Heliyon 10 (2024) e34103

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

COVIDHealth: A novel labeled dataset and machine learning-based web application for classifying COVID-19 discourses on Twitter

Mahathir Mohammad Bishal [a], Md. Rakibul Hassan Chowdory [a], Anik Das [b], Muhammad Ashad Kabir [c,*]

[a] Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram, 4349, Bangladesh
[b] Department of Computer Science, St. Francis Xavier University, Antigonish, B2G 2W5, NS, Canada
[c] School of Computing, Mathematics, and Engineering, Charles Sturt University, Bathurst, 2795, NSW, Australia

ARTICLE INFO

Dataset link: https://github.com/Bishal16/COVID19-Health-Related-Data-Classification-Website

Keywords:
COVID-19 discussions
Twitter dataset
Deep learning
Machine learning
Classification
Web application

ABSTRACT

The COVID-19 pandemic has sparked widespread health-related discussions on social media platforms like Twitter (now named ‘X’). However, the lack of labeled Twitter data poses significant challenges for theme-based classification and tweet aggregation. To address this gap, we developed a machine learning-based web application that automatically classifies COVID-19 discourses into five categories: health risks, prevention, symptoms, transmission, and treatment. We collected and labeled 6,667 COVID-19-related tweets using the Twitter API, and applied various feature extraction methods to extract relevant features. We then compared the performance of seven classical machine learning algorithms (Decision Tree, Random Forest, Stochastic Gradient Descent, Adaboost, K-Nearest Neighbor, Logistic Regression, and Linear SVC) and four deep learning techniques (LSTM, CNN, RNN, and BERT) for classification. Our results show that the CNN achieved the highest precision (90.41%), recall (90.4%), F1 score (90.4%), and accuracy (90.4%). The Linear SVC algorithm exhibited the highest precision (85.71%), recall (86.94%), and F1 score (86.13%) among classical machine learning approaches. Our study advances the field of health-related data analysis and classification, and offers a publicly accessible web-based tool for public health researchers and practitioners. This tool has the potential to support addressing public health challenges and enhancing awareness during pandemics. The dataset and application are accessible at https://github.com/Bishal16/COVID19-Health-Related-Data-Classification-Website.

1. Introduction

Social media platforms, including Twitter (now named ‘X’), Facebook, Whatsapp, Weibo, and others, have evolved into powerful channels for real-time communication during natural disasters and disease outbreaks across the globe [1]. These platforms have become primary mediums for individuals to communicate, share their experiences, and exchange thoughts [2]. It holds the potential to serve as a valuable public health tool for scientists to promptly convey accurate information during pandemics, efficiently collecting

* Corresponding author.
E-mail addresses: mahathirbishal@gmail.com (M.M. Bishal), alvihasan361@gmail.com (M.R.H. Chowdory), x2021gmg@stfx.ca (A. Das), akabir@csu.edu.au (M.A. Kabir).

https://doi.org/10.1016/j.heliyon.2024.e34103
Received 4 November 2023; Received in revised form 27 June 2024; Accepted 3 July 2024

Available online 8 July 2024

reliable data [3]. Today, researchers harness the wealth of unstructured data from social media to construct effective frameworks for healthcare applications [4,5].

Twitter, a microblogging and long-distance informal communication service, allows users to send "tweets" limited to 280 characters. With over 368 million monthly active users worldwide [6], it has become an essential platform for sharing ideas, data, and experimentation among medical experts for more than a decade [7,8]. It has emerged as a rapid and direct communication tool for disseminating COVID-19 information to the general public. People have turned to Twitter to share discussions related to the pandemic [9].

User-generated content from social media platforms has gained substantial recognition for syndromic surveillance during global health emergencies, such as the 2009 H1N1 pandemic [10–16], the 2014 Ebola outbreak [17–22], the 2003 SARS epidemic [23], and recently COVID-19 pandemic [24,25].

COVID-19 presents a significant global health threat [26,27], with particular severity observed in individuals with weakened immune systems, diabetes, or pre-existing conditions like lung or heart disease [28]. This virus is highly contagious [29,30], and the analysis of its transmission identifies both direct modes, such as person-to-person contact [31], and indirect pathways, including transmission via contaminated surfaces [32]. Despite the pandemic's impact, there has been a notable absence of studies that focused on automatically classifying social media discourse related to COVID-19. Several initiatives have emerged to create COVID-19 Twitter datasets [33–35] which primarily collect and provide tweet IDs. These datasets offer an opportunity to create labeled datasets and train machine learning models for classifying COVID-19 discourse.

In this paper, we present a comprehensive classification of social media discourse related to COVID-19 on Twitter. We utilised tweet IDs from a pre-existing COVID-19 Tweet dataset [33] and meticulously labeled the tweets into five distinct categories: health risks, prevention, symptoms, transmission, and treatment. To address dataset imbalances, we applied preprocessing techniques and data augmentation methods. We then extracted features using three distinct methods, and employed both classical machine learning and deep learning techniques to classify the tweets. Our key contributions are as follows:

- We have developed and curated a novel labeled COVID-19 Twitter dataset, building upon existing tweet ID datasets by adding meticulously labeled data, enabling the analysis and classification of COVID-19-related discussions across five essential categories: health risks, prevention, symptoms, transmission, and treatment. This dataset provides a valuable resource for researchers to explore public perceptions, concerns, and behaviors related to COVID-19.
- We have performed a thorough empirical evaluation, employing both classical machine learning and deep learning techniques, to establish a baseline classification performance for our novel COVID-19 Twitter dataset. This study provides a foundation for future research, enabling the comparison and improvement of classification models on this dataset.
- To demonstrate the real-world applicability and utility of our approach, we have designed and developed a functional web application prototype in the form of a Chrome extension, integrating the optimal classification model. This extension enables users to classify COVID-19-related Twitter discourses in real-time, showcasing the potential for practical implementation and impact.

The remaining sections of the paper are organized as follows: Section 2 discusses related works, and Section 3 presents the workflow of our proposed methodology. Section 4 describes the details of collecting the dataset, labeling, preprocessing and the description of the dataset. Section 5 explains data sampling, feature extraction, and various classification methods. Section 6 presents the outcomes of our experimental evaluation. After a brief discussion in Section 8, Section 9 concludes the paper by summarizing the key findings and future work in the field.

2. Related works

Online social media has played a pivotal role in infectious disease monitoring, prevention, and control for several years [36]. Ng et al. [37] used Twitter for public perception research during the monkeypox outbreak. Khatua et al. [38] classified Twitter discourses about the 2014 Ebola and 2016 Zika outbreaks, categorized tweets into five perspectives for both Ebola and Zika outbreaks: health risks, prevention, symptoms, transmission, and treatment. Inspired by this, we explored the same five health perspectives for COVID-19.

Preventive measures are crucial in curbing the spread of infectious diseases like COVID-19 [39,40]. However, only a limited number of studies have specifically addressed preventive measures, such as handwashing, social distancing, and face shields. Many of these studies have focused on opinion mining related to mask-wearing from tweets [41–43]. An exception is the work by Doogan et al. [44], which concentrates on nonpharmaceutical interventions (NPIs) encompassing seven categories, including gathering restrictions, lockdowns, personal protection, social distancing, workplace closures, testing and tracing, and travel restrictions.

Identifying various combinations of symptoms is essential for characterizing infectious diseases [45,46]. Mackey et al. [25] have explored COVID-19 symptom self-reporting through bi-term topic modeling, an unsupervised machine learning approach applied to Twitter data. Shen et al. [24] have introduced a supervised machine-learning approach that leverages diagnosis reports and symptoms from Weibo posts to predict COVID-19 case counts. Other studies have focused on extracting prevalent symptoms from tweets [47] and sentiment analysis [48–51].

Analyzing treatments for COVID-19 is vital [52]. Several studies have examined classes directly or indirectly related to treatment, encompassing public perception and opinion mining of COVID-19 vaccines [53,54], anti-vaccination sentiment identification [55,56], detection of vaccine misinformation [57], and conspiracies [58] from social media. These studies used Twitter data to understand

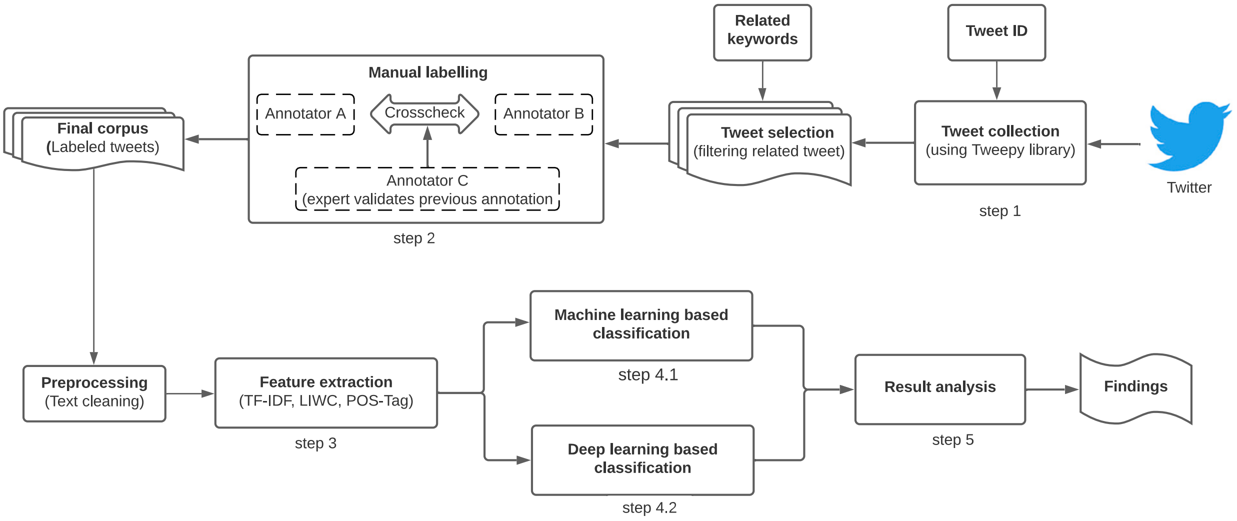


Fig. 1. Workflow of our proposed methodology.

public perceptions and sentiments towards COVID-19 vaccines, revealing a prevalence of negative sentiments, misinformation, and conspiracy theories. Machine learning and natural language processing techniques were employed to identify and classify tweets as positive, neutral, or negative, and to detect misinformation and anti-vaccination content. However, these studies do not comprehensively cover the five perspectives of COVID-19 discourses mentioned above, highlighting the need for more thorough investigation.

