Date of publication JUN-30, 2024, date of current version APR-21, 2024.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.2.3547

Ensemble-based Disease Outbreak Detection: Comparative Analysis of Health News Information Retrieval Techniques

MANJU JOY[1], M. KRISHNAVENI [2]
[1]Department of Computer Applications, FISAT, Kerala, India
[2]Department of Computer Science, Avinashilingham University, Coimbatore, Tamil Nadu, India

Corresponding author: Manju Joy (e-mail: 19phcsp010@ avinuty.ac.in).

ABSTRACT In India, Kerala is the first state to report a COVID-19 infection case, in January 2020, in a medical student, who returned from Wuhan, China. More recently, in June 2022, Kerala also reported India's first case of monkeypox disease. News websites often publish articles dedicated to reporting disease occurrences and live updates of outbreaks. Through the utilization of data gathered from online digital resources, early detection of outbreaks is possible, and this potential is already identified by the research community. As webpages give a comprehensive collection of reports covering a wide range of themes through hyperlinks, precisely categorizing news articles based on their headlines and retrieving health news is a tedious operation. Hence, this paper proposes a novel and efficient news retrieval technique grounded on an ML-based classification method with an ensemble learning approach to identify reports of disease occurrences from web pages by focusing specifically on the health context of Kerala and a comparison with baseline methods for information retrieval such as keyword-based, phrase-based, and content-based latent semantic analysis method is made.

KEYWORDS Ensemble learning; Epidemic surveillance; Outbreak detection; Text mining; Natural language processing.

I. INTRODUCTION

THE category of biothreats has historically been neglected, and downplaying disease outbreaks can incur significant costs. Weaknesses in healthcare systems were exposed by the COVID-19 pandemic, underscoring the significance of robust bio-surveillance systems for real-time monitoring of disease outbreaks. 'Health Map' is a renowned global health monitoring system that gives advance alerts of infectious diseases occurring globally, by gathering outbreak data from PROMED mail, Official alerts of the WHO, and online news collectors. Other popular surveillance systems are PROMED, GPHIN, Epispider, BioCaster, BioPak flasher, etc. Studies claim that all outbursts of infectious diseases declared by WHO are first sensed through these casual online resources [1]. A Canadian software company named 'BlueDot' provided warnings about the unusual increase in pneumonia cases in Wuhan, China on December 31, 2019, by analyzing text data available in various digital resources such as news reports, and social media. WHO declared it as a 'Covid-19 Pandemic' on March 11, 2020.

In January 2020, Kerala was the first Indian state to report a COVID-19 infection for the first time, in a medical student who returned from Wuhan, China. In June 2022, Kerala also reported India's first case of monkeypox. Experts attribute this trend to the sensitivity and effectiveness of the established conventional surveillance mechanism existing in Kerala which helped the state to identify the country's primary instances of infectious diseases. The distribution of individuals from Kerala worldwide, urbanization, and shifts in climate patterns are additional contributing factors [2]. An efficient Bio surveillance system that can access epidemic information from online sources and measure the levels of spread and severity is the need of the hour in the state, which motivated the study to be conducted in this direction.

The proposed study aims to develop a machine learning-based news categorization system to accurately detect occurrences of diseases happening in different regions of Kerala, so that timely and reliable information on disease outbreaks can be provided to concerned authorities and public health agencies. By leveraging the power of ensemble learning techniques, the performance of the classification algorithms can be enhanced [3]. Detecting outbreaks from unstructured text data is a challenging task in NLP. Overcoming legal, ethical, and privacy barriers associated with data sharing is also

crucial in this regard. To develop strategies to avert outbreaks, regional or locality-specific data is crucial. Major contributions of the paper are as follows:

1. In the existing methods, obtaining health-related data is hard and time-consuming. Due to this limitation, a web scraping technique is proposed for data collection from web pages of news websites.

2. Text data available in news articles are unstructured, noisy and inconsistent. Understanding the context and semantics of news articles which are available as short text data, is critical for accurate outbreak detection. Hence the data is pre-processed using standard natural language processing techniques and to address the challenges in short text classification, effective feature engineering or representation learning method is used to capture the essence of short texts.

3. Retrieving information that aligns with a particular user requirement can be a laborious process. Consequently, an exploration of baseline information retrieval methods such as term-based, phrase-based, and concept-based methods has been conducted to examine their limitations. An efficient Machine learning-based classification method with an ensemble learning approach is proposed to retrieve health news information with improved accuracy.

