Open Access Article

A FRAMEWORK FOR THE COVID MEDICATION RECOMMENDATION SYSTEM BASED ON THE COLLABORATIVE SEARCH TECHNIQUE

Mrs.U.Hemamalini

Research Scholar & Assistant Professor, Department of Information Technology, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai.
hemababu2501@gmail.com

Dr.S.Perumal

Associate Professor, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai. perumal.scs@velsuniv.ac.in

Abstract **-** The fundamental component that contributes to a civilization's progression toward development is the improvement in its standard of living, particularly in terms of health and welfare. The recent pandemic has brought to light the excessively stressed situation of the health sector as well as the immediate demand to investigate the degree to which individuals are satisfied with the services that they receive. Nevertheless, the human happiness component is an important indicator in the healthcare industry, just as it is in other service sectors. This requires very little effort and focuses on intangible metrics. In order to analyse the feedback and the processing, a powerful Natural Language Processing model is developed to characterise the sentiment of the feedback. Additionally, an analysis of the patient's current symptoms is performed, and the analysis is run in order to output the overall results and suggestion. Here, we have utilised the various preprocessing techniques in order to reduce the number of words that have been removed from the incoming unprocessed text. The majority of the text is in an unstructured format; hence, the model will be enhanced by applying pre-processing techniques such as stemming, lemmatization, normalisation, and so on. After employing techniques such as the n-bag of words, tokenization, and noisy entity removal, feature extraction is carried out in order to make it possible to selectively select the words that denote the sentiment or emotion that is being described by the words. This makes it possible for the words to be chosen in such a way that they denote the sentiment or emotion that is being described by the words. The combined dataset is requested by the system so that it can be analysed for sentiment labelling purposes. Since the dataset is a collection of twitter conversation during COVID time along with geographical indicators and concentrated on conversation around COVID-19, this will enable the model to pick up words and analysis of the user sentiment with respect to healthcare. Panacea Lab COVID-19 dataset is used to train using. The method that was utilised to make observations is listed.

Keywords— *Sentiment analysis, twitter sentiment, n-gram, feature extraction, covid data, data optimization.*

I. INTRODUCTION

Received: October 30, 2022 / Revised: November 18, 2022 / Accepted: December 03, 2022 / Published: December 30, 2022

About the authors：Mrs.U.Hemamalini

Email: hemababu2501@gmail.com

The Novel corona epidemic has triggered an extraordinary crisis for humanity in general, the likes of which have not been seen in several decades. It didn't take long for a great number of nations to be engulfed by this pandemic, which swept across the globe like a raging forest fire, wreaking havoc in particular on the healthcare industry, which was already in a precarious position, and prompting lockdowns and holds in business and public life as a halt in an effort to slow down the increasing number of cases. In regard to the first wave of COVID 19, we have seen a massive toll that was caused by a lack of tracking and testing. In addition, a large number of tools and analytics are required in order to determine the hotspots, the prioritization of treatment to patients, and by the government to manage public health, planning, and management. The fact that there have been around 6 million deaths and 520 million confirmed instances of the disease as of May 2022 [1] has sparked a lot of changes in the way that we used to do things. The various sectors, particularly the healthcare sector and the administration, were put under a significant amount of stress.

This tension has contributed to harm the already overworked and pressured sector, producing a potential loss of reputation and also causing the public to suffer in huge consequence as a result due to the bad healthcare that is supplied. Analysis of emotions is one of the primary focuses of natural language processing (NLP). The extraction of opinions and, as a result, the determination of the emotion across a platform where a great deal of human communication occurs, such as social media, offers a great deal of untapped potential. Mining the emotional state and evaluation of the opinions from social media due to the fact that they are quite vast and in real time this is undeniably the case appealing as the datas that is available is in real world and vast and it can be used to contribute to the forecasting of future events, evaluation of the state of the public's emotional state, identification of sales targets. NLP to make use of the algorithms that extract and evaluate the contextual meaning of the words.