Multi-class classifications are inherently complex due to potential feature overlap between classes, in contrast to binary classification [59], and recognizing minority class features can be challenging [60]. To date, no study has collectively focused on the mentioned classes in a single framework, with two of them lacking related works, and only a limited number of studies addressing the other three classes. The development of an automatic recognition system, such as a web application, for these significant topics holds potential to enhance the usability of classification outputs for both healthcare facilities and the general population.

3. Methodology

Fig. 1 provides a high-level overview of the proposed methodology employed in this study. The construction of the COVIDHealth dataset involved a series of key steps:

Data Collection. In the first step, we gathered tweets by utilizing tweet IDs obtained from the COVID-19 Tweets dataset [33]. To ensure the relevance of the collected data, we employed predefined keywords as detailed in Section 4.

Data Annotation. Following data collection, we conducted a two-step annotation process. Initially, two independent annotators labeled the collected tweets based on the criteria outlined in Section 4. Subsequently, a third expert meticulously reviewed the annotations, addressing any discrepancies through consensus to ensure the accuracy and consistency of the final dataset.

Preprocessing and Feature Extraction. As raw text data is not directly compatible with various machine learning and deep learning algorithms, we performed text preprocessing in the third step. We implemented three distinct feature extraction techniques to derive meaningful features from the text data, outlined in Section 5.2.

Classification. In steps 4.1 and 4.2, we fed the extracted features into a variety of machine learning and deep learning algorithms for classification purposes (described in Section 5.3). This step encompassed the model training and evaluation phase to assess the performance of these algorithms.

Performance Evaluation and Application Development. In the final step (step 5), we conducted a comprehensive analysis to evaluate the performance of the various machine learning and deep learning classifiers. Based on the results (reported in Section 6, we proceeded to develop a web application prototype as a Chrome extension, using the best-performing model (presented in Section 7).

The subsequent subsections of this article will provide a detailed description of each of these steps, offering insight into the methods and techniques employed in the creation of the COVIDHealth dataset and the subsequent analysis and application development.

4. Building the COVIDHealth dataset

4.1. Data collection and labeling

We have used a publicly available dataset, COVID-19 Tweets Dataset [33], consisting of an extensive collection of 1,091,515,074 tweet IDs, and continuously expanding. The dataset was compiled by tracking over 90 distinct keywords and hashtags commonly associated with discussions about the COVID-19 pandemic. From this massive dataset, we focused on a specific time frame, encompassing data from August 05, 2020, to August 26, 2020, to meet our research objectives. As this dataset contains only tweet IDs, we

Table 1
Classification type with their definition.

Class name	Situation in COVID-19	Related keywords
Health risks	People aged over their sixties or people with heart disease, lung problems, weak immune systems, or diabetes are more at risk of being affected by COVID-19.	Lung disease, heart disease, diabetes, weak immunity, front line heroes
Prevention	Avoiding close contact, covering sneezes and coughs, covering nose and mouth around others with face shields or covers, disinfecting and cleaning more often, washing hands frequently, and routine health monitoring.	Wash hands, homeschooling, close contact, cover mouth, cover nose, coughs, sneezes, clean and disinfect, face shields
Symptoms	Common COVID-19 symptoms, e.g., cough or cold, congestion or runny nose, breathing issues, fever, muscle or body aches, sore throat, diarrhea, nausea or vomiting, loss of taste or smell, headache, fatigue	Shortness of breath, cough, fever, chills, fatigue, vomiting, nausea, diarrhea, headache, sore throat
Transmission	Person-to-person spread, the virus spreads easily between people, touching a surface or object that has the virus on it, spread between animals, people	Person-to-person spread, spreads easily between people, touching a surface, touching an object, spread between animals, spread between people
Treatment	Probable vaccine development and drugs used for COVID-19 treatment	Vaccine, drugs, paracetamol, herd immunity

Table 2
Tweets distribution in dataset.

Class	Tweets count
Health risk	978
Prevention	2,046
Symptoms	1,402
Transmission	802
Treatment	1,439
Total	6,667

have used the Twitter developer API to retrieve the corresponding tweets from Twitter. This retrieval process involved searching for tweet IDs and extracting the associated tweet texts, and it was implemented using the Twython library.[1] In total, we successfully collected 21,890 tweets during this data extraction phase.

Following guidelines set by the CDC and WHO, we categorized tweets into five distinct classes for classification: health risks, prevention, symptoms, transmission, and treatment, as detailed in Table 1. Specifically, individuals aged over sixty, or those with pre-existing health conditions such as heart disease, lung problems, weakened immune systems, or diabetes, are at higher risk of severe COVID-19 complications. Therefore, tweets categorized as 'health risks' pertain to the elevated risks associated with COVID-19 due to age or specific health conditions. 'Prevention' related tweets encompass discussions on preventive and precautionary measures regarding the COVID-19 pandemic. Tweets discussing common COVID-19 symptoms, including cough, congestion, breathing issues, fever, body aches, and more, are classified as 'symptoms' related tweets. Conversations pertaining to the spread of COVID-19 between individuals, between animals and humans, and contact with virus-contaminated objects or surfaces are categorized as 'transmission' related tweets. Lastly, tweets indicating vaccine development and drugs used for COVID-19 treatment fall under the 'treatment' related category.

We determined specific keywords for each of the five classes (health risks, prevention, symptoms, transmission, and treatment) based on the definitions provided by the CDC and WHO on their official websites. These definitions, along with their associated keywords, are detailed in Table 1. For instance, the CDC and WHO indicate that individuals over the age of sixty with conditions like heart disease, lung problems, weak immune systems, or diabetes face a higher risk of severe COVID-19 complications. In accordance with this definition, we selected relevant keywords such as "lung disease", "heart disease", "diabetes", "weak immunity", and others to identify tweets related to health risks within the larger tweet dataset. This approach was consistently applied to define keywords for the remaining four classes. Subsequently, we filtered the initial dataset of 21,890 tweets to extract tweets relevant to our predefined classes, resulting in a total of 6,667 tweets based on the selected keywords.

To ensure the accuracy of our dataset, two separate annotators individually assigned the 6,667 tweets to the five classes. A third annotator, a natural language expert, meticulously cross-checked the dataset and provided necessary corrections. Subsequently, the two annotators resolved any discrepancies through mutual agreement, resulting in the final annotated dataset. Table 2 provides an overview of the distribution of tweets among the five classes, with 978, 2046, 1402, 802, and 1439 tweets annotated as 'health risk', 'prevention', 'symptoms', 'transmission', and 'treatment', respectively (as presented in Table 2). Additionally, Table 3 offers a selection of example tweets from each of the defined categories, providing insights into the nature of the annotated content.

[1] https://github.com/ryanmcgrath/twython.

Table 3
Example tweets from the dataset for each class.

Class	Sample tweet
Health risk	COVID went after people with diabetes, obesity, high blood pressure. If you have these, you’re at higher risk of being infected.
Prevention	Use mask, avoid gathering, wash hand.#stay_home #stay_safe.
Symptoms	Cough, sore throat, shortness breath, runny nose and loss of smell are primary symptoms of covid19.
Transmission	Right now, about 100 students are in quarantine because of close contact with a positive #Covid_19 individual.
Treatment	Great reminder: until we have definitive pharmacological interventions for COVID, it’s down to masks, ventilation, testing.

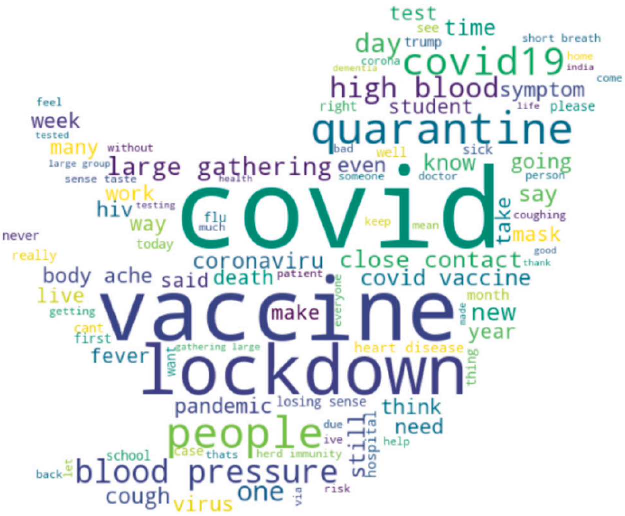


Fig. 2. Word cloud representation of twitter dataset.

