as reddit, enable healthcare service providers to collect people/patient experience data. These forums are valuable sources of people's opinions, which can be examined for knowledge discovery and user behaviour analysis. In a typical sub-reddit forum, a user can use keywords and apply search tools to identify relevant questions/answers or comments sent in by other reddit users. Moreover, a registered user can create a topic or post a new questions to start discussions with other community members. Other users can reflect and share their views and experiences in response to each of the questions. In these online forums, people may express their positive and negative comments, or share questions, problems, and needs related to health issues. By analysing these comments, we can identify valuable recommendations for improving health-services and understanding the problems of users.

In late December 2019, the outbreak of a novel coronavirus causing COVID-19 was reported [1]. Due to the rapid spread of the virus, the World Health Organization declared a state of emergency. In this paper, we used automated extraction of COVID-19–related discussions from social media and a natural language process (NLP) method based on topic modeling to uncover various issues related to COVID-19 from public opinions. Moreover, we also investigate how to use LSTM recurrent neural network for sentiment classification of COVID-19 comments. Our findings shed light on the importance of using public opinions and suitable computational techniques to understand issues surrounding COVID-19 and to guide related decision-making. Our investigation was guided by the following specific research questions (RQ):

RQ1) How can important concepts in NLP methods such as topic modeling be applied in online discussions to uncover various issues related to COVID-19 from public opinions?

RQ2) How can we obtain the sentiment polarity of the COVID-19 comments posted by users reflecting their opinions?

RQ3) What is the comparative performance of various machine-learning algorithms for sentiment classification of COVID-19 online discussions, and which classification algorithm performs better?

To address the above questions, we focused on analysing COVID-19–related comments to detect sentiment and semantic ideas relating to COVID-19 based on the public opinions of people on reddit. Specifically, we used automated extraction of COVID-19-related discussions from social media and a natural language process (NLP) method based on topic modeling to uncover various issues related to COVID-19 from public opinions. The main contributions of this paper are as follows:

- We present a systematic framework based on NLP that is capable of extracting meaningful topics from COVID-19–related comments on reddit.
- We propose a deep learning model based on Long Short-Term Memory (LSTM) for sentiment classification of COVID-19–related comments, which produces better results compared with several other well-known machine-learning methods.
- We detect and uncover meaningful topics that are being discussed on COVID-19–related issues on reddit, as primary research.
- We calculate the polarity of the COVID-19 comments related to sentiment and opinion analysis from 10 sub-reddits.

Our findings shed light on the importance of using public opinions and suitable computational techniques to understand issues surrounding COVID-19 and to guide related decision-making. Overall, the paper is structured as follows. First, we provide a brief introduction to online healthcare forums. Discussion of COVID-19–related issues and some similar works are provided in section II. In section III, we describe the data pre-processing methods adopted in our research, and the NLP and deep-learning methods applied to the COVID-19 comments database. Next, we present the results and discussion. Finally, we conclude and discuss future works based on NLP approaches for analysing the online community in relation to the topic of COVID-19.

II. Related Work

Machine and deep-learning approaches based on sentiment and semantic analysis are popular methods of analysing text-content in online health forums. Many researchers have used these methods on social media such as Twitter, reddit [2]–[7], and health information websites [8], [9]. For example; Halder and colleagues [10] focused on exploring linguistic changes to analyse the emotional status of a user over time. They utilized a recurrent neural network (RNN) to investigate user-content in a huge dataset from the mental-health online forums of healthboards.com. McRoy and colleagues [11] investigated ways to automate identification of the information needs of breast cancer survivors based on user-posts of online health forums. Chakravorti and colleagues [12] extracted topics based on various health issues discussed in online forums by evaluating user posts of several subreddits (e.g., r/Depression, r/Anxiety) from 2012 to 2018. VanDam and colleagues [13] presented a classification approach for identifying clinic-related posts in online health communities. For that dataset, the authors collected 9576 thread-initiating posts from WebMD, which is a health information website.

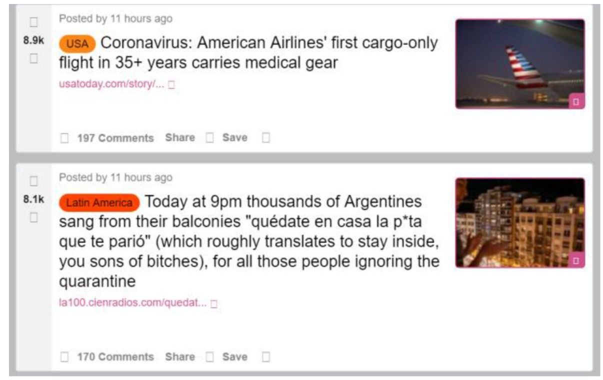


Fig. 1. Example of user-questions about “COVID-19”on reddit.

The COVID-19–related comments from an online healthcare-oriented group can be considered potentially useful for extracting meaningful topics to better understand the opinions and highlight discussions of people/users and improve health strategies. Although there are similar works regarding various health issues in online forums, to the best of our knowledge, this is the first study to utilize NLP methods to evaluate COVID-19–related comments from sub-reddit forums. We propose utilizing the NLP technique based on topic modeling algorithms to automatically extract meaningful topics and design a deep-learning model based on LSTM RNN for sentiment classification on COVID-19 comments and to understand the positive or negative opinions of people as they relate to COVID-19 issues to inform relevant decision-making.

III. Framework Methodology

This section clarifies the methods used to investigate the main contributions to this study, which proposes the use of an unsupervised topic model, with a collaborative deep-learning model based on LSTN RNN to analyse COVID-19–related comments from sub-reddits. The developed framework, shown inFig. 2, uses sentiment and semantic analysis for mining and opinion analysis of COVID-19–related comments.

