manuscript No.
(will be inserted by the editor)

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

Hamed Jelodar [1] · **Yongli Wang** [1] · **Rita Orji**[2] · **Hucheng Huang**[3]

Received: date / Accepted: date

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.

Hamed Jelodar
jelodar@njust.edu.cn

Yongli Wang
yongliwang@njust.edu.cn

Rita Orji
rita.orji@dal.ca

Hucheng Huang
schuang6@126.com

[1] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China
[2] Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada
[3] School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China

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

1 Introduction

Online 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 question to start discussions with other community members. In answering the questions, users reflect and share their views and experiences. 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 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 is provided in section 2. In section 3, 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.

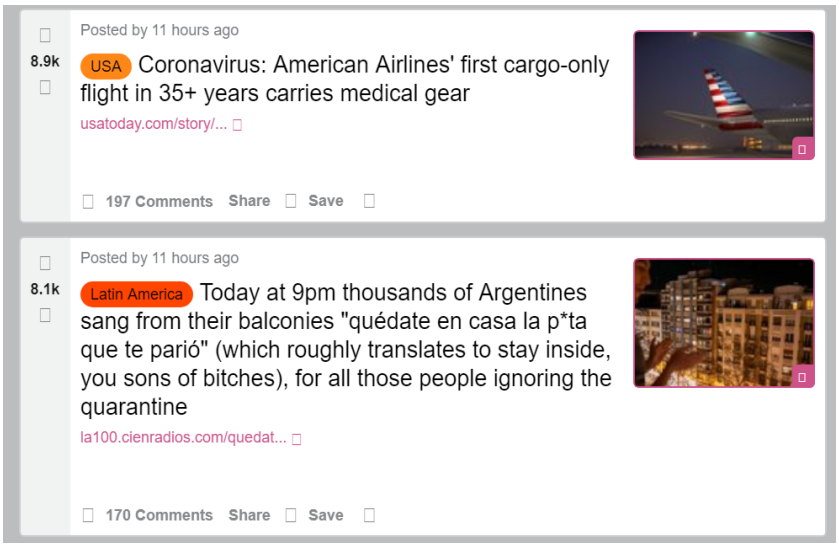

Fig. 1: Example of user-questions about ”COVID-19”on reddit

2 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.

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 classi-

fication 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.

3 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 in Fig. 2, uses sentiment and semantic analysis for mining and opinion analysis of COVID-19–related comments.

3.1 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.

3.2 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 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.

3.3 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 based on Latent Dirichlet Allocation (LDA) [14] is one type of document modelling approach. As a third step, we utilized topic modeling based on an LDA Topic model and Gibbs sampling [15] 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