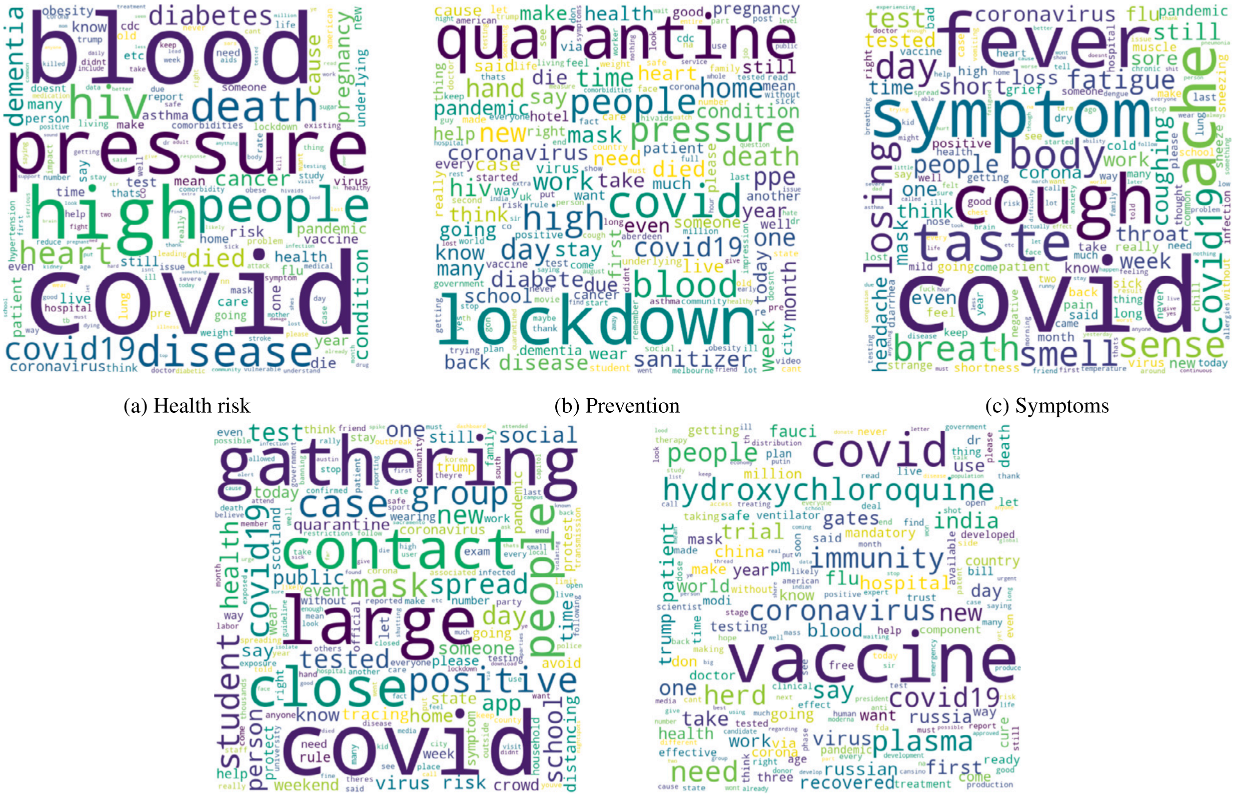
Table 4
Top-10 frequent words from tweets in the COVIDHEALTH dataset.

Rank	Health risk		Prevention		Symptoms		Transmission		Treatment		Whole dataset	
	Word	Count	Word	Count	Word	Count	Word	Count	Word	Count	Word	Count
1	COVID	508	quarantine	609	Covid	945	covid	631	vaccine	927	COVID	2957
2	blood	276	lockdown	603	fever	335	large	428	covid	480	Vaccine	1001
3	pressure	244	covid	393	cough	281	gathering	357	vaccines	240	People	684
4	high	217	people	170	aches	228	close	267	hydroxychloroquine	123	Quarantine	656
5	people	153	pressure	161	taste	225	contact	257	plasma	107	Lockdown	645
6	hiv	137	high	159	body	216	people	162	coronavirus	105	Blood	507
7	heart	116	like	104	losing	188	cases	83	immunity	103	High	444
8	disease	99	home	91	breath	162	gatherings	82	herd	90	Large	440
9	diabetes	84	get	82	smell	152	groups	77	need	90	Pressure	410
10	dementia	77	work	75	fatigue	143	spread	68	people	81	Gathering	363

4.2. *Dataset visualization*

Our dataset comprises a total of 6,667 data points categorized into five classes. A Word cloud representation of the entire dataset is depicted in Fig. 2. Notably, words such as ‘COVID’, ‘vaccine’, and ‘lockdown’ prominently feature in this word cloud, signifying their prevalence within the dataset.

Furthermore, individual word clouds were generated for each of the five classes, as presented in Figs. 3a, 3b, 3c, 3d, and 3e. To provide a more detailed insight, we identified the top 10 most frequent words within each of these word clouds, which are summarized in Table 4 for better comprehension. Interestingly, the term ‘Covid’ is a common occurrence across all classes. However, upon its exclusion from the *health risk* class, we observed that words related to ‘blood pressure’ and ‘heart disease’ prominently characterize this class, aligning with the typical concerns and focus of the health risk category. In the *prevention* class, terms such as ‘quarantine’ and ‘lockdown’ dominate the word cloud, reflecting the emphasis on preventive measures. In the *symptoms* class, words like ‘fever’, ‘symptom’, and ‘cough’ exhibit higher frequency compared to other terms. Within the *transmission* class, the phrase ‘large gathering’ and the concept of ‘close contact’ are the most prominent. Lastly, in the *treatment* class, ‘COVID vaccine’ and ‘vaccines’ emerge as the dominant terms, underscoring the focus on treatment and vaccination within this category.



(a) Health risk

(b) Prevention

(c) Symptoms

(d) Transmission

(e) Treatment

Fig. 3. Word cloud representation of five different classes of the COVIDHEALTH dataset.

4.3. Data preprocessing

After labeling all tweet data, we applied a series of preprocessing techniques to clean the unstructured and non-categorized dataset. This step involved the systematic application of various data preprocessing methods to refine the raw text data, including the removal of unnecessary elements such as mentions, hashtags, URLs, repeated characters, punctuation, stopwords, and text in other languages.

To begin, we utilized regular expressions to remove mentions (words preceded by the ‘@’ symbol) and hashtags (denoted by the ‘#’ symbol) from all tweet data. Subsequently, we employed regular expressions to eliminate URLs present within the text. Following this, we proceeded to remove irrelevant words, such as ‘rt’, ‘brt’, and newline characters (‘\n’) from the tweets. Additionally, if a word contained more than two repeated characters (e.g., ‘tooooo muuuuuch’), we reduced the repetitions to a maximum of two characters. While not a perfect solution due to altered spellings (e.g., ‘muuch’), it effectively reduces the feature space by consolidating variations like ‘muuuch’ and ‘muuuuuch’ into a common form, ‘muuch’. Furthermore, we removed all punctuation marks, as they are typically unnecessary for text classification purposes. Stopwords, which are frequently occurring words that do not contribute unique information for classification, were also removed from the text using the Natural Language Toolkit (NLTK). To maintain consistency, we retained only English characters in the dataset, filtering out any text in other languages. This preprocessing stage aimed to enhance the quality of the text data and prepare it for subsequent analysis and classification tasks.

5. Machine learning techniques for text classification

5.1. Dataset sampling techniques

The dataset exhibits varying tweet counts across each class, leading to a data imbalance challenge. In such scenarios, conventional classifiers may struggle when dealing with classes of unequal sizes, favoring the larger class. To address this issue, we employed three oversampling techniques – SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), and random oversampling – as well as one under-sampling technique. This enabled us to generate a balanced dataset while retaining the original imbalanced dataset. As a result, we worked with a total of five distinct datasets.

5.2. Feature extraction

The number of features extracted through the feature extraction techniques is detailed in Table 5, and they are described below.

Table 5
List of extracted features with brief description.

Scope	Feature Name	Description	Feature no.	Output Type
Linguistic measure	LIWC	Measures textual features	69	Real
Word frequency	TF-IDF	Measures the importance of a word in a document	12,649	Real
Word-category disambiguation	POS tag	Counts the number of parts of speech in a document	35	Integer

Linguistic Inquiry and Word Count (LIWC) [61], a transparent text analytics software, utilizes the original news text in the dataset to extract a comprehensive array of psychological and linguistic features. We used LIWC to extract features from 13 different dimensions: (i) summary dimension – consists of 8 features, including word count and word per sentence, (ii) punctuation mark – consists of 12 features, including comma, semicolon, quote, and hyphen, (iii) function words – consists of 15 features, including pronoun, article, and conjunction, (iv) perceptual process – consists of 4 features, including seeing, hearing, and feeling, (v) biological process – consists of 5 features, including body, sexuality, and health, (vi) other grammar – consists of 6 features, including interrogatives and numbers, (vii) time orientation – consists of 3 features, such as past, present, and future, (viii) relativity – consists of 4 features, including motion, space and time, (ix) affect – consists of 6 features, including positive emotion, negative emotion, and anxiety, (x) personal concerns – consists of 6 features, including achievement, leisure, and home, (xi) social – consists of 5 features, including human, family, and friend, (xii) informal language, consists of 6 features, including filler, and swear, (xiii) cognitive process – consists of 7 features, including certainty, insight, and inhibition.

Term frequency and inverse document frequency (TF-IDF) [62] weighting method is applied to the dataset to assess the importance of a term within a document. Term frequency (TF) quantifies the frequency of a term within a specific document. Conversely, the inverse document frequency (IDF) is a metric for determining the significance of a term across the entire dataset. The TF and IDF product is the weight of the TF-IDF in a given word. The higher the value of TF-IDF, the rarer it is. The TF-IDF weight is frequently used in text mining and information retrieval.

Part-of-speech tagging (POST) [63] also known as word-category disambiguation, is used to annotate a word with a corresponding part of the speech based both on its definition and on its context for resolving lexical ambiguities. The Stanford Part of Speech Tagger was utilized for this task, and the resulting feature extracted from this process is the count of POS tags. Within the corpus, 33 distinct tagsets (a list of part-of-speech tags) were identified. To obtain the POS tag count, each tweet's words were tallied according to their assigned POS tags, resulting in the derivation of 33 individual features.

5.3. Classification methods

Classification is a machine learning technique wherein the algorithm learns from the provided data and subsequently categorizes new observations based on this learned knowledge. This subsection covers both classical machine learning and deep learning algorithms used in the classification process.

5.3.1. Classical machine learning

Below, we will briefly discuss the classical machine learning techniques employed in this study.

Decision tree (DT) [64] is a powerful machine learning algorithm that is widely used for both classification and regression tasks. It is a graphical representation of a decision-making process that resembles a tree structure with nodes and branches. Each node represents a feature or attribute, and each branch represents a decision or outcome based on that feature. The decision tree works by recursively splitting the data into subsets based on the most informative features at each node, effectively partitioning the data into categories or predicting numerical values. This process continues until a stopping criterion is met, such as a predefined depth or a minimum number of data points in a node. Decision trees are known for their interpretability and ease of understanding. They are used in a wide range of applications, from finance to healthcare, and are often a fundamental building block in more complex machine learning models.

