International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 2579-2584 www.ijaem.net ISSN: 2395-5252

Clinical Text Classification for Covid-19: An Empirical Evaluation

Dr. H M Keerthi Kumar, Advi K S, Navami Hegde, Rakshitha S Mendon, Swathi G

Department of Information Science And Engineering,
Yenepoya Institute Of Technology, Moodbidri

Submitted: 05-07-2021 Revised: 17-07-2021 Accepted: 20-07-2021

ABSTRACT-Technology improvements have a fast impact oneach subject of life, be it clinical subject or some other subject. The COVID-19 pandemic was a health emergency that required a rapid response by the National Health System. It is transmitted by inhaling or having contact with droplets. In this system, clinical reports of the patients are collected. The dataset consist of clinical reports in the form of text. In this work, supervised machine learning techniques are used for classifying the text into four different categories. In this project, clinical text reports are classified into four different categories of diseases such that it can help in detecting corona virus from earlier clinical symptoms. Empirical evaluation of various machine learning techniques such as supervised learning algorithm will be used in this project work.

I INTRODUCTION

COVID-19 had been first said with the aid of using officers in Wuhan City, China, in December 2019. Retrospective investigations with the aid of using Chinese government have diagnosed human instances with onset of signs in early December 2019.The novel corona virus changed into labeled corona virus sicknesses through World health organization(WHO). COVID-19 is a large family of viruses that cause illness ranging from common cold to extra drastic diseases.The four classes of viruses, COVID, ARDS, SARS and both (consists a person that is having both corona virus as well as ARDS).COVID-19,spreads through breathing drops when coughing, silent or needles can be infected, from and touching the contaminated surfaces with the virus and touching their eyes, mouth or nose. The corona virus (COVID-19) affects all aspects of society and all dimensions of sustainable development.

This epidemic puts intensive pressure on healthcare, economic, and social structures. As a result, part of the emergency communication takes place through social media systems, which can also be helpful tools to seize social change. Social media platforms are broadly used at the moment to obtain and share different types of information about this crisis. The lockdown has pressured many academic establishments to cancel their classes. Examinations, internship etc. to pick out the web modes. Initially,the educators and the college students had been stressed and didn't recognize how to cope up with the scenario of this unexpected disaster that forced closure of the academic activities. Health and social systems across the global are struggling to cope. The situation is especially challenging in humanitarian, fragile and low income country contexts, where health and social systems are already weak. Health facilities in many places are closing or limiting services

The dataset consist of clinical reports in the form of text. In this work, supervised machine learning techniques are used for classifying the text into four different categories COVID, SARS, ARDS and both (COVID, SARS). In this project, clinical text reports are classified into four different categories of diseases such that it can help in detecting corona virus from earlier clinical symptoms

II LITERATURE SURVEY

Mark &Helena [1] proposed the reviews and compares several machine-learning approaches that were developed in addition to a rule-basedsystem that was developed. The machine learning approachesincluded naïve Bayesian learner, decision trees, and SupportVector Machines (SVMs). Vikas et al. [2] discusses adetailed survey on the text classification process and various algorithms used in this field Textclassification is the process of classifying text documents intofixed number of predefined classes. Krishna et al. [3] they used the social media platformTwitter for analysis they Looking at the statistics of COVID19 infected, recovered, and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 2579-2584 www.ijaem.net ISSN: 2395-5252

deathcases of Italy and other countries, Indians knew that drastic measureswere needed in India to stop the numbers from rising exponentially.Forman et al. [4]presents an empiricalcomparison of twelve feature selection methods (e.g. Information Gain) evaluated on a benchmarkof 229 text classification problem instances that were gathered from Reuters. The results are analyzed from multiple goal perspectives—accuracy, F-measure, precision,and recall—since each is appropriate in different situations.Akib et al. [5]In this paper, they classified textual clinical reports into four classes by usingclassical and ensemble machine learning algorithms. Featureengineering was performed using techniques like Termfrequency/inverse document frequency (TF/IDF), Bag ofwords (BOW) and report length.Kumar and Harish [6] conducted a series of experimentation toverify the effectiveness of various preprocessing techniques on Stanford TwitterSentiment Dataset. They demonstratedthe role of various preprocessing techniques in Twitter sentimentclassification.Nenad [7]the proposed solution showssatisfiab prediction performance and enables efficient resourceexchange by reducing the trading costs.Sciandra et al.[8]In this paper we focused on Italian social media communication about COVID-19. The analysis of the conversations going on Twitter, through the odds ratios and the similarities of word embeddings,managed to capture events, topics, and personalities of the COVID-19 emergency.Basu et al. [9] In this paper we are interested in examining whetherautomatic classification of news texts can be improved bya pre-filtering the vocabulary to reduce the feature set usedin the computations. First we compare artificial neuralnetwork and support vector machine algorithms for use astext classifiers of news items. Secondly, we identify areduction in feature set that provides improved results.