However, the feedback and quality measurement methods that are currently in use have a primary emphasis on the medical element, while the human pleasure part is frequently overlooked and is frequently intangible.

Natural Language Processing (NLP), often known as "natural language analysis," is an improved method that can be used to mine sentiment or general impression. Text mining is another term for this process, which involves generating and analysing enormous quantities of texts by first structuring the raw texts in a way that makes them compatible with NLP algorithms. [2]. The NLP techniques have the benefits of being able to consume and analyse enormous amounts of unformatted text, as well as being speedier than traditional methods. [3] [4]. There are a variety of strategies and tools available for the aforementioned extractions of texts [5]. The extraction and processing of materials can involve a wide variety of methods [6]. [7].

Given that the questionnaire and various other feedback slips, in addition to the reports, are the means by which the sentiments are extracted at the present time, natural language processing models can be used effectively for conducting sentiment analysis and obtaining contextual information as well. We also investigate a significant application domain known as sentimental analysis, which is also known as the application of artificial intelligence to the study of opinion or feeling. The process of utilizing various Natural Language Processing strategies are combined with the appropriate subsequent

techniques such as those used in linguistics, machine learning, and other related fields to conduct sentiment analysis is known as sentiment analysis. [8] In many cases, The classification of a text's emotions, derived through SA, can then be analyzed and interpreted by separating the opinions that were looked at according to whether they were positive, neutral, or negative. The goal of sentimental analysis is to extract the contextual information, and to derive the emotions, intent, the behavior of the source of the textual information. For SA, a variety of methods are utilized, including using an approach that is based on dictionaries as well as an n-grams approach. A good technique to handle the right productivity of resources is to handle them in an efficient manner while also managing the issue before it reaches a crucial period. It is more wise and efficient to keep track on morale and the general feeling of the people than it is to do reactive answers to the circumstance.

II. STRUCTURE OF THE WORK

The following is a description of the various sections of the paper: The literature review is summarised in the third section; the methods or system design, including a description of every component of the system, is covered in the fourth section; the findings and findings interpretation are discussed in the following section; and in the fifth and final section, a conclusion is offered.

III. LITERATURE SURVEY

When it comes to the medical field, there has been a lot of research done that focuses on bringing NLP approaches for real-time applications, and when we look specifically at the healthcare domain, we see a lot of papers providing methodologies as well as datasets. The following gives a description of some of the more significant parts of the work. [9] have utilised around six methods, such as N-grams and TF-IDF, which is used for differentiating the sentiment, and worked on the model for training on the dataset, which classifies the tweets based on their sentiment, and have come to the conclusion that ML algorithms are superior in terms of suitability and exhibited substantial results when utilised with TF-IDE features as compared to the extraction of N-Gram features. Nevertheless, the authors have observed significant outcomes when logistic regression was applied to the whole in a variety of metrics.

[10] [11] have presented the study they've done using NLP techniques for Twitter for a variety of applications. The datasets, which were restricted to the English language, were analysed with several natural language processing (NLP) techniques in order to draw conclusions on opinions and feelings. However, while Twitter is available in 35 different languages, the possible data sources have been reduced as a result, and it is now important to discover ways to incorporate the languages that are not supported by Twitter.

[12] The author suggested using three levels of different approaches for feature extraction. They have attempted to classify the data using SVM, J48, and Naive Bayes. Using either an AI-based strategy or a jargon-based system, the evaluative assessment in [6] determines the location on a given material that is the farthest away from the origin. Classifiers such as Naive Bayes(NB), Support Vector Machine(SVM), and K-Nearest Neighbor(KNN (k=10)) were used on the datasets. SVM provided the most imperative precision, while KNN provided the most vital survey.

[13] Sohan and colleagues have noted that reviews of datasets have been listed for the benefit of the academic community, with a particular emphasis on the dataset that is available for use in COVID-19 investigations. The study also included a listing of the dataset, which had a variety of reports, including pathological and radiographic tests, the final test report, and a patient summary, among other things. The purpose of the review paper was to compile, and a good number of datasets are presented here.