Adaboost [65]. Boosting is a machine-learning technique based on a combination of several relatively weak and inexact rules for constructing a highly accurate Prediction Law. AdaBoost, unlike boost-by-majority, combines the weak hypotheses by summing their probabilistic predictions. In a real-valued neural network summing the outcomes of the networks and then selecting the best prediction performs better than selecting the best prediction of each network and then combining them with a majority rule [66].

Random forests (RFs) [67] are a collection of tree predictors in which the values of a random vector sampled independently and with the same distribution for all trees in the forest are used to predict the behavior of each tree. If the number of trees in a forest grows larger, the generalization error converges to a limit. The intensity of individual trees in the forest and the correlation between them determine the generalization error of a forest of tree classifiers. When a random set of features is used to separate each node, the error rates are comparable to AdaBoost, which theoretically reduces any learning algorithm error that consistently generates classifiers whose performance is a little better than random assumption but more robust in terms of noise. Internal estimates are used to track error, power, and correlation [68].

Stochastic gradient descent (SGD) [69] is an iterative approach for optimizing an objective function with sufficient smoothness properties (e.g., differentiable or sub-differentiable). It replaces the actual gradient which is calculated from the entire data set by an estimated gradient which is calculated from a randomly selected subset of the data. So it can be regarded as a stochastic approximation of gradient descent optimization. This reduces the computational burden and achieves quicker iterations in trade for a lower convergence rate, particularly in the case of high-dimensional optimization problems.

K-nearest neighbor [70]. The intuition behind the k-nearest neighbor (kNN) classification is very simple: examples are categorized by their closest neighbor's class. More than one neighbor must also be taken into account so that the method is more generally called kNN classification, where k closest neighbors are used in the class determination. Because the training examples are required during run time, i.e., they must be in memory during run time, often they are referred to as memory-based classification. Since the induction is delayed, a lazy learning technique is considered.

Logistic regression [71] is a supervised classification algorithm used to estimate a target variable's probability. The type of objective or dependent variable is dichotomous, meaning that only two classes are possible. The dependent variable, in simple words, is binary in nature with either 1 (this means success/yes) or 0 (this indicates failure/no). A logistic regression model predicts $P(Y = 1)$ as a function of X mathematically.

Linear SVC [72]. The aim of the Linear SVC is to fit the data that is provided, returning a hyperplane that is "best fit" and divides or classifies data. From there, some features can be fed to the classifier after receiving the hyperplane to see what the predicted class is. This makes this particular algorithm more acceptable for use, although this can be used for many situations.

5.3.2. Deep learning

Deep learning makes it possible to learn data representation with multiple abstraction levels through computational models that consist of various processing layers. This increases the state-of-the-art in speech identification, the recognition of visual artifacts, the detection of objects, and many other domains including drug discoveries and genomics dramatically. Deep learning detects complex structures in large data sets with the use of the backpropagation algorithm to tell how the computer changes its membership functions, which are used to calculate the representation in each layer from the representation in the previous layer [73].

Convolutional Neural Networks (CNNs) [74] are similar to conventional artificial neural networks (ANNs) since they consist of self-optimizing neurons [75]. Each neuron still receives input and carries out an action (such as a scalar product followed by a non-linear function) — the basis for countless ANNs. The entire network will still express a single perceptive score feature from input raw text through to the final output of the class score (the weight). The final layer contains class-related loss functions. CNNs are comprised of three types of layers. These are convolutional layers, pooling layers, and fully-connected layers. When these layers are stacked, a basic CNN architecture has been formed. By calculating the scalar product between its weights and the region linked to the input volume, the convolution layer determines the output of neurons that are connected to local regions of the input. The pooling layer will then simply perform downsampling along the spatial dimensionality of the given input, further reducing the number of parameters within that activation. The fully connected layers try to generate class scores from the activations for classification purposes.

We have used an embedding dimension of size 100 in our CNN architecture. The optimizer used in our CNN is RMSProp. Loss accuracy is measured by binary cross-entropy. We have used one dimensional convolution layer for this work. We add two layers to our CNN architecture. We have used relu activation in the first layer of the architecture and sigmoid activation in the second layer. Sixty epochs have been considered for the training dataset.

Recurrent Neural Network (RNN) [76] is a feedforward neural network with an internal memory that is a generalization of the feedforward neural network. RNN is recurrent in nature since it executes the same function for each data input, and the current input's outcome is dependent on the previous computation. The output is replicated and transmitted back into the recurrent network when it is created. It evaluates the current input as well as the output it has learned from the prior input when making a decision.

In the architecture of RNN, we have used 0.3 dropouts between the nodes of the convolution layer. We have used rmsprop optimizer for layers in RNN architecture. We have used sigmoid activation for the RNN layer. We have used an embedding dimension of size 100 for RNN architecture.

Long Short-Term Memory Networks (LSTMs) [77] represent a type of Recurrent Neural Network (RNN) designed to capture long-term dependencies. LSTMs gained widespread popularity and have proven effective in various applications. Unlike ordinary RNNs composed of simple repeating modules, LSTMs utilize a more complex structure, particularly in their cell state, depicted as a horizontal line traversing the diagram. The cell state in LSTMs functions akin to a conveyor belt, allowing data to flow unchanged with minimal linear interactions. Notably, LSTMs regulate the information flow through structures known as gates. Gates act as a selective mechanism, enabling the controlled addition or deletion of information from the cell state. Comprising a sigmoid neural net layer and a point-wise multiplication operation, gates play a crucial role in determining which information is allowed to pass through. To update the cell state, LSTMs employ an input gate layer (sigmoid) that selects values for updating and a tanh layer generating a vector of new candidate values [78]. These components are combined in the subsequent phase to create a comprehensive state update.

In this study, a sequential LSTM model is utilized. The architecture employs a softmax activation function and the Adam optimizer. Categorical cross-entropy is used to measure loss accuracy. The LSTM layer incorporates a dropout rate of 0.2, and recurrent dropout is also set at 0.2.

The *Bidirectional Encoder Representations from Transformers (BERT)* [79] model is structured in two distinct stages: pre-training and fine-tuning. During pre-training, the model is trained on a wide, unlabeled corpus. Subsequently, in the fine-tuning phase, all parameters are further adjusted using labeled data for specific tasks, building upon the knowledge gained during pre-training. The initial parameters for fine-tuning are derived from the pre-trained model. The BERT architecture is rooted in a bidirectional transformer multi-layer encoder design, which eliminates the need for recurrence and instead leverages a mechanism based on attention for establishing global dependencies between input and output [80].

Two primary types of pre-training are employed: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, a portion of input tokens is randomly masked, and the model learns to predict these masked tokens, thus fostering deep bidirectional representations. The NSP task can be generated from any monolingual corpus, facilitating the training process. BERT models can then be fine-tuned for various downstream tasks using the transformer's self-attention mechanism [79].

Table 6
Dataset setting for classical machine learning algorithm.

Dataset	Class					Total
	Symptoms	Treatment	Health risk	Transmission	Prevention	
Original	1,402	1,439	978	802	2,046	6,667
Oversampling	2,046	2,046	2,046	2,046	2,046	10,230
Undersampling	800	800	800	800	800	4,000

In this study, the BERT model is trained using a batch size of 32 for training and 8 for evaluation. The learning rate utilized in the architecture is set to $1e-5$. A warm-up proportion of 0.5 is applied. The maximum sequence length is limited to 50. The BERT model is trained for a total of 3 epochs.

5.4. Evaluation metrics

The following assessment measures were used to assess the classification performance: accuracy, precision, recall, and F1 score. The definitions of accuracy, precision, recall, and F1 score are as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = 2\times\frac{Precision \times Recall}{Precision + Recall}$$

Here TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

6. Experiments and results

In this section, we will present the findings of our study, focusing on dataset settings and the results of both classical machine learning and deep learning algorithms. In this study, we employed seven different classical machine learning algorithms: Decision Tree, Random Forest, Stochastic Gradient Descent, K-Nearest Neighbor, Adaboost, Logistic Regression, and Linear SVC. For deep learning algorithms, we exclusively utilized the TF-IDF feature extraction method, excluding the LIWC and POS tag feature extraction methods. This exclusion was due to the poor performance of LIWC and POS tag features in the context of deep learning, resulting in significantly lower accuracy scores. The TF-IDF method yielded 12,649 features, while LIWC and POS tag methods produced only 69 and 35 features, respectively. When compared to the vast number of features from TF-IDF, the features extracted by LIWC and POS tag methods were minimal. Consequently, these two feature extraction methods did not yield superior results; instead, they introduced additional computational costs without significant benefits.

6.1. Dataset settings for machine learning algorithm

The initial pre-processed dataset comprised a total of 6,667 data points, with an inherent class imbalance. To address the class imbalance, we generated one under-sampled dataset and three oversampled datasets, employing random oversampling, ADASYN, and SMOTE techniques. Table 6 provides a detailed overview of these datasets. Specifically, the undersampled dataset comprised 800 samples per class, totaling 4,000 samples, while the oversampled dataset consisted of 2,046 samples per class, totaling 10,230 samples. To ensure robust model evaluation, we employed 10-fold cross-validation during training with classical machine learning techniques.

6.2. Results for machine learning algorithm

Table 7 presents the results of a 10-fold cross-validation using classical machine learning algorithms, including classification accuracy, precision, recall, and F1 score. Notably, the Linear SVC algorithm achieved the highest accuracy of 86.23% on the ADASYN-based oversampling dataset. However, the Linear SVC algorithm on the SMOTE-based oversampling dataset yielded the highest precision (85.71%), recall (86.94%), and F1 score (86.13%), respectively. These results highlight the effectiveness of the Linear SVC algorithm in conjunction with SMOTE-based oversampling for achieving optimal classification performance. The Linear SVC algorithm also achieved the best performance on the original dataset, with results (precision 85.56%, recall 86.44%, F1 85.89%, and accuracy 85.94%) closely matching those obtained with SMOTE. Conversely, the lowest accuracy scores were observed when using the under-sampling dataset, across all algorithms. This is due to the smaller amount of sample data available in this particular dataset. However, for the other datasets, excluding under-sampling, every dataset exhibited moderate performance.

