Evolving Systems
https://doi.org/10.1007/s12530-023-09558-1

ORIGINAL PAPER

Sensecor: A framework for COVID-19 variants severity classification and symptoms detection

T. K. Balaji[1] · **Annushree Bablani**[1] · **S. R. Sreeja**[1] · **Hemant Misra**[2]

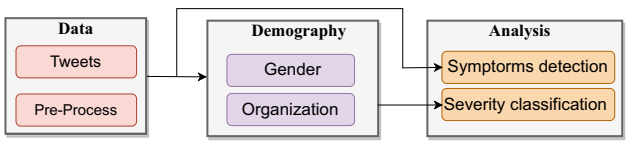
Received: 20 April 2023 / Accepted: 23 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Social media platforms, such as Twitter, allow users to share their thoughts and opinions on various topics, including pandemics like COVID-19. This data can be used to analyze public sentiment and assess the severity of different Corona variants. In this study, a new framework called SENSECOR is proposed to perform opinion mining on Twitter data. SENSECOR uses natural language processing techniques to identify the severity levels and most common symptoms associated with Corona variants. The dataset includes over 160,000 tweets related to COVID-19. SENSECOR is evaluated against several deep learning models, including RoBERTa, BERT, ELECTRA, XLNet, LSTM, and BiLSTM, as well as traditional machine learning models. The results show that SENSECOR achieves the highest accuracy rate of 91%, surpassing all other methods. This suggests that SENSECOR is a promising tool for assessing the severity of Corona variants and identifying the most common associated symptoms.

Graphical abstract

Graphical abstract of SENSECOR framework



Keywords Machine learning · Tweet analysis · Opinion mining · COVID-19 classification

1 Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has had a profound impact on the world, leading to millions of deaths, disrupted economies, and strained healthcare systems (Mohan et al. 2022). The pandemic has also posed multifaceted challenges to governments, businesses, and individuals worldwide. As the pandemic progressed, the virus evolved, giving rise to a variety of variants, such as Beta, Delta, and Omicron (Markov et al. 2023). These variants have introduced new challenges and uncertainties, such as increased transmissibility and reduced vaccine effectiveness.

Social media platforms have played a vital role in the COVID-19 pandemic, providing channels for sharing information, connecting people, and coordinating responses. However, the emergence of the Omicron variant led to

✉ Annushree Bablani
annushree.bablani@iiits.in

T. K. Balaji
balaji.tk@iiits.in

S. R. Sreeja
sreeja.sr@iiits.in

Hemant Misra
hemant.misra@swiggy.in

[1] Computer Science and Engineering Department, Indian Institute of Information Technology Sri City, Sri City, AP, India

[2] Applied research, Swiggy, Bangalore, Karnataka, India

Published online: 23 December 2023

widespread anxiety, partly due to the rapid spread of misinformation and infodemics on social media.

The COVID-19 pandemic witnessed an unprecedented surge in social media usage, creating an invaluable opportunity for researchers to tap into extensive user-generated data. Social media platforms effectively transformed into virtual laboratories for scrutinizing public sentiment, information propagation, and behavioural reactions to the crisis. Prior research on past epidemics highlights the potential of social media analysis in shaping public health interventions, bolstering disease surveillance, and improving risk communication. Many studies have analyzed the behaviour of COVID-19 spread using diverse datasets (clinical and non-clinical). These studies employed different NLP, ML, and DL techniques for analysis. While clinical data offers crucial insights into the virus's behaviour and patient outcomes, non-clinical sources like social media content provide an insightful lens into public sentiment and emerging pandemic-related trends.

To the best of current knowledge, research concerning the detection of COVID severity or symptoms has primarily relied on clinical data. Social media data has been utilized for various disease analyses, including depression intensity estimation, anxious depression prediction, mental disorder forecasting, pulmonary disease analysis, and even suicidal activity prediction. A substantial research gap exists in utilizing this data source for evaluating the severity and symptoms of COVID-19 variants. To bridge the gap, this study adopts NLP techniques to analyse COVID-19 data sourced from Twitter.

Twitter is a popular micro-blogging platform where people share their opinions in various formats, such as text, images, links, and videos. Researchers can use the Twitter API to analyze data from a wide range of Twitter users (Cantini et al. 2021). This type of analysis can be especially helpful for governments during pandemics, as it can help them to make informed decisions. For example, Twitter mood analysis techniques (Naseem et al. 2021) can be used to identify people who are feeling depressed or anxious. This information can help officials in providing timely assistance to people in specific geographical regions who need it most. The flexibility and richness of Twitter data make it a valuable tool for assessing the severity of COVID-19 variants. This motivates us to use Twitter data for this research.

A new framework called SENSECOR is proposed to perform Twitter-based opinion analysis to SENSE the CORona. It is designed to identify the severity of COVID-19 variants and highlight its most prevalent symptoms by analyzing user tweets. This framework is developed by using a hybrid approach that combines rule-based and machine learning (ML) methods. Previous studies have primarily focused on assessing COVID-19 disease severity or identifying symptoms using clinical and laboratory data. For instance, in(Benito-León et al. 2021), researchers applied unsupervised ML algorithms to clinical and laboratory reports to ascertain COVID-19 severity. However, existing studies have typically concentrated on either severity or symptom identification, neglecting the importance of considering both aspects when dealing with potential pandemics. Hence, this research employs NLP techniques to examine social media data and develop the SENSECOR framework. This framework facilitates the quick assessment of symptoms and severity associated with any variant of COVID-19.

The key contributions of this paper are:

1. **Extensive Data Collection**: An extensive dataset comprising over 0.16 million tweets related to the Omicron variant of COVID-19 in India is collected. These tweets are gathered during a crucial period, the peak of daily cases (February 3-9, 2022). This dataset forms the foundation for subsequent analysis.
2. **SENSECOR Framework Development**: Introduced the SENSECOR framework, a hybrid approach for classifying tweets based on the severity level of the Omicron variant. Additionally, it identifies common symptoms associated with the variant. This framework holds promise for advancing the understanding of the impact of COVID-19 variants and enhancing public health responses.
3. **Demographic Analysis**: Demographic analysis of the collected tweets is conducted. This analysis offers valuable insights into the characteristics of the users who posted these tweets. Understanding users' demographics can provide contextual information that enriches the interpretation of the data.
4. **Performance Evaluation**: A comprehensive statistical performance analysis of the ML and deep learning algorithms and the SENSECOR framework is presented. This evaluation assesses the framework's effectiveness and reliability in classifying severity and identifying symptoms.

These contributions collectively contribute to a deeper comprehension of the Omicron variant's implications by harnessing social media data, employing NLP and ML techniques, and providing demographic insights. This research aids in more informed decision-making in public health and pandemic management.

The rest of this paper is structured as follows: Sect. 2 discusses the literature review. Section 3 explains the methods used to analyze the data. Section 4 presents the results and discussions of this study. Finally, Sect. 5 gives the conclusion of this study.

2 Literature review

Numerous studies on COVID-19 employ NLP, ML, or DL techniques, leveraging a diverse range of data sources, including clinical and non-clinical data, to obtain critical insights and results. These analyses provide valuable information for understanding the virus's impact, predicting outcomes, and informing public health decisions.

2.1 Studies employing clinical data

Researchers utilize various methods to extract valuable insights from non-image clinical data, such as electronic health records (EHRs), enabling risk assessment, patient stratification, optimized treatment plans and more. For example, (Bhatia et al. 2022; Kukar et al. 2021) presented ML-based COVID-19 diagnosis based on routine blood parameters. In these, the first study has conducted research on data specific to India and in the second study, data specific to Slovenia. In both studies, the XG Boost model has been applied to predict the diagnosis. Authors (Zoabi et al. 2021) applied ML technique to predict COVID-19's existence and symptoms in a patient. RT-PCR test and symptoms based questionnaire data has been used to train the model.In (Lee et al. 2021) proposed a model that can predict whether a patient infected with COVID-19 will develop severe outcomes using only patients historical EHR before hospital admission. The model uses recurrent neural networks(RNN) to predict a risk score that represents the probability for a patient to progress into severe status. The model achieved an area under the receiver operating characteristic curve (AUROC) of 0.846 in predicting patients' outcomes, averaging over 5-fold cross-validation.

Additionally, clinical data modalities such as X-rays, CT scans, and ultrasounds are commonly used for image-based diagnostics because they are reliable(Dance et al. 2014). However, manually identifying COVID-19 from these clinical scans is a time-consuming, labour-intensive process and prone to human error. As a result, deep learning (DL) in radiography imaging has emerged as a consistent and effective approach to achieve significant improvements. For instance,(Zhao et al. 2021) Used CT images and applied transfer learning to analyse the impact of COVID-19. The model was pre-trained on ImageNet21k and demonstrated strong generalizability with an impressive accuracy of 99.2% in detecting COVID-19 cases. Using the Grad-CAM visualization technique, the authors enhanced the model's interpretability, aiding clinical doctors in manual screening. In another study, (Alshazly et al. 2021) presented deep learning techniques for automated COVID-19 detection using chest CT images. The visual explanations for model decisions, feature visualizations, and accurate localization of COVID-19-associated regions are mentioned in the study. (Ulloa et al. 2022) conducted a retrospective study in Ontario, Canada, comparing Omicron and Delta variants from late 2021 and found that Omicron cases had a 60% lower risk of hospitalization. Although the fatality risk is also lower for Omicron, it wasn't statistically significant. The study suggests that Omicron may be less severe, but factors like high vaccination rates need consideration. Further research is needed to validate these findings and assess long-term implications. The summary of studies using clinical data on COVID-19 is presented in the Table 1.