III METHODOLOGY

The proposed methodology consists of five steps. In first step data collection is being performed, second stepdefines the refining of data, third step gives an overview of preprocessing, fourth step provides a mechanism for feature extraction and the last step gives an overview of ensemble machine learning algorithms.

3.1 Data collection

As W.H.O announced infectious corona virus pandemic as health crisis. Concerning this epidemic the researchers and hospitals give open access to the data related to COVID-19.About 212 patients's data is fetched from an open source data repository Git Hub and stored which have revealed syndrome of corona virus and other viruses.Data consists of about 24 attributes namely patient id, offset, sex, age, finding, survival, incubated, went_icu, needed_supplemental_O2, extubated, temperature, pO2_saturation, leukocyte_count, neutrophil count, lymphocyte count, view, modality, date, location, folder, filename, DOI, URL, License, Clinical notes and other notes.

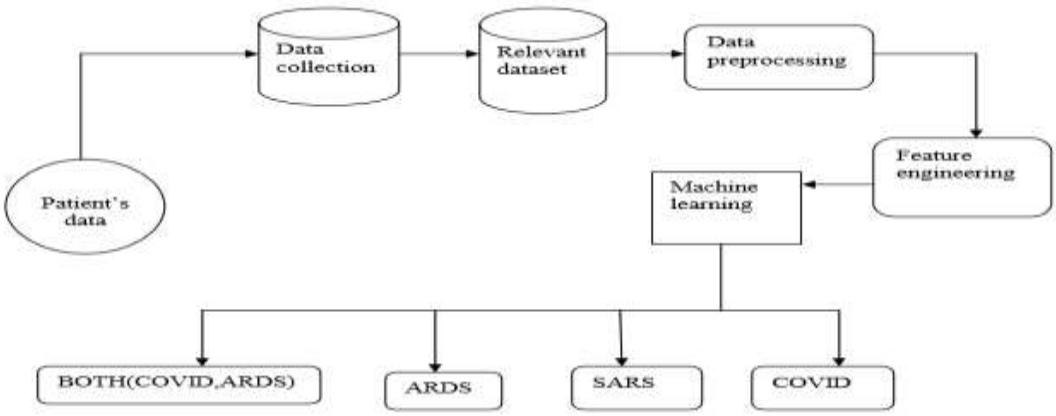
3.2 Relevant dataset

Considering our work is about text mining bring out the clinical notes and findings.Clinical notes consist of text while as the attribute finding consist label of the corresponding text. By considering those report that are written in English language only.

3.3 Preprocessing

The text is unordered so it needed to be purified such that machine learning methods can be applied.Different steps are being followed in this stage; the text is being cleaned by removing irrelevant or unnecessary text. Stop wards, Punctuation and lemmatization are being done such that the data is refined in a greater way.

Fig.1 Block Diagram



International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 2579-2584 www.ijaem.net ISSN: 2395-5252