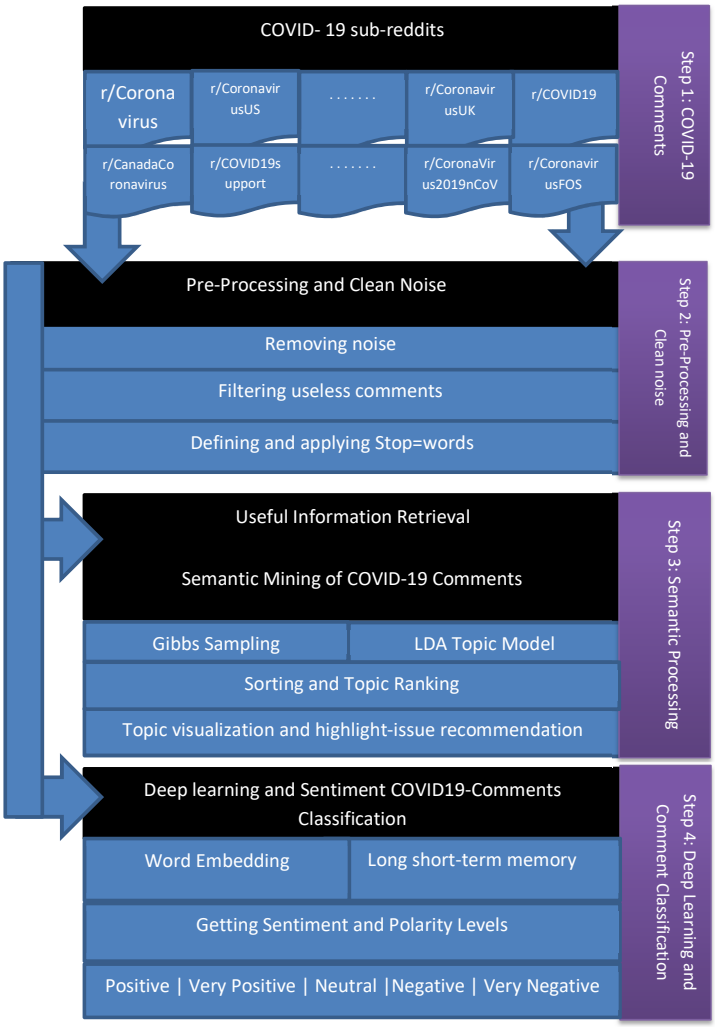

Fig. 2: An overview of the research framework for obtaining meaningful results of COVID-19 comments

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:

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

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_d, 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 1 Pre-processing and removing the noise to prepare the input data

Input : A bunch of COVID-19 comments as main document context
Output : A bunch of text document in string.

```
d_i= Get_data(); getting COVID-19 comments as pure data.
for d_i.row (all_record) != last_record do
    d_i_2= d_i.cleanData(d_i); removing stop-words, clean noise
    d_i_2=d_i_2.arranged(); processing to arrange dataset.
end for
return d_i_2 as a string
```

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

```
Pre-process and removing noise and clean data by Algorithm 1.
for each topic k ∈ {1, 2, . . . , k} do
    word-probability under the topic of sampling —— or the word distribution for topic k among
    COVID-19–related comments
    ϕ ∼ Dirichlet(β)
end for
for each COVID-19–related comments d ∈ {1, . . . , D} do
    The topic distribution for document m
    dθ ∼ Dirichlet(α)
    for Per word in COVID-19–related content-document d do
        sampling the distribution of topics in the COVID-19–related comments-documents to obtain the
        topic of the word:
        .Z_d ∼ Mul(θ)
        word-sampling undert the topic, W_d ∼ Mul(ϕ)
    end for
end for
```

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 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 θ_m, w_n 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.

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

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

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

3.4 Deep-Learning and 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 [16] - [21]. 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 [22] - [24]. 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 [25], [26] 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:

$$f_t = \sigma(W_{fz}z_{t-1} + W_{fx}x_t + b_f) \tag{2}$$

$$i_t = \sigma(W_{iz}z_{t-1} + W_{ix}x_t + b_i) \tag{3}$$

$$o_t = \sigma(W_{oz}z_{t-1} + W_{ox}x_t + b_o) \tag{4}$$

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. 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:

$$\widetilde{C}_t = \phi(W_{cz}z_{t-1} + W_{cx}x_t + b_c) \tag{5}$$

$$C_t = i_t \odot \widetilde{C}_t + f_t \odot C_{t-1} \tag{6}$$

$$z_t = o_t \odot \phi(C_t) \tag{7}$$

In which, C_i is the cell state, z_t is the hidden output, and x_t is an input vector. W and b are the weight matrix and the bias term respectively. σ is sigmoid and ϕ is tanh. $\odot$ 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 [1], which were trained over a large corpus of COVID-19–related comments (Figure 3). 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 [27].

4 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 website[2]). We used MALLET[3] to implement the inference and capture the LDA topic model to retrieve latent topics. We used the Python library Keras [4] to implement our deep-learning model.

