An improved machine learning technique for identify informative COVID-19 tweets

Sreejagadeesh Malla1 · P. J. A. Alphonse1

Received: 11 October 2021 / Revised: 28 December 2021 / Accepted: 4 June 2022
© The Author(s) under exclusive licence to The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2022

Abstract Twitter users are increasingly using the platform to share information, particularly in the case of disease outbreaks such as COVID-19. It’s difficult to find informative tweets about coronavirus on Twitter. Recognizing tweets associated with disease evaluation in social media is a critical endeavour because it is a subset of associated data. Existing works rely solely on subject identification, vocabulary construction, idea extraction, polarity detection, descriptive Terms, and disease-related statistical characteristics, resulting in a lack of precision in detecting tweet content. To solve this problem, this study used parts of speech tags and high-resolution graphics. To address this issue, we proposed an IPSH (Informative POS statistical High Frequency) model for predicting COVID-19 tweet content that incorporates parts of speech tags and high-frequency words as features into the existing machine learning model. The model was found to be more efficient when compared to baseline machine learning models using the Twitter COVID-19 disease dataset.

Keywords Health care analysis · Ensemble technique · Pandemic · COVID-19

1 Introduction

One of the most important aspects of a health emergency is disease prediction (Hirose and Wang 2012). It’s difficult to distinguish between disease-related and non-disease-related symptoms in the COVID-19 pandemic from various sources. Respiratory sound, X-Ray image, and social media content analysis are the primary sources for the current COVID-19 outbreak. In this paper, social media content has been used for analyzing corona disease posts. People have been using Twitter to share their feelings via tweets related to coronavirus disease and related issues. Sample informative and uninformative Tweets about the COVID-19 pandemic are shown in Table 1. As previously noted, most past research has indicated tweets related to a low-grade sickness. The detection of tweets related to health situations is the subject of this article (COVID-19).

All online social networking sites are trending with Corona-related topics during the current COVID-19 pandemic outbreak. Many people continue to post and communicate with COVID-19 messages on these websites. Twitter is a popular online media microblogging and social networking service in which users post and interact with messages known as "tweets." Tweets are in various categories of information in a massive amount of data at a significant speed, consisting of both disease-related and non-disease-related tweets (Goodwin et al. 2009). This information includes symptoms of illness such as a cold, fever, headache, runny nose, body pains, and so on (Quincey et al. 2016). Various social and government organisations’ or departments’ resources, such as social media, should be highlighted in order to raise awareness and deliver medical kits and treatments for various ailments. The classified tweet information contains one-of-a-kind resources to assist the suffering and suspected users in learning the status of the pandemic (Loey

1 Department of Computer Applications, National Institute of Technology, Tiruchirapalli, Tamilnadu, India

Published online: 07 July 2022
In this case, we must categorise tweets related to various diseases in order to provide medical facilities for the authorities in order to keep them from becoming ill.

To diagnose these types of situations, we need a standard automated system to collect high-level tweeting as an individual reaches the final stage of breathing. Disease prediction is a critical component in health emergencies (Rudra et al. 2017). Automatic systems (Chew and Eysenbach 2010) are required in pandemic situations to detect informative tweets from Twitter posts. It is difficult to distinguish between INFORMATIVE and UNINFORMATIVE (Jagadeesh and Alphonse 2020) tweets from a source involved in the COVID-19 outbreak. According to previous research, the majority of previous studies focused on identifying disease-related tweets using low-grade methods. In this paper, we look at the issue of disease-related tweets during a health emergency. To that end, we propose an IPSH (Information of words, parts of speech, statistical, and high-frequency terms as features) method for solving the aforementioned problem. For the classification of tweets, the existing method uses two types of features: informative words and statistical features (Chen et al. 2016). In order to improve classification accuracy, we use quality and efficient features in our proposed method IPSH. For base-level and second-level classifiers, we used Random Forest and Adaboost classifiers, respectively. The proposed method is associated with two base methods, SVM and BOW (Han et al. 2013).