Table 7
Precision, recall, F1-score and accuracy score of different classical machine learning algorithm (10 fold cross validation).

Dataset	ML Technique	Precision	Recall	F1	Accuracy
Original	Decision Tree	83.15	82.57	82.71	82.88
	Random Forest	84.58	83.47	83.95	83.86
	Stochastic Gradient Descent	84.79	84.79	84.72	84.90
	K-nearest Neighbor	74.91	73.18	73.81	73.96
	Adaboost	82.38	78.05	77.42	77.57
	Logistic Regression	84.64	84.10	54.25	54.32
	Linear SVC	85.56	86.44	85.89	85.94
SMOTE	Decision Tree	82.54	82.54	82.47	82.50
	Random Forest	84.38	83.56	83.89	83.94
	Stochastic Gradient Descent	84.49	86.10	84.95	84.98
	K-nearest Neighbor	67.96	69.48	57.73	57.54
	Adaboost	82.90	79.54	77.58	77.74
	Logistic Regression	84.83	84.55	84.92	84.94
	Linear SVC	**85.71**	**86.94**	**86.13**	86.12
ADASYN	Decision Tree	82.59	82.16	82.27	82.33
	Random Forest	83.90	82.97	83.36	83.32
	Stochastic Gradient Descent	81.66	83.26	82.98	82.90
	K-nearest Neighbor	64.88	25.91	15.40	15.46
	Adaboost	81.40	77.32	76.64	76.72
	Logistic Regression	82.06	83.90	83.38	83.50
	Linear SVC	85.33	86.40	**86.13**	**86.23**
Random Over Sampling	Decision Tree	82.57	82.55	82.47	82.54
	Random Forest	84.66	84.65	84.47	84.54
	Stochastic Gradient Descent	84.64	86.26	85.04	85.50
	K-nearest Neighbor	70.65	72.23	70.32	70.35
	Adaboost	81.60	75.67	74.88	74.95
	Logistic Regression	84.64	86.11	85.03	85.08
	Linear SVC	85.26	86.50	85.69	85.75
Under Sampling	Decision Tree	41.00	42.00	43.00	43.20
	Random Forest	56.00	57.00	56.00	56.15
	Stochastic Gradient Descent	57.00	58.00	58.00	58.02
	K-nearest Neighbor	46.00	50.00	50.00	50.07
	Adaboost	52.00	53.00	54.00	54.29
	Logistic Regression	50.00	51.00	51.00	51.20
	Linear SVC	45.00	47.00	47.00	47.14

Table 8
Dataset setting for deep learning.

Dataset		Class					Total
		Symptoms	Treatment	Health risk	Transmission	Prevention	
Original	Training (70%)	971	1009	690	554	1442	4666
	Validation (20%)	293	285	191	167	397	1333
	Testing (10%)	138	145	97	81	206	667
Balanced (not augmented)	Training	640	640	640	640	640	3200
	Validation	80	80	80	80	80	400
	Testing	80	80	80	80	80	400
Balanced (augmented)	Training	1600	1600	1600	1600	1600	8000
	Validation	200	200	200	200	200	1000
	Testing	200	200	200	200	200	1000

6.3. Dataset settings for deep learning techniques

For deep learning techniques, we have prepared two types of datasets: a balanced dataset (not augmented) and a balanced dataset (augmented), both derived from the original imbalanced dataset. The dataset was divided into training (80%), validation (10%), and testing (10%) subsets. In the balanced (not augmented) dataset, the training dataset includes 640 samples for each class, while the validation and testing datasets consist of 80 samples for each class. On the other hand, the balanced (augmented) dataset features 1,600 samples for each class in the training set, with 200 samples for each class in both the validation and testing sets. Detailed dataset settings for deep learning algorithms can be found in Table 8.

Table 9
Result table for testing dataset of deep learning algorithms.

Dataset	Model name	Precision	Recall	F1	Accuracy
Original	LSTM	89.32	89.86	89.26	89.20
	CNN	73.36	73.24	73.38	73.29
	RNN	77.40	77.40	77.33	77.25
	BERT	83.40	82.16	82.38	82.29
Balanced (not augmented)	LSTM	78.96	79.34	79.23	79.28
	CNN	83.34	83.32	83.26	83.20
	RNN	79.39	79.38	79.41	79.36
	BERT	87.83	88.54	88.20	88.09
Balanced (augmented)	LSTM	89.47	89.40	89.37	89.40
	CNN	**90.41**	**90.40**	**90.40**	**90.40**
	RNN	87.42	87.60	87.46	87.60
	BERT	88.62	90.33	89.16	89.36

6.4. Results for deep learning

Four deep learning architectures – LSTM, CNN, RNN, and BERT – were employed for each dataset, with results reported in Table 9. On the original dataset, LSTM yielded the most promising results, achieving a precision of 89.32%, recall of 89.86%, F1 score of 89.26%, and accuracy of 89.2%. In the balanced (not augmented) dataset, BERT achieved the best performance, with a precision of 87.83%, recall of 88.54%, F1 score of 88.2%, and accuracy of 88.09%. Notably, LSTM performance significantly dropped in this dataset, likely due to limited data. In contrast, CNN performance improved significantly compared to the original dataset, attributed to the balanced dataset. RNN performance remained consistently low in both datasets. In the balanced (augmented) dataset, LSTM performance significantly improved compared to the balanced (not augmented) dataset, highlighting the importance of sample size for LSTM model performance. BERT demonstrated consistent performance across all three datasets. CNN delivered the highest performance among all deep learning architectures and datasets, achieving a precision of 90.41%, recall of 90.1%, F1 score of 90.4%, and accuracy of 90.4%.

6.5. Best result

In this subsection, we present the most notable results achieved by both classical machine learning and deep learning algorithms.

Within the realm of classical machine learning algorithms, the Linear SVC algorithm delivered the best results. The corresponding confusion matrix is shown in Fig. 4, with an F1 score of 86.13%. Linear SVC's superior performance can be attributed to its ability to maximize the margin between classes, effectively handling high-dimensional data and noise. By leveraging the kernel trick, Linear SVC efficiently transforms the data into a higher-dimensional space, enabling it to capture complex relationships and improve classification accuracy. Additionally, its robust regularization mechanism helps to prevent overfitting, making it well-suited for the dataset and contributing to its outstanding performance.

Among the deep learning algorithms, the CNN model performed exceptionally well, yielding the best results when applied to the balanced augmented dataset. The corresponding confusion matrix is presented in Fig. 5, with an impressive precision of 90.41%, recall of 90.4%, F1 score of 90.4%, and accuracy of 90.4%, respectively. The exceptional performance of CNN in text classification tasks can be attributed to its unique strengths. Firstly, its convolutional layers excel at extracting local features from text data, allowing it to capture subtle patterns and nuances. Additionally, its pooling layers enable the extraction of position-invariant features, making it robust to variations in word order. Furthermore, CNN's hierarchical representation of text data enables it to capture complex contextual relationships and long-range dependencies, providing a deeper understanding of the text. Its ability to learn robust features and tolerate noise in the data also contributes to its outstanding performance.

7. Web application

We have leveraged our top-performing CNN model to create a user-friendly web application,[2] which can greatly assist individuals in classifying various COVID-19 related text data into the aforementioned five distinct categories. This web application employs the CNN model we developed as its backend processing engine. Users can input text, which the application then sends to the model for analysis and classification.

Using the application is straightforward: users need to navigate to the website, input or paste the text they wish to classify, and click the "Predict" button to receive the output. To enhance user convenience, we have also developed a Google Chrome browser extension,[3] making it even easier to access the website. Upon clicking the extension, a popup labeled "Go for classification" will appear, and clicking this button opens a new tab with the classification website.

[2] https://github.com/Bishal16/COVID19-Health-Related-Data-Classification-Website.

[3] https://github.com/Bishal16/Google-Chrome-Extension_Covid19-Health-Related-Data-Classifier.

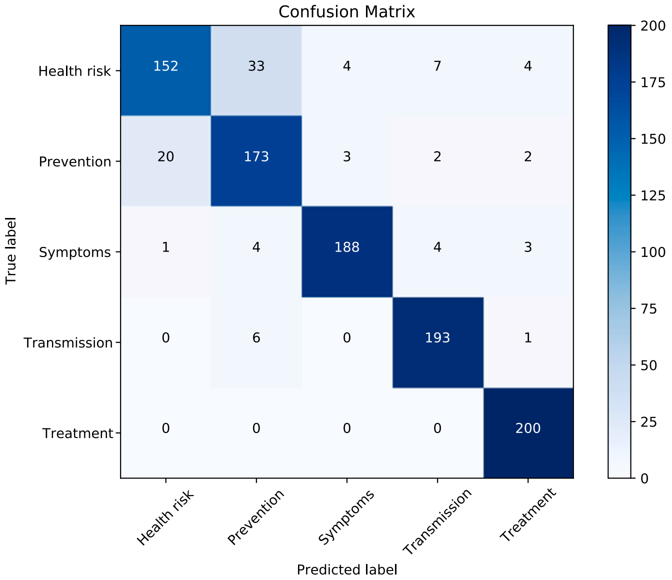


Fig. 4. Confusion matrix for Linear SVC algorithm.