[1] https://www.kaggle.com/watts2/glove6b50dtxt
[2] https://www.kaggle.com/khalidalharthi/coronavirus-posts-in-reddit-platform
[3] http://mallet.cs.umass.edu/
[4] https://pypi.org/project/Keras/

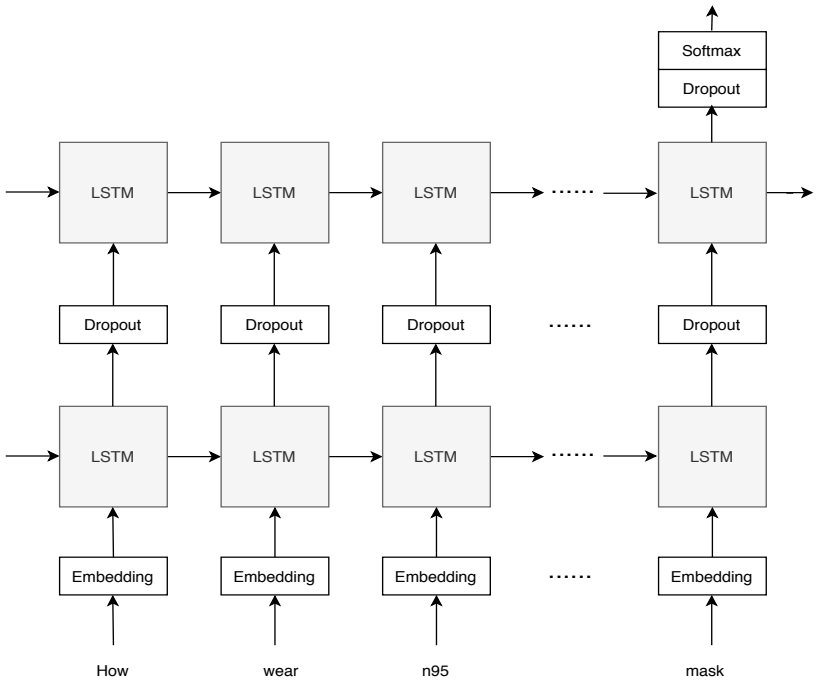

Fig. 3: Structure of the LSTM designed for COVID-19 sentiment classification.

Table 1: 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	
Propration : 12.7966		Propration : 6.90415		Propration : 5.72494		Propration : 5.5769		Propration : 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	government	symptoms
bad	start	hard	clear	fuck	lives	infection	long	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 numbers
understand	coming	average	simply	fucking	imagine	infected	article	pandemic	
Topic 9		Topic 30		Topic 58		Topic 76		Topic 63	
Rank 6		Rank 7		Rank 8		Rank 9		Rank 10	
Propration : 5.03657		Propration : 4.75303		Propration : 4.62488		Propration : 4.41009		Propration : 0.36916	
good		good	friend	home	taking	health	matter	hospital	patient
thinking	group	hope	wife	stay	open	idea	correct	medical	workers
working	home	feel	healthy	health	public	medical	thread	hospitals	staff
stuff	worried	house	times	italy	day	months	science	healthcare	case
bit	month	started	kind	today	face	wrong	kids	patients	cities
happen	expect	safe	hit	cases	yesterday	true	result	care	sick
small	support	fine	doctor	weeks	food	positive	majority	public	room
works	side	hard	person	risk	confirmed	travel	effective	city	beds
experience	heard chance bring	months	coming starting	days	social	edit	scale	health	states
future		live		hope	pretty	disease	specifically	person	emergency

Table 2: Topics ranking 11 to 25 from COVID-19–related comments on reddit.