2.2 Studies employing social media data

Social media platforms serve as repositories of non-clinical data that can be leveraged to gain valuable insights. By applying NLP techniques, researchers can analyze vast amounts of text data from platforms like Twitter, gaining insights into public sentiment, monitoring health behaviours, and detecting potential outbreaks. For example,(Mathur et al. 2020) discussed the assessment of emotional states expressed on Twitter during the COVID-19 pandemic. The study used the NRC Word-Emotion Association Lexicon (EmoLex) to classify tweets into fundamental emotions, providing insights into public sentiment. The study highlighted the potential usefulness of this analysis for authorities and policymakers in addressing mental health and social well-being. Another study by(Ogbuokiri et al. 2022) focused on analyzing the variations in sentiment towards COVID-19 vaccines in three major South African cities using geo-tagged Twitter posts. The study found significant correlations between sentiment intensity, vaccine-related topics, and key metrics such as vaccination rates and COVID-19 cases. The research highlighted the importance of local context in shaping sentiment and the potential of Twitter data for informed health policy and decision-making in community-based infectious disease discussions.

(Bhat et al. 2020) analyzed the sentiment of over 80,000 tweets related to COVID-19. Authors found that 51.97% of the tweets expressed positive sentiment, 34.05% expressed neutral sentiment, and 13.96% expressed negative sentiment. The authors concluded that sentiment analysis can be used to track public opinion on pandemics and to inform public health interventions.(Li et al. 2020) explored the application of lexical and ML methods in the fight against COVID-19. NLP is used for sentiment analysis to categorize social media content on Weibo as a case study. The study involved analyzing 367,462 posts to extract relevant features. These features are then used to train ML algorithms, including support vector machine (SVM), naive Bayes (NB), and random

Table 1 Summary of COVID-19 applicational studies leveraging clinical data

Study	Application	Model	Performance	Approach	Dataset	Data type
Zhao et al. (2021)	COVID-19 detection	CNN	Auc−99.2%	DL	CT scan images and ImageNet21k Tan and Le (2021)	Image
Alshazly et al. (2021)	COVID-19 detection in patient	ResNet101 DenseNet201	Auc−99.4% Auc−92.9%	DL	SARS-CoV-2 CT Scan dataset Soares et al. (2020), COVID19-CT dataset He et al. (2020)	Image
Dastider et al. (2021)	Predicting COVID-19 patient severity condition	Hybrid CNN-LSTM	Auc-79%	DL	Lung ultrasound data Roy et al. (2020)	Image
Zoabi et al. (2021)	Predicting COVID-19 and its symptoms in a patient	Gradient boost	auROC-82%	ML	RT-PCR and Questionaire	Clinical text
Bhatia et al. (2022)	Severity and mortality prediction models to triage COVID-19 patients	XGBoost Chen and Guestrin (2016)	auROC-92%	ML	Blood parameters	Clinical text
Kukar et al. (2021)	COVID-19 diagnosis by routine blood tests	XGBoost	auROC-97%	ML	Blood parameters	Clinical text
Lee et al. (2021)	Severity prediction for COVID-19 patients	RNN	auROC- 84.6%	DL	EHR	Clinical text
Ulloa et al. (2022)	Severity identification between Omicron and Delta variants of COVID-19	whole-genome sequencing (WGS) S-gene target failure(SGTF)	–	–	COVID-19 patients hospital records	Clinical text

forest (RF). The goal is to classify unlabeled data based on previously labeled data. Specifically, the COVID-19 related information is categorized into seven distinct types of situational data, including emotional, perceptual, and affiliation factors. The most promising results are obtained using the high-performing RF classifier.

(Nuser et al. 2022) proposed a hybrid model that combines convolutional neural networks (CNN) and long short-term memory (LSTM) to analyze user sentiment towards the COVID-19 vaccine. This model is applied to a dataset of 13,190 tweets and achieved an accuracy of 83%, which outperformed the accuracy of both CNN and LSTM algorithms individually. The research contribution of the paper is to assist medical staff and the government in analyzing people's reactions towards the COVID-19 vaccine. (Zhou et al. 2021) discussed the impact of the COVID-19 pandemic on community depression dynamics using user-generated content on Twitter. The authors propose a new approach based on multimodal features from tweets and term frequency-inverse document frequency (TF-IDF) with logistic regression (LR) to build depression classification models. Multimodal features capture depression cues from emotion, topic, and domain-specific perspectives. The study results showed that people became more depressed after the outbreak of COVID-19. The measures implemented by the government, such as the state lockdown, also increased depression levels.

The summary of studies using non-clinical data, such as social media data on COVID-19, is presented in table 2.

2.3 Studies on diseases symptoms extraction from text

Multiple researchers have applied NLP and ML techniques to extract symptoms from clinical notes within EHR's for clinical research purposes, as well as from non-clinical texts like social media data. Various NLP models have been employed to detect and identify signs and symptoms of disease viz-a-viz heart failure(Vijayakrishnan et al. 2014), mental illness(Jackson et al. 2017), and urine infections (Gundlapalli et al. 2017).

Combining rule-based and ML/DL methods in hybrid approaches provides a thorough strategy for COVID-19 analysis. Rule-based systems organize and extract data from unstructured text, while ML and DL models deal with intricate patterns and predictions. This multi-modal approach offers a comprehensive view of the pandemic, assists in risk assessment, and aids in evidence-based decision-making for clinical and public health purposes. For instance, (Luo et al. 2021) developed a framework for detecting symptoms of COVID-19 within tweets. Authors initially trained a rule-based named entity recognition (NER) system for symptom identification using deep learning (DL) models with

Table 2 Summary of COVID-19 applicational studies leveraging social media data

Study	Application	Model	Performance	Approach	Dataset
(Mathur et al. 2020)	Emotional analysis on COVID-19	EMOLEX	-	Lexicon / Rule based	Twitter
(Ogbuokiri et al. 2022)	Sentiments analysis toward COVID-19 vaccines in South African cities	Vader, LDA, SVM, Decision Tree, Naive Bayes, LR	NB (0.68), LR (0.75), SVM (0.70), DT (0.62)	Hybrid (Lexicon/ Rule-based + ML)	Twitter
(Bhat et al. 2020)	Sentiment analysis of SM response on the COVID-19 Outbreak	–	–	Lexicon / Rule based	Twitter
(Li et al. 2020)	Exploring the Impacts of COVID-19 on People's Mental Health to Assist Policy and Provide Timely Services to Infected Populations	RF	Auc-65%	Hybrid Based Models	Weibo
(Nuser et al. 2022)	Sentiment analysis of COVID-19 vaccine	CNN+LSTM	Auc-83%	DL	Twitter
(Priyadarshini et al. 2022)	Sentiments and psychology of twitter users during COVID-19	Polarity based	–	Lexicon / Rule based	Twitter
(Zhou et al. 2021)	Depression detection due to COVID-19	TF-IDF+LR	Auc-90%	ML	Twitter

Electronic Health Record (EHR) data. Authors subsequently applied this NER system to tweets, enabling the detection of COVID-19 symptoms using social media data.

Hence, the studies on COVID-19 severity and symptom detection have primarily relied on clinical data. This means that there is a need for more research using non-clinical data like social media data to study the different strains of the COVID-19 virus severity and symptoms. This study fills this gap by employing NLP techniques on Twitter data.

3 Methodology

The emergence of new variants of COVID-19 has posed a significant challenge to researchers seeking to understand its impact. The most common way to identify symptoms is through collecting data from patients admitted to healthcare centres. But, this method is time-consuming and requires many hours of human effort. To address this challenge, researchers have turned to social media data to study the novel variants of COVID-19. For this study, Twitter is used as a source of data. To achieve the goal of identifying common symptoms and classifying the severity level of the variant, a new framework called SENSECOR is proposed to sense the corona. SENSECOR utilizes a hybrid approach, combining rule-based and ML methods, to detect symptoms and classify severity levels. Each tweet is categorized into one of four severity levels: *Severe*, *Mild*, *No-symptoms*, or *Undefined* to represent the COVID-19 variant's severity accurately.

3.1 Data collection

A total of 204,729 tweets related to the Omicron variant of the Coronavirus are collected for this study using search keywords including *#omicron, #omicron symptoms, #omicronVariant, and Omicron*. The data collection is facilitated by using a package called TWARC,[1] a command-line utility and Python library designed for collecting tweets and returning them as JSON objects. TWARC offers user-friendly commands to acquire tweets from Twitter in various ways,

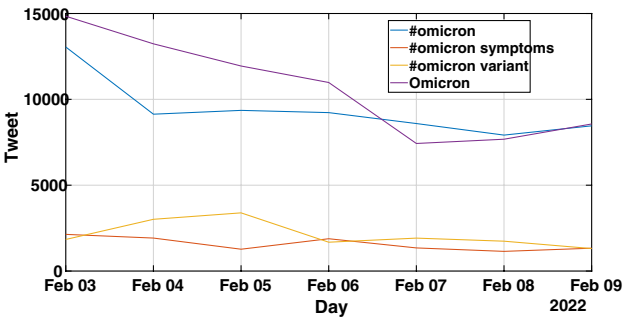


Fig. 1 Day-wise data collection

[1] https://github.com/DocNow/twarc.

making it a versatile tool for data retrieval. The day-wise data collection process for the COVID-19 Omicron dataset is visualized in Fig. 1.