The rest of the paper is organized as follows: The second section includes related works and the third section describes the proposed methodology and its working process. Section four discusses the evaluation of the proposed model and its comparison with baseline methods. A conclusion and future work recommendations are given in section five of the paper.

II. RELATED WORKS

NLP is a prominent field of AI that has witnessed significant advancements in recent years and has opened up new possibilities for numerous applications in the domain of public health. One such critical area of research is the detection and monitoring of disease outbreaks through the analysis of text data. Accurate and real-time identification of disease outbreaks is essential to ensure effective public health response and mitigation strategies. Traditional surveillance methods have proven to be valuable, but they often suffer from delays and limitations in coverage. Several techniques are proposed in the literature for information retrieval. A dictionary of patterns and text processing algorithms can be used to extract relevant information from internet sources. Only those diseases and locations already known and available in the dictionary are identified in this approach [1]. Expression-based queries are proposed to retrieve relevant disease-related pages on the web and automatic classification of documents is done using Supervised ML approaches such as Naïve Bayes and SVM [4]. To retrieve online resources for detecting occurrences of infectious diseases in Korea, the authors proposed two DL algorithms such as ConvNet and BiLSTM, and the documents are collected from websites of International bodies such as WHO, NCDC, and SAMOH. Their results show that Bidirectional LSTM performed better than ConvNet [5]. A Deep-learning-based ranking algorithm is proposed and a data filtering method is used for collecting disease data from online news articles such as Bing News, SNS data from Twitter, and web search queries from Google for six countries for developing Eagle eye- a worldwide disease-related topic extraction system [6]. Disease names and symptoms are used as keywords in Eagle Eye and its performance is compared with HealthMap, taking Covid-19 as a case study. Various ML algorithms are proposed to classify verses from the Qur'an based on their fundamental aspects of Islam. The algorithms are tested with English and Indonesian translations of Al-Baqarah verses. Results show that the SVM has the best accuracy of the Indonesian translation, achieving 81.443%, and the Naïve Bayes classifier has the best accuracy of the English translation, achieving 78.35% [7]. Information retrieval models based on classification algorithms are proposed by authors to develop a model for automated online news text classification in the Sindhi language, where LSVC and MLP classifiers gave a maximum accuracy of 84%. A dataset comprising news articles belonging to five categories collected from newspapers such as Awami Awaz and Daily Jhoongar, are utilized for the study [8]. The authors proposed an ensemble ML model by combining predictions from MLP, KNN, and Random Forest to identify spam product reviews based on a majority vote of the contributing models [9]. LSA algorithm is used to cluster news articles automatically from Czech newspapers by leveraging an agglomerative clustering approach. Cosine distance is the similarity metric used and the synonymy problem is reduced [10]. From the analysis, it is clearly noted that short text classification is a challenging task due to the limited amount of information available in the text. This limited context may lead to ambiguity and difficulty in accurate classification [11]. The sparsity of features or fewer words in the short text can affect the performance of traditional feature-based classifiers, as there may be insufficient information to discriminate between different classes [12],[22]. By combining the capabilities of multiple classification models, the ensemble learning approach allows the creation of an improved predictive model compared to a single model, thus enhancing the results [13]. Ensemble learning algorithms have proven to outperform individual classifiers used. Sufficient training data is crucial for building accurate classification models. However, obtaining a large and diverse labeled dataset for short text classification is also challenging, particularly for the health domain. Deep learning models are popular in text classification tasks due to their ability to automatically learn hierarchical representations of data from raw input. Lack of sufficient samples can substantially reduce the performance of a DL model [14]. Another major problem to be addressed is class imbalance, where some classes have more examples than others. This may result in biased models that perform well on the majority class but struggle with classes with minimum samples. Employing techniques like data augmentation could mitigate the effects of class imbalance problems.

III. MATERIALS AND METHODS

The objective of this research is to explore and identify an effective strategy, utilizing ML based classification method, to extract information about both familiar and unfamiliar disease outbreaks from highly unstructured text data. Diverse news retrieval approaches are experimented on the data collected to determine the most efficient method for news retrieval. While previous research studies have concentrated on news classification through a full-text approach, the majority have overlooked short-text classification centered on headlines. The lack of sufficient context and background information in news articles makes its understanding a tedious task. To address this gap, a novel machine-learning model based on an ensemble approach is proposed in this paper for short text classification. The model aims to automatically retrieve news articles relevant to the health domain. The proposed methodology involves

several phases such as collecting data from news websites using Web scraping technique, preprocessing, feature weighting, implementation of various classification algorithms, and finally, performance evaluation.