This paper’s dye points are as follows:

1. Identify Insightful COVID-19 Tweets, such as symptoms, prevention, and precautions, and so on.
2. Using four features to identify INFORMATIVE tweets with greater accuracy.
3. Observing improved results with latest COVID-19 data set.
4. Using a combination of Adaboost, Random Forest, and other models to compare the results.

The following is how this paper map is organized:-The relevant work is explained in Sect. 2, the proposed approach is presented in Sect. 3, the experimental findings are presented in Sect. 4, and the conclusion is presented in Sect. 5.

2 Related work

During a health crisis, a massive amount of short critical messages were posted on Twitter (Agarwal et al. 2011). Victims are identified by detecting the target tweets on Twitter at the time of crisis. The automated detection of target tweets is a challenging task.

Examination of social media user messages can measure different population services, including public health interventions. The tweet content review reveals that users record typical, often extreme signs as well as medicinal uses. By applying two existing methods for extracting clinical words from natural language created for the analysis of clinical records (MetaMap, cTakes) to a real-world collection of medical blog posts (Denecke 2014). Such tools can also accurately extract medical concepts which are stated explicitly in texts from medical 21 Code check failed social media data, but the abstraction lacks essential details captured in paraphrase or formulated in popular language. An interactive framework to identify blogs using advanced features to describe their knowledge content is put in place to illustrate the medical and efficient content of blog posts. The findings show that the qualities of various health-related Web services are substantially different (Denecke and Nejdl 2009). Weblogs and reaction portals mainly deal with diseases and medicines. More information on anatomy and procedures is provided by the Wiki and encyclopedia. While patients...
and nurses discuss personal aspects of their lives, their blog posts tend to convey health-related information to doctors. The data and resources are made accessible to the research community to allow Twitter’s richer text analysis and related social media data sets (Fox 2011). A modern way of answering TREC-CDS medical questions is by first revealing the reply and then choosing and evaluating research documents containing the reply (Gimpel et al. 2010). AIDR is used to classify messages posted during disasters in a set of user-defined categories of information (Imran et al. 2014). Post first word 2 with word embedding focused on 52 million tweets linked to the crisis. We present human-annotated standardized lexical tools for different lexical variations to fix tweet language issues (Imran et al. 2016).

The identification and classification of tweets related to health (Madichetty and Sridevi 2019) and current disease pandemic COVID-19 have been discussed; (Malla and Alphonse 2021); (Babu and Eswari 2020). Text mining, Chest X-Ray image classification and respiratory techniques based on AI techniques are discussed (Aggarwal and Zhai 2012, Wiysobunri et al. 2020 and Kumar and Alphonse 2021). Text classification has been performed by Machine Learning models like Naive Bayes and k-NN classifier for with the help of the top-most frequency word features and low-level lexical features.

3 Disease tweets detection

The working methodology for train and test the model are shown in Fig. 1 and Fig. 2. Different phases in INFORMATIVE tweets detection are (Bethard et al. 2016) tweet collection, pre-processing, feature extraction, training, and testing. The following sections explain the flow of the IPSH method.

3.1 Collection of health-related tweets

INFORMATIVE Tweets are collected from the COVID-19 data set, as shown in Table 2.
3.1.1 Informative tweets

Tweets relevant to symptoms like, runny nose, cough, sore throat, possibly a headache, and maybe a fever attack last a couple of days. Messages having public death information also come under this category.

3.1.2 Uninformative tweets

Tweets that are not dealing with symptoms of the disease are irrelevant to UNINFORMATIVE tweets. This type of tweet may point to the tweets related to the non-symptomatic tweets related to diseases or different.

3.2 Tweet data pre-processing

Pre-processing of tweets is a very important phase for tweets analysis (Krouska et al. 2016). The following steps involved pre-processing of tweets.

1. Normalization: Processes of transfer of all upper case letters to lower case letters(Kouloumpis et al. 2011).
2. Tokenization: convert tweets into small parts(tokens).
3. Stop-Word Removal: Not useful words like an, are, a, of, from, etc.
4. @ (user-mentions), URLs, # (hash-tags), three or less than characters, and numerical values are removed from tweet data.