A. Preparing the Input Data

Reddit is an American social media, a discussion website for various topics that includes web content ratings. In this social media, users are able to post questions and comments, and to respond to each other regarding different subjects, such as COVID-19. The posts are organised by subjects created by online users, called “sub-reddits”, which cover a variety of topics like news, science, healthcare, video, books, fitness, food, and image-sharing. This website is an ideal source for collecting health-related information about COVID-19–related issues. This paper focuses on COVID-19–related comments of 10 sub-reddits based on an existing dataset as the first step in producing this model.

B. Removing Noise and Stop-Words

One of the most important steps in pre-processing COVID-19–related comments is removing useless words/data, which are

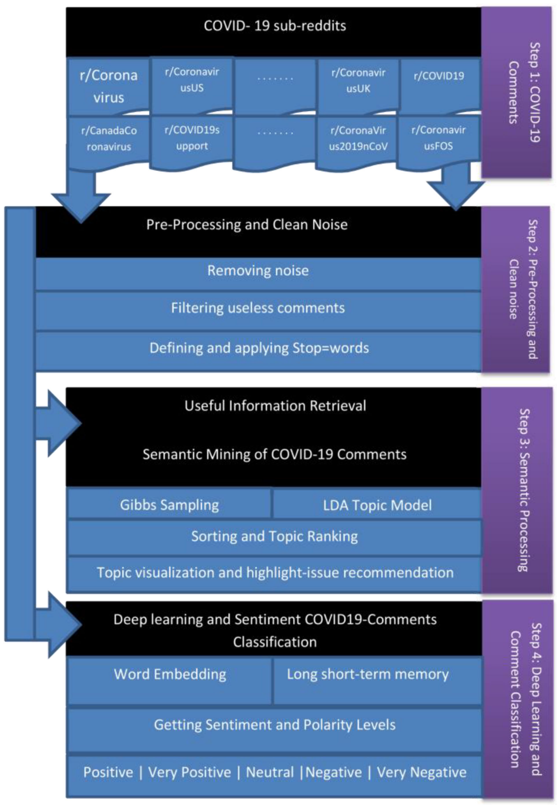


Fig. 2. An overview of the research framework utilized to obtain meaningful results from COVID-19–related comments.

defined as stop-words in NLP, from pure text. Moreover, we also decreased the dimensionality of the features space by eliminating stop-words. For example, the most common words in the text comments are words that are usually meaningless and do not effectively influence the output, such as articles, conjunctions, pronouns, and linking verbs. Some examples include: am, is, are, they, the, these, I, that, and, them.

C. Semantic Extraction and COVID-19 Comment Mining

Text-document modeling in NLP is a practical technique that represents an individual document and the set of text-documents based on terms appearing in the text-documents. Topic modeling is one type of document modeling approach to semantic extraction in natural language processing. Latent Dirichlet Allocation (LDA) [14] and Probabilistic Latent Semantic Analysis (PLSA) are popular methods of topic modeling. One of the main strengths of the LDA is that it has a rich internal structure and can use the probabilistic algorithm to train the model. LDA can have the effect of dimensionality reduction, suitable for large-scale corpus. LDA is a probabilistic model where each document in a corpus is described by a random mixture over hidden topics. Each of the hidden topics is described by a distribution over terms. The most important advantage of LDA against pLSI is that it considers that the text documents in a huge corpus have several hidden topics which, by-turn, are distributions over terms created in the documents of the huge corpus. Another benefit of LDA is that straightforward inference approaches can be supplied on

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

formerly unseen documents, [15]–[17]. In this section, the aim of implementing the LDA model is to extract semantic aspects.

For learning LDA, there are various methods, such as Variational Bayes and Gibbs Sampling [18], [19], which are two popular techniques based on approximate inference methods to estimate the parameters of the model. Most researchers, however, prefer to consider Gibbs sampling methods for learning LDA models because they are more efficient and simpler than the other methods [17], [18].

As a third step, we utilized topic modeling based on an LDA Topic model and Gibbs sampling [20] for semantic extraction and latent topic discovery of COVID-19–related comments. COVID-19 comments, however, can depend on various subjects that are discussed by reddit users. In this step, we can detect and discover these meaningful subjects or topics. Therefore, based on the LDA model, we considered a collection of documents, such as COVID-19–related comments and words, as topics (K), where the discrete topic distributions are drawn from a symmetric Dirichlet distribution. The probability of observed data D was computed and obtained from every COVID-19–related comment in a corpus using the following equation:

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

Determined α parameters of topic Dirichlet prior and also considered parameters of word Dirichlet prior as β. *M* is the number of text-documents, and *N* is the vocabulary size. Moreover, (α, θ) was determined for the corpus-level topic distributions with a pair of Dirichlet multinomials. (β, φ) was also determined for the topic-word distributions with a pair of Dirichlet multinomials. In addition, the document-level variables were defined as θ_{dd}, which may be sampled for each document. The word-level variables z_{dn}, w_{d_n}, were sampled in each text-document for each word [14].