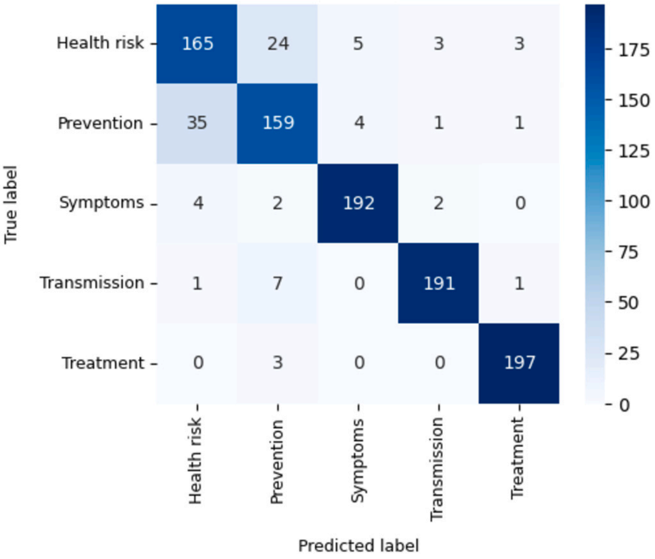


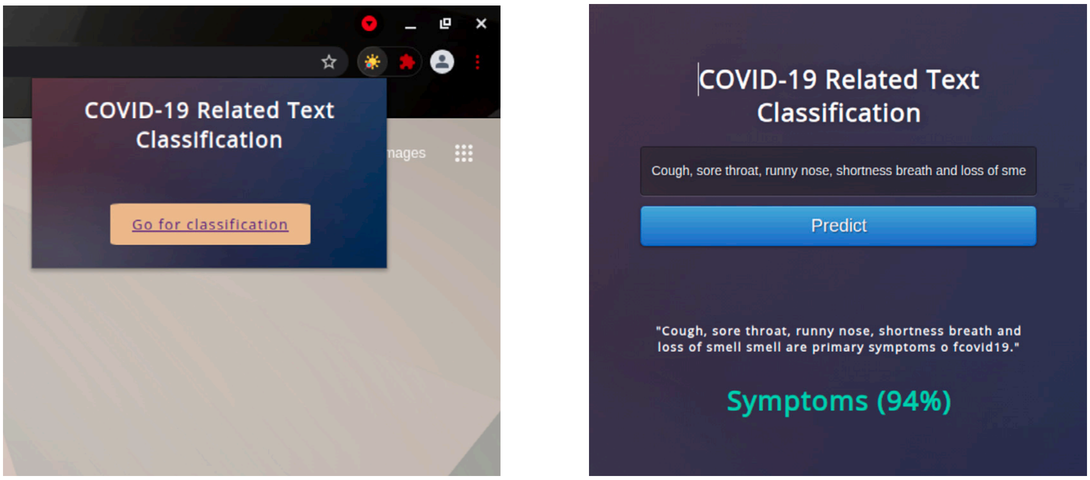
Fig. 5. Confusion matrix for CNN model.

As illustrated in Fig. 6a, the extension simplifies the process of accessing the classification website. When users click the extension, they are presented with the “Go for classification” option, allowing them to swiftly access the website. Fig. 6b shows the predicted class name along with its corresponding class accuracy. The deep learning model responsible for these predictions consistently achieves an impressive accuracy rate of 90.50%.

8. Discussion

The findings of this research hold significant implications for understanding and utilizing social media data in the context of public health, particularly during global health emergencies like the COVID-19 pandemic. The study’s focus on classifying COVID-19-related discourses within Twitter data contributes to the growing body of knowledge surrounding the role of social media in disseminating information and shaping public discourse during crises.

Numerous studies have concentrated on tweet classification related to COVID-19, covering a range of topics such as vaccine misinformation [81], fake news [82], stances toward online education [83], sentiment analysis [84] and emotion classification [85]. The significance of this research lies in its attempt to bridge the gap in the existing literature [86] by specifically analyzing COVID-19-related discussion on Twitter. By categorizing discussions into five distinct classes – health risks, prevention, symptoms, transmission,


COVID-19 Related Text Classification
Go for classification
mages
COVID-19 Related Text Classification
Predict

Cough, sore throat, runny nose, shortness breath and loss of sme

"Cough, sore throat, runny nose, shortness breath and loss of smell smell are primary symptoms o fcovid19."

Symptoms (94%)

(a) Google Chrome extension

(b) Demo of the web application

Fig. 6. Screenshot of chrome extension and real-time COVID-19 related text classification website.

and treatment – the study provides a nuanced understanding of the diverse aspects of health-related conversations during the pandemic. In the broader context of relevant studies, this research aligns with the trend of leveraging social media platforms for syndromic surveillance during global health emergencies, as seen in studies related to previous pandemics such as H1N1 [87], Ebola [88], and SARS [89]. The findings complement existing literature by specifically addressing the dearth of studies focusing on health risks and transmission-related content in the context of COVID-19 on social media.

The empirical study's comparison of classical machine learning and deep learning approaches is noteworthy [90]. The superior performance of the CNN algorithm, especially in comparison to classical ML methods, highlights the potential of advanced techniques in extracting meaningful insights from social media data. This finding aligns with the evolving landscape of deep learning applications, emphasizing the importance of considering more sophisticated algorithms for analyzing complex and dynamic datasets [91].

The introduction of a new COVID-19 Twitter dataset is a practical contribution, enabling researchers and public health professionals to explore and analyze pandemic-related discussions comprehensively. The development of a web application prototype further underscores the practical applicability of the research, providing a tangible tool for real-time monitoring and analysis of health-related content on Twitter [92,93].

However, it is crucial to acknowledge the limitations inherent in this study. Our sample size of 6,667 tweets may be limited in representing the vast and diverse range of COVID-19-related discussions on Twitter. The tweets were collected for a specific time period, from August 05, 2020, to August 26, 2020, and since then, the public discourse may have evolved significantly. However, we argue that our study aimed to explore specific aspects of COVID-19 discourse, namely, health risks, prevention, symptoms, transmission, and treatment. While a larger sample size would have been ideal, it would also have been a laborious, resource-intensive, and time-consuming task. Our study's focus on specific aspects of the conversations, selected from a large COVID-19 dataset [33] of over 2.2 billion tweet IDs as of now [94], allowed us to delve deeper into the nuances of the topic and identify key themes and patterns that may have been missed in a larger, more general study. Nevertheless, future studies should strive to collect and analyze larger, more diverse datasets to provide a more comprehensive understanding of COVID-19-related discussions on Twitter.

The exclusive training of the classifier on 140-character tweets raises questions about its adaptability to longer-form content and potentially affects prediction accuracy. Additionally, the focus on Twitter data may limit the generalizability of the findings to other social media platforms, as different platforms may exhibit distinct communication patterns and content structures. To address those limitations, one could consider expanding the training dataset to include longer-form content, allowing the classifier to adapt to diverse text lengths and potentially improving prediction accuracy across various content types [95]. Additionally, to enhance the generalizability of the findings, incorporating data from multiple social media platforms and adjusting the model to account for distinct communication patterns and content structures inherent to each platform would provide a more comprehensive understanding of health-related discussions in the broader digital landscape [96].

The challenges associated with accurately detecting health-related information highlight the complexities of analyzing social media data for public health research. The misclassifications and nuances underscore the need for ongoing refinement and adaptation of advanced models to capture the subtleties of user-generated content accurately. In particular, the integration of large language models, such as GPT [97], could offer a promising avenue for advancing the accuracy and efficiency of health-related content classification on social media [98,99]. These models possess the capability to comprehend context, decipher nuanced language, and adapt to varying lengths of text, addressing some of the challenges associated with reported speech and the character limitations of tweets [100]. Leveraging such advanced language models in conjunction with the methodologies presented in this study could potentially enhance the classification accuracy and broaden the applicability of the system across diverse social media platforms and communication styles. To further advance the field, future research should consider exploring the transferability of the model to other social media platforms. Additionally, comparative studies across different regions and demographic groups could provide valuable insights into the variations in health-related discussions on social media [101].

While this research makes significant contributions in analyzing and classifying health-related discussions on Twitter during the COVID-19 pandemic, it also highlights the evolving nature of the digital landscape and the ongoing need for refinement and adaptation in methodologies. The findings contribute not only to the academic discourse but also offer practical tools for public health practitioners and policymakers to monitor and respond to health-related conversations in real-time. This ability to interpret social media data quickly and accurately can enhance public health communication strategies, allowing more targeted and effective dissemination of information. Moreover, our work provides a framework that can be adapted to other health crises, improving preparedness and response capabilities for future pandemics. As the field continues to evolve, future studies should build upon these findings to enhance the effectiveness of utilizing social media data for public health surveillance and intervention strategies.

Furthermore, our approach can be extended to identify bots and other malicious entities in big data analytics. By leveraging the features extracted from text data, we can train machine learning and deep learning models to detect and classify bots, enabling more effective data filtering and cleaning. This application has significant implications for improving data quality and reducing the impact of bots on big data analytics. Future work includes exploring the application of our approach to bot detection in big data analytics, including the development of more advanced machine learning and deep learning models to improve detection accuracy and efficiency.

Nevertheless, the recent changes at Twitter, including the restriction of API access and introduction of paid tiers, pose significant challenges for Twitter-based social media research studies [102]. These changes limit the availability of data, increasing the barrier to entry for researchers and potentially stifling innovation in the field. Furthermore, the costs associated with the paid API tiers may disproportionately affect early-career researchers, students, and researchers from under-resourced institutions, exacerbating existing inequalities in the research community.

9. Conclusion

In this study, we harnessed machine learning algorithms to categorize health-related expressions within COVID-19 tweets. Our primary goal was to classify COVID-19 related tweets into five distinct classes: health risks, prevention, symptoms, transmission, and treatment. We curated a dataset comprising 6,667 tweets and meticulously annotated each one. This dataset underwent a comprehensive data refinement process, encompassing multiple data pre-processing steps. Additionally, we applied three distinct feature extraction techniques. Our study leveraged a combination of seven classical machine learning algorithms, including Decision Tree, Random Forest, Stochastic Gradient Descent, K-nearest Neighbor, Adaboost, Logistic Regression, and Linear SVC, alongside four deep learning algorithms—LSTM, CNN, RNN, and BERT. Among the machine learning models, Stochastic Gradient Descent yielded the highest F1 score of 86.34%, while the deep learning approach saw CNN delivering an impressive F1 score of 90%.