3.3 Feature extraction from tweets

The model mainly depends on four features, namely,

1. Informative words
2. POS tagging
3. Statistical features
4. High-frequency words.

3.3.1 Informative words

TF_IDF technique (Sreenivasulu and Sridevi 2017) is used for informative word extraction for better results. This technique depends on Term Frequency(T_F) and Inverse Document Frequency(ID_F), as shown in Eq. 1.

3.3.2 POS tagging

Parse trees help create NERs((most named entities are Nouns) and obtaining relations among the words, but these parse trees are built by parts of speech(POS) Tag. POS Tags

Table 2 Detecting Informative Features in the Disease-Related Tweets

| S.No | Information category | Informative words (Features) | Examples |
|------|----------------------|------------------------------|----------|
| 1    | Place                | Italy, New York              | As the number of #COVID19 deaths surpassed 100,000 worldwide, NY surpassed Italy on number of recorded deaths, New York state has more coronavirus cases than any country worldwide HTTPURL |
| 2    | Analysis             | Survey                       | US has over 2600 active cases of #COVID19 with more than 60 deaths. US might eventually be the first country to find a drug, but as the old saying goes, prevention is indeed better than cure. Here, its commendable how proactive the govt has been in dealing with the situation |
| 3    | Disease              | COVID-19, Swine flu, H1N1, Swine influenza | In 2009 1.4 billion people were infected by Swine Flu, out of which around 500,000 people died. Scientists say that Swine Flu was less contagious than the new #COVID and that is what makes it more dangerous. #COVID19outbreak #Covid19india |
| 4    | Human impact         | Recovered, Active, effected, People, Confirmed, Death and Dead terms | @USER Plan A does not work. Time to use plan B. What is plant B? How about antiviral drug Remdesivir? COVID-19 treatment trial is promising. 125 COVID-19 patients, 113 with severe illness, 123 recovered within 1 week, 2 died HTTPURL |
| 5    | Financial impact     | Stock market, Import and Export Terms | CDFA developed the COVID-19 Resource Center to serve as a central hub of everything the development finance world is doing to mitigate the impacts of #COVID19. CDFA’s Weekly Development Finance COVID-19 Briefings highlight these recovery efforts. HTTPURL |
| 6    | Other                | Strong and Effective terms   | We will all stay strong and fight this #COVID19 epidemic together. One of the finest ways to fight this together is to follow Medical Practitioners’ advise. If you are suspecting anything, get tested, the virus is not a death sentence. #COVID19NIGERIA #LetUsEndCOVID19Together |

© Springer
are used to reduce a word to its root form. POS tagging is a method of labeling a term in a corpus, based on its meaning and description, into a corresponding part of speech tag. This function is not straightforward because a specific word can have another part of speech depending on the situation in which the word is worn. So POS Tagging is a significant step.

### 3.3.3 Statistical features

This Statistical feature extraction process includes all steps of data pre-processing except the third step. Statistical features (as shown in Table 2) like frequency of numbers, hashtags (#), user mentions (@), URLs, wh-words are extracted in this stage. Algorithm 1 describes the procedure involved in the statistical features extraction. Some example tweets are tabulated in Tables 2 and 3, respectively.

High-frequency words.

High-frequency words (Bharti et al. 2015) are repeating words in the corpus. To analyze the frequency of words from the tweets, tweets.

1. URL free.
2. Words should be in unique case (generally lower case)
3. Contains valid words (remove not useful words)

### 4 Proposed IPSH method

We proposed the IPSH method for detecting tweets associated with the current pandemic outbreak (COVID-19). This method aims to improve the quality of training by two levels of classifiers (base-level and second-level). Firstly base-level classifier takes the vector as input from four different features such as informative words, POS tagging, statistical feature, and high-frequency words and produces output to the second-level classifier. The second-level classifier is used to correct the improper training of the base-level classifier. Performance mainly depends on the model’s diversified value. This IPSH model gains a high diversifying value by using four features, which are used in base-level classifiers individually.