Topic 31		Topic 17		Topic 61		Topic 1		Topic 96	
Rank 11		Rank 12		Rank 13		Rank 14		Rank 15	
Propration : 3.81556		Propration : 3.53217		Propration : 3.34602		Propration : 3.15658		Propration : 2.92918	
cases	mortality	coronavirus	shut	virus	test	pretty	food	weeks	yesterday
rate	weeks	quarantine	corona	flu	confirmed	shit	literally	measures	action
flu	confirmed	years	national	country	control	america	month	stop	economy
deaths	population	wait	words	china	spread	american	guys	panic	spread
numbers	wuhan	stupid	mind	countries	singapore	stay	real	school	absolutely
death	good	happening	bet	sars	human	love	live	link	closed
days	patients	shit	heard	pandemic	epidemic	reddit	life	close	country
china	reported	word	source	chinese	global	americans	full	early	expect
case	fatality	watch	sound	infected	remember	usa	realize	public	moment
infected	spread	dangerous	common	bad	subreddit	death	dude	lockdown	fine
Topic 93		Topic 4		Topic 40		Topic 43		Topic 83	
Rank 16		Rank 17		Rank 18		Rank 19		Rank 20	
Propration : 2.22481		Propration : 2.1929		Propration : 2.09387		Propration : 1.91949		Propration : 1.73082	
covid	elderly	pay	tax	fucking	lmao	trump	states	home	stores
young	older	money	years	fuck	supposed	cdc	corona	sick	essential
risk	years	companies	employees	government	treatment	president	federal	working	company
fever	healthy	company	leave	guy	ass	government	pandemic	workers	businesses
immune	long	insurance	bill	man	tested	covid	united	store	told
age	die	paid	taxes	stupid	dumb	country	pence	stay	grocery
sick	younger	paying	healthcare	fucked	sell	coronavirus	americans	business	jobs
cough	coronavirus	free	job	sick	canada	administration	hoax	employees	paid
life	asthma	cost	business	rate	damn	response	national	job	restaurants
cold	conditions	afford	income	takes	idiots	america	million	weeks	customers
Topic 86		Topic 12		Topic 44		Topic 99		Topic 88	
Rank 21		Rank 22		Rank 23		Rank 24		Rank 25	
Propration : 1.59421		Propration : 1.54827		Propration : 1.53468		Propration : 1.48268		Propration : 1.44486	
masks	medical	test	market	market	war	school	closed	food	store
mask	sick	testing	confirmed	stock	global	kids	weeks	buy	days
wearing	public	tested	numbers	economy	buying	schools	care	water	stuff
wear	supply	tests	free	buy	parts	flu	husband	stock	supply
face	shortage	symptoms	negative	years	chain	home	travel	buying	panic
surgical	hands	positive	cases	supply	stocks	parents	students	supplies	stocked
protect	prevent	flu	day	economic	manufacturing	family	cdc	rice	family
protection	workers	cdc	doctor	deaths	goods	close	closing	bought	canned
effective	gloves	kits	cough	production	panic	sick	told	weeks	prepping
infected	doctors	fever	labs	markets	run	children	stay	eat	prepared