3.2 Data pre-processing

The primary step of text analysis is to prepare the data so that accurate information can be extracted from it; this process is called data pre-processing. The text pre-processing will improve the classification algorithm's performances. Authors in (Jianqiang and Xiaolin 2017) explained the effects of text pre-processing methods used for sentiment classification. The authors presented many valuable ways to improve the accuracy while extracting sentiment, like removing stopwords and URLs from text.

To ensure the study's accuracy, the collected tweets are meticulously processed to include unique tweets in the analysis. Every tweet has a unique number called TweetID. The duplicate tweets are removed using this TweetID. Eliminating duplicate information is a crucial step in improving the quality of the study. In addition, converting all upper case letters to lowercase, removing links, punctuation, and numbers to optimize information extraction. The resulting dataset comprised of 1,60,413 tweets, which are used to develop the SENSECOR framework.

If a Tweet has a slight change in wording or is retweeted with different accompanying text, using TweetID for duplicate elimination may not be sufficient. In such cases, a similarity metric like Levenshtein distance (Levenshtein et al. 1966) as presented in Eq. (1), which measures the minimum number of single-character edits needed to transform one string into another, can be useful. A threshold distance can be defined to deduplicate tweets using Levenshtein distance, beyond which the two tweets are considered distinct. For example, in this study, a threshold value of 6 is considered, meaning tweets with a Levenshtein distance of 6 or less are considered duplicates.

$$L_{a,b}(m,n) = \begin{cases} \max(m,n) & \text{if } \min(m,n) = 0 \\ \min \begin{cases} L_{a,b}(m-1,n)+1 \\ L_{a,b}(m,n-1)+1 \\ L_{a,b}(m-1,n-1)+1_{(a_m \neq b_n)} \end{cases} & \text{Otherwise} \end{cases} \tag{1}$$

In the Eq. (1) the distance between two strings a and b can be defined by using a function $L_{a,b}\,(m,n)$ where m and n represent the prefix length of string a and b respectively. Consider two strings, "kitten" and "sitting" and the goal is to find the Levenshtein distance between them.

a = "kitten"
b = "sitting"
m = 6 (length of "kitten")
n = 7 (length of "sitting")

$L_{a,b}(m-1,n)+1$ represents deleting a character from string a.
$L_{a,b}(m,n-1)+1$ represents inserting a character into string a.
$L_{a,b}(m-1,n-1)+1_{(a_m \neq b_n)}$ represents substituting a character in string a with a character from string b. The term $1_{(a_m \neq b_n)}$ is an indicator function that returns 1 if a_m and b_n are not equal (i.e. if characters at the current positions are different), and 0 otherwise.

The algorithm considers all these possibilities and chooses the minimum value among them as the Levenshtein distance. It iterates through the characters of both strings, calculating the distance incrementally until it reaches the final value. In this example, the Levenshtein distance between "kitten" and "sitting" would be calculated using this recursive algorithm to be 3 because you can transform "kitten" into "sitting" by substituting 'k' with 's', deleting 'e', and adding 'i' and 'g'.

3.3 Data Annotation

To evaluate the effectiveness of the proposed framework, a manually annotated dataset[2] comprising 4,000 tweets has been created. These manually labelled tweets are curated and selected from the collected dataset. The data is categorised into four classes, i.e. *No-symptoms, Mild, Severe*, and *Undefined*, and each class comprises 1,000 tweets. Three volunteers with post-graduate level and knowledge in NLP are selected for manual labelling. Clear guidelines, inter-annotator agreements (IAA), policies and criteria to ensure accurate and consistent labelling are provided to the annotators. Regular checks for inter-rater reliability among the volunteers are conducted to ensure consistency in labelling.

Guidelines for annotating tweets for severity classification:

1. Read the entire tweet carefully and understand its context before deciding.
2. Determine the overall severity of the illness based on the symptoms mentioned in the tweet. The severity can be classified as *severe*, *No-symptoms*, *Mild*.
3. If the tweet does not clearly indicate the severity of the illness or if it contains conflicting or unclear information, classify it as *Undefined*.
4. Use objective criteria to classify the severity of the illness, such as the type and duration of symptoms, the

[2] https://github.com/balajitk7/sampledata.

need for medical intervention, or any other relevant factors.
5. Avoid making assumptions or interpretations not supported by the tweet.
6. If you are unsure about the severity of the illness, seek clarification from a senior annotator.
7. Be consistent in your classification decisions and follow the guidelines closely.
8. Maintain confidentiality and avoid sharing any personal or sensitive information in the tweets.

To ensure a reliable inter-annotator agreement (IAA) among annotators when classifying the severity of tweets, the following steps are taken:

1. Provide clear and concise guidelines. The guidelines define what constitutes *Severe*, *No-symptoms*, *Mild*, and *Undefined* illness, and provide examples for each category.
2. Conduct training sessions. These sessions ensure that all annotators are familiar with the guidelines and are consistent in their interpretations.
3. Use multiple annotators. Each tweet is classified independently by multiple annotators. This allows for any annotation discrepancies to be identified and resolved through discussion.
4. Continuously monitor IAA. The IAA is monitored throughout the annotation process. This allows for feedback to be provided to annotators to improve the quality of their annotations.

By following those guidelines, IAA policies are implemented to ensure the annotations are accurate, consistent, and high-quality. This high level of consistency and reliability in the annotation process has ultimately improved the accuracy of the severity classification.

3.4 Proposed method

The study has two primary objectives: firstly, to identify the prevalent symptoms associated with the Omicron variant of COVID-19, and secondly, to evaluate the disease's severity. Each tweet will be allocated to one of four specified categories to accomplish this. The resulting classification report will provide an overview of the severity levels related to the new Omicron variant.

The proposed SENSECOR framework is a hybrid approach that combines rule-based and ML techniques. The SENSECOR uses a rule-based approach in two phases:

1. Symptom detection phase: To identify symptoms in COVID-19 tweets.
2. Severity classification phase: To provide the initial severity class labels for COVID-19 tweet data.

The labeled data with predefined severity classes, such as *"Severe", "Mild"*, or *"No-symptoms"* is unavailable. The rule-based system efficiently identifies and categorizes tweets based on predetermined criteria and patterns. The labeled tweets are then used to train ML models, which can capture complex patterns and relationships within the data. This improves the accuracy of predicting the severity of COVID-19 cases. The framework provides the flexibility of employing either rule-based classification or a DL-based approach to label tweet severity.

In essence, SENSECOR synergizes the strengths of rule-based and ML methodologies, improving its ability to classify and understand the severity of COVID-19 cases. A detailed architecture of the proposed framework is shown in Fig.2. The proposed framework architecture takes tweets as input and identifies the demographics of the tweeters using the M3model, such as their gender and whether they are individuals or organizations. Then, the framework analyzes the tweets to identify the symptoms of the Omicron variant. Finally, the framework classifies the tweets based on the severity of the Omicron variant that the users mentioned in their tweets. For the severity analysis, organizational tweets are omitted because they do not belong to individuals. The severity classification phase can use either a rule-based or ML-based model to classify the severity of the tweets.

3.4.1 Symptom detection phase

To identify symptoms of COVID-19 using the framework, it is essential to provide the system with an initial set of known symptoms associated with the virus. The framework then employs a rule-based approach to analyze a dataset of COVID-19 related tweets, looking for the presence of these symptoms in the text of each tweet. This methodology allows the framework to systematically scan and identify instances where individuals on social media discuss or report COVID-19 symptoms. The base symptoms for the framework are adopted based on the studies (Gallo Marin et al. 2021) and (CDC 2021). The symptoms include fever or chills, cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting, and diarrhoea.