1. Informative word features (Wang et al. 2011) are converted to vector by using the "CountVectorizer" function with handmade features and gives its value to one of the ensemble classifiers in a Base-level classifier.
2. POS features (Xia et al. 2011) are converted to vector by using the "POS" function and gives their value to one of the ensemble classifiers in the base-level classifier.
3. Statistical features (Guo et al. 2010) are converted to vector by using the "Extract_features" function and provides its value to one of the ensemble classifiers in the base-level classifier.
4. High-frequency words (Jin et al. 2009) are converted to vectors using the "CountVectorizer" function with max features set to 50 and gives its value to one of the ensemble classifiers in the base-level classifier.

In case there are no informative features in a tweet, they are not helpful for target tweet detection. So, the remaining features are heaped with the aid of the base-level classifier to alert the target tweet. Even in the absence of anyone among the four features, Tweet can be identified with the remaining ones. The choosing of an accurate classifier for the selected features is essential for the ensemble model. Therefore, different modern-day (state-of-art) models such as AdaBoost, Random Forest, Bagging, and SVM (Kim et al. 2003) are

| Table 3 Tweets related to statistical features |
|-----------------------------------------------|
| **Statistical Features** | **Example of statistical featured** | **Tweets** |
|-----------------------------------------------|
| Frequency of hashtags (#) | Seven #Coronavirus #COVID19 patients et al.-Furat General Hospital in #Baghdad, #Iraq have reportedly tested negative for the virus, and they will be discharged soon. More details when available.#CoronavirusUpdates | #Baghdad #Iraq #COVID19 #CoronavirusUpdates |
| Frequency of user-mention (@) | Senate President Tito Sotto says he believes he was not exposed to ACT-CIS Party-list Rep. Eric Yap, who tested positive for COVID-19. @USER @covid19 | @covid19 @USER |
| Frequency of URLs | Our @USER details reasons why Black population in #NC faces higher #Coronavirus risk. In @USER more than one-third of #COVID deaths so far are African-Americans. @USER @USER leader offers perspective. HTTPURL | HTTPURL |
| Frequency of \textbf{wh}-words | Wow France had a bad hit yesterday. 1355 new deaths in 24 h. The most a country has had so far. I knew they was in the eye of the storm. Poor souls. RIP Germany looks to be the next casualty of this invisible war. #France #COVID19 #coronavirus @USER @Germany | wow |
| Frequency of Numerals | NEW: 298 new coronavirus cases have been confirmed in Washington, bringing total cases to 8,682, including 394 deaths, state Health Officer Kathy Lofy said | 298, 8682, 394 |
used in base-level classifiers and also in second-level classifiers in the IPSH model. The combinations of classifiers used for the experiment are shown in Table 9.

Algorithm 1: IPSH model proposed by four features

**Input:** N number of tweets  
**Classifiers (ensemble models):**  
AdaBoost- base-level classifier  
Random Forest- second-level classifier  
**Output:** Tweets classification(disease or non-disease).

**Step 1**: Total Tweets(N) in a data set are given to the model as input.  
**Step 2**: Informative words, POS tagging words, statistical features, and High-frequency words are abstracted from the tweets separately.  
**Step 3**: The base-level classifier takes all extracted features individually as input.  
**Step 4**: The second-level classifier takes input from the base-level classifier(its output) for the classification of disease tweets.

4.1 Dataset

COVID-19 data set is used to classify. Subsections 4.1.1 to 4.1.3 explain the details of the data set as shown in Table 4.

4.2 COVID-19 labeled data set

The COVID-19 tweets are collected from Nguyen et al. (2020) informative COVID-19 English Tweets. This data set is labeled as INFORMATIVE and UNINFORMATIVE. The Data set is a collection of Training, validation, and test sub-data set in the CSV format.

4.3 Training phase

As shown in Fig. 2, the selected four features from the data set are trained individually using one of the ensemble classifiers at the base level, and their output is forwarded to a second-level classifier for a better score.