Fig.2Preprocessed data set

	clinical_notes	finding	punctuation	tokenization	stopwords	Stemming
2	On January 22, 2020, a 65	COVID	On January 22 2020 a 6	['on', 'january', '22', '2	['january', '22', '2020', '6	['januari', '22', '2020', '65yearold', 'man'
3	On January 22, 2020, a 65	COVID	On January 22 2020 a 6	['on', 'january', '22', '2	['january', '22', '2020', '6	['januari', '22', '2020', '65yearold', 'man'
4	On January 22, 2020, a 65	COVID	On January 22 2020 a 6	['on', 'january', '22', '2	['january', '22', '2020', '6	['januari', '22', '2020', '65yearold', 'man'
5	On January 22, 2020, a 65	COVID	On January 22 2020 a 6	['on', 'january', '22', '2	['january', '22', '2020', '6	['januari', '22', '2020', '65yearold', 'man'
6	diffuse infiltrates in the	COVID	diffuse infiltrates in th	['diffuse', 'infiltrates',	['diffuse', 'infiltrates', 't	['diffus', 'infiltr', 'bilater', 'lower', 'lung
7	progressive diffuse inte	COVID	progressive diffuse in	['progressive', 'diffus	['progressive', 'diffuse',	['progress', 'diffus', 'interstiti', 'opac', 'c
8	Severe ARDS. Person is i	ARDS	Severe ARDS Person is	['severe', 'ards', 'pers	['severe', 'ards', 'person	['sever', 'ard', 'person', 'intub', 'og', 'pla
9	Case 2: chest x-ray obtai	COVID	Case 2 chest xray obta	['case', '2', 'chest', 'xra	['case', '2', 'chest', 'xray'	['case', '2', 'chest', 'xray', 'obtain', 'jan',
10	Case 2: chest x-ray obtai	COVID	Case 2 chest xray obta	['case', '2', 'chest', 'xra	['case', '2', 'chest', 'xray'	['case', '2', 'chest', 'xray', 'obtain', 'jan',
11	SARS in a 74-year-old ma	SARS	SARS in a 74yearold m	['sars', 'in', 'a', '74year	['sars', '74yearold', 'mar	['sar', '74yearold', 'man', 'develop', 'syn
12	SARS in a 74-year-old ma	SARS	SARS in a 74yearold m	['sars', 'in', 'a', '74year	['sars', '74yearold', 'mar	['sar', '74yearold', 'man', 'develop', 'syn
13	SARS in a 74-year-old ma	SARS	SARS in a 74yearold m	['sars', 'in', 'a', '74year	['sars', '74yearold', 'mar	['sar', '74yearold', 'man', 'develop', 'syn
14	SARS in a 29-year-old wo	SARS	SARS in a 29yearold w	['sars', 'in', 'a', '29year	['sars', '29yearold', 'wor	['sar', '29yearold', 'woman', 'present', '
15	SARS in a 29-year-old wo	SARS	SARS in a 29yearold w	['sars', 'in', 'a', '29year	['sars', '29yearold', 'wor	['sar', '29yearold', 'woman', 'present', '
16	SARS in a 42-year-old wo	SARS	SARS in a 42yearold w	['sars', 'in', 'a', '42year	['sars', '42yearold', 'wor	['sar', '42yearold', 'woman', 'present', '
17	SARS in a 46-year-old wo	SARS	SARS in a 46yearold w	['sars', 'in', 'a', '46year	['sars', '46yearold', 'wor	['sar', '46yearold', 'woman', 'present', '
18	SARS in a 46-year-old wo	SARS	SARS in a 46yearold w	['sars', 'in', 'a', '46year	['sars', '46yearold', 'wor	['sar', '46yearold', 'woman', 'present', '
19	SARS in a 73-year-old wo	SARS	SARS in a 73yearold w	['sars', 'in', 'a', '73year	['sars', '73yearold', 'wor	['sar', '73yearold', 'woman', 'present', '
20	SARS in a 73-year-old wo	SARS	SARS in a 73yearold w	['sars', 'in', 'a', '73year	['sars', '73yearold', 'wor	['sar', '73yearold', 'woman', 'present', '

3.3.1Preprocessing method

3.3.1.1 Removing Punctuation

Text data contains various words with the wrong spelling, special symbols, emojis, etc. we already cleaned this kind of noisy text data inFig. 2

3.3.1.2 Tokenization

Tokenization is splitting an entire text document into smaller units' sows in Fig. 2 such as individual words or terms.

3.3.1.3Stopwords

Stop words are the words in any language which does not add much meaning to a sentence. These are some of the most common, short function words, such as the, a, of, for, was etc.

3.3.1.4 Stemming

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes. For example" diabetes" is a word and its prefix is "es" if we remove "es" from "diabetes" then we will get base word which is "diabet".

3.4 Feature engineering

From the pre-processed clinical reports, various features are extracted as per the semantics and are converted into probabilistic values. TF/IDF technique for extracting relevant features will be used. Unigrams, bigrams is also extract.