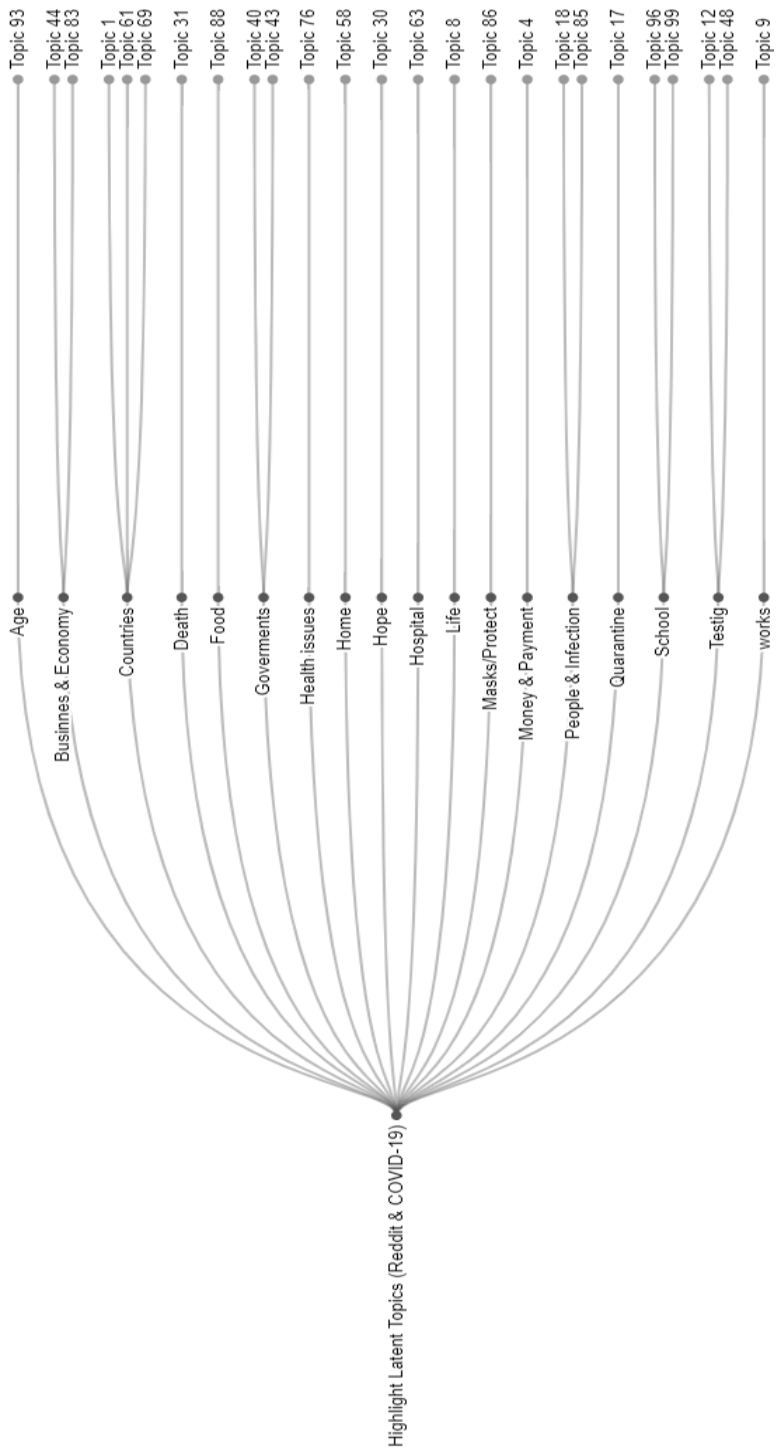

Fig. 4: Cluster dendrogram of highlight latent topics generated in a COVID-19–related discussion

(a) Topic Word 85

(b) Topic Word 18

Fig. 5: Word cloud visualisation based on the word-weight of the topics.

According to Table 1 and 2 and Figures 4-8, 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 1 and Figure 5 (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 suggest that negative words were more highlighted than positive words.

(a) Topic Word 63

(b) Topic Word 4

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

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

(a) Topic Word 30

(b) Topic Word 93

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

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”. 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”,

(a) Topic Word 48

(b) Topic Word 17

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

”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.

4.1 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 [28] - [30]. This technique has many applications in various disciplines, such as opinion mining in online healthcare communities [31] - [33]. We obtained the sentiment of the COVID-19–related comments using the SentiStrength algorithm [34] - [36]. 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

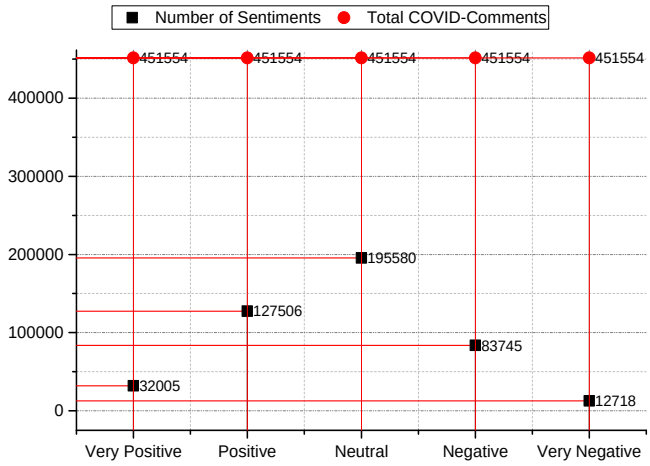

Fig. 9: Distribution of COVID-19 comments with positive, negative or neutral sentiments of reddit Data

of the COVID-19–related comments. We calculated the average sentiment of all comments as negative, positive, or neutral. Figure 9 shows the sentiment of all comments in the database along with the average sentiment of comments containing the terms COVID-19.