5 IPSH model analysis

This section deals with experimental results of proposed model with different combinations of the base-level classifiers and the Second-level classifiers and also compares IPSH model’s performance (F1-score, recall, precision, AUC_ROC score, and accuracy) with existing baseline models. The IPSH model is implemented using python language with the machine learning package “Scikit” (Pedregosa et al. 2011). The experiments are conducted in two different phases (in-domain and cross-domain) as shown in the following sections. The detailed results of IPSH model are analyzed and discussed in next section.

### 5.1 Informative words as features

As features, a collection of informative words from tweets is used. These characteristics are used as input parameters to train and test the various machine learning models, as shown in the Table 5 and Fig. 3., in order to select the best baseline model.

### 5.2 POS tags

A collection of parts of speech from tweets is used as a feature. These characteristics, as shown in the Table 6, are used

| Table 5 | Various classifiers performance (for informative words as features) on COVID-19 dataset |
|---------|------------------------------------------------------------------------------------------|
| Classifier       | Accuracy | F1-score | Execution Time |
| Logistic Regression | 0.75     | 0.71     | 0.23           |
| AdaBoost Classifier | 0.78     | 0.78     | 0.03           |
| Random Forest Classifier | 0.75     | 0.7     | 0.15           |
| BernoulliNB     | 0.75     | 0.77     | 0.04           |
| Calibrated Classifier CV | 0.72     | 0.75     | 2.59           |
| Linear SVC      | 0.74     | 0.78     | 1.35           |
| Linear Discriminant Analysis | 0.77     | 0.76     | 0.08           |
| Ridge Classifier CV | 0.73     | 0.72     | 0.04           |
| Ridge Classifier  | 0.77     | 0.75     | 0.04           |
| NuSVC            | 0.77     | 0.77     | 5.44           |
| LGBM Classifier  | 0.71     | 0.73     | 0.25           |
as input parameters to train and test the various machine learning models in order to select the best baseline model as well as Fig. 4.

5.3 Statistical features

A collection of statistical terms from tweets is employed as a feature. These features are utilised as input parameters to train and evaluate the various machine learning models in order to identify the optimal baseline model, as indicated in the Table 7 as well as Fig. 5.

5.4 High-frequency words as features

A collection of high-frequency words from tweets is used as a feature. In order to select the best baseline model, these characteristics are used as input parameters to train and test the various machine learning models, as shown in the Table 8 as well as Figs. 6, 7.

According to Tables 6, 7, 8, and 9, the proposed model is trained with these features (Informative Words, POS Tags, Statistical Features, and High-Frequency Words) as inputs using 10 classifiers, and it is demonstrated that among all of them, the AdaBoost classifier outperformed in predicting tweet content.

5.5 Performance measures

The IPSH model used the following parameters to identify tweet data in both scenarios (in-domain and cross-domain).

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

(1)

where TP (True Positive)—the number of correctly detected disease tweets.
FP (False Positive)—number of wrongly detected disease tweets.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

(2)

The recall is calculated using the Eq. 2.
F1-Score can be computed using the Eq. 3

Accuracy is computed by using Eq. 4.

AUC_ROC score describes in Eq. 5.

where $T_0$—a sum of the ranks of the disease tweets, $n_0$—number of disease tweets, and $n_1$—number of non-disease tweets.

The tests are achieved with different base-level classifiers such as Bagging, SVM, AdaBoost, and Random Forest and their values can be found in Table 5 and 6. Random Forest dominates other variations such as AdaBoost, SVM, and Bagging in the base-Level classifier. The performance is mainly based on the base-level classifiers but remains unchanged when the Second-classifier changes. The Random Forest provides huge achievement when compared with Bagging, SVM, and AdaBoost. The model’s efficiency is mainly based on the base-level classifiers and by their extreme properties. Random Forest is the Outperform Ensemble algorithm for selected features. However, the IPSH method surpasses both the SVM and BOW. Informational words relevant to disease-related shall be explored by using TF-IDF technology to classify disease tweets. Features of tweets are extracted based on POS tagging. Disease-related tweets contain features of statistical (count of affected or dead), web links (represents URLs), etc. A large amount of vocabulary causes a sparsity problem to base-line-1 (SVM with the BOW model) model. Baseline-2 (BOW with informative and statistical features) did not classify the targeted tweets. Besides, cross-domain tests for the data set are carried out to verify the reliability of the IPSH process.