[14] In their study, Barbara Calabrese and her colleagues emphasised the possibilities offered by social media. The benefits of using social networks are taken advantage of. The researchers have provided a list of the relevant literature. and the procedures that are stated in it since they found this to be a very helpful source of data. The methodologies with regard to data extraction, feature extraction, normalisation, modelling for contextual and emotion detection, etc., as well as specified upcoming fields of additional research such as the behavioural analysis, computing tools, etc.

[15] In their recent article, Chiara Zucco et al. provide a comprehensive evaluation of the numerous strategies that have been put into practise employing NLP techniques for SA. They have organised and categorised the numerous NLP strategies, such as lexicon-based approaches and bag of words being effective.

[16] The practise of examining the reviews of medical drugs that have been written by people or patients who have taken certain drugs is known as sentiment analysis, and it is a process that is carried out in the field of medicine. The authors have given considerable consideration to the prospect of making use of TF-IDF feature extraction in conjunction with Fast Text word embedding. In addition, some preprocessing in order to add linguistic filters to the embedded data that has been pretrained has also been given considerable thought. The purpose of this project is to attain the contextual of the reviews that were obtained, and the outcomes have showed outperformance in comparison to the models that were utilised for comparison.

[17] Yue Han and his colleagues have worked on similar uses of natural language processing in the medical area. SentiDrugs is the name they gave to the new dataset that they provided as part of their overall body of work. Its primary focus is on the analysis and interpretation of people's feelings regarding various medications. This dataset was developed with drug reviews in mind, which enables for more nuanced and accurate evaluations to be produced. Additionally, a model known as BiGRU was built, and it is utilised for the purpose of gaining semantic context. This is accomplished by the utilisation of a pretrained model known as "BiGRU," which is capable of doing analysis in both directions. In addition, a variety of instances of pertinent benchmark data were presented.

[18] Character n-grams and Twitter data have been used by the authors Hasan et al. to create a mixture of linear support vector machine learning (Linear SVM). They have brought to our attention the benefits of modelling and classifying the text by making use of n-grams, as well as the usefulness that comes along with applying these tools.

III. SYSTEM DESIGN

The System can be broken down into the following, which is depicted in figure 1. We have the pre-trained model that was trained using the PANACEA dataset of twitter, and it is used to evaluate the sentimental analysis of the twitter conversation of the covid-19 timeline. In order to obtain the model,

we used a variety of feature extractions, and by using the techniques that were explained in detail, we were able to obtain the model that is used to evaluate the sentiment of the text input.

After that, the trained model is added to the pipeline that is in charge of the categorization of the incoming data. Lastly, the results of the classification are displayed. Prior, the patient's medical history, the test results that are relevant to our assessment, and the patient's current problems are compiled and then incorporated as input data. This information is obtained by the use of a questionnaire. This data is then preprocessed using the same techniques that are given in the training process. Additionally, some techniques such as lemmatization and extra features are used to evaluate the sentiment. Finally, the model analysis is repeated for training in order to achieve better results with tuning the output.

A. DATASET

In the domain of natural language processing, a learning model contains an essential component called a dataset. The relevance of this component cannot be overstated. The development of a natural language processing model will be aided by the use of a good and diverse model, and the dataset that is used for modelling is from panacea dataset. This dataset is sourced from Twitter in order to obtain the data as an indication to the measure of people's current emotion. It is essential to keep in mind that the dataset that is utilised for modelling is the panacea dataset since it is free of any bias. When you take into consideration that social media is a real-time medium that has a massive user base of over 229 million people who are actively engaged in its many activities, The dataset is a rich source of corpus information that can be utilised in modelling of NLP [19].