Web scraping is the process of gathering information from the World Wide Web and storing it in a file or a database for further analysis. There are three stages of the web scraping process such as fetching phase (getting web pages from a specific website using a URL), the extraction of textual information from a web page, and then the transformation phase where the extracted data is turned into a structured format for storage. Data pre-processing is done to convert raw data into a comprehensible and consistent form by removing noise, and to enable the machine to discover hidden patterns present in it. An efficient NLP package in Python, Neattext, is used for preprocessing. Since text data is designed for humans and not machines, it must be transformed into a machine-readable format by a process called indexing or vectorization which converts words into numerical vectors. Count Vectorization and TF-IDF Vectorization are commonly used indexing methods, that produce a matrix representing the occurrence of terms in documents. For effective classification and to eliminate irrelevant features, the bag-of-words technique is applied. Additionally, the TF-IDF approach is employed for feature selection based on the importance of words. The TF-IDF score reflects the significance of a word by considering its frequency in a document and its importance in the entire dataset. Words that are unique to a specific document or a limited set of documents tend to have higher TF-IDF scores. Conversely, commonly occurring words receive lower scores and are pruned [1],[16],[25]. Dealing with the high-dimensional feature space is crucial in news classification. Irrelevant, noisy (infrequent words), and redundant features in documents can reduce the accuracy of the system. By implementing proper feature extraction and selection techniques, unwanted features are eliminated during preprocessing. The TF-IDF value for a term *tm* in a specific document DT is calculated using the following equation:

$$\text{TF-IDF}(tm,\text{DT}) = \text{TF}(tm,\text{DT}) * \text{IDF}(tm), \quad (1)$$

here, TF(*tm*, DT) or term frequency computation is specific to a document DT which specifies the number of occurrences of the term *tm* in a specific document DT, while IDF(*tm*) or inverse document frequency is computed by considering the entire corpus or collection of documents, i.e., it is computed by dividing the total number of documents in the corpus by the number of occurrences of the term *tm* in the entire corpus. The inverse document frequency value decreases when a term is present in all documents. These computed weights play a significant role in training a classification model. Following feature selection, appropriate classification algorithms are applied to the feature vectors. Algorithms such as LR, Support Vector Machines, K-Nearest Neighbors, Naïve Bayes, Random Forest, Bernoulli Naive Bayes, and XGBoost are implemented for text categorization. The proposed ensemble classifier is developed by combining the top performing classifiers such as SVM, BernouliNB and Random Forest classifier. Performance evaluation of various text classification algorithms is given in Table 1 of section 4 where experimental results are discussed.

In the proposed work, a dataset having N number of documents is used, where each document DT consists of a collection of *m* features. A set of class labels $C = \{c_1, c_2, ..., c_n\}$ are associated with each document DT, where c_i can be either +1 (indicates that the document belongs to the health news category) or 0(non-health news category). SVM is an efficient text classification algorithm for both binary class and multiclass classification problems. In binary classification problems, it aims to invent a hyperplane that is optimal in separating the two classes. The equation for representing the hyperplane is as follows:

$$w^{\mathrm{T}}x + b = 0\,, \quad (2)$$

where x is the vector corresponding to an input feature, w is the weight vector and b is the bias term. The SVM classification rule depends on the function

$$f(x) = w^{\mathrm{T}}x + b\,, \quad (3)$$

which can be positive or negative. If $f(x) \geq 0$, x is classified as +1; otherwise 0.

Bernoulli NB is based on Bayes' theorem and is a probabilistic classifier. The algorithm assumes that features are binary, and the class conditional probabilities are estimated using the occurrences of features [18]. For each feature, its occurrence in a feature vector is denoted by a binary value, where 1 indicates the existence of the feature and 0 indicates its absence. First, we estimate the prior probabilities of the two classes. Let $P(\text{H} = +1)$ and $P(\text{H} = 0)$ represent the prior probabilities of class, +1(health news) and class 0 (non-health news) respectively. These probabilities can be computed as the proportions of data points belonging to each class in the training set. Next, we estimate the class conditional probabilities for each feature. For each feature, we compute the probability of its occurrence given a specific class. Let us consider a binary feature X_j. We estimate the probability $P(X_j = 1 \mid \text{H} = +1)$ as the proportion of data points in class +1, where feature X_j is present. Similarly, we estimate $P(X_j = 1 \mid \text{H} = 0)$ as the proportion of data points in class 0, where feature X_j is present.