### Table 7 Various classifiers performance (for statistical words as features) on COVID-19 dataset

| Classifier(statistical) | Accuracy | F1-Score | Execution Time |
|-------------------------|----------|----------|----------------|
| Logistic Regression     | 0.77     | 0.77     | 0.04           |
| AdaBoost Classifier     | 0.80     | 0.79     | 0.12           |
| RandomForest Classifier | 0.76     | 0.76     | 0.15           |
| BernoulliNB             | 0.75     | 0.76     | 0.04           |
| Calibrated ClassifierCV | 0.75     | 0.76     | 2.59           |
| LinearSVC               | 0.74     | 0.76     | 1.35           |
| LinearDiscriminant Analysis | 0.74 | 0.76 | 0.08 |
| RidgeClassifierCV       | 0.73     | 0.76     | 0.04           |
| Ridge Classifier        | 0.73     | 0.76     | 0.04           |
| NuSVC                   | 0.73     | 0.76     | 5.44           |
| LGBM Classifier         | 0.74     | 0.76     | 0.25           |

$$\text{Recall} = \frac{TP}{TP + FN}$$  
$$F1-Score$$

F1-Score can be computed using the Eq. 3

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$  

$$\text{Accuracy}$$

Accuracy is computed by using Eq. 4.

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

AUC_ROC score.

![Machine Learning Classifiers Performance for POS words as Features](image)
6 Results

In this experiment, COVID-19 data set is used for training and testing and the proposed IPSH model is compared against state-of-art machine learning models like Bag of Words, AdaBoost, SVM, and Random Forest in terms of accuracy, F1-score, Precision and Recall as shown in Table 10. Figure 3 shows that the proposed ensemble machine learning model is exceptional when compared with the other machine learning models. The proposed IPSH model aims to achieve 83.06% accuracy, 83.73% F1-score, 83.84% precision and 83.67% recall compared with other profound learning techniques, as shown in Table 10.

Referring to Table 10, it is well understood that our proposed model has achieved an better accuracy and F1 score compared with existing models as mentioned. This gives a clear indication that the model has succeeded in differentiating the informative tweets related to COVID-19 disease outbreaks from combined tweets.

7 Conclusion

Users often discuss disease information via Twitter. A massive volume of data is shared on Twitter before and after a pandemic situation. The proposed IPSH method recognizes INFORMATIVE tweets effectively by using the AdaBoost in the base level and the Random Forest as a second-level classifier. The experimental results convey that the recommended IPSH method succeeds in the ensemble technique. This paper emphasizes the importance of combining informative words, POS features, statistical properties, and high-frequency words for detecting INFORMATIVE tweets during the COVID-19 pandemic. The proposed research is for identifying other disease outbreaks through microblog sites and for developing methods to identify different diseased symptoms. Further, we plan to use the deep learning models to predict Insightful tweets and then determine if the organizations or offenders will benefit from the discovered tweets or not.

![Machine Learning Classifiers Performance for Statistical words as Features](image-url)

Table 8 Various classifiers performance (for high-frequency words as features) on COVID-19 dataset

| Classifier               | Accuracy | F1-Score | Execution Time |
|-------------------------|----------|----------|----------------|
| Logistic Regression     | 0.78     | 0.78     | 0.05           |
| AdaBoost Classifier     | 0.79     | 0.79     | 0.13           |
| RandomForest Classifier | 0.77     | 0.76     | 0.15           |
| BernoulliNB             | 0.75     | 0.76     | 0.04           |
| CalibratedClassifierCV  | 0.75     | 0.76     | 2.59           |
| LinearSVC               | 0.74     | 0.76     | 1.35           |
| LinearDiscriminant Analysis | 0.74 | 0.76     | 0.08           |
| RidgeClassifierCV       | 0.73     | 0.76     | 0.04           |
| RidgeClassifier         | 0.73     | 0.76     | 0.04           |
| NuSVC                   | 0.73     | 0.76     | 5.44           |
| LGBMClassifier          | 0.74     | 0.76     | 0.25           |
Machine Learning Classifiers performance for High-Frequency words as Features

![Graph showing performance of different classifiers](image)

Results Analysis

![Graph comparing IPSH model with other machine learning models](image)

Fig. 6 Machine Learning Classifiers performance for High-Frequency words as Features

Fig. 7 Results Comparison of IPSH Model with Different Machine Learning models
Acknowledgements
We are extremely thankful to NIT Trichy which has supported all the way to execute this study.