Algorithm 2 describes a general process as part of our framework for extracting latent topics and semantic mining. The input data consists of the number of COVID-19–related comments as the context of the document: Line 1 processes the pure-data to eliminate noise and stop-words based on Algorithm 1. Lines

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: $d\theta$~*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**

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

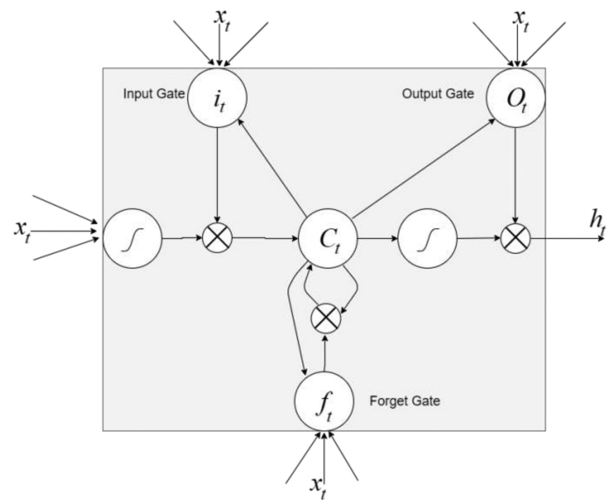


Fig. 3. The framework of a simple LSTM memory cell. Here, as shown, this structure includes three gates $(\mathrm{f_t}, \mathrm{i_t}, \mathrm{o_t})$, and a memory cell $(\mathrm{c_t})$.

2-5 compute the probability of the word distribution from Topic K[*i*]. Lines 6-11 compute the probability of the topic distribution from the COVID-19-Content-Document m [*i*]. As highlighted in Equation 1, the variables $\theta,\ w$ are computed for document-level and word-level of the framework. In more detail, the LDA handles topics as multinomial distributions in documents and words as a probabilistic mixture of a pre-determined number from latent topics. Lines 1-3 of Algorithm 3 show the semantic mining to extract the latent topics. We then used a sorting function to determine the recommended highlighted topics. Because the Gibbs sampling method is used in this step, the time requested for model inference can be specified as the sum of the time for inferring LDA. Therefore, the time complexity for LDA is O(N K), where N denotes the total size of the corpus (COVID-19–related comments) and K is the topic number.

D. Deep Learning and COVID-19–Sentiment Classification

Deep neural networks have been successfully employed for different types of machine-learning tasks, such as NLP-based methods utilizing sentiment aspects for deep classification [21]–[26]. Deep neural networks are able to model high-level abstractions and to decrease the dimensions by utilizing multiple processing layers based on complex structures or to be combined with non-linear transformations. RNNs are popular models with demonstrated importance and strength in most NLP works [27]–[29]. The purpose of RNNs is to use consecutive information, and the output is augmented by storing previous calculations. In fact, RNNs are equipped with a memory function that saves formerly calculated information. Basic RNNs, however, have some challenges due to gradient vanishing or exploding, and they are unable to learn long-term dependencies. LSTM [30], [31] units have the benefit of being able to avoid this challenge by adjusting the information in a cell state using 3 different gates.

The formula for each LSTM cell can be formalized as:

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

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

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

Where W, U, b are the parameters in the gates and the cell states. The forget (f_t), input (i_t), and output (o_t) gates for each LSTM cell are determined by these 3 equations, eqs. 2–4, respectively. Based on Figure 3, in an LSTM layer, the forget gate determines which previous information from the cell state is forgotten. The input gate controls or determines the new information that is saved in the memory cell. The output gate controls or determines the amount of information in the internal memory cell to be exposed. The cell-memory/input block equations are:

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

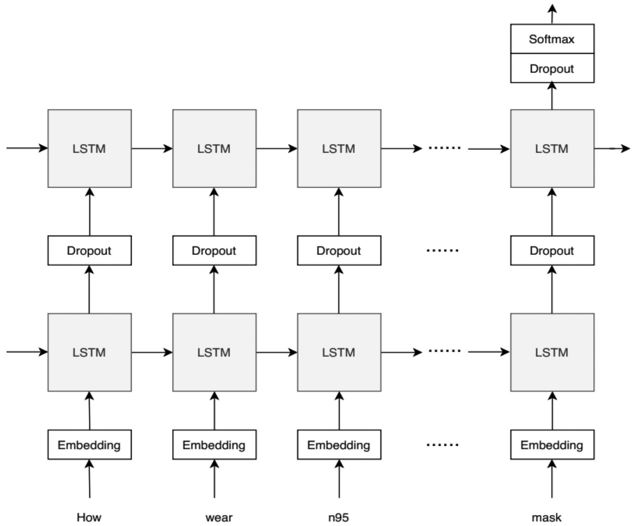


Fig. 4. Structure of the LSTM designed for COVID-19 sentiment classification.

$$\mathrm{C_t} = \mathrm{i_t} \odot \tilde{\mathrm{C}}_\mathrm{t} + \mathrm{f_t} \odot \mathrm{C_{t-1}} \tag{6}$$

$$\mathrm{h_t} = \mathrm{o_t} \odot \tanh\ (\mathrm{C_t}) \tag{7}$$

In which, $\mathrm{C_i}$ is the cell state, $\mathrm{h_t}$ is the hidden output, and x_t is an input vector. σ is sigmoid and $\odot$ is element-wise multiplication.

As the last step of this framework, an LSTM model was utilised to assess the COVID-19–related comments of online users who posted on reddit, in order to recognize the emotion/sentiment elicited from these comments. We designed two LSTM-layers and for pre-trained embeddings, considered the Glove-50 dimension, which were trained over a large corpus of COVID-19–related comments (Figure 4). The processed text from the COVID-19–related comments, however, is changed to vectors with a fixed dimension by converting pre-trained embeddings. Moreover, COVID-19 comments can also be described as a characters-sequence with its corresponding dimension creating a matrix [32].

IV. Experiment Details

In this section, we provide a detailed description of the data collection and experimental results followed by a comprehensive discussion of the results. We assessed 563,079 COVID-19–related comments from reddit. The dataset was collected between January 20, 2020 and March 19, 2020 (the full dataset is available at Kaggle.[1] We used MALLET[2] to implement the inference and capture the LDA topic model to retrieve latent topics. We used the Python library Keras[3] to implement our deep-learning model.