Bernoulli NB follows the assumption that the features are conditionally independent when class labels are given. This means that the occurrence of features does not affect each other, given the class. Based on this assumption, we can estimate the joint probability of all features given a class as the product of the individual feature probabilities. To classify a document or new feature vector X, which is a set of features $(x_1, x_2, x_3, \ldots\ldots x_m)$, we compute the posterior probabilities for each class using Bayes' theorem given below:

$$P(\text{X}|\ \text{H} = +1) = (P(X_1 = x_1\ |\text{H} = +1) * P(X_2 = x_2\ |\text{H} = +1) * \ldots * P(X_m = x_m\ |\text{H} = +1), \quad (4)$$

$$P(\text{X}|\ \text{H} = 0) = (P(X_1 = x_1|\text{H} = 0) * P(X_2 = x_2\ |\text{H} = 0) * \ldots * P(X_m = x_m\ |\ \text{H} = 0)). \quad (5)$$

Class label which maximizes the posterior probability will determine the class of the feature vector X.

Random Forest is an ensemble method that constructs a collection of decision trees using bootstrap aggregating or bagging. A randomly selected feature subset is used to build each decision tree and the concept of bagging is employed, which involves creating multiple bootstrap samples from the given training dataset. Randomly selected data points from the

original dataset with replacement are used to create each bootstrap sample. These bootstrap samples are then used to build individual decision trees. For each decision tree in the Random Forest, a feature subset is randomly selected from the available features. This feature subset is typically smaller than the total number of features. The purpose of using a feature subset is to introduce diversity among the decision trees and reduce correlation. The splitting criterion used is entropy. To make a prediction using an individual decision tree, an individual tree is traversed starting from the root node, and the feature values of the input data point are compared with the decision rules at each node. This traversal continues until a leaf node is reached. In the leaf node, the prediction is made based on the majority class of the training examples that ended up in that leaf. Each tree independently generates its own prediction by following its set of rules. Then, through a process called majority voting, the class label that receives the most votes across all the decision trees is chosen as the ultimate prediction.

In the proposed method, results of SVM, BernoulliNB, and Random Forest classifiers are combined to develop a stacking classifier for efficient retrieval of health news. Each base classifier within the ensemble contributes its own votes or predictions, but there is no direct notion of feature weights at the ensemble level. The ensemble learner combines the predictions of its constituent classifiers based on a voting approach to make a conclusion.

Algorithm for the proposed ensemble classifier method

Step 1: Collect news articles from websites using web scraping techniques and store them in a file.

Step 2: Perform preprocessing operations such as tokenization, stop word removal, removal of special symbols and numbers.

Step 3: Convert pre-processed data into numeric format.

Step 4: Split the sample dataset into training and test sets.

Step 5: Define individual classifiers such as SVM, Bernoulli Naive Bayes, and Random Forest.

Step 6: Create a Voting Classifier by specifying the individual classifiers as estimators.

Step 7: Train the Voting Classifier using the training set and retrieve the feature names and count the occurrences of features in the training set by iterating over the samples in the training set.

Step 8: Evaluate the performance of the trained model using test samples.

A diagrammatic representation of the proposed methodology is given in Figure 1.