Funding
There was no funding or grants received that assisted in this study.

Declarations

Conflict of interest
The authors declare they have no conflicts of interest.

Human and animals rights
In our work, no animals or human are involved.

Informed consent
Not applicable as no human or animal sample was involved in this study.

References
Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ (2011) Sentiment analysis of twitter data. In: Proceedings of the workshop on language in social media (LSM 2011), pp. 30–38

Aggarwal CC, Zhai C (2012) A survey of text classification algorithms. Mining text data. Springer, Boston, pp 163–222

Babu YP, Eswari R (2020) CIA_NITT at WNUT-2020 Task 2: classification of COVID-19 tweets using pre-trained language models. arXiv preprint arXiv:2009.05782, pp. 471–474

Bethard S, Carpuat M, Cer D, Jurgens D, Nakov P, Zesch T (2016) Proceedings of the 10th International workshop on semantic evaluation (SemEval-2016). In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp. 1–46

Bharti SK, Babu KS, Jena SK (2015) Parsing-based sarcasm sentiment recognition in twitter data. In: 2015 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM), pp. 1373–1380

Chen C, Wang Y, Zhang J, Xiang Y, Zhou W, Min G (2016) Statistical features-based real-time detection of drifted twitter spam. IEEE Trans Inf Forensics Secur 12(4):914–925

Chew C, Eysenbach G (2010) Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one 5(11):e14118

Denecke K (2014) Extracting medical concepts from medical social media with clinical NLP tools: a qualitative study. In: Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing, pp. 54–60

Denecke K, Nejdl W (2009) How valuable is medical social media data? Content analysis of the medical web. Inf Sci 179(12):1870–1880

Fox S (2011) The social life of health information, California Healthcare Foundation, pp. 1–45

Gimpel K, Schneider N, O’Connor B, Das D, Mills D, Eisenstein J, Heilman M, Yogatama D, Flanigan J, Smith NA (2010) Part-of-speech tagging for twitter: annotation, features, and experiments. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science

Goodwin R, Haque S, Neto F, Myers LB (2009) Initial psychological responses to Influenza A, H1N1 (“ Swine flu”). BMC Infect Dis 9(1):1–6

Guo Z, Zhang L, Zhang D, Zhang S (2010) Rotation invariant texture classification using adaptive LBP with directional statistical features. In: IEEE international conference on image processing, pp. 285–288

Han Q, Guo J, Schuetze H (2013) Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In: Second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), pp. 520–524

Hirose H, Wang L (2012) Prediction of infectious disease spread using Twitter: a case of influenza. In: IEEE International symposium on parallel architectures, algorithms and programming, pp. 100–105

Imran M, Castillo C, Lucas J, Meier P, Vieweg S (2014) AIDR: artificial intelligence for disaster response. In: Proceedings of the 23rd international conference on world wide web, pp. 159–162

Imran M, Mitra P, Castillo C (2016) Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. arXiv preprint arXiv:1605.05894, pp. 1–6.