Table 3: 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][+1 MultiplePositiveWords] ones[0] remain[0] safe[1] healthy[0] [[Sentence=-1,5=word max, 1-5]][[[5,-1 max of sentences]]]
	Ah yes manbaby magnificent immune system. better luck next time covid-19	Ah[0] yes[0] manbaby[0] magnificent[3] immune[0] system[0] [[Sentence=-1,4=word max, 1-5]] better[0] luck[2] next[0] time[0] covid[0] 19[0] [[Sentence=-1,3=word max, 1-5]][[[4,-1 max of sentences]]]
	I really hope whole eu follows italy. they shut everything pandemic .	I[0] really[0] hope[2][1LastWordBoosterStrength] whole[0] eu[0] follows[0] italy[0] [[Sentence=-1,4=word max, 1-5]] they[0] shut[0] everything[0] pandemic[0] [[Sentence=-1,1=word max, 1-5]][[[4,-1 max of sentences]]]
Negative	Greed prejudice racism hate kill faster covid-19 .	greed[-2] prejudice[-2][-1 MultiplePositiveWords] racism[-1] hate[-3] kill[-1] faster[0] covid[0] 19[0] [[Sentence=-4,1=word max, 1-5]][[[1,-4 max of sentences]]]
	" Deeply concerned by inaction over the virus "- because you refused to identify this as a pandemic fucking fucks .	deeply[0] concerned[-1] by[0] inaction[0] over[0] the[0] virus[0] because[0] you[0] refused[-1] to[0] identify[0] this[0] as[0] a[0] pandemic[0] fucking[0] fucks[-2][-2 LastWordBoosterStrength] [[Sentence=-5,1=word max, 1-5]][[[1,-5 max of sentences]]]
	So much bullshit one thread alone. scary times .	so[0] much[0] bullshit[-2] one[0] thread[0] alone[0] [[Sentence=-3,1=word max, 1-5]] scary[-3] times[0] [[Sentence=-4,1=word max, 1-5]][[[1,-4 max of sentences]]]
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] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
	Everyone wear mask case unintentionally spreading everyone .	everyone[0] wear[0] mask[0] case[0] unintentionally[0] spreading[0] everyone[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
	Would kind enough link one studies stories ?	would[0] kind[1][-1 LastWordBoosterStrength] enough[0] link[0] one[0] studies[0] stories[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]

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.

4.2 Deep classification and Feature Analysis

To 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 (Senti-ML1), Naive Bayes (Senti-ML2), Logistic Regression (Senti-ML3), K Nearest Neighbors (Senti-ML4) techniques. Figure 4 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.

5 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 [37] - [39]; 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.

Overall, we extended the analysis to check whether we could find a dependency of semantic aspects of user-comments for different issues on COVID-19–related topics. In this case, 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 1 and 2 and Figures 4-8. 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

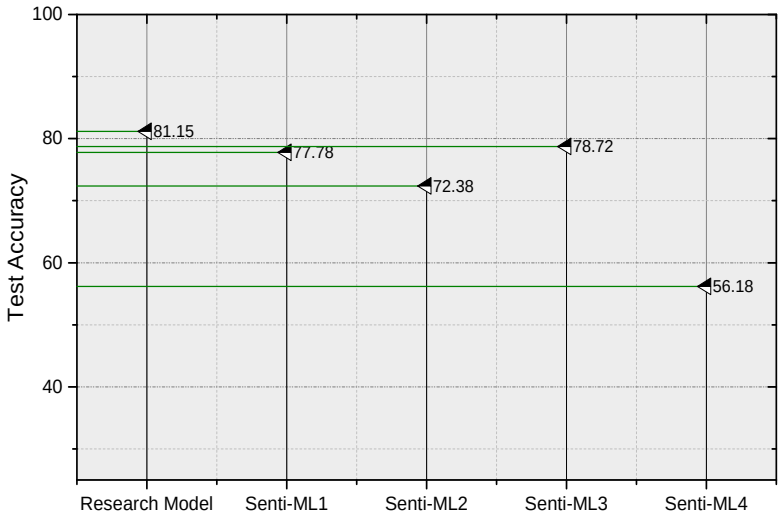

Fig. 10: Accuracy performance of the methods for COVID-19 sentiment-classification using various features

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.