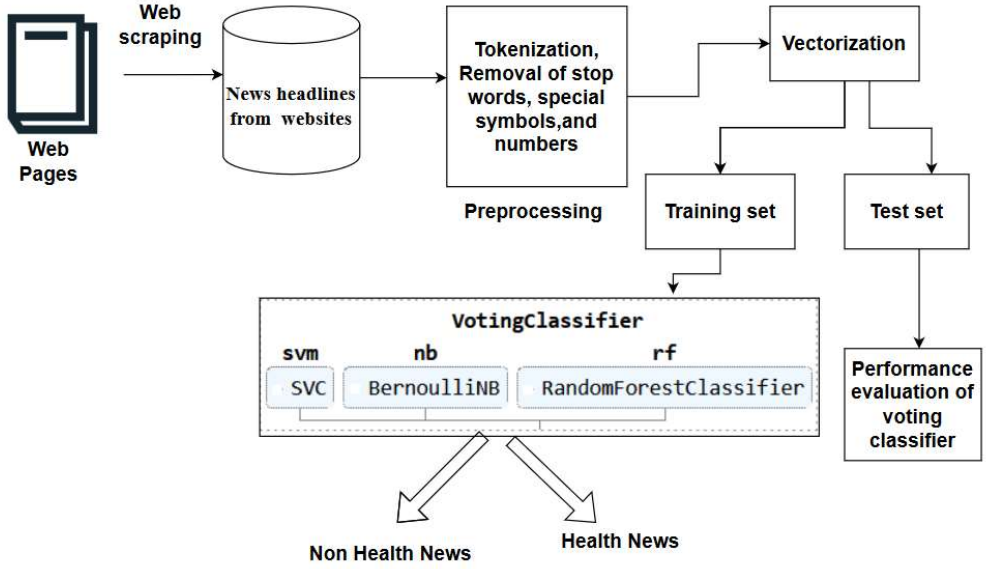

Figure 1. Proposed Methodology

IV. EXPERIMENTAL RESULTS

The dataset used for the study comprises news articles gathered from news portals covering Kerala news, associated hyperlinks of detailed news stories and type of news articles. The type of news article is labeled as either 1(means health-related news) or 0(non-health-related news). Both the local and national level news websites are considered and data is collected for the period from 01/01/2012 to 31/07/2023. The dataset has 3550 records with 1751 health-related and 1799 other news samples. Python programming language and Beautiful Soup library are used for web scraping. Infectious diseases such as Dengue, Leptospirosis, Japanese Encephalitis, Hepatitis, H1N1, Meningitis, Malaria, Measles, and Covid- 19, are reported in Kerala most often in the last ten years [19].

Classical methods for automatic text classification and retrieval, categorized into rule-based method and machine learning (data-driven) method are studied. The most commonly used rule-based information retrieval methods such as keyword-based method, phrase-based method, and Concept based method are primarily used in some epidemic intelligence systems for information recovery. In the term-based or keyword-based method, analysis of documents is based on terms/keywords whereas in the phrase-based method analysis of news articles is based on phrases that are clear and more discriminative than keywords [20].

Classification accuracy obtained for phrase-based method is lesser than term-based method. Fewer Computations and faster response are the main advantage of these two methods. However, if a novel type of illness is reported, the term-based approach might fail to gather the required information. To ensure precision in outcomes, updates to the dictionary become essential. The phrase-based technique tends to retrieve a significant amount of inaccurate and irrelevant data. In

contrast, the concept-based strategy operates on the principle that terms capturing textual semantics hold greater significance in distinguishing pertinent from non-relevant documents, tailored to a specific user need. This approach facilitates the categorization of web documents into distinct clusters, each denoting a specific topic [21]. The concept-based approach is operationalized by employing the Latent Dirichlet Allocation (LDA) algorithm.

ML methods are efficient in retrieving information which is of specific interest to the users. So, various text classification algorithms are experimented on the dataset. The ensemble learning approach combines the predictions of its constituent classifiers to make better prediction on unseen data. To decide on the combination of ensemble learners, the performance evaluation of various ML-based classifiers is done and the results obtained are shown in Table 1. Three top performing algorithms, SVM, Bernoulli NB and Random Forest, are combined in the proposed method to ensure improved accuracy.

Table 1. Performance evaluation of various text classification algorithms

Classification Models	Training Accuracy %	Precision	Recall	F1-Score
Logistic Regression	93.01	0.97	0.9	0.93
SVM	**93.05**	**0.96**	**0.9**	**0.93**
Adaboost	90.39	0.94	0.86	0.9
Random Forest	**94.48**	**0.96**	**0.93**	**0.95**
XGBoost	91.43	0.94	0.89	0.91
KNN	89.36	0.87	0.93	0.9
Naïvebayes(NB)	80.14	0.73	0.97	0.84
BernoulliNB	**95.62**	**0.97**	**0.94**	**0.96**
Stacking Classifier	**96.28**	**0.97**	**0.95**	**0.96**

The proposed method is very efficient in binary classification tasks involving short text. In comparison to other news retrieval methods, the proposed approach has a notably high level of accuracy.

A comparison of the baseline methods with the proposed ensemble method is given in Table 2 and its diagrammatic illustration is given in Figure 2.

Table 2. Comparison of the proposed method with baseline methods

Information retrieval techniques	Accuracy	Precision	Recall	F1-Score
Term based method	84%	96%	84%	90%
Phrase-based method	55%	92%	52%	70%
Concept-based method	79%	72%	90%	81%
Proposed ensemble method	**96%**	**97%**	**95%**	**96%**

The accuracy of the proposed model is high whereas the accuracy of the phrase-based method is very low.