To carry out the symptom identification task, the N-grams technique based on NLP is employed. N-grams refer to continuous sequences of words, symbols, or tokens and are defined as the adjacent sequences of words in a text (Brown et al. 1992). Unigrams, bigrams, and trigrams are generated for the entire dataset, along with its respective frequencies at the word level. Symptom-wise frequency is then calculated

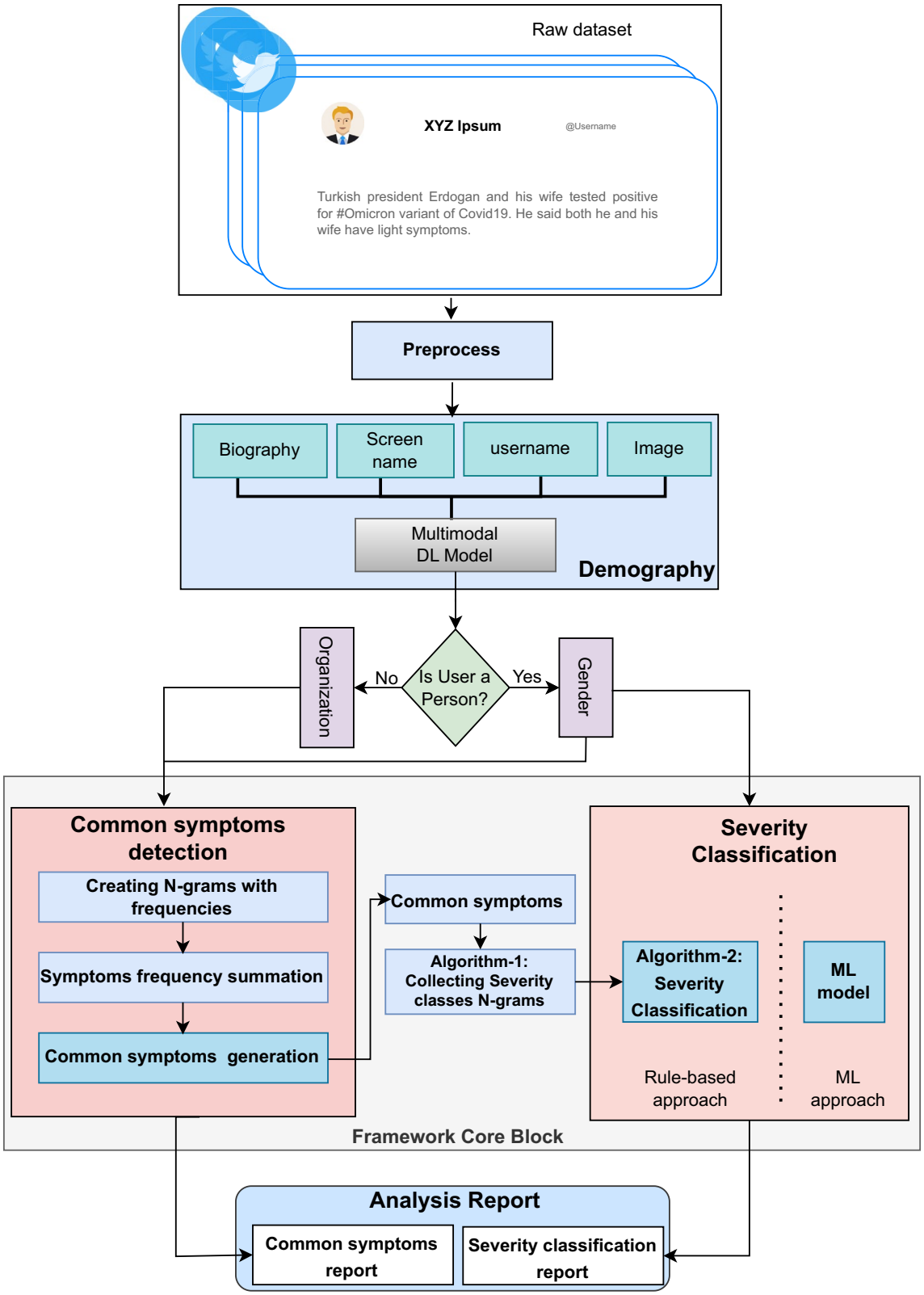


Fig. 2 Detailed flow diagram of SENSECOR framework

Table 3 Synonyms of words used for this study

Word	Synonyms
Mild	*'mild', 'low', 'light', 'modest', 'minimal', 'fewer', and 'less'.*
Severe	*'vital', 'terrible', 'nasty', 'strong', 'heavy', 'severe', 'very bad', 'critical', 'chest pain', 'difficulty breathing', 'very high', 'hard', 'very heavy', 'extreme', 'difficulty in breathing', 'breathing issues', and 'shortness of breath', 'hospitalized', 'icu', 'intensive care', 'ventilator', 'respiratory distress', 'respiratory failure', 'death', 'dying', 'fatal', 'organ failure', 'heart attack', 'multi organ failure'*
No-Symptoms	*'asymptomatic', 'no symptoms', 'feeling good', 'healthy', 'fine'*
No	*'won't', 'not', 'Neither', 'unlikely', 'dont have', 'none', 'didnt', 'didn't', 'no', 'null', 'zero', 'haven't' and 'without'.*
Symptoms	*'sign', 'effect', and 'symptoms'.*

by searching for the unigram, bigram, and trigram word combinations for each symptom and summing its respective frequencies. For example, to determine the frequency of the symptom tired', the N-gram list is searched for all possible synonyms of tired', including tired', tiredness', weak', and weakness', and its individual word frequencies are added together to calculate the overall frequency of the tired' symptom. Hence, if the frequencies of tired', tiredness', weak', and weakness' are 16, 10, 23, and 7, respectively, the total frequency of the tired' symptom would be 56. Similarly, the frequencies of the CDC-recognized symptoms are determined from the dataset using the N-gram technique. The most frequently occurring symptoms are then compiled into a list, which will be used in the second task of the SENSECOR framework.

3.4.2 Severity classification phase

The second task of this study is to identify the severity level of the Omicron variant from the tweets. The severity level is classified into four classes: *"Severe", "Mild", "No-symptoms",* and *"Undefined"* using the N-grams. For example, the N-grams considered for *"Severe"* are *'vital', 'terrible', 'nasty', 'strong', 'heavy', 'severe', 'very bad', 'critical', 'chest pain', 'difficulty breathing', 'heavy fever', 'very high', 'hard', 'very heavy', 'extreme', 'difficulty in breathing', 'breathing issues', 'shortness of breathe'*, and so on. If a tweet possesses characteristics from multiple classes during classification, it can cause ambiguity in the tweet's classification. As a result, these tweets are designated as *"Undefined"*. For instance, the tweet

"Got the PCR results in for our three children from their testing this past Friday. Two tested positive, one had mild symptoms, and the other was asymptomatic. One tested negative and never had symptoms. That's the one that usually gets severe fevers"

This tweet describes various symptoms without clearly indicating severity, resulting in its classification as *"Undefined"*. The *"Undefined"* label is dropped from the analysis to give accurate severity of illness.

The tweet classification process extends to include demographic insights, such as determining the user's gender using the M3model. The M3model is a multi-modal deep neural architecture designed for gender identification and user type classification (person, organization, or bot) based on Twitter profiles. This model leverages both the user's profile picture and username to make these determinations (Wang et al. 2019). It operates as a multi-modal architecture, using Densenet (Huang et al. 2017) for image classification and LSTM (Hochreiter and Schmidhuber 1997) for text classification.

Tweets from non-organizational accounts are then subcategorized based on gender, with labels such as 'male' and 'female' applied. This approach enhances the granularity of the tweet classification process by incorporating demographic attributes.

To assign the label *"Mild"* to a given tweet, the framework examines the tweet's content for the presence of relevant N-grams or lexicons, such as 'mild symptoms', 'mild fever', 'less symptoms', 'fewer symptoms', 'low fever', 'affected low fever', 'less severe', 'light symptoms', 'light fever', 'minimal symptoms', 'fewer symptoms', 'not severe', and similar expressions. If the match is found, the tweet is labelled *"Mild"*. The synonyms used to collect N-grams for this study are collected based on the symptoms detected. The synonyms used in this study are presented in Table 3.

The process of selecting these synonyms involved multiple methods. Firstly, cosine similarity is used to calculate the similarity between words in the four predefined classes on the COVID-19 tweet dataset. Words with the highest

similarity scores were recognized as potential synonyms for each class. Secondly, user tweet analysis is used to identify similar words by analyzing user tweets that revolve around discussions of the COVID-19 Omicron variant. Additionally, frequently used words from the top ten most tweeted messages related to Omicron severity and a curated list of terms from dictionaries are considered. This thorough approach to word selection is employed to enhance the framework's precision in classification, ultimately aiming to reduce the number of tweets categorized as *"Undefined"*.

Two rule-based algorithms are used to classify the severity of each tweet. The First algorithm gathers N-grams associated with each class, while the second algorithm is tasked with processing the tweets for classification. This classification process utilizes the class-specific N-grams obtained from the first algorithm as input.

3.4.3 Algorithm-1: Collecting N-grams

The algorithm 1 is designed to collect N-grams from a set of tweets (ψ) to help classify the severity of COVID-19 symptoms. It uses predefined sets of synonyms for the classes *"Mild", "Severe", and "No-symptoms"* as well as a set of common symptoms (β) and a set of negation words (no_{grams}).

Algorithm 1 Algorithm for collecting N-grams for severity classification

Begin

Input: ψ, τ, β, M_{syn}, S_{syn}, NS_{syn}, no_{grams}

// ψ is, set of all tweets
// τ is a tweet in set ψ
// β is a set of top ten common symptoms
// $M_{syn}, S_{syn}, NS_{syn}$ are sets with synonyms for *Mild, Severe, and No-symptoms*

Function *getLabel-Ngrams* (γ_{syn} , β)**:**
    $\theta_{grms} = \gamma_{syn} \times \beta^*$
    return θ_{grms}

// θ_{grams} is N-grams set for given label
M_{grms}=getLabel-Ngrams(M_{syn}, β)
S_{grms}=getLabel-Ngrams(S_{syn}, β)
NS_{grms}=getLabel-Ngrams(no_{grams}, β)+NS_{syn}

// M_{grms} is set of N-grams for *Mild* class
// S_{grms} is set of N-grams for *Severe* class
// NS_{grms} is N-grams set of *No-symptoms* class

End

Initializing sets:

- β is a set of common symptoms that is selected from the Symptoms detection phase.
- M_{syn} is a set of synonyms for *"Mild"*.
- S_{syn} is a set of synonyms for *"Severe"* symptoms.
- NS_{syn} is a set of synonyms for *"No-symptoms"*.
- no_{grams} is a set of negation words.