[1][Online]. Available: https://www.kaggle.com/khalidalharthi/coronavirus-posts-in-reddit-platform

[2][Online]. Available: http://mallet.cs.umass.edu/

[3][Online]. Available: https://pypi.org/project/Keras/

RQ1) How can important concepts in NLP methods such as topic modeling be applied in online discussions to uncover various issues related to COVID-19 from public opinions?

To address the first research question above, in this section, we discovered how we extracted meaningful topics based on semantic-comment-mining and topic modeling in different issues on COVID-19–related topics, as considered in steps of the proposed framework. According to Table I and Figures 5–9, the following observations were made: Topics 85 and 18 had a similar concept in “People/Infection”. Topic 85 included words referring to people, such as “people”, “virus”, “day”, “bad”, “stop”, “news”, “worse”, “sick”, “spread”, and “family”. This topic is the first ranked topic discovered from the generated latent topics, in which most users express their opinion and comment on this issue. Based on Table I and Figure 6(a)in this topic, the terms “people” and “virus” were the most highlighted words, with word-weights of 0.1295% and 0.0301%, respectively. Also, we can see the importance of the term “family” from this topic. In addition, Topic 18 contains the telling words “virus”, “people”, “symptoms”, “infection”, “cases”, “disease”, “pneumonia”, “coronavirus”, and “treatment”. Other revealing words in Topic 18 included “people”, “infection”, and “treatment”. These terms initially suggest a set of user comments about treatment issues. Moreover, the sentiment analysis of the terms suggests that negative words were more highlighted than positive words.

Topic 63 also addresses healthcare and hospital issues with the most frequent term being “hospital”. Words such as “hospital”, “medical”, “healthcare”, “patients”, “care”, and “city” were included. The terms “hospital”, “medical”, and “healthcare” were the most highlighted words, with word-weights of 0.0561%, 0.0282%, and 0.0278%, respectively. Other words worth mentioning that were seen for this topic were “person”, “patient“, “staff”, “workers”, and “emergency”. Topic 63 was assigned as medical staff issues. Topic 4 included words relating to money, such as ”pay“, ”money“, ”companies“, ”insurance“, ”paid“, ”free“, ”cost“, ”tax“, ”years“, and ”employees". Moreover, the sentiment analysis of the terms suggested that negative words were more highlighted than positive words.

Topic 30 covers user’s comments concerning issues related to “feelings and hopes” and highlight words such as “good”, “hope”, “feel”, “house”, “safe”, “hard”, “months”, “fine”, “live”, and “friend”. Moreover, sentiment analysis of terms suggested that positive words were more highlighted than negative words. Positive words such as “good”, “hope”, “safe”, “fine”, “kind”, and “friend”, thus pertain to the phenomenon of “positive feelings“. For Topic 93, we can see that there was a clear focus on ”people, age, and COVID issues" with the top words being “covid”, “young”, “risk”, “fever”, “immune”, “age”, “sick”, “cough”, “life”, “cold”, “elderly”, and “older”. The terms “covid”, “young”, and “risk” were the most highlighted words, with word-weights of 0.0299%, 0.0222%, and 0.0218%, respectively, and this topic had negative polarity.

Topic 48 also addresses “COVID-19 testing issues” and contains words like “people”, “testing”, “government”, “country”, “tested”, “test”, “infected”, “home”, “covid”, and “pandemic”.

TABLE I
TOP 10 TOPICS FROM COVID-19–RELATED COMMENTS ON REDDIT

Topic 85		Topic 69		Topic 8		Topic 18		Topic 48	
Rank 1		Rank 2		Rank 3		Rank 4		Rank 5	
Proportion: 12.7966		Proportion: 6.90415		Proportion: 5.72494		Proportion: 5.5769		Proportion: 5.28395	
people	sick	china	normal	people	free	virus	severe	people	countries
virus	great	population	found	die	times	people	risk	testing	care
day	spread	means	general	shit	happen	symptoms	source	governme	symptoms
bad	start	hard	clear	fuck	lives	infection	long	t country	tests
stop	person	years	takes	long	hate	cases	pretty	tested	spread
news	told	question	real	life	save	disease	infections	test	situation
worse	contact	place	start	care	governments	pneumonia	treatment	infected	south
days	family	comment	single	wrong	economy	case	viruses	home	social
big	spreading	kind	similar	money	dying	coronavirus	information	covid	shut
understand	coming	average	simply	fucking	imagine	infected	article	pandemi(	numbers
Topic 9		Topic 30		Topic 58		Topic 76		Topic63	
Rank 6		Rank 7		Rank 8		Rank 9		Rank 10	
Proportion: 5.03657		Proportion: 4.75303		Proportionn: 4.62488		Proportion: 4.41009		Proportion: 0.36916	
good	group	good	friend	home	taking	health	atter	hospital	patient
thinking	home	hope	wife	stay	open	idea	correct	medical	workers
working	worried	eel	healthy	health	public	medical	thread	hospitals	staff
stuff	month	house	times	italy	day	months	science	healthcare	case
bit	expect	started	kind	today	face	wrong	kids	patients	cities
happen	support	safe	hit	cases	yesterday	true	result	care	sick
small	side	fine	doctor	weeks	food	positive	majority	public	room
works	heard	hard	person	risk	confirmed	travel	effective	city	beds
experience	chance	months	coming	days	social	edit	scale	health	states
future	bring	ive	starting	hope	pretty	disease	specifically	person	emergency