[20] The data started being collected on March 11th, 2020, and it has so far yielded around 3.3 million tweets in total. The data that was gathered is multilingual, but it is still displayed in a variety of languages, as can be seen in figure 2, which can be found below. The final dataset contained a total of 990,198,297 unique tweets, while a cleaned version of the dataset contained 252,342,227 unique tweets after retweets and other duplicates were removed. Even the bigrams and trigrams necessary for the process are provided for you. We have completed a frequency analysis using the dataset that we have, and information from the live API is being used for correlation analysis.

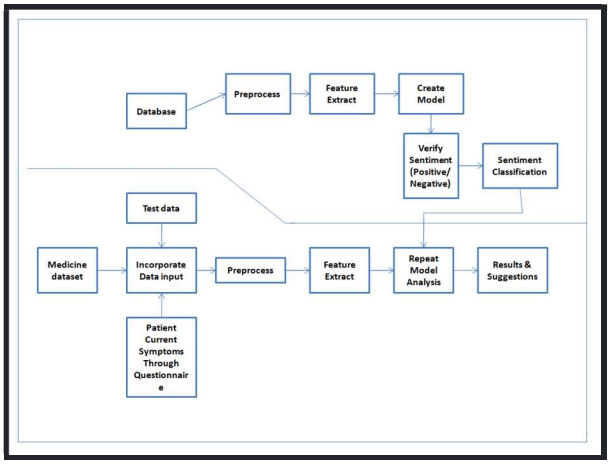


FIG 1: SYSTEM DESIGN FOR OM

B. DATA PREPROCESSING

It is vital and critical to preprocess the data in order to guarantee that the data acquired will be incorporated into the model in a way that is both unbiased and dependable. Data preprocessing refers to the act of combining several methods in order to make the data emphasise the features of the data that are significant and to ensure that the model produces data that is impartial and not skewed. The following is an explanation of the many methods utilised in the data processing.

TOKENIZATION

Tokenization is the first step in the process of breaking lengthy text or whole words down or digesting them into smaller or more identifiable components called "token." These portions can then be used in subsequent steps. The token can be anything from the entire statement to only a few words. Better tokenization is a good approach to guarantee that contextualization is carried out by the subsequent analysis and algorithms. Tokens are also a crucial component of the process used to carry out semantic and lexical analysis, and broken tokens might result in a breakdown in comprehension. Moreover, better tokenization is a good way to ensure that contextualization is performed by the subsequent analysis and algorithms.

NORMALIZATION

In order to get the processing and analysis of the data set up, we need to either normalise or standardise the data. This is done to ensure that the preparation of each text is handled in an identical manner. This ensured that the data words and lexical phrases were not given an excessive amount of weightage, which maintained a level playing field. In order to normalise data, it is common practise to remove punctuation marks and capital letters, as well as convert numbers into their word-format equivalents. This is done in order to ensure that the data is processed in a manner that is consistent throughout.

STEMMING

There is a possibility of confusion between the words that are formatted in different tenses and the words that are in a different tense. We therefore need to convert the words that have a different tense format to their base verb in order to even things out and extract the context. This will allow us to make the context clearer. For example, the verb achieve can stand in for both running and ran, as well as achieve and achievements.

LEMMATIZATION

In the process of lemmatization, an attempt is made to combine words that have a meaning that is analogous to one another or that originate from the same root or token. It can be thought of as being comparable to the process of stemming in certain respects. In addition, the Stanford Natural Language Processing Group describes lemmatization as "normally referring to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma." This definition was provided by the Stanford Natural Language Processing Group. In common parlance, "doing things properly with the use of a vocabulary and morphological analysis of words" is what's meant when someone talks about lemmatization. [21]

Lemmatization is the process of detecting whether a word is being used as a noun or a verb based on the context in which the word is being used. This determination is made based on the meaning of the

word. This brings us one step closer to our ultimate aim of increasing the number of distinct phrases that are represented while decreasing the number of individual components that make up the same job.

USAGE OF VARIOUS PREPROCESSING TECHNIQUES

In order to do trustworthy context analysis, in addition to pre-processing techniques such as stemming and lemmatization, we require the following conversions.