Define the function getLabel-Ngrams():

- This function takes two parameters: γ_{syn} (synonyms for a label) and β (common symptoms).
- It calculates θ_{grms}, which is the set of N-grams of a required class, which is generated by combining synonyms with common symptoms using $\gamma_{syn} \times \beta^*$.
- For example, if getLabel-Ngrams(M_{syn}, β) called, it will return N-grams for the class *"Mild"* like 'mild fever', 'slight cough'

Generate N-grams sets for classes:

- M_{grms} is the set of N-grams for "Mild" symptoms, generated by calling getLabel-Ngrams(M_{syn}, β).
- S_{grms} is the set of N-grams for "Severe" symptoms, generated by calling getLabel-Ngrams(S_{syn}, β).
- NS_{grms} is the set of N-grams for "No-symptoms" generated by calling getLabel-Ngrams(no_{grams}, β) and adding NS_{syn}.

The algorithm calls the function getLabel-Ngrams( ) three times, once for each severity level. The results of these calls are stored in the variables M_{grms}, S_{grms}, and NS_{grms}. These variables contain the N-grams for the mild, severe, and no symptoms severity levels, respectively. Finally, The outputs of the algorithm 1 i.e. the three sets of N-grams: M_{grms}, S_{grms}, and NS_{grms} are used as inputs to algorithm2.

Here are some examples of N-grams in the set M_{grms} (*Mild* severity level): mild fever, moderate cough, not bad sore throat, few symptoms, and feeling better.

Some examples of N-grams in the set S_{grms} (*Severe* severity level): severe fever, difficulty breathing, chest pain, hospitalized, and critical condition.

Some examples of N-grams in these NS_{grms} (*No-symptoms* severity level): asymptomatic, no symptoms, feeling good, healthy, and fine.

The algorithm gives dataset-rich N-grams to classify the given text sentence into one of the four specified classes.

3.4.4 Algorithm-2: Severity classification

Algorithm 2 is designed for classifying the severity of COVID-19 symptoms based on the content of a given tweet

(τ) using a predefined set of N-grams associated with different severity levels: *"Mild", "Severe", "No-symptoms"* and *"Undefined"*. The pre requisite for the algorithm is M_{grams}, S_{grams}, and NS_{grams} as an input to classify τ. It examines each τ in ψ and checks the N-grams with different patterns to determine the tweet's classification. If the tweet matches predefined word sequences patterns, then it will be classified with the corresponding class label.

Assuming M_{grams} set of *"Mild"* symptoms having ['slight fever', 'headache'] words in the set. S_{grams} set of *"Severe"* symptoms having ["severe cough", "difficulty breathing"] in it. no_{grams} is N-grams for negation set having words ['no','not']. Lets consider an example tweet τ as "I have a slight fever and a headache". In this, the tweet mentions "slight fever" and "headache" which match the N-grams in M_{grams} (associated with *"Mild"* symptoms). There are no negation words, and no words from S_{grams} are present. Therefore, the algorithm assigns the severity label Sv_l as *"Mild"* to the τ.

Additionally, the algorithm is able to process sentences with inverted severity representations. For example, when presented with a tweet like, "I tested positive for omicron, but I'm not experiencing any severe symptoms" the algorithm correctly assigns it the label *"Mild"* It achieves this even when the sentence does not explicitly contain words or phrases that are commonly associated with *"Mild"* symptoms. Lines 9 and 15 of Algorithm2 handle the case of inverted severity representation in the tweet. The algorithm checks if any words from the set M_{grms}, or S_{grms} appear after a word from the set no_{grms}. If so, the severity is shifted from *"Severe"* to *"Mild"* or *''Mild"* to *"Severe"*.

Algorithm 2 Algorithm for severity classification

```
Begin
    Input: τ, no_grams, M_grams, S_grams, NS_grams
    Output: Sv_l
                                        // Sv_l output label
6   for τ != ∅ do
7       if τ has any (M_grms) then
8           for words in τ do
                // handling inverted severity sentences
9           if words is any(no_grams) + any(S_grams) then
10              Sv_l is 'Mild'
                goto step-6
11          Sv_l is "Mild"  // labelled as ''Mild''
12          goto step-6
13      else if τ is any (S_grms) then
14          for word in τ do
                // handling inverted severity sentences
15          if words is any(no_grams) + any(M_grams) then
16              Sv_l is "Severe"
                goto step-6
17          Sv_l = "Severe"          // labelled as 'Severe'
18      else if τ is any (NS_grms) then
19          Sv_l is "No-symptoms"    // labelled as 'No-symptoms'
20          goto step-6
21      else
22          Sv_l is "Undefined"      // labelled as 'undefined'
23      Write Sv_l
24  return to 6
25  End
```

When building an ML model, the framework labelled dataset is used to train it. The development data is set aside to fine-tune the model's hyper-parameters, which are adjustable settings that can affect the model's performance. By adjusting the hyper-parameters using the development data, the model's accuracy can be improved, resulting in more accurate predictions when it is deployed in real-world scenarios. In this study, for training deep learning, the AdamW optimizer is used with a learning rate of 0.00005 for BERT-based models and 0.001 for LSTM models. The baseline methods used in this study are:

Table 4 Sample classification results generated by the proposed SENSECOR framework

S.No	Tweet	Class
1	Turkish president Erdogan and his wife tested positive for #Omicron variant of Covid19. He said both he and His wife have **light symptoms**.	Mild
2	@KimCrayton1 I was tested positive with Omicron, feeling **severe fever, uncomfortable in breathing**. Was admitted in hospital 3 days ago.	Severe
3	31 days and still tested positive of Covid without **zero symptoms no fevers no chilling nothing**.	No-Symptoms

RoBERTa: A large language model pre-trained on a massive dataset of text and code (Liu et al. 2019).

ELECTRA: A language model that is trained to distinguish between real and generated text. (Clark et al. 2020)

XLNet: A language model that is trained on a masked language modelling task (Yang et al. 2019).

LSTM: A recurrent neural network that is used to process sequential data (Hochreiter and Schmidhuber 1997).

BiLSTM: A bidirectional LSTM that can process sequential data in both directions (Graves and Schmidhuber 2005).

Random forest: A ML algorithm that builds multiple decision trees to make predictions (Ho 1995).

Naive Bayes: A ML algorithm that assumes that the features of a data point are independent of each other (McCallum and Nigam 1998).

Support vector machine: A ML algorithm that finds the best hyperplane to separate two classes of data points (Cortes and Vapnik 1995).

The sample of tweet classifications using the SENSECOR framework is presented in Table 4.

4 Results and discussions

The framework has identified the topmost common symptoms of Omicron and also provides valuable insights into the diverse manifestations of this COVID-19 variant. Cold/flu-like symptoms are the most prevalent, accounting for 33.86% of cases. Fever is also a common symptom, affecting 30.02% of people with COVID-19 Omicron. Other symptoms such as tiredness, cough, sore throat, headache, and more exhibit varying but notable percentages, underlining the diverse range of manifestations in COVID-19 patients. While some symptoms like chest pain are relatively rare at 0.44%, the data illustrates the multifaceted nature of COVID-19 symptoms, which can assist healthcare professionals in diagnosis and monitoring. The topmost Omicron symptoms identified by the framework are listed in Table 5.

Table 5 The most common symptoms identified by the framework

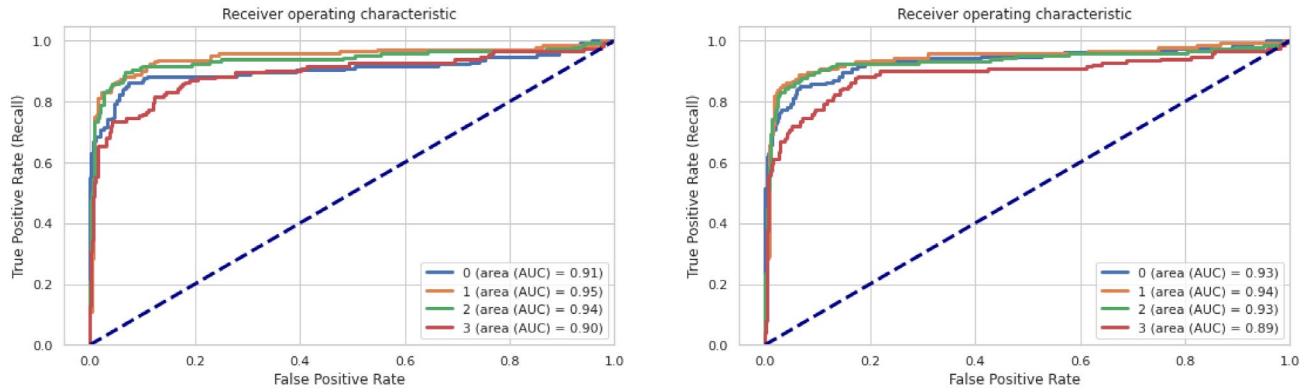
Symptom	Share%
Cold/Flu	33.86
Fever	30.02
Tiredness	9.90
Eyes	3.68
Cough	3.49
Sore throat	3.33
Headache	3.09
Running nose	2.49
Taste loss	2.44
Breathing	2.24
Fatigue	2.16
Smell loss	1.31
Skin rash	1.20
Chest pain	0.44

The performance of the rule-based severity detection models of the SENSECOR is evaluated against ML models, including BERT, ELECTRA, RoBERTa, XLNet, BiLSTM, as well as traditional classifiers like SVM, Naïve Bayes, and Random Forest. The manually labeled dataset is used for testing the models performances. The performance assessment results for various severity classification models are presented in Table 6. The scores for the *"Undefined"* class have been excluded from the table, as they are not utilized in severity analysis.