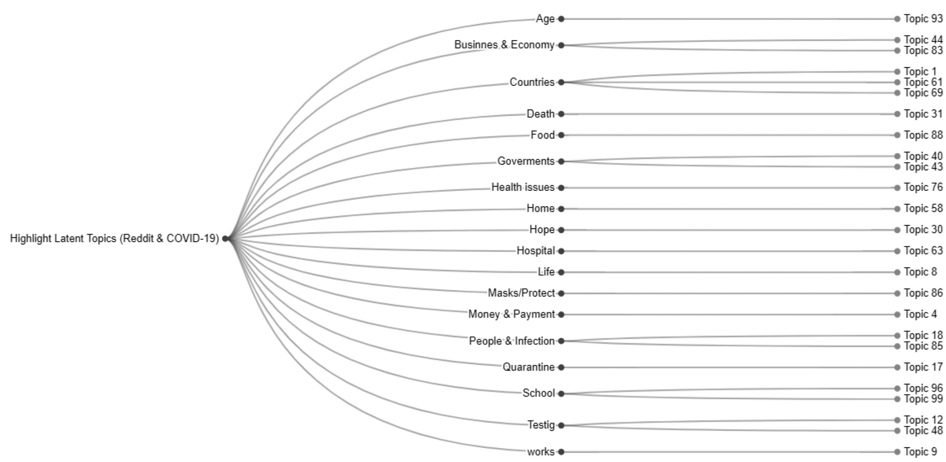


Fig. 5. Cluster dendrogram of highlight latent topics generated in a COVID-19–related discussion

Based on the results, the terms “people” and “testing” were the most highlighted words with word weights of 0.0447% and 0.0337%, respectively. Moreover, the opinion words based on sentiment analysis scored high in negative polarity for Topic 17. The top terms of this topic were “coronavirus“, ”quarantine“, ”stupid“, ”happening“, ”shit“, ”watch“, and ”dangerous“, thus pertaining to the phenomenon ”quarantine issues“. The terms ”coronavirus" and “quarantine” were the most highlighted words, with word-weights of 0.0353% and 0.0346%, respectively.

A. Sentiment and Polarity Results

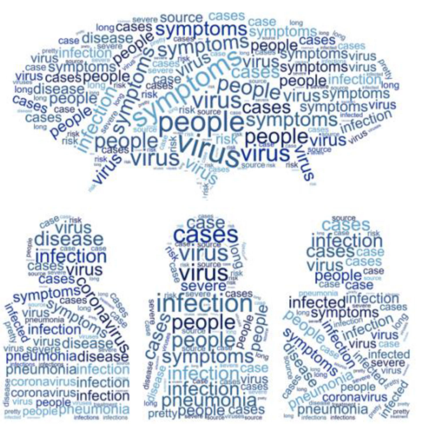
Sentiment analysis is a practical technique in NLP for opinion mining that can be used to classify text/comments based on word polarities [33]–[35]. This technique has many applications in various disciplines, such as opinion mining in online healthcare communities [36]–[38].

RQ2) How can we obtain the sentiment polarity of the COVID-19 comments posted by users reflecting their opinions?

To address the second important question, we obtained the sentiment of the COVID-19–related comments using the SentiStrength algorithm [39]–[41]. However, SentiStrength is a free sentiment analysis method with 2310 sentiment words and word stems obtained from the Linguistic Inquiry to classify social web texts. An example is shown to determine the sentiment scores of the COVID-19 comments by SentiStrength in Table II. SentiStrength includes a number of rules [39], which we used in this research to cope with special cases for sentiment analysis. The following rules are incorporated into SentiStrength:

– If there are repetitive letters in a term, it is determined as a strength boost sentiment word and the score is increased by 1. For example, ‘haaaappy’ is more positive than ‘happy’. Moreover, neutral words are determined to have a positive sentiment strength of 2.

(a) Topic Word 85



(b) Topic Word 18

Fig. 6. Word cloud visualisation based on the word-weight of the topics..

(a) Topic Word 63

(b) Topic Word 4

Fig. 7. Word cloud visualisation based on the word-weight of the topics.

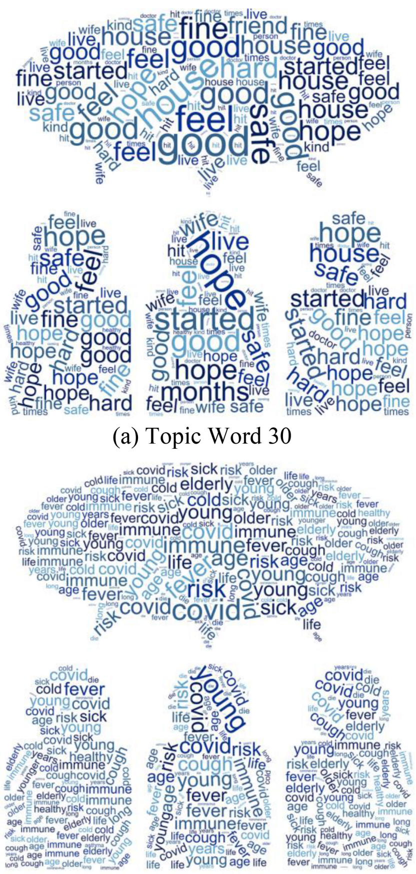
– A list of negative words is considered to neutralize the sentiment words. For example; “I do not hate him” is not classified as a negative sentiment.
– The term “miss” is a special word with a negative strength of -2 and a positive strength of 2. It is frequently considered to state love and sadness at the same time, as in the common phrase, “I miss you”.
– A list of idioms is considered to identify the emotions of a few common phrases, which helps to override a particular emotional word strength. The idiom list is updated with phrases that show word senses for common sentiment words. For example, ’wuts good’.
– A list of booster words that are considered to weaken or strengthen the sentiment of the words. For example; the term ‘very’ increases the positive strength of the score by $+1$.
– A list of emoticon words with polarities considered to determine additional sentiment. For example, ‘(^ ^)’ is positive, and also ’)-: ’ is negative.