Table 1: Unigram method using chi-square(x^2)

CHI-SQUARE(x^2)				
Algorithm	**Precision**	**Recall**	**F1 score**	**Accuracy (%)**
Multinomial Naı¨ve Bayes	0.51	0.71	0.60	60
Random Forest (RF)	0.84	0.84	0.84	84
Support Vector Machine(SVM)	0.85	0.84	0.83	84
Decision tree	0.88	0.87	0.87	87
K-Nearest Neighbours(KNN)	0.90	0.90	0.89	90

Table 2: Unigram method using Entropy

ENTROPY				
Algorithm	**Precision**	**Recall**	**F1 score**	**Accuracy (%)**
Multinomial Naı¨ve Bayes	0.61	0.69	0.60	69
Random Forest (RF)	0.80	0.79	0.78	79
Support Vector Machine(SVM)	0.87	0.86	0.84	86
Decision tree	0.90	0.90	0.90	90
K-Nearest Neighbours(KNN)	0.85	0.83	0.84	83

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 2579-2584 www.ijaem.net ISSN: 2395-5252

Table 3: Unigram method using Mutual Information

MUTUAL INFORMATION				
Algorithm	**Precision**	**Recall**	**F1 score**	**Accuracy (%)**
Multinomial Naı¨ve Bayes	0.64	0.77	0.69	77
Random Forest (RF)	0.89	0.86	0.85	86
Support Vector Machine(SVM)	0.91	0.90	0.89	90
Decision tree	0.83	0.83	0.83	83
K-Nearest Neighbours(KNN)	0.87	0.87	0.86	87

Table 4: bigrammethod using chi-square(x^2)

CHI-SQUARE(x^2)				
Algorithm	**Precision**	**Recall**	**F1 score**	**Accuracy (%)**
Multinomial Naı¨ve Bayes	0.82	0.78	0.79	78
Random Forest (RF)	0.81	0.81	0.80	81
Support Vector Machine(SVM)	0.92	0.92	0.91	92
Decision tree	0.81	0.80	0.80	80
K-Nearest Neighbours(KNN)	0.85	0.84	0.83	84

Table 5: bigram method using Entropy

ENTROPY				
Algorithm	**Precision**	**Recall**	**F1 score**	**Accuracy (%)**
Multinomial Naı¨ve Bayes	0.81	0.81	0.81	81
Random Forest (RF)	0.85	0.84	0.84	84
Support Vector Machine(SVM)	0.93	0.93	0.93	93
Decision tree	0.83	0.82	0.81	82
K-Nearest Neighbours(KNN)	0.84	0.83	0.82	83

Table 6: bigram method using Mutual Information

MUTUAL INFORMATION				
Algorithm	**Precision**	**Recall**	**F1 score**	**Accuracy (%)**
Multinomial Naı¨ve Bayes	0.82	0.78	0.79	78
Random Forest (RF)	0.76	0.76	0.75	76
Support Vector Machine(SVM)	0.82	0.76	0.73	76
Decision tree	0.81	0.80	0.79	80
K-Nearest Neighbours(KNN)	0.83	0.83	0.82	83

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 2579-2584 www.ijaem.net ISSN: 2395-5252

3.5 Machine learning classification

The classification is performed to classify the given text into four different types of viruses. The four classes of viruses, COVID, ARDS, SARS and both (consists a person that is having both corona virus as well as ARDS). Various supervised madchine learning algorithms will be used to classify the text into these categories.

3.5.1 Feature Selection method

3.5.1.1 Chi Square

A chi-square(x^2) statistics is a test that measures how a model compares to actual observed data. The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, and drawn from a large enough sample.The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship.

3.5.1.2 Entropy

Entropy is a measure of the randomness in the information being processed. Entropy is a measure of disorder or uncertainty and the goal of machine learning models and data scientists in general is to reduce uncertainty.

3.5.1.3 Mutual Information

Mutual Information is one of many quantities that measures how much one random variable tells us about another. It is a dimensionless quantity with units of bits, and can be thought of as the reduction in uncertainty about one random variable given knowledge of another.

3.5.2 Machine learning algorithms

3.5.2.1 Support Vector Machine(SVM)

SVM is a supervised machine learning model which uses classification algorithms for two group classification problems. After giving an SVM model set of labeled training data for each category, they are able to categories new text. It can be applied to any kind of vectors which encode any kind of data. This means that in order to leverage the power of SVM text classification, texts have to be transformed into vectors.