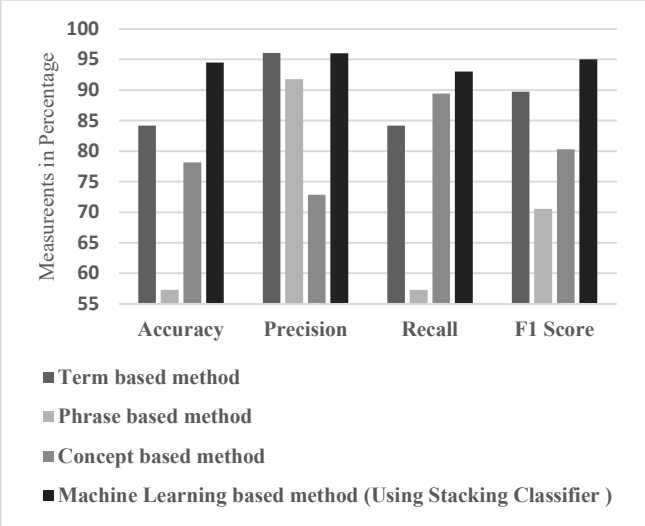

Figure 2. Comparison of baseline methods with the proposed method

The accuracy of the proposed model obtains the value of 96% whereas the accuracy of Term based method, Phrase-based method, and Concept based method are 84%, 55%, and 79% respectively.

The top-scoring domain-specific features are identified, and only these characteristics are employed for training, resulting in increased classification performance. The top 20 features and their assigned weights for high-accuracy models such as SVM, Random Forest, BernoulliNB and the proposed Stacking model are given in Figure 3.

SVM	
features	Weights
monkeypox	3.846325892
Ist	3.604271575
cases	3.431548685
tnn	3.092823351
dengue	2.994763464
nipah	2.968683952
fever	2.660895582
covid	2.377169619
covid19	2.258967315
diseases	2.17882665
zika	2.134803177
positive	2.066976557
diphtheria	2.022711108
outbreak	1.999840932
chikungunya	1.99909832
shigella	1.969792635
malaria	1.956513073
tpr	1.903499275
hepatitis	1.722986034
coronavirus	1.660227758

Random Forest	
features	Weights
cases	0.067651302
kerala	0.045668839
monkeypox	0.033195223
1st	/0.032347471
dengue	0.028758846
tnn	0.022692139
covid	0.018131482
virus	0.016798432
reported	0.016742559
outbreak	0.015923821
fever	0.015853181
covid19	0.012470203
nipah	0.012204617
health	0.010333301
positive	0.010223988
disease	0.00937248
zika	0.00869269
chikungunya	0.00846576
alert	0.008367221
deaths	0.00818779

BernoulliNB	
features	Weights
kerala	-0.860891952
cases	-1.185853842
Ist	-1.44270845
tnn	-1.6264596
dengue	-1.947599339
monkeypox	-2.241216147
reported	-2.265313698
covid	-2.414058959
health	-2.423537703
virus	-2.442769065
fever	-2.492530575
new	-2.566638547
state	-2.577688383
outbreak	-2.634846797
case	-2.670778806
covid19	-2.733692632
deaths	-2.77343296
positive	-2.903110784
nipah	-2.91861497
disease	-2.966624189

Stacking Classifier	
features	Weights
kerala	684
cases	384
ist	279
tnn	232
dengue	168
monkeypox	125
case	125
state	124
reported	123
covid	119
new	118
health	112
virus	106
fever	99
outbreak	86
covid19	82
deaths	77
india	73
says	72
reports	70

Figure 3. The top 20 features and their assigned weights for high-accurate classification models

While analyzing the feature weights, we could observe that weight values are in different ranges for different ML models, such as for SVM from 3.8 to 1.6, and for Random forest from 0.06 to 0.01. In SVM and Random forest, larger absolute values suggest greater importance, indicating that the corresponding features have a stronger influence on the classification decision. Negative weights occur in Bernoulli Naive Bayes due to the logarithmic transformation performed during the training process. In the algorithm, the logarithm is applied to the probability estimates to avoid numerical underflow and expand computational efficiency. In stacking classifier, the relative importance of the predictions from the base classifiers are captured as feature weights rather than the original features.