The findings and analyses from this study signify that COVID-19 health-related phrases within prepared datasets can be effectively classified using a spectrum of machine learning and deep learning algorithms. Given the distinct nature of our research, the proposed model could potentially serve as a standardized framework for the classification of COVID-19 discourses within Twitter data. The outcomes of this research hold promise for global healthcare efforts against COVID-19 and offer valuable insights to researchers in this field. Furthermore, the publicly available dataset and web application provide a foundational resource for further academic research and practical application, encouraging continued innovation in the use of social media data for public health surveillance.

However, it is important to acknowledge certain limitations within our study. We employed data from a limited timeframe and did not incorporate the entire available dataset. Our dataset relied on manual labeling, which restricted the volume of labeled data. To address these limitations in future work, we intend to expand the dataset's size significantly. We are also exploring automated dataset labeling approaches to replace manual annotation. Furthermore, we plan to experiment with modified neural networks integrated with transfer learning techniques to enhance the study's robustness, accuracy, and overall outcomes.

CRediT authorship contribution statement

Mahathir Mohammad Bishal: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Md. Rakibul Hassan Chowdory:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Anik Das:** Writing – review & editing, Validation, Methodology, Conceptualization. **Muhammad Ashad Kabir:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The corresponding author, Associate Professor Ashad Kabir, holds the position of Associate Editor at Heliyon Journal.

Data availability statement

The data that support the findings of this study are openly available at https://github.com/Bishal16/COVID19-Health-Related-Data-Classification-Website.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] M. Reveilhac, A. Blanchard, The framing of health technologies on social media by major actors: prominent health issues and covid-related public concerns, Int. J. Inf. Manag. Data Insights 2 (1) (2022) 100068.

[2] D. Schillinger, D. Chittamuru, A.S. Ramírez, From "infodemics" to health promotion: a novel framework for the role of social media in public health, Am. J. Publ. Health 110 (9) (2020) 1393–1396.

[3] J. Liu, T. Singhal, L. Blessing, K.L. Wood, K.H. Lim, Epic: an epidemics corpus of over 20 million relevant tweets, arXiv preprint, arXiv:2006.08369, 2020.

[4] S.R. Rufai, C. Bunce, World leaders' usage of Twitter in response to the covid-19 pandemic: a content analysis, J. Public Health 42 (3) (2020) 510–516.

[5] X. Lin, R. Kishore, Social media-enabled healthcare: a conceptual model of social media affordances, online social support, and health behaviors and outcomes, Technol. Forecast. Soc. Change 166 (2021) 120574.

[6] Statista Research Department, Leading countries based on number of X (formerly Twitter) users as of January 2023, https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/, 2013. (Accessed 10 November 2023).

[7] H. Rosenberg, S. Syed, S. Rezaie, The Twitter pandemic: the critical role of Twitter in the dissemination of medical information and misinformation during the covid-19 pandemic, Can. J. Emerg. Med. 22 (4) (2020) 418–421.

[8] S.Y. Arafat, S. Hakeem, S.K. Kar, R. Singh, A. Shrestha, R. Kabir, Communication during disasters: role in contributing to and prevention of panic buying, in: Panic Buying and Environmental Disasters: Management and Mitigation Approaches, Springer, 2022, pp. 161–175.

[9] E. Chen, K. Lerman, E. Ferrara, Covid-19: the first public coronavirus Twitter dataset, arXiv preprint, arXiv:2003.07372, 2020.

[10] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, B. Liu, Predicting flu trends using Twitter data, in: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2011, pp. 702–707.

[11] E.H. Chan, V. Sahai, C. Conrad, J.S. Brownstein, Using web search query data to monitor Dengue epidemics: a new model for neglected tropical disease surveillance, PLoS Negl. Trop. Dis. 5 (5) (2011) e1206.

[12] C. Chew, G. Eysenbach, Pandemics in the age of Twitter: content analysis of tweets during the 2009 h1n1 outbreak, PLoS ONE 5 (11) (2010) e14118.

[13] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in: Proceedings of the First Workshop on Social Media Analytics, 2010, pp. 115–122.

[14] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (7232) (2009) 1012–1014.

[15] V. Lampos, T. De Bie, N. Cristianini, Flu detector-tracking epidemics on Twitter, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2010, pp. 599–602.

[16] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google flu: traps in big data analysis, Science 343 (6176) (2014) 1203–1205.

[17] C. Alicino, N.L. Bragazzi, V. Faccio, D. Amicizia, D. Panatto, R. Gasparini, G. Icardi, A. Orsi, Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google trends-based query volumes, Infect. Dis. Poverty 4 (1) (2015) 54.

[18] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, N. Ramakrishnan, Misinformation propagation in the age of Twitter, Computer 12 (2014) 90–94.

[19] J. Kalyanam, S. Velupillai, S. Doan, M. Conway, G. Lanckriet, Facts and fabrications about Ebola: a Twitter based study, arXiv preprint, arXiv:1508.02079, 2015.

[20] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, R. Maciejewski, Visualizing social media sentiment in disaster scenarios, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 1211–1215.

[21] M. Odlum, S. Yoon, What can we learn about the Ebola outbreak from tweets?, Am. J. Infect. Control 43 (6) (2015) 563–571.

[22] E. Yom-Tov, Ebola data from the internet: an opportunity for syndromic surveillance or a news event?, in: Proceedings of the 5th International Conference on Digital Health 2015, 2015, pp. 115–119.

[23] B.P. Ehrenstein, F. Hanses, B. Salzberger, Influenza pandemic and professional duty: family or patients first? A survey of hospital employees, BMC Public Health 6 (1) (2006) 1–3.

[24] C. Shen, A. Chen, C. Luo, J. Zhang, B. Feng, W. Liao, et al., Using reports of symptoms and diagnoses on social media to predict covid-19 case counts in mainland China: observational infoveillance study, J. Med. Internet Res. 22 (5) (2020) e19421.

[25] T. Mackey, V. Purushothaman, J. Li, N. Shah, M. Nali, C. Bardier, B. Liang, M. Cai, R. Cuomo, et al., Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with covid-19 on Twitter: retrospective big data infoveillance study, JMIR Public Health Surveill. 6 (2) (2020) e19509.

[26] Z. Chen, J. Guo, Y. Jiang, Y. Shao, High concentration and high dose of disinfectants and antibiotics used during the covid-19 pandemic threaten human health, Environ. Sci. Eur. 33 (1) (2021) 1–4.

[27] A.K. Das, N. Islam, M. Billah, A. Sarker, Covid-19 pandemic and healthcare solid waste management strategy–a mini-review, Sci. Total Environ. (2021) 146220.

[28] WHO, Covid-19 high risk groups, URL, https://www.who.int/westernpacific/emergencies/covid-19/information/high-risk-groups, 2021. (Accessed 6 December 2021).

[29] M.A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique, Covid-19 infection: origin, transmission, and characteristics of human coronaviruses, J. Adv. Res. 24 (2020) 91.

[30] Y.-R. Guo, Q.-D. Cao, Z.-S. Hong, Y.-Y. Tan, S.-D. Chen, H.-J. Jin, K.-S. Tan, D.-Y. Wang, Y. Yan, The origin, transmission and clinical therapies on coronavirus disease 2019 (covid-19) outbreak–an update on the status, Mil. Med. Res. 7 (1) (2020) 1–10.

[31] I. Ghinai, T.D. McPherson, J.C. Hunter, H.L. Kirking, D. Christiansen, K. Joshi, R. Rubin, S. Morales-Estrada, S.R. Black, M. Pacilli, et al., First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) in the USA, Lancet 395 (10230) (2020) 1137–1144.

[32] W.H. Organization, et al., Getting your workplace ready for covid-19: how covid-19 spreads, 19 March 2020, tech. rep., World Health Organization, 2020.

[33] R. Lamsal, Design and analysis of a large-scale covid-19 tweets dataset, Appl. Intell. 51 (5) (2021) 2790–2804.

[34] C.E. Lopez, C. Gallemore, An augmented multilingual Twitter dataset for studying the covid-19 infodemic, Soc. Netw. Anal. Min. 11 (1) (2021) 102.

[35] E. Chen, K. Lerman, E. Ferrara, et al., Tracking social media discourse about the covid-19 pandemic: development of a public coronavirus Twitter data set, JMIR Public Health Surveill. 6 (2) (2020) e19273.

[36] E. Hagg, V.S. Dahinten, L.M. Currie, The emerging use of social media for health-related purposes in low and middle-income countries: a scoping review, Int. J. Med. Inform. 115 (2018) 92–105.

[37] Q.X. Ng, C.E. Yau, Y. Lim, L. Wong, T. Liew, Public sentiment on the global outbreak of monkeypox: an unsupervised machine learning analysis of 352,182 Twitter posts, Publ. Health 213 (2022) 1–4.

[38] A. Khatua, A. Khatua, E. Cambria, A tale of two epidemics: contextual word2vec for classifying Twitter streams during outbreaks, Inf. Process. Manag. 56 (1) (2019) 247–257.

[39] S. Omer, S. Ali, et al., Preventive measures and management of covid-19 in pregnancy, Drugs Ther. Perspect. 36 (6) (2020) 246–249.

[40] I. Ali, O.M. Alharbi, Covid-19: disease, management, treatment, and social impact, Sci. Total Environ. 728 (2020) 138861.

[41] L.-A. Cotfas, C. Delcea, R. Gherai, I. Roxin, Unmasking people's opinions behind mask-wearing during covid-19 pandemic—a Twitter stance analysis, Symmetry 13 (11) (2021) 1995.

[42] M. Al-Ramahi, A. Elnoshokaty, O. El-Gayar, T. Nasralah, A. Wahbeh, Public discourse against masks in the covid-19 era: infodemiology study of Twitter data, JMIR Public Health Surveill. 7 (4) (2021) e26780.

[43] L. He, C. He, T.L. Reynolds, Q. Bai, Y. Huang, C. Li, K. Zheng, Y. Chen, Why do people oppose mask wearing? A comprehensive analysis of us tweets during the covid-19 pandemic, 2021.