3.5.2.2 Multinomial Naı¨ve Bayes

MNB computes class probabilities of a given text by using Bayes rule. It is a bayes theorem based classification technique with an assumption of independence among predictors. The classifiers Naive Bayes believe that the inclusion of a specific feature in a class is irrelevant to any other function.

3.5.2.3 Random Forest (RF)

RF is a classification algorithm that consists of several tree for decisions. While constructing each individual tree, it uses bagging and features variability to try to construct an uncorrelated forest of trees whose prediction by committee is more reliable than that of any individual tree.

3.5.2.4 K-Nearest Neighbours(KNN)

KNN is a supervised machine learning algorithms used for regression and classification problems in machine learning. This takes the data and classifies new data points based on measures of similarity. Classification to its neighbor's is achieved by majority vote.

3.5.2.5 Decision tree

Decision tree builds a tree structure in context of classification or regression models. This breaks down a collection of data into smaller and smaller subsets while at the same time incrementally creating a related decision tree. There are two or more branches of a decision node. Leaf node reflects a ranking or judgment.

IV RESULT AND DISCUSION

We used a windows system with 4 GB Ram and 1.70 GHz processors for performing this work.For improving the accuracy of the entire machine learning algorithms pipeline is being used. After performing the statistical computation, deeper insights about the data were achieved. In unigramthe data is being split into 90:10 ratios were 90% data is being used for training the model and 10% is used for testing the model and in bigramthe data is being split into 80:20 ratios were 80% data is being used for training the model and 20% is used for testing the model. We have clinical text reports of patients that are labeled into four classes.The classification was done using machine learning algorithms by supplying them features that were extracted in the feature engineering step. In order to explore the generalization of our model from training data to unseen data and reduce the possibility of overfitting, we split our initial dataset into separate training and test subsets. In order to explore the generalization of our model from training data to unseen data and reduce the possibility of overfitting, we split our initial dataset into separate training and test subsets.

V CONCLUSION

COVID-19 has shocked the world due to spreading of the virus.Various researchers are working for conquering this deadly virus. We usedclinical reports which are labeled in four classes namely COVID, SARS, ARDS and both (COVID, SARS). Various features like TF/IDF, Count Vectorizations are being extracted from these clinical reports.The machine learning algorithms are used for classifying clinical reports into four different classes.After performing

classification, it was revealed that in unigram contains 90% accuracy score in KNN, decision tree, SVM and bigram contains 93% accuracy score in SVM(Support Vector Machine).Various other machine learning algorithms that showed better results were random forest, Multinomial Naı¨ve Bayes, KNN, decision tree.The efficiency of models can be improved by increasing the amount of data. Also, the disease can be classified on the gender-based such that we can getinformation about whether male are affected more or females. More feature engineering is needed for better results and deep learning approach can be used in future.

REFERENCES

[1]. http://www.iaeng.org/publication/WCE2007/WCE2007_pp1072-1075
[2]. https://ieeexplore.ieee.org/document/8125990
[3]. https://pubmed.ncbi.nlm.nih.gov/32305035/
[4]. https://www.jmlr.org/papers/volume3/forman03a/forman03a
[5]. https://link.springer.com/article/10.1007/s41870-020-00495-9
[6]. https://link.springer.com/chapter/10.1007/978-981-10-8633-5_3
[7]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9232908
[8]. https://ieeexplore.ieee.org/document/9219595L
[9]. https://www.researchgate.net/publication/221183673_Support_Vector_Machines_for_Text_Categorization
[10]. https://www.researchgate.net/publication/221183673_Support_Vector_Machines_for_Text_Categorization
[11]. https://link.springer.com/article/10.1007/s41870-020-00495-9
[12]. https://ieeexplore.ieee.org/document/9219595
[13]. https://www.jmlr.org/papers/volume3/forman03a/forman03a_full.pdf
[14]. https://pubmed.ncbi.nlm.nih.gov/32305035/
[15]. https://link.springer.com/chapter/10.1007/978-981-10-8633-5_3
[16]. http://www.iaeng.org/publication/WCE2007/WCE2007_pp1072-1075.pdf
[17]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9232908
[18]. https://ieeexplore.ieee.org/document/8125990