In this study, it is observed that the deep learning model BERT ELECTRA (Clark et al. 2020) outperformed other extended models of BERT (Devlin et al. 2018). The study showed that the RoBERTa model achieved an accuracy of 82%, slightly lower than the BERT ELECTRA model. On the other hand, XLNet achieved an accuracy of 82%, which is better than RoBERTa but lower than ELECTRA. Compared to the BiLSTM model, both RoBERTa and XLNet models performed better with an accuracy of 82%. The BiLSTM model is a sequence processing model that comprises two LSTMs (Hochreiter and Schmidhuber 1997), one for processing input in one direction and the other for processing it in the opposite direction. The SENSECOR framework achieves the highest accuracy of 91%, followed by BERT ELECTRA at 83%. RoBERTa attains an accuracy of 82%, while the remaining models have accuracies ranging from 76% to 81%. These results indicate that SENSECOR and BERT ELECTRA exhibit good accuracy scores among all

Table 6 Performance evaluation of models: presenting F-score, Precision, and Recall for each class along with the model's accuracy

Model	F-score			Precision			Recall			Accuracy
	No symptom	Mild	Severe	No symptom	Mild	Severe	No symptom	Mild	Severe	
SENSECOR	0.89	0.92	0.92	0.85	0.94	0.93	0.88	0.92	0.93	**0.91**
ELECTRA	0.83	0.85	0.94	0.85	0.83	0.86	0.84	0.82	0.86	0.83
RoBERTa	0.80	0.83	0.82	0.82	0.89	0.87	0.83	0.87	0.78	0.82
SVM	0.83	0.85	0.85	0.81	0.88	0.86	0.84	0.82	0.84	0.82
XLNet	0.81	0.87	0.77	0.83	0.86	0.85	0.84	0.82	0.84	0.82
Random forest	0.74	0.86	0.80	0.86	0.83	0.87	0.80	0.84	0.87	0.81
Naïve Bayes	0.80	0.84	0.87	0.74	0.86	0.88	0.86	0.83	0.87	0.81
BiLSTM	0.76	0.77	0.81	0.77	0.81	0.78	0.75	0.74	0.85	0.77
LSTM	0.81	0.78	0.73	0.84	0.81	0.67	0.77	0.76	0.81	0.76



(a) Naïve Bayes (b) SVM

Fig. 3 Multiclass one-vs-rest (OVR) ROC curves for the ML models. Here class 0 (No symptoms),class 1(Mild), class 2(Severe), and class 3(Undefined)

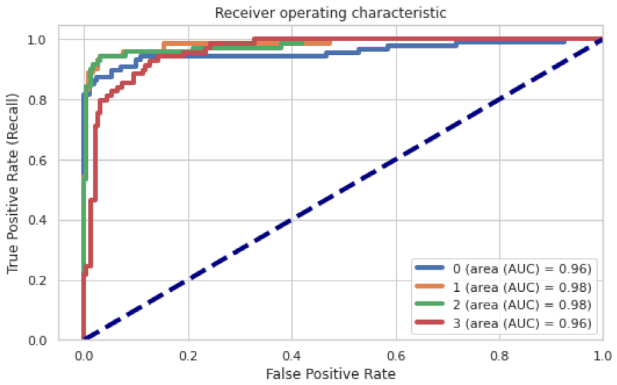
models, while the others exhibit slightly lower performance on the given dataset in terms of accuracy.

The ROC response for the models is presented, with ROC curves generated for class 0 (No symptoms), class 1 (Mild), class 2 (Severe), and class 3 (Undefined). The ROC curve of the trained models is shown in Figs. 3 and 4. The SENSECOR framework severity classification phase is initially developed using a rule-based approach. As a result, it does not generate a Receiver Operating Characteristic (ROC) curve, which is typically associated with ML models and illustrates the trade-off between true positive and false positive rates. This approach ensures that the framework classifies severity without the need for probabilistic thresholds or ML models, making it a reliable tool for severity detection using social media data.

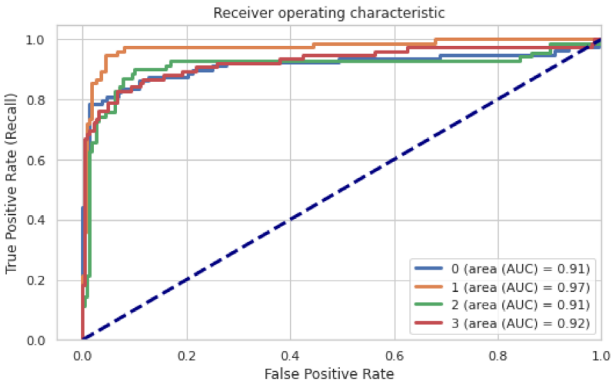
The proposed framework can also identify the gender, organization, or non-organization of the tweets. This information can be used better to understand the distribution of tweets from different groups and to identify potential sources of bias. The dataset used in this study contains 68% tweets from individuals and 32% tweets from non-individuals or organizations. This suggests that the majority of tweets about the Omicron variant are coming from individuals.

The pie chart in Fig. 5a shows the distribution of tweets from individuals and organizations. As you can see, individuals account for the majority of tweets, while organizations account for a smaller minority. Figure 5b, Fig. 5c, and d show the Omicron variant severity distribution for females, males, and both males and females, respectively. As you can see, males are more likely to report severe symptoms than females. This is consistent with the study of (Bechmann et al. 2022) shown that men are more likely to experience severe illness from COVID-19 than women.

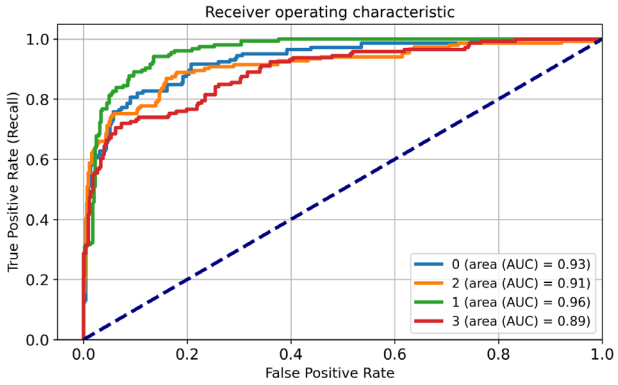
Organizational tweets are typically from news channels, groups, and other sources. These tweets are not always assertive tweets from individuals. For example, a news channel might tweet about a study that found that the Omicron variant is more likely to cause severe symptoms in females. This tweet would not be considered an assertive tweet because it



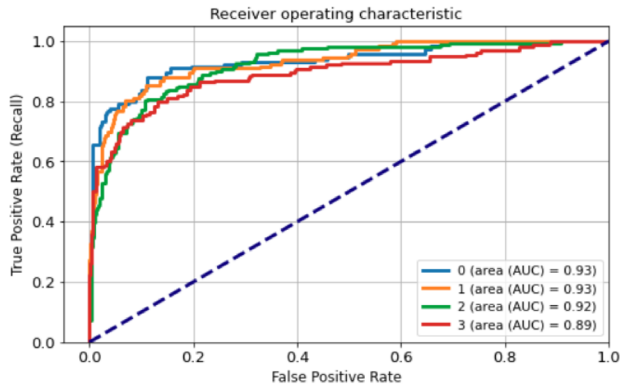
(a) BERT ELECTRA



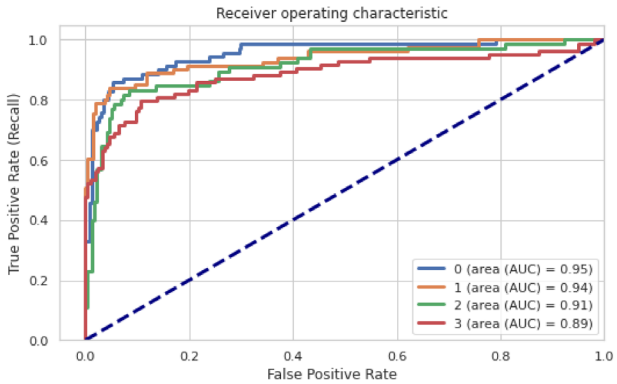
(b) RoBERTa



(c) BiLSTM



(d) LSTM



(e) XLNet

Fig. 4 Multiclass one-vs-rest (OVR) ROC curves for the trained deep learning models. Here class 0 (No symptoms), class 1(Mild), class 2(Severe), and class 3(Undefined)

is simply reporting on the findings of a study. Hence, such tweets are removed from the analysis.