- Changing all capital letters to lowercase
- Getting rid of punctuation marks
- Getting rid of stop words that are already programmed in as normal English stop words
- Finally getting rid of filler phrases

NOISE REMOVAL

In almost all cases, the extraction of raw data will require the addition of noise, which is a term that refers to unwelcome elements that are coupled with the data that is used in the training process. We got rid of the noise by utilising a few of the mechanical procedures and also with the assistance of the regular expression(RE), which denote the series of characters that characterise the look for or selection of a particular character from a collection of characters, such as removal of blank spaces. This allowed us to do things like get rid of the extra spaces that were there between the words. [22] The stages involved in the preprocessing are broken down into tables I and II below.

EXTRACTING FEATURES AND BUILDING A MODEL

[23] In order to facilitate text contextual characterization and analysis, the n-gram approaches are one of the ways that have been made available. It is believed that the method is both flexible and rapid, and can deal with a number of inconsistencies in the raw data. The n-gram will be a portion of a more extensive sentence or a series of characters when it is completed. Therefore, n represents the number of characters that are anticipated to be present in that particular portion of the string. The alphabet is represented at the most fundamental level of n-gram analysis, which occurs when n equals one. This method is appropriate and effective for classification purposes. In addition, the method requires a low amount of memory to function and offers the benefits of auto stemming.

Let's have a look at an example of an N-gram using the text "The Coronavirus Pandemic" as our sample text. The n-gram can be represented in a variety of ways, including those displayed below

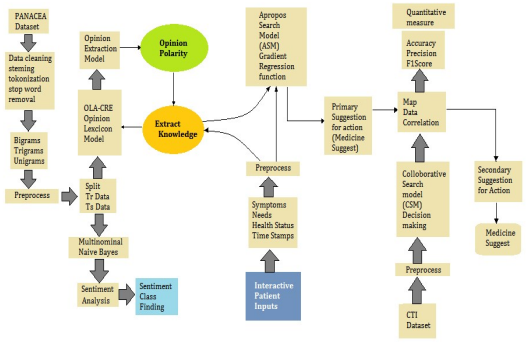


Fig2. System Architecture of proposed Novel COVID medicine Suggestion model

This method can be used to categorize the language, and it also be used to utilise the same approaches

When n = 1, it is called unigram: the, coronavirus, pandemic
When n = 2, it is, called unigram: the coronavirus, coronavirus pandemic
When n = 3, it is called unigram: the, coronavirus pandemic

by comparing the lexicon of the language with ngrams , however this strategy may require a substantial investment of resources. [18]

Fig. 2 depicts the unique opinion mining framework that was merged with COVID drug recommendation box for the purpose of this evaluation. The following is a formulation of the implementation summary of the Novel COVID suggestion box that has been proposed.

OPINION MINING

The PANACEA database serves as the basis for the formulation of opinion extractions in the proposed method, which places primary emphasis on the implementation of a novel opinion mining framework. Conversations are built up in the supplied database through the use of standard datasets. A model of the opinion lexicon is built. The fundamental purpose of the system is to offer an accurate opinion on the inputs that have been provided by using the various lexical data that is included with the database. Mining opinions is a method of data analysis that focuses on the relativity between different facts in order to establish links. Opinions are nothing more than suggestions that are formed as a result of the systematic study of a variety of already existing discussions that are very similar to one another and the formation of a similar pattern on the basis of offering beneficial suggestions to help make decisions.

OLA-CRE

Perform an initialization of the lab libraries for SCIPY, NUMBY, and MATPLOT, as well as the libraries for linear regression and the Tree tokenize tools. Compute OLM (Opinion lexicon model) to get Sentiment Class(Positive, Negative, Neutral). Calculate the CRE model by first separating the COVID dataset into the training set and the testing set. Data Compare the findings of the obtained Lexicon to serve as the initial opinion, and make use of the CRE model to study the COVID-related sentiment. Investigate the secondary opinion by viewing it through the prism of the CRE model. In order to reach a decision, it is necessary to first examine both groups' findings and then pick one to represent each attitude. The OLA-CRE model has resulted in the production of a number of outputs, which may be summed up as the ideas on conversations that were made in reference to the COVID sentiments.