V. CONCLUSION

The objective of the study is to detect outbreaks of diseases based on raw text data available on the internet sources so that health officials can be alerted at an early stage and immediate actions can be taken. The study is carried out by collecting information from the health domain of Kerala, a state in India, which reported the first case of most of the infectious disease outbreaks that happened in the country recently. Classical information retrieval techniques, such as keyword-based, phrase-based, and concept-based methods are studied and drawbacks are identified. An ensemble-based classification method is proposed which can analyze huge volumes of text data and effectively retrieve health news from the public domain with an accuracy of 96%. After identifying and extracting health-related news links, detailed stories of outbreaks are retrieved from associated web pages. The proposed method based on ensemble learning has proven to outperform all other information retrieval methods. The ensemble learners can capture different aspects of the data and exploit complementary strengths of individual classifiers, resulting in more accurate and reliable predictions. Transfer learning-based techniques or deep learning-based models such as RNN or LSTM also can be experimented with in the future for efficient information retrieval if a large corpus is available for the study.

References

[1] Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, John S. Brownstein, "HealthMap: Global infectious diseases monitoring through automated classification and visualization of internet media reports," *Journal of the American Medical Informatics Association: JAMIA*, vol. 15, issue 2, pp. 150-157, 2008. https://doi.org/10.1197/jamia.M2544

[2] S. Jayesh, S. Sreedharan, "Analysing the Covid-19 cases in Kerala: A visual exploratory data analysis approach," *SN Comprehensive Clinical Medicine*, vol. 2, pp. 1337-1348. https://doi.org/10.1007/s42399-020-00451-5

[3] A. Jain, J. Mandowara, "Text classification by combining text classifiers to improve efficiency of classification," *International Journal of Computer Applications*, vol. 6, no. 2, pp. 126-129, 2016.

[4] E. Arsevska, S. Valentin, J. Rabatel, J. de Goër de Hervé, S. Falala, R. Lancelot, M. Roche, "Web monitoring of emerging animal infectious diseases integrated in the French animal health epidemic intelligence system," *PLOS one,* pp. 1-25, 2018. https://doi.org/10.1371/journal.pone.0199960

[5] M. Kim, K. Chae, S. Lee, H. J. Jang and S. Kim, "Automated classification of online sources for infectious disease occurrences using machine-learning-based natural language processing approaches," *International Journal of Environmental Research and Public Health*, vol. 17, no. 24, 2020. https://doi.org/10.3390/ijerph17249467

[6] B. Jang, M. Kim, I. Kim and J. Kim, "Eagle eye: A worldwide disease-related topic extraction system using deep learning based ranking algorithm and internet-source data," *Sensors*, vol. 21, no. 14, 2021. https://doi.org/10.3390/s21144665

[7] R. Hidayat and S. Minati, "Comparative analysis of text mining classification algorithms for English and Indonesian Qur'an translation," *International Journal on Informatics for Development*, vol. 8, no. 1, pp. 47-51, 2019. https://doi.org/10.14421/ijid.2019.08108

[8] I. A. Kandhro , S. Z. Jumani, A. A. Lashari, S. S. Nangraj, Q. A. Lakhan, M. T. Baig and S. Guriro, "Classification of Sindhi headline news documents based on TF-IDF text analysis scheme," *Indian Journal of Science and Technology*, vol. 12, no. 33, pp. 1-10, 2019. https://doi.org/10.17485/ijst/2019/v12i33/146130

[9] M. Fayaz, A. Khan, J. Ur Rahman, A. Alharbi, M. Irfan Uddin, B. Alouffi, "Ensemble machine learning model for classification of spam product reviews," *Hindawi*, vol. 2020, Article ID 8857570, pp. 1-10. https://doi.org/10.1155/2020/8857570

[10] M. Rott and P. Cerva, "Investigation of latent semantic analysis for clustering of Czech news articles," *Proceedings of the 25th IEEE International Workshop on Database and Expert Systems Applications*, 2014, pp. 223-227. https://doi.org/10.1109/DEXA.2014.54

[11] M. I. Rana, S. Khalid, M. U. Akbar, "News classification based on their headlines: A review," *Proceedings of the IEEE 17th International Multi-Topic Conference*, 2014, pp. 211-216. https://doi.org/10.1109/INMIC.2014.7097339

[12] X. Luo, "Efficient English text classification using selected machine learning techniques," *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401-3409, 2021. https://doi.org/10.1016/j.aej.2021.02.009

[13] U. Suleymanov, S. Rustamov, "Automated news classification using machine learning methods," *Proceedings of the IOP Conference Series: Materials Science and Engineering*, 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* 459 012006. https://doi.org/10.1088/1757-899X/459/1/012006