Therefore, with all COVID-19–related comments tagged with sentiment scores, we calculated the average sentiment of the entire dataset along with comments mentioning only 10 COVID-19 sub-reddits. The main objective of this analysis was to identify the overall sentiment of the COVID-19–related comments. We calculated the average sentiment of all comments as negative, positive, or neutral. Figure 10 shows the sentiment of all comments in the database along with the average sentiment of comments containing the terms COVID-19. For each of the polar comments in our labelled dataset, we assigned negative and positive scores utilizing SentiStrength, and employed the various scores directly as rules for building inference about the polarity/sentiment of the COVID-19 comments.

Based on SentiStrength, we determined that a comment was positive if the positive sentiment score was greater than the negative sentiment score, and also considered a similar rule for determining a positive sentiment. For example, a score of $+5$ and -4 indicates positive polarity and a score of $+4$ and -6 indicates negative polarity. Moreover, If the sentiment scores were equal (such as -1 and $+1$, $+4$ and -4), we determined that the comment was neutral.

B. Deep Classification and Feature Analysis

RQ3) What is the comparative performance of various machine-learning algorithms for sentiment classification of COVID-19 online discussions, and which classification algorithm performs better?



(a) Topic Word 30

(b) Topic Word 93

Fig. 8. Word cloud visualisation based on the word-weight of the topics.

TABLE II
EXAMPLES OF COVID-19 COMMENTS FROM THE REDDIT CORPUS

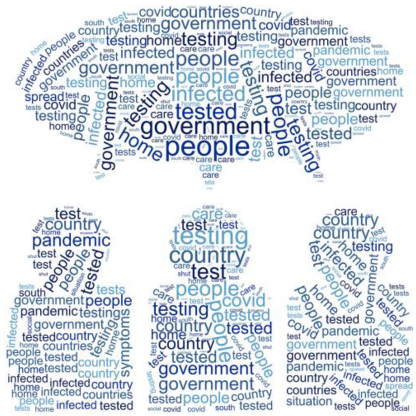
Polarity	People's Comment	Score of the words
Positive	I hope loved ones remain safe healthy.	I[0] hope[2] loved[3] ones[0] remain[0] safe[1] healthy[0]
	Ah yes manbaby magnificent immune system. better luck next time covid-19	Ah[0] yes[0] manbaby[0] magnificent[3] immune[0] system[0] better[0] luck[2] 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]

To address our third important research question, we prepare the dataset to automatically classify the sentiment of the COVID-19 comments for all of the data, we labelled each of the comments as very positive, positive, very negative, negative, and neutral based on the sentiment score obtained using the Sentistrength method. The training set had 338,666 COVID-19–related comments and the testing set had 112,888 comments. In this experiment, we evaluated the proposed LSTM-model and also supervised machine-learning methods using the Support Vector Machine (SVM), Naive Bayes, Logistic Regression, K Nearest Neighbors (KNN) techniques. To deploy the ML algorithms to sentiment classification, however, we utilized the Scikit-learn package,[4] which is a Python library supporting many machine-learning methods for Python language. We selected these methods for COVID-19 sentiment classification because their high accuracy and effectiveness are well-known for sentiment classification in natural language processing [42]. Figure 11 shows the accuracy of the best model for classifying a COVID-19 comment as either a very positive, positive, very negative, negative, or neutral sentiment. Our approach based on the LSTM model, which classified all COVID-19 comments in the majority class achieved 81.15% accuracy, which was higher than that of traditional machine-learning algorithms. We believe that the sentiment and semantic techniques can provide meaningful results with an overview of how users/people feel about the disaster.

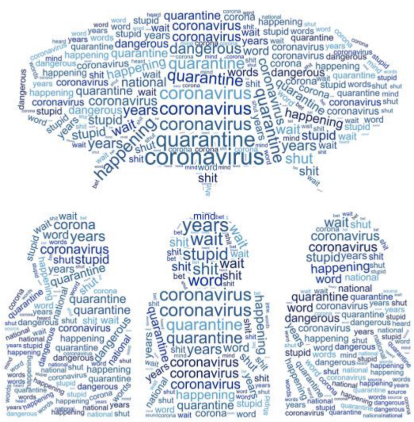
C. Discussion and Practical Findings

Analysing social media comments on platforms such as reddit could provide meaningful information for understanding people's opinions, which might be difficult to achieve through traditional techniques, such as manual methods. The text content on reddit has been analysed in various studies [43]–[45]; to the best of our knowledge, this is the first study to analyse comments by considering semantic and sentiment aspects of COVID-related comments from reddit for online health communities. In this research, we investigated three important research questions and proposed a systematic framework that appropriately addresses the questions. To answer RQ1, we considered an existing dataset that included 563,079 comments from 10 sub-reddits. We found and detected meaningful latent topics of terms about COVID-19 comments related to various issues. Thus, user comments proved to be a valuable source of information, as shown in Tables I and Figures 5–9. A variety of different visualisations was used to interpret the generated LDA results. As mentioned, LDA is a probabilistic model that, when applied to documents, hypothesises that each document from a collection has been generated as a mixture of unobserved (latent) topics, where a topic is defined as a categorical distribution over words. Regarding the top-ranked topics for the COVID-19 comments, it is possible to recognise many words probably related to needs and highlight discussions of the people or users on reddit. In summary, to address RQ2 and RQ3, we obtained the polarity words for each comment (Fig 10) and designed a two-layer LSTM to detect meaningful latent-topics and sentiment-comment-classification on COVID-19–related issues from healthcare forums on reddit. We demonstrated that our deep-learning model based on LSTM produces better results than several other well-known machine-learning methods for sentiment classification (Fig 11).

[4][Online]. Available: https://scikit-learn.org/



(a) Topic Word 48



(b) Topic Word 17

Fig. 9. Word cloud visualisation based on the word-weight of the topics.