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. 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 [40] - [41] that can be used in the future for sentiment level classification as a novel method of retrieving meaningful latent topics from public comments.

6 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.

Acknowledgements

We acknowledge SciTechEdit International, LLC (Highlands Ranch, CO, USA) for providing pro bono professional English-language editing of this article. This work has been awarded by the National Natural Science Foundation of China (61941113, 81674099, 61502233), the Fundamental Research Fund for the Central Universities (30918015103, 30918012204), Nanjing Science and Technology Development Plan Project (201805036), and "13th Five-Year" equipment field fund (61403120501), China Academy of Engineering Consulting Research Project(2019-ZD-1-02-02).

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Declaration of Conflict of Interest : All authors declare no conflict of interest directly related to the submitted work.

References

1. Malta, Monica, Anne W. Rimoin, and Steffanie A. Strathdee. "The coronavirus 2019-nCoV epidemic: Is hindsight 20/20?." EClinicalMedicine 20 (2020).
2. Thomas, J., Prabhu, A. V., Heron, D. E., & Beriwal, S. (2019). Reddit and Radiation Therapy: A Descriptive Analysis of Posts and Comments Over 7 Years by Patients and Health Care Professionals. Advances in radiation oncology, 4(2), 345-353.
3. Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106, 92-104.
4. Barros, J. M., Buitelaar, P., Duggan, J., & Rebholz-Schuhmann, D. (2019, November). Unsupervised Classification of Health Content on Reddit. In Proceedings of the 9th International Conference on Digital Public Health (pp. 85-89).
5. Roy, M., Moreau, N., Rousseau, C., Mercier, A., Wilson, A., & Atlani-Duault, L. (2020). Ebola and localized blame on social media: analysis of Twitter and Facebook conversations during the 2014–2015 Ebola epidemic. Culture, Medicine, and Psychiatry, 44(1), 56-79.

6. Rong, J., Michalska, S., Subramani, S., Du, J., & Wang, H. (2019). Deep learning for pollen allergy surveillance from twitter in Australia. BMC medical informatics and decision making, 19(1), 208.
7. Batbaatar, E., & Ryu, K. H. (2019). Ontology-Based Healthcare Named Entity Recognition from Twitter Messages Using a Recurrent Neural Network Approach. International Journal of Environmental Research and Public Health, 16(19), 3628.
8. Naderi, Hamid, Sina Madani, Behzad Kiani, and Kobra Etminani. "Similarity of medical concepts in question and answering of health communities." Health informatics journal (2019): 1460458219881333.
9. Vydiswaran, V. V., & Reddy, M. (2019). Identifying peer experts in online health forums. BMC medical informatics and decision making, 19(3), 68.
10. Halder, K., Poddar, L., & Kan, M. Y. (2017, September). Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 127-135).
11. McRoy, S., Rastegar-Mojarad, M., Wang, Y., Ruddy, K. J., Haddad, T. C., & Liu, H. (2018). Assessing unmet information needs of breast cancer survivors: Exploratory study of online health forums using text classification and retrieval. JMIR cancer, 4(1), e10.
12. Chakravorti, D., Law, K., Gemmell, J., & Raicu, D. (2018, November). Detecting and Characterizing Trends in Online Mental Health Discussions. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 697-706). IEEE.
13. 1VanDam, C., Kanthawala, S., Pratt, W., Chai, J., & Huh, J. (2017). Detecting clinically related content in online patient posts. Journal of biomedical informatics, 75, 96-106.
14. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
15. Mimno, D., Wallach, H., & McCallum, A. (2008, December). Gibbs sampling for logistic normal topic models with graph-based priors. In NIPS Workshop on Analyzing Graphs (Vol. 61).
16. Alshemali, B., & Kalita, J. (2020). Improving the reliability of deep neural networks in NLP: A review. Knowledge-Based Systems, 191, 105210.
17. Giménez, M., Palanca, J., & Botti, V. (2020). Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis. Neurocomputing, 378, 315-323.
18. Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., ... & Zhang, A. (2020). Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing. Journal of Machine Learning Research, 21(23), 1-7.
19. Park, H. J., Song, M., & Shin, K. S. (2020). Deep learning models and datasets for aspect term sentiment classification: Implementing holistic recurrent attention on target-dependent memories. Knowledge-Based Systems, 187, 104825.
20. Abualigah, L., Alfar, H. E., Shehab, M., & Hussein, A. M. A. (2020). Sentiment Analysis in Healthcare: A Brief Review. In Recent Advances in NLP: The Case of Arabic Language (pp. 129-141). Springer, Cham.
21. Balamurali, Anumeera, and Balamurali 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, pp. 122-142. IGI Global, 2020.
22. Unanue, I. J., Borzeshi, E. Z., & Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of biomedical informatics, 76, 102-109.
23. Luo, Y. (2017). Recurrent neural networks for classifying relations in clinical notes. Journal of biomedical informatics, 72, 85-95.
24. Huang, J., & Feng, Y. (2019, October). Optimization of Recurrent Neural Networks on Natural Language Processing. In Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition (pp. 39-45).
25. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
26. Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4580-4584). IEEE.
27. Meisheri, H., Ranjan, K., & Dey, L. (2017, November). Sentiment extraction from Consumer-generated noisy short texts. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 399-406). IEEE.