Jagadeesh MS, Alphonse PJ (2020) NIT_COVID-19 at WNUT-2020 Task 2: deep learning model RoBERTa for identify informative COVID-19 english tweets. In W-NUT@ EMNLP, pp. 450–454

Table 9 Selection of base and second-level classifiers by their combined performance

| Base Classifier | Second Classifier | Accuracy | AUC_ROC |
|-----------------|-------------------|----------|---------|
| AdaBoost        | AdaBoost          | 80.51    | 80.78   |
| AdaBoost        | Bagging           | 79.02    | 80.21   |
| AdaBoost        | Random Forest     | 83.06    | 83.73   |
| AdaBoost        | SVM               | 79.51    | 80.78   |
| Bagging         | AdaBoost          | 72.56    | 72.59   |
| Bagging         | Bagging           | 67.16    | 60.00   |
| Bagging         | Random Forest     | 73.56    | 72.13   |
| Bagging         | SVM               | 72.56    | 71.13   |
| Random Forest   | AdaBoost          | 81.51    | 80.87   |
| Random Forest   | Bagging           | 79.51    | 79.8    |
| Random Forest   | Random Forest     | 80.59    | 80.8    |
| Random Forest   | SVM               | 77.03    | 77.8    |
| SVM             | AdaBoost          | 77.02    | 60.32   |
| SVM             | Bagging           | 67.02    | 60.32   |
| SVM             | Random Forest     | 77.02    | 70.32   |
| SVM             | SVM               | 77.02    | 70.32   |

Table 10 Results comparison of the proposed model with existing methods on COVID-19 dataset using performance metrics

| Models               | Accuracy | F1-Score | Precision | Recall |
|----------------------|----------|----------|-----------|--------|
| BoW                  | 61.55    | 58.32    | 50.94     | 68.18  |
| AdaBoost             | 76.1     | 75.9     | 76.2      | 76.2   |
| SVM                  | 82.04    | 82.71    | 81.12     | 84.19  |
| Random Forest        | 82.06    | 81.58    | 81.92     | 83.57  |
| Proposed Method (IPSH) | 83.06    | 83.73    | 83.84     | 83.67   |
Jin W, Ho HH, Srihari RK (2009) OpinionMiner: a novel machine learning system for web opinion mining and extraction. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1195–1204
Kim HC, Pang S, Je HM, Kim D, Bang SY (2003) Constructing support vector machine ensemble. Pattern Recogn 36(12):2757–2767
Kouloumpis E, Wilson T, Moore J (2011) Twitter sentiment analysis: The good the bad and the omg!. In: Fifth International AAAI conference on weblogs and social media
Krouska A, Troussas C, Virvou M (2016) The effect of preprocessing techniques on Twitter sentiment analysis. In: 2016 7th International conference on information, intelligence, systems and applications (IISA), pp. 1–5
Kumar LK, Alphonse PJA (2021) Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: cough, voice, and breath. Alex Eng J 61(2):1319–1334
Loey M, Manogaran G, Taha MHN, Khalifa NEM (2021) A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. Measurement 167:1–19
Madichetty S, Sridevi M (2019) Disaster damage assessment from the tweets using the combination of statistical features and informative words. Soc Netw Anal Min 9(1):1–11
Malla S, Alphonse PJA (2021) COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. Appl Soft Comput 107:1–15
Nguyen DQ, Vu T, Rahimi A, Dao MH, Nguyen LT, Doan L. WNUT-2020 task 2: identification of informative COVID-19 english tweets. arXiv preprint arXiv:2010.08232, pp. 1–5
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
Quincey de E, Kyriacou T, Pantin T (2016) #hayfever; A longitudinal study into hay fever related tweets in the UK. In: Proceedings of the 6th international conference on digital health conference, pp. 85–89.
Rudra K, Sharma A, Ganguly N, Imran M (2017) Classifying information from microblogs during epidemics. In: Proceedings of the 2017 international conference on digital health, pp. 104–108.
Sreenivasulu M, Sridevi M (2017) Mining informative words from the tweets for detecting the resources during disaster. In: International conference on mining intelligence and knowledge exploration, pp. 348–358
Wang Z, Zhao Q, Chu D, Zhao F, Guibas LJ (2011) Select informative features for recognition. In: 2011 18th IEEE international conference on image processing, pp. 2477–2480
Wiysobunri BN, Erden HS, Toreyin BU (2020) An ensemble deep learning system for the automatic detection of COVID-19 in X-ray images, pp. 1–9
Xia R, Zong C, Li S (2011) Ensemble of feature sets and classification algorithms for sentiment classification. Inf Sci 181(6):1138–1152

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