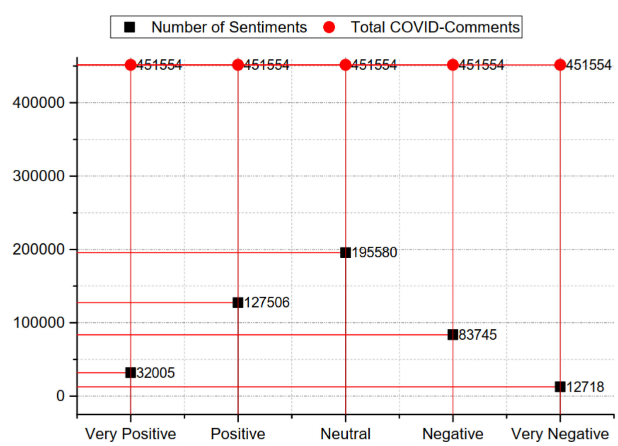


Fig. 10. Distribution of COVID-19 comments with positive, negative or neutral sentiments of reddit Data

This research was limited to English-language text, which was considered a selection criterion. Therefore, the results do not reflect comments made in other languages. In addition, this study was limited to comments retrieved from January 20, 2020 and March 19, 2020. Therefore, the gap between the period in which the research was being completed and the time-frame of our study may have somewhat affected the timeliness of our results.

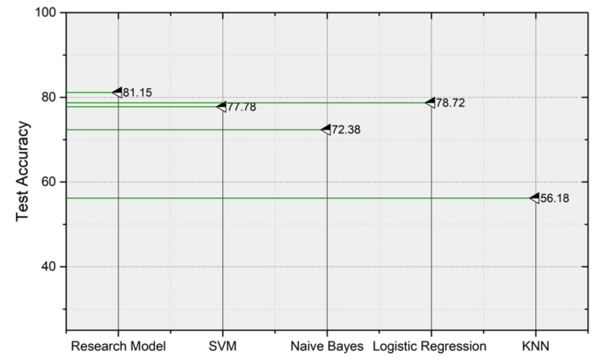


Fig. 11. Accuracy performance of the methods for COVID-19 sentiment-classification using various features.

Overall, the study suggests that the systematic framework by combining NLP and deep-learning methods based on topic modelling and an LSTM model enabled us to generate some valuable information from COVID-19–related comments. These kinds of statistical contributions can be useful for determining the positive and negative actions of an online community, and to collect user opinions to help researchers and clinicians better understand the behaviour of people in a critical situation. Regarding future work, we plan to evaluate other social media, such as Twitter, using hybrid fuzzy deep-learning techniques [46], [47] that can be used in the future for sentiment level classification as a novel method of retrieving meaningful latent topics from public comments.

V. Conclusion

To our knowledge, this is the first study to analyse the association between COVID-19 comments' sentiment and semantic topics on reddit. The main goal of this paper, however, was to show a novel application for NLP based on an LSTM model to detect meaningful latent-topics and sentiment-comment-classification on COVID-19–related issues from healthcare forums, such as sub-reddits. We believe that the results of this paper will aid in understanding the concerns and needs of people with respect to COVID-19–related issues. Moreover, our findings may aid in improving practical strategies for public health services and interventions related to COVID-19.

Acknowledgment

The autors would like to thank SciTechEdit International, LLC (Highlands Ranch, CO, USA) for providing pro bono professional English-language editing of this article.

References

[1] M. Malta, A. W. Rimoin, and S. A. Strathdee, "The coronavirus 2019-nCoV epidemic: Is hindsight 20/20?," *EClinical Medicine*, 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 years by patients and health care professionals," *Adv. Radiat. Oncol.*, vol. 4, no. 2, pp. 345–353, 2019.

[3] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Gener. Comput. Syst.*, vol. 106, pp. 92–104, 2020.

[4] J. M. Barros, P. Buitelaar, J. Duggan, and D. 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] J. Rong, S. Michalska, S. Subramani, J. Du, and H. Wang, "Deep learning for pollen allergy surveillance from twitter in Australia," *BMC Med. Inform. Decis. Making*, vol. 19, no. 1, pp. 228–221, 2019.

[7] E. Batbaatar, and K. H. Ryu, "Ontology-Based healthcare named entity recognition from twitter messages using a recurrent neural network approach," *Int. J. Environ. Res. Public Health*, vol. 16, no. 19, 2019, Art. no. 3628.

[8] H. Naderi, S. Madani, B. Kiani, and K. Etminani, "Similarity of medical concepts in question and answering of health communities," *Health Inf. J.*, to be published, doi: 10.1177/1460458219881333.

[9] V. V. Vydiswaran and M. Reddy, "Identifying peer experts in online health forums," *BMC Med. Inf. Decis. Making*, vol. 19, no. 3, pp. 42–114, 2019.

[10] K. Halder, "Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach," in *Proc. 8th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2017, pp. 127–135.

[11] S. McRoy, M. Rastegar-Mojarad, Y. Wang, K. J. Ruddy, T. C. Haddad, and H. Liu, "Assessing unmet information needs of breast cancer survivors: Exploratory study of online health forums using text classification and retrieval," *JMIR Cancer*, vol. 4, no. 1, pp. 1–15, 2018.

[12] D. Chakravorti, "Detecting and characterizing trends in online mental health discussions," in *IEEE Int. Conf. Data Mining Workshops,* 2018, pp. 697–706.

[13] C. VanDam, S. Kanthawala, W. Pratt, J. Chai, and J. Huh, "Detecting clinically related content in online patient posts," *J. Biomed. Inform.*, vol. 75, pp. 96–106, 2017.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[15] S. Doumit and A. Minai, "Semantic knowledge inference from online news media using an LDA-NLP approach," in *Int. Joint Conf. Neural Netw.*, Jul. 2011, pp. 3068–3071.