28. Rajput, Adil. ”Natural Language Processing, Sentiment Analysis, and Clinical Analytics.” In Innovation in Health Informatics, pp. 79-97. Academic Press, 2020.
29. Sharma, T., Bajaj, A., & Sangwan, O. P. (2020). Deep Learning Approaches for Textual Sentiment Analysis. In Handbook of Research on Emerging Trends and Applications of Machine Learning (pp. 171-182). IGI Global.
30. Habimana, Olivier, Yuhua Li, Ruixuan Li, Xiwu Gu, and Ge Yu. ”Sentiment analysis using deep learning approaches: an overview.” Science China Information Sciences 63, no. 1 (2020): 1-36.
31. Marin, Iuliana, Nicolae Goga, and Andrei Doncescu. ”[WiP] Sentiment Analysis Electronic Healthcare System Based on Heart Rate Monitoring Smart Bracelet.” In 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA), pp. 99-104. IEEE, 2018.
32. Yang, C. C., & Jiang, L. (2018). Enriching user experience in online health communities through thread recommendations and heterogeneous information network mining. IEEE Transactions on Computational Social Systems, 5(4), 1049-1060.
33. Goeuriot, L., Na, J. C., Min Kyaing, W. Y., Khoo, C., Chang, Y. K., Theng, Y. L., & Kim, J. J. (2012, January). Sentiment lexicons for health-related opinion mining. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (pp. 219-226).
34. Thelwall, M. (2017). The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. In Cyberemotions (pp. 119-134). Springer, Cham.
35. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.
36. Thelwall, M., & Buckley, K. (2013). Topic-based sentiment analysis for the Social Web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology, 64(8), 1608–1617.
37. Okon, E., Rachakonda, V., Hong, H. J., Callison-Burch, C., & Lipoff, J. (2019). Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. Journal of the American Academy of Dermatology.
38. Park, A., & Conway, M. (2017). Tracking health related discussions on Reddit for public health applications. In AMIA Annual Symposium Proceedings (Vol. 2017, p. 1362). American Medical Informatics Association.
39. Pandrekar, S., Chen, X., Gopalkrishna, G., Srivastava, A., Saltz, M., Saltz, J., & Wang, F. (2018). Social media based analysis of opioid epidemic using Reddit. In AMIA Annual Symposium Proceedings (Vol. 2018, p. 867). American Medical Informatics Association.
40. Zhou, S., Chen, Q., & Wang, X. (2014). Fuzzy deep belief networks for semi-supervised sentiment classification. Neurocomputing, 131, 312-322.
41. Ramasamy, B., & Hameed, A. Z. (2019). Classification of healthcare data using hybridised fuzzy and convolutional neural network. Healthcare technology letters, 6(3), 59-63.