Overall, the framework can be used to gain valuable insights into the distribution of tweets from different groups and the severity of the Omicron variant in different populations.

4.1 Discussion on performance evaluation

The effectiveness of models in classifying data is evaluated using a confusion matrix. This matrix comprises four computations that include true positives (tp), which represent correctly identified class samples. True negatives (tn) indicate the

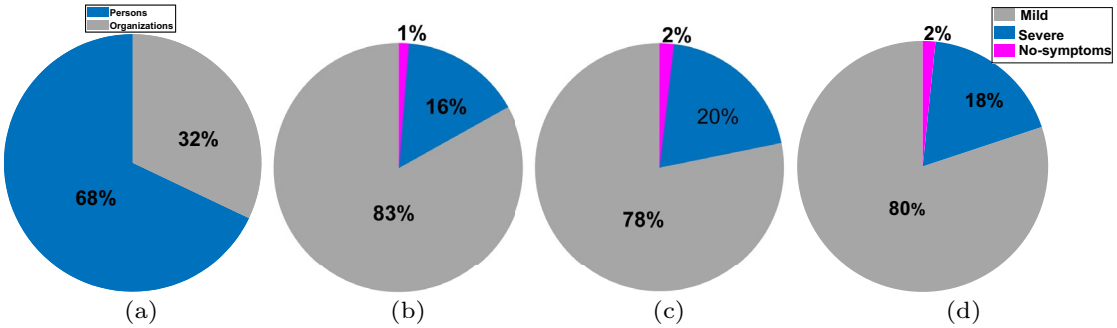


Fig. 5 Pie chart for **a** Participation of Persons with organizations and Severity classification of **b** Female, **c** Male, **d** All male and female

presence of a sample that is correctly not linked to the intended class. False positives (fp) occur when a sample is incorrectly predicted, resulting in a Type 1 error. Finally, false negatives (fn) occur when a sample is not recognized as belonging to the class it represents.

The performance measurements of multi-class classifications are calculated using multiple mathematical equations given from Eq. (2) to Eq. (8). These metrics assess the performance across multiple classes, denoted as C_i, where tp_i are true positive for C_i, and fp_i is false positive, fn_i is false negative for class i, and tn_i is true negative for class i respectively. The equations account for various classes, and the resulting indices are averaged using either the micro M or macro μ approach, incorporating a constant l from i. The weightage given to recall compared to precision is influenced by the value of Υ. Where Υ is chosen such that recall is considered Υ times as important as precision. Specifically, Υ is selected to reflect that recall is considered Υ times as significant as precision in the evaluation process.

The multiclass F-score can be calculated using Eq. (8). Equation (9) and Eq. (10) are used to calculate a ROC curve (receiver operating characteristic curve) for showing the classification performance of models in a graph format.

$$Accuracy = \frac{\sum_{i=1}^{l} \frac{(tp_i+tn_i)}{(tp_i+fn_i+fp_i+tn_i)}}{l} \tag{2}$$

$$Precision_{\mu} = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fp_i)} \tag{3}$$

$$Recall_{\mu} = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fn_i)} \tag{4}$$

$$Fscore_{\mu} = \frac{((\Upsilon^2 + 1)Precision_{\mu} * Recall_{\mu})}{(\Upsilon^2 Precision_{\mu} + Recall_{\mu})}e \tag{5}$$

$$Precision_M = \frac{\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}}{l} \tag{6}$$

$$Recall_M = \frac{\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}}{l} \tag{7}$$

$$Fscore_M = \frac{((\Upsilon^2 + 1)Precision_M Recall_M)}{(\Upsilon^2 Precision_M + Recall_M)} \tag{8}$$

$$TPR = \frac{tn}{tp + fn} \tag{9}$$

$$FPR = \frac{fp}{fp + tn} \tag{10}$$

The degree of steepness in the ROC curve is used to enhance the true positive rate while decreasing the false positive rate. Hence, the performance evaluation of different classifiers on the severity dataset shows that the classifiers have given satisfactory performance in detecting the severity.

The severity classification phase of the SENSECOR framework has concluded that the Omicron variant of COVID-19 exhibits mild characteristics. To confirm this finding, a cross-reference is performed with studies relying on conventional data sources, including patient records and surveys. According to the proposed study's results, 80% of the classified tweets indicated mild symptoms, 18% exhibited severe symptoms, and 2% reported no symptoms. The tweets falling under the undefined category are excluded from the analysis, as they did not discuss severity. The wordcloud for each severity class is presented in Fig. 6.

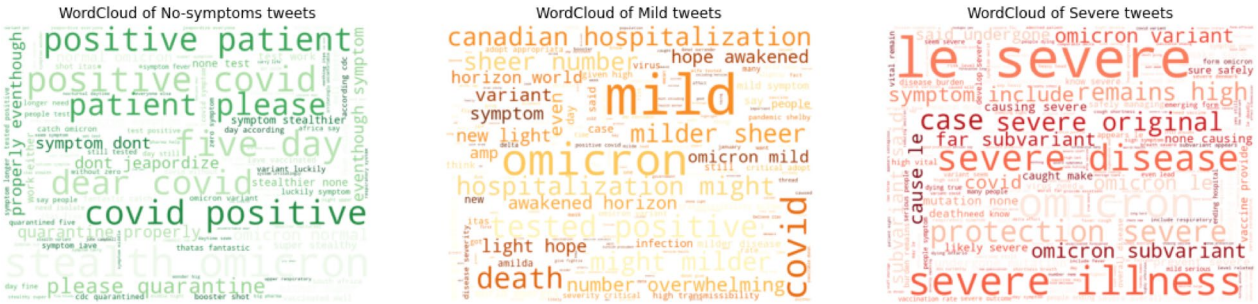


Fig. 6 Word cloud of each class

4.2 Comparative study

The validity of the COVID-19 Omicron severity results produced by the proposed framework has been confirmed by comparing the results with other studies conducted by researchers. These studies, which relied on patient hospital records rather than social media data, encompass references such as studies of (Wrenn et al. 2022; Nealon and Cowling 2022; Maslo et al. 2022; Mayr et al. 2022) and others.

Authors (Wrenn et al. 2022) discovered that infections attributed to the Omicron variant resulted in significantly lower morbidity, including reduced hospital admissions and the need for oxygen supplementation, as well as substantially lower mortality rates compared to those caused by the Delta variant. Authors in (Nealon and Cowling 2022) studied the COVID-19 Omicron variant using laboratory and genomic data of hospital surveillance data linked to COVID-19 cases. The authors confirmed that the Omicron variant may cause less severe illness. The authors in (Maslo et al. 2022) compared the characteristics and outcomes of hospitalized patients in South Africa during the Omicron wave of COVID-19 to previous waves of the pandemic. The authors analyzed data from 5 hospitals in the Western Cape Province of South Africa. Authors in this study suggest that the Omicron variant of COVID-19 may be less severe regarding mortality. The authors in (Mayr et al. 2022) conducted a retrospective analysis using a matched cohort of US patients to compare the disease severity of subjects infected during the SARS-CoV-2 Omicron and Delta predominant periods. The results showed that during the Omicron period, a higher percentage of patients had mild disease, and a lower percentage had moderate or severe disease than the Delta predominant period.

There is a prevailing belief among users that the Omicron variant of COVID-19 is comparatively less severe than its predecessor, the Delta variant as per the study (Christensen et al. 2022). While the Delta variant has been linked to severe symptoms like shortness of breath, fatalities, black fungus infections, and intense coughing, the common symptoms associated with Omicron, including fever, runny nose, cough, and cold, appear to be milder in nature. Reports also suggest that Omicron has resulted in significantly fewer fatalities when compared to the Delta variant, as highlighted in studies like (Johnson 2022; Ulloa et al. 2022). Moreover, data gathered from various countries indicates that Omicron carries a reduced risk of hospitalization when contrasted with Delta, as outlined in (Nyberg et al. 2022).

These findings are encouraging, suggesting that the Omicron variant is less severe than initially thought. Hence the framework is consistent with the results of other studies. However, it's important to note that despite its seemingly milder symptomatology and lower severity in terms of hospitalization and fatalities, the World Health Organization has identified that Omicron exhibits a notably higher growth rate and spreads with exceptional rapidity when compared to the Delta variant.

The SENSECOR framework is well-suited for healthcare applications, especially those involves rapidly evolving diseases like COVID-19. It can effectively process unlabeled data, particularly useful in healthcare where labeled data may be limited. It is customizable to specific diseases or populations. However, the SENSECOR framework also has some limitations. It relies on a limited set of synonyms from top users, which may limit its accuracy. It may struggle to generalize to new or unseen data sources or to variations in how people express symptoms. It does not learn from data, so it may need frequent rule updates to stay accurate. Despite these limitations, the SENSECOR framework is a promising tool for agile healthcare applications, where rapid deployment and explainable decision-making are essential.

Overall, the evidence suggests that the SENSECOR framework is an accurate and reliable method for classifying the severity of diseases. The framework can be applied for real-time public health monitoring, early warnings, public awareness campaigns, research, misinformation detection, epidemiological studies, global health monitoring, risk assessment, and government policy adaptation. Hospitals and healthcare facilities can use the framework to anticipate and allocate resources more efficiently.