[16] S. Kim, and J. Yoon, "Link-topic model for biomedical abbreviation disambiguation," *J. Biomed. Inform.*, vol. 53, pp. 367–380, 2015.

[17] H. Jelodar, Y. Wang, and C. Yuan, "Latent dirichlet allocation (LDA) and Topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019.

[18] A. Plangprasopchok and K. Lerman, "Modeling social annotation: A bayesian approach," *Assoc. Comput. Machinery Trans. Knowl. Discovery Data*, vol. 5, no. 1, pp. 1–32, 2010.

[19] S. Gao, X. Li, Z. Yu, Y. Qin, and Y. Zhang, "Combining paper cooperative network and topic model for expert topic analysis and extraction," *Neurocomputing*, vol. 257, pp. 136–143, 2017.

[20] D. Mimno, H. Wallach, and A. McCallum, "Gibbs sampling for logistic normal topic models with graph-based priors," in *Proc. NIPS Workshop Analyzing Graphs*, Dec. 2008, vol. 61, pp. 1–8.

[21] B. Alshemali, "Improving the reliability of deep neural networks in NLP: A review," *Knowl.-Based Syst.*, vol. 191, 2020, Art. no. 105210.

[22] M. Giménez, J. Palanca, and V. Botti, "Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis," *Neurocomputing*, vol. 378, pp. 315–323, 2020.

[23] J. Guo, H. He, T. He, L. Lausen, and M. A. Li, "Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing," *J. Mach. Learning Res.*, vol. 21, no. 23, pp. 1–7, 2020.

[24] H. J. Park, M. Song, and K. S. Shin, "Deep learning models and datasets for aspect term sentiment classification: Implementing holistic recurrent attention on target-dependent memories," *Knowl.-Based Syst.*, vol. 187, 2020, Art. no. 104825.

[25] L. Abualigah, H. E. Alfar, M. Shehab, and A. M. A. Hussein, "Sentiment analysis in healthcare: A brief review," in *Recent Advances in NLP: The Case of Arabic Language*. Berlin, Germany: Springer, 2020, pp. 129–141.

[26] A. Balamurali and B. Ananthanarayanan, "Develop a neural model to score bigram of words using bag-of-words model for sentiment analysis," in *Neural Networks for Natural Language Processing*, Pennsylvania, United States: IGI Global, 2020, pp. 122–142.

[27] I. J. Unanue, E. Z. Borzeshi, and M. Piccardi, "Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition," *J. Biomed. Inform.*, vol. 76, pp. 102–109, 2017.

[28] Y. Luo, "Recurrent neural networks for classifying relations in clinical notes," *J. Biomed. Inform.*, vol. 72, pp. 85–95, 2017.

[29] J. Huang and Y. Feng "Optimization of recurrent neural networks on natural language processing," in *Proc. 8th Int. Conf. Comput. Pattern Recognit.*, 2019, pp. 39–45.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE Int. Conf. Acoustics, Speech Signal Process.*, Apr. 2015, pp. 4580–4584.

[32] H. Meisheri, K. Ranjan, and L. Dey, "Sentiment extraction from Consumer-generated noisy short texts," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2017, pp. 399–406.

[33] A. Rajput, "Natural language processing, sentiment analysis, and clinical analytics," in *Innovation in Health Informatics*.. Cambridge, MA, USA: Academic Press, 2020, pp. 79–97.

[34] T. Sharma and A. Bajaj, "Deep learning approaches for textual sentiment analysis," in *Handbook of Research on Emerging Trends and Applications of Machine Learning*. Pennsylvania, United States: IGI Global, 2020, pp. 171–182

[35] 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.

[36] I. Marin, N. Goga, and A. Doncescu, "[WiP] Sentiment analysis electronic healthcare system based on heart rate monitoring smart bracelet," in *Proc. IEEE 11th Conf. Service-Oriented Comput. Appl.*, 2018, pp. 99–104.

[37] C. C. Yang, and L. Jiang, "Enriching user experience in online health communities through thread recommendations and heterogeneous information network mining," *IEEE Trans. Computat. Soc. Syst.*, vol. 5, no. 4, pp. 1049–1060, Dec. 2018.

[38] L. Goeuriot *et al.*, "Sentiment lexicons for health-related opinion mining," in *Proc. 2nd ACM SIGHIT Int. Health Inform. Symp.*, Jan. 2012, pp. 219–226.

[39] M. Thelwall, "The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength," in *Cyberemotions*. Berlin, Germany: Springer, pp. 119–134.

[40] M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inform. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.

[41] M. Thelwall, "Topic-based sentiment analysis for the Social Web: The role of mood and issue-related words," *J. Amer. Soc. Inform. Sci. Technol.*, vol. 64, no. 8, pp. 1608–1617, 2013.

[42] 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.

[43] E. Okon, V. Rachakonda, H. J. Hong, C. Callison-Burch, and J. Lipoff, "Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics," *J. Amer. Acad. Dermatol.*, 2019, to be published.

[44] 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.

[45] S. Pandrekar and X. Chen, "Social media based analysis of opioid epidemic using Reddit," in *Proc. Amer. Med. Inform. Assoc. Annu. Symp.*, 2018, vol. 2018, pp. 867–876.

[46] S. Zhou, Q. Chen, and X. Wang, "Fuzzy deep belief networks for semi-supervised sentiment classification," *Neurocomputing*, vol. 131, pp. 312–322, 2014.

[47] B. Ramasamy and A. Z. Hameed, "Classification of healthcare data using hybridised fuzzy and convolutional neural network," *Healthcare Technol. Lett.*, vol. 6, no. 3, pp. 59–63, 2019.