[44] C. Doogan, W. Buntine, H. Linger, S. Brunt, Public perceptions and attitudes toward covid-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of Twitter data, J. Med. Internet Res. 22 (9) (2020) e21419.

[45] X. Zhou, J. Menche, A.-L. Barabási, A. Sharma, Human symptoms–disease network, Nat. Commun. 5 (1) (2014) 1–10.

[46] K. Emmett, Nonspecific and atypical presentation of disease in the older patient, Geriatrics 53 (2) (1998) 50–52.

[47] E. Alanazi, A. Alashaikh, S. Alqurashi, A. Alanazi, Identifying and ranking common covid-19 symptoms from tweets in Arabic: content analysis, J. Med. Internet Res. 22 (11) (2020) e21329.

[48] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, T. Zhu, Public discourse and sentiment during the covid 19 pandemic: using latent Dirichlet allocation for topic modeling on Twitter, PLoS ONE 15 (9) (2020) e0239441.

[49] S. Srivastava, M.K. Sarkar, C. Chakraborty, Machine learning approaches for covid-19 sentiment analysis: unveiling the power of bert, in: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2024, pp. 0092–0097.

[50] S. Srivastava, M.K. Sarkar, C. Chakraborty, Sentiment analysis of Twitter data using machine learning: Covid-19 perspective, Int. J. Data Anal. Tech. Strateg. 16 (1) (2024) 1–16.

[51] S. Srivastava, C. Chakraborty, M.K. Sarkar, A graph neural network-based machine learning model for sentiment polarity and behavior identification of covid patients, Int. J. Data Sci. Anal. (2023) 1–10.

[52] Z. Khan, Y. Karataş, A. Ceylan, H. Rahman, Covid-19 and therapeutic drugs repurposing in hand: the need for collaborative efforts, Pharm. Hosp. Clin. 56 (1) (2021) 3–11.

[53] A.A. Mir, S. Rathinam, S. Gul, Public perception of covid-19 vaccines from the digital footprints left on Twitter: analyzing positive, neutral and negative sentiments of twitterati, Libr. Hi Tech (2021).

[54] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D.S. Gherai, F. Tajariol, The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement, IEEE Access 9 (2021) 33203–33223.

[55] Q.X. Ng, S.R. Lim, C.E. Yau, T.M. Liew, Examining the prevailing negative sentiments related to covid-19 vaccination: unsupervised deep learning of Twitter posts over a 16 month period, Vaccines 10 (9) (2022) 1457.

[56] Q.G. To, K.G. To, V.-A.N. Huynh, N.T. Nguyen, D.T. Ngo, S.J. Alley, A.N. Tran, A.N. Tran, N.T. Pham, T.X. Bui, et al., Applying machine learning to identify anti-vaccination tweets during the covid-19 pandemic, Int. J. Environ. Res. Public Health 18 (8) (2021) 4069.

[57] M.A. Weinzierl, S.M. Harabagiu, Automatic detection of covid-19 vaccine misinformation with graph link prediction, J. Biomed. Inform. 124 (2021) 103955.

[58] D. Gerts, C.D. Shelley, N. Parikh, T. Pitts, C.W. Ross, G. Fairchild, N.Y.V. Chavez, A.R. Daughton, "Thought I'd share first" and other conspiracy theory tweets from the covid-19 infodemic: exploratory study, JMIR Public Health Surveill. 7 (4) (2021) e26527.

[59] M. Koziarski, M. Woźniak, B. Krawczyk, Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise, Knowl.-Based Syst. 204 (2020) 106223.

[60] M. Sahare, H. Gupta, A review of multi-class classification for imbalanced data, Int. J. Adv. Comput. Res. 2 (3) (2012) 160.

[61] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: liwc and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54.

[62] H. Ahmed, I. Traore, S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques, in: International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Springer, 2017, pp. 127–138.

[63] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: a survey on identification and mitigation techniques, ACM Trans. Intell. Syst. Technol. 10 (3) (2019) 1–42.

[64] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Syst. Man Cybern. 21 (3) (1991) 660–674.

[65] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.

[66] H. Drucker, R. Schapire, P. Simard, Boosting performance in neural networks, in: Advances in Pattern Recognition Systems Using Neural Network Technologies, World Scientific, 1993, pp. 61–75.

[67] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann, 1996, pp. 148–156.

[68] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[69] L. Bottou, O. Bousquet, 13 the Tradeoffs of Large-Scale Learning, Optimization for Machine Learning, 2011, p. 351.

[70] P. Cunningham, S.J. Delany, k-nearest neighbour classifiers, arXiv preprint, arXiv:2004.04523, 2020.

[71] S. Menard, Applied Logistic Regression Analysis, vol. 106, Sage, 2002.

[72] F. Pérez-Cruz, A. Navia-Vázquez, P.L. Alarcón-Diana, A. Artés-Rodrǵuez, Svc-based equalizer for burst tdma transmissions, Signal Process. 81 (8) (2001) 1681–1693.

[73] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[74] K. O'Shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint, arXiv:1511.08458, 2015.

[75] O.I. Abiodun, A. Jantan, A.E. Omolara, K.V. Dada, N.A. Mohamed, H. Arshad, State-of-the-art in artificial neural network applications: a survey, Heliyon 4 (11) (2018).

[76] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, Phys. D: Nonlinear Phenom. 404 (2020) 132306.

[77] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[78] H. Jelodar, Y. Wang, R. Orji, H. Huang, Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: nlp using lstm recurrent neural network approach, arXiv preprint, arXiv:2004.11695, 2020.

[79] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv preprint, arXiv:1706.03762, 2017.

[81] M.A. Weinzierl, S.M. Harabagiu, Automatic detection of covid-19 vaccine misinformation with graph link prediction, J. Biomed. Inform. 124 (2021) 15, https://doi.org/10.1016/j.jbi.2021.103955.

[82] D. Warman, M.A. Kabir, Covidfakeexplainer: an explainable machine learning based web application for detecting covid-19 fake news, in: 10th IEEE Asia-Pacific Conference on Computer Science and Data Engineering, IEEE, 2023.

[83] O. Hamad, A. Hamdi, S. Hamdi, K. Shaban, Steducov: an explored and benchmarked dataset on stance detection in tweets towards online education during covid-19 pandemic, Big Data Cogn. Comput. 6 (3) (2022) 88.

[84] M.E. Basiri, S. Nemati, M. Abdar, S. Asadi, U.R. Acharrya, A novel fusion-based deep learning model for sentiment analysis of covid-19 tweets, Knowl.-Based Syst. 228 (2021) 21.

[85] F.B. Oliveira, A. Haque, D. Mougouei, S. Evans, J.S. Sichman, M.P. Singh, Investigating the emotional response to covid-19 news on Twitter: a topic modeling and emotion classification approach, IEEE Access 10 (2022) 16883–16897.

[86] A. Sanaullah, A. Das, A. Das, M.A. Kabir, K. Shu, Applications of machine learning for covid-19 misinformation: a systematic review, Soc. Netw. Anal. Min. 12 (1) (2022) 94, https://doi.org/10.1007/s13278-022-00921-9.

[87] A. Signorini, A.M. Segre, P.M. Polgreen, The use of Twitter to track levels of disease activity and public concern in the US during the influenza a h1n1 pandemic, PLoS ONE 6 (5) (2011) e19467.

[88] E.H.-J. Kim, Y.K. Jeong, Y. Kim, K.Y. Kang, M. Song, Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news, J. Inf. Sci. 42 (6) (2016) 763–781.

[89] X. Zhu, S. Wu, D. Miao, Y. Li, Changes in emotion of the Chinese public in regard to the sars period, Soc. Behav. Pers. Int. J. 36 (4) (2008) 447–454.

[90] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: a new paradigm to machine learning, Arch. Comput. Methods Eng. 27 (2020) 1071–1092.

[91] F. Barbieri, J. Camacho-Collados, L.E. Anke, L. Neves, Tweeteval: unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1644–1650.

[92] P.P. Morita, I. Zakir Hussain, J. Kaur, M. Lotto, Z.A. Butt, Tweeting for health using real-time mining and artificial intelligence–based analytics: design and development of a big data ecosystem for detecting and analyzing misinformation on Twitter, J. Med. Internet Res. 25 (2023) e44356.

[93] L. Sinnenberg, A.M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, R.M. Merchant, Twitter as a tool for health research: a systematic review, Am. J. Publ. Health 107 (1) (2017) e1–e8.

[94] R. Lamsal, Coronavirus (covid-19) tweets dataset (2020), https://doi.org/10.21227/781w-ef42, https://dx.doi.org/10.21227/781w-ef42.

[95] K. Jones, J.R. Nurse, S. Li, Are you Robert or Roberta? Deceiving online authorship attribution models using neural text generators, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, 2022, pp. 429–440.

[96] A. Khattar, S. Quadri, Generalization of convolutional network to domain adaptation network for classification of disaster images on Twitter, Multimed. Tools Appl. 81 (21) (2022) 30437–30464.

[97] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[98] P.E. Christensen, S. Yadav, S. Belongie, Prompt, condition, and generate: classification of unsupported claims with in-context learning, arXiv preprint, arXiv: 2309.10359, 2023.

[99] P. Törnberg, Chatgpt-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning, arXiv preprint, arXiv:2304.06588, 2023.

[100] K.S. Kalyan, A. Rajasekharan, S. Sangeetha, Ammu: a survey of transformer-based biomedical pretrained language models, J. Biomed. Inform. 126 (2022) 103982.

[101] Y. Li, X. Wang, X. Lin, M. Hajli, Seeking and sharing health information on social media: a net valence model and cross-cultural comparison, Technol. Forecast. Soc. Change 126 (2018) 28–40.

[102] B.I. Davidson, D. Wischerath, D. Racek, D.A. Parry, E. Godwin, J. Hinds, D. van der Linden, J.F. Roscoe, L. Ayravainen, 2023, Social media apis: a quiet threat to the advancement of science.