5 Conclusion

Social media data is a valuable resource for understanding public sentiment and behaviour related to various topics, including health and disease. Additionally, social media data is readily accessible and could be analyzed quickly and at a large scale using ML and natural language processing (NLP) techniques. These characteristics made social media data a cost-effective and valuable tool for monitoring and responding to pandemics and other public health crises.

Proposed SENSECOR framework that uses the power of ML and NLP to scrutinize large quantities of social media data. The framework senses the severity level and most common symptoms of the Omicron variant. The severity level conclusions of the Omicron variant drawn from the social media data by the SENSECOR framework matched the studies that drew results on the Omicron variant using traditional data sources such as patient records or surveys. It was found that the Omicron variant is less severe than initially anticipated. The proposed framework identified that cold, flu, fever, and tiredness are highly discussed symptoms among users. The proposed framework processed the results in a shorter amount of time. Moreover, it could analyze any unlabelled data and conclusions. The proposed framework has generated the results with 91% accuracy on the test dataset. BERT ELECTRA follows next.

With a little update to the framework rules, it could be a valuable resource for public health officials and researchers in addressing future pandemics and epidemics. The framework approach presented a promising opportunity to understand the severity and identify symptoms of not only the COVID-19 pandemic but also other public health crises.

Funding Funding This work was supported by the Indian institute of information technology Sri City, India.

Data Availability Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest The authors have not disclosed any competing interests.

Ethical approval This article does not contain any studies conducted by any of the authors on human participants or animals.

Informed consent Informed consent is obtained from all individual participants included in the study.

References

Alshazly H, Linse C, Barth E et al (2021) Explainable covid-19 detection using chest ct scans and deep learning. Sensors 21(2):455

Bechmann N, Barthel A, Schedl A et al (2022) Sexual dimorphism in covid-19: potential clinical and public health implications. Lancet Diabetes Endocrinol 10(3):221–230

Benito-León J, Del Castillo MD, Estirado A et al (2021) Using unsupervised machine learning to identify age-and sex-independent severity subgroups among patients with covid-19: Observational longitudinal study. J Med Internet Res 23(5):e25988

Bhat M, Qadri M, Kundroo M et al (2020) Sentiment analysis of social media response on the covid19 outbreak. Brain Behav Immun 87:136

Bhatia S, Makhija Y, Jayaswal S et al (2022) Severity and mortality prediction models to triage indian covid-19 patients. PLOS Digital Health 1(3):e0000020

Brown PF, Della Pietra VJ, Desouza PV et al (1992) Class-based n-gram models of natural language. Comput Linguist 18(4):467–480

Cantini R, Marozzo F, Bruno G et al (2021) Learning sentence-to-hashtags semantic mapping for hashtag recommendation on microblogs. ACM Trans Knowl Discov Data (TKDD) 16(2):1–26

CDC (2021) Symptoms of covid-19. https://wwwcdcgov/coronavirus/2019-ncov/symptoms-testing/symptomshtml

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Christensen PA, Olsen RJ, Long SW et al (2022) Signals of significantly increased vaccine breakthrough, decreased hospitalization rates, and less severe disease in patients with coronavirus disease 2019 caused by the omicron variant of severe acute respiratory syndrome coronavirus 2 in houston, texas. Am J Pathol 192(4):642–652

Clark K, Luong MT, Le QV, et al (2020) Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

Dance D, Christofides S, Maidment A, et al (2014) Diagnostic radiology physics: A handbook for teachers and students. endorsed by: American association of physicists in medicine, asia-oceania federation of organizations for medical physics, european federation of organisations for medical physics

Dastider AG, Sadik F, Fattah SA (2021) An integrated autoencoder-based hybrid cnn-lstm model for covid-19 severity prediction from lung ultrasound. Comput Biol Med 132:104296

Devlin J, Chang MW, Lee K, et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Gallo Marin B, Aghagoli G, Lavine K et al (2021) Predictors of covid-19 severity: a literature review. Rev Med Virol 31(1):1–10

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw 18(5–6):602–610

Gundlapalli AV, Divita G, Redd A et al (2017) Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing. J Biomed Inform 71:S39–S45

He X, Yang X, Zhang S, et al (2020) Sample-efficient deep learning for covid-19 diagnosis based on ct scans. medrxiv pp 2020–04

Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, IEEE, pp 278–282

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
Huang G, Liu Z, Van Der Maaten L, et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
Jackson RG, Patel R, Jayatilleke N et al (2017) Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project. BMJ Open 7(1):e012012
Jianqiang Z, Xiaolin G (2017) Comparison research on text preprocessing methods on twitter sentiment analysis. IEEE Access 5:2870–2879
Johnson AG (2022) Covid-19 incidence and death rates among unvaccinated and fully vaccinated adults with and without booster doses during periods of delta and omicron variant emergence-25 us jurisdictions, april 4–december 25, 2021. MMWR Morbidity and mortality weekly report 71
Kukar M, Gunčar G, Vovko T et al (2021) Covid-19 diagnosis by routine blood tests using machine learning. Sci Rep 11(1):10738
Lee J, Ta C, Kim JH et al (2021) Severity prediction for covid-19 patients via recurrent neural networks. AMIA Summit Trans Sci Proceed 2021:374
Levenshtein VI, et al (1966) Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady, Soviet Union, pp 707–710
Li S, Wang Y, Xue J et al (2020) The impact of covid-19 epidemic declaration on psychological consequences: a study on active weibo users. Int J Environ Res Public Health 17(6):2032
Liu Y, Ott M, Goyal N, et al (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
Luo X, Gandhi P, Storey S et al (2021) A deep language model for symptom extraction from clinical text and its application to extract covid-19 symptoms from social media. IEEE J Biomed Health Inform 26(4):1737–1748
Markov PV, Ghafari M, Beer M et al (2023) The evolution of sars-cov-2. Nat Rev Microbiol 21(6):361–379
Maslo C, Friedland R, Toubkin M et al (2022) Characteristics and outcomes of hospitalized patients in south africa during the covid-19 omicron wave compared with previous waves. JAMA 327(6):583–584
Mathur A, Kubde P, Vaidya S (2020) Emotional analysis using twitter data during pandemic situation: Covid-19. In: 2020 5th international conference on communication and electronics systems (ICCES), IEEE, pp 845–848
Mayr FB, Talisa VB, Castro AD et al (2022) Covid-19 disease severity in us veterans infected during omicron and delta variant predominant periods. Nat Commun 13(1):3647
McCallum A, Nigam K et al (1998) A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorization. Madison, WI, pp 41–48
Mohan S, Solanki AK, Taluja HK et al (2022) Predicting the impact of the third wave of covid-19 in india using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. Comput Biol Med 144:105354
Naseem U, Razzak I, Khushi M et al (2021) Covidsenti: a large-scale benchmark twitter data set for covid-19 sentiment analysis. IEEE Tran Comput Soc Syst 8(4):1003–1015
Nealon J, Cowling BJ (2022) Omicron severity: milder but not mild. Lancet 399(10323):412–3
Nuser M, Alsukhni E, Saifan A et al (2022) Sentiment analysis of covid-19 vaccine with deep learning. J Theor Appl Inf Technol 100(12):4513–4521
Nyberg T, Ferguson NM, Nash SG et al (2022) Comparative analysis of the risks of hospitalisation and death associated with sars-cov-2 omicron (b.11.529) and delta (b1617.2) variants in england: a cohort study. Lancet 399(10332):1303–12
Ogbuokiri B, Ahmadi A, Bragazzi NL et al (2022) Public sentiments toward covid-19 vaccines in south african cities: An analysis of twitter posts. Front Public Health 10:987376
Priyadarshini I, Mohanty P, Kumar R et al (2022) A study on the sentiments and psychology of twitter users during covid-19 lockdown period. Multimedia Tools and Applications 81(19):27009–27031
Roy S, Menapace W, Oei S et al (2020) Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. IEEE Trans Med Imaging 39(8):2676–2687
Soares E, Angelov P, Biaso S, et al (2020) Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. MedRxiv pp 2020–04
Tan M, Le Q (2021) Efficientnetv2: Smaller models and faster training. In: International conference on machine learning, PMLR, pp 10096–10106
Ulloa AC, Buchan SA, Daneman N et al (2022) Estimates of sars-cov-2 omicron variant severity in ontario, canada. JAMA 327(13):1286–1288
Vijayakrishnan R, Steinhubl SR, Ng K et al (2014) Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. J Cardiac Fail 20(7):459–464
Wang Z, Hale S, Adelani DI, et al (2019) Demographic inference and representative population estimates from multilingual social media data. In: The world wide web conference, pp 2056–2067
Wrenn JO, Pakala SB, Vestal G et al (2022) Covid-19 severity from omicron and delta sars-cov-2 variants. Influenza Other Respir Viruses 16(5):832–836
Yang Z, Dai Z, Yang Y, et al (2019) Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neural Inform Process Syst 32
Zhao W, Jiang W, Qiu X (2021) Deep learning for covid-19 detection based on ct images. Sci Rep 11(1):14353
Zhou J, Zogan H, Yang S et al (2021) Detecting community depression dynamics due to covid-19 pandemic in Australia. IEEE Trans Comput Soc Syst 8(4):982–991
Zoabi Y, Deri-Rozov S, Shomron N (2021) Machine learning-based prediction of covid-19 diagnosis based on symptoms. Npj Digital Med 4(1):3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.