[14] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu and J. Gao, "Deep learning based text classification: A comprehensive review," *ACM Computing Surveys (CSUR),* vol. 54, no. 3, pp. 1-40, 2021. https://doi.org/10.1145/3439726

[15] R. Singh, S. A. Chun, V. Atluri, "Developong machine learning models to automate news classification," *Proceedings of the 21st Annual International Conference on Digital Goverment Research*, 2020. https://doi.org/10.1145/3396956.3397001

[16] T. Xia, Y. Chai, "An improvement to TF-IDF: Term distribution based term weight algorithm," *Journal of Software,* vol. 6, no. 3, pp. 413-420, 2011. https://doi.org/10.4304/jsw.6.3.413-420

[17] M. Nasir , M. Bakhtyar, J. Baber, S. Lakho, B. Ahmed and W. Noor, "BIOPAK flasher: Epidemic disease monitoring and detection in Pakistan using text mining," *arXiv:2106.06720*, 2021. https://doi.org/10.48550/arXiv.2106.06720

[18] M. A. Fauzi, A. Z. Arifin, S. C. Gosaria, "Indonesian news classification using Naive Bayes and two-phase feature selection model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 8, no. 3, pp. 610 - 615, 2017. http://doi.org/10.11591/ijeecs.v8.i3.pp610-615

[19] T. Jacob John, K. Rajappan, K. K. Arjunan, "Communicable diseases monitored by disease surveillance in Kottayam District, Kerala state," *Indian J Med,* vol. 120, no. 2, pp. 86-93, 2004.

[20] S. V. Gaikwad, A. Chaugule, P. Patil, "Text mining methods and techniques," *International Journal of Computer Applications*, vol. 85, pp. 42-45, 2014. https://doi.org/10.5120/14937-3507

[21] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, O. A. Alomari, "Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering," *Expert System with Applications*, vol. 84, pp. 24-36, 2017. https://doi.org/10.1016/j.eswa.2017.05.002

[22] L.-M. Chen, B.-X. Xiu, Z.-Y. Ding, "Multiple weak supervision for short text classification," *Applied Intelligence*, vol. 52, pp. 9101-9116, 2022. https://doi.org/10.1007/s10489-021-02958-3

[23] C. Dreisbach, T. A. Koleck, P. E. Bourne and S. Bakken, "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data," *International Journal of Medical Informatics*, vol. 125, pp. 37-46, 2019. https://doi.org/10.1016/j.ijmedinf.2019.02.008

[24] L. Yao, Z. Pengzhou and Z. Chi, "Research on news keyword extraction technology based on TF-IDF and TextRank," *Proceedings of the*

IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), 2019, pp. 452-455. https://doi.org/10.1109/ICIS46139.2019.8940293

[25] D. Wang, H. Zhang, "Inverse-category-frequency based supervised term weighting schemes for text categorization," *Journal of Information Science and Engineering*, vol. 29, no. 2, pp. 209-225, 2013.

[26] M. B. Khan, "Urdu news classification using application of machine learning algorithms on news headline," *International Journal of Computer Science and Network Security*, vol. 21, no. 2, pp. 229-237, 2021. https://doi.org/10.22937/IJCSNS.2021.21.2.27

***Ms. Manju Joy** is an Assistant Professor of Department of Computer Applications, Federal Institute of Science and Technology (FISAT), Angamaly, Kerala, India. She is a research scholar at Avinashilingam University for Women, Coimbatore, Tamil Nadu, India. Her research interests include Data mining, Natural Language Processing, Machine learning, and Deep learning.*

Ms. Manju Joy received her Master's Degree in Computer Applications from MES College of Engineering, Kuttipuram. She has more than 15 years of teaching experience.

***Dr.M.Krishnaveni** is an Assistant Professor of Department of Computer Science, Avinashilingam University for Women, Coimbatore, Tamil Nadu, India. She has research experience under Defence projects and worked on disciplines like Artificial Intelligence, IoT, Image Processing, Speech Processing, Data Mining, and Computational Intelligence. She has published 5 Books, 9 Book chapters and 95 research papers in both national and international level. She has research projects under various funding agencies and acts as an active member of Centre for Machine Learning and Artificial Intelligence. She has received awards such as best presenter award, best paper award in IEEE world AI IoT congress 2021, best young teacher award IASTE 2017, best NSS Programme officer award, NYLP 2016, Government of India.*