INTERACTIVE WINDOW

The social concern that served as the impetus for the establishment of the suggested system and provided the framework for the public's formation of awareness and conduct of self-analysis. Through the utilisation of a software-integrated questionnaire that was designed, information regarding the user's present health state, as well as their symptoms pertaining to the existing sickness, their uncertainties, time frames on the matter, and existing doubts are gathered. In addition, these questions are produced on the backend using the most frequently asked questions and answers, and the Opinion Lexicon model is used to analyse the users' proposed answers to these questions. The interactive window is possible to collect a range of vital information on the symptoms that continue to occur and

then modify those values before sending them to the backend. These facts, which are derived from the responses, are saved as numeric information..

KNOWLEDGE EXTRACTION

This section concludes the pattern of data that must be established in order to derive judgements and construct a knowledge base from the OLA-CRE model. In order to ensure that the limits on user interactions do not have a direct effect on decision making, the unique strategy that was developed includes the incorporation of a framework for opinion analysis.

Fig 4. Knowledge extraction framework

The knowledge extraction framework is depicted in Figure, and it focuses on exploring the database for data patterns that have occurred similarly before using chosen keywords. It has been determined that the knowledge formulation is the most reliable approach to the detection of comparable occurrences. The knowledge store will record the repeated patterns in their own distinct database. Repeat the procedure until the maximum number of epochs have passed after the call request is sent to the knowledge base at each cycle of opinion extraction.

ASM ARCHITECTURE

Figure demonstrates the design of the Apropos search mechanism, which is used to carry out search operations that match up with pertinent inquiries. The questions are being generated through an integrated questionnaire framework provided by Google Collaboratory. These inquiries are taken into consideration as the input for the analysis module that continues to be experienced by the user. In the modern era of the internet, there are many interactive medication suggestion windows that are analyzed to give answers to various diseases.

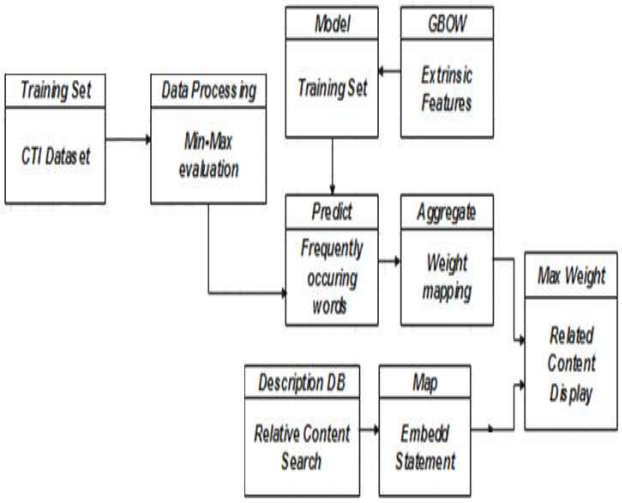


Fig.5 ASM architecture

People started having conversations with doctors and other medical professionals while they were on the move. It is not always safe to trust the recommendations that are provided by websites. There are a lot of automated bots that respond to those user interactions in order to take immediate actions. Despite the fact that there are knowledge considerations with the OLA-CRE Knowledge Base, the action recommendation is being modelled with the proposed ASM module.

- ASM is an innovative optimization model that is used to search for relevant data based on the key points that are presented.
- The model takes into account the CTI database and divides the data into a training set that accounts for 75% and a testing set that accounts for 25%.
- The COVID-19 sentiment-related keywords are accompanied by them in the generated text. The resultant text also displays pharmaceutical suggestions that come up frequently in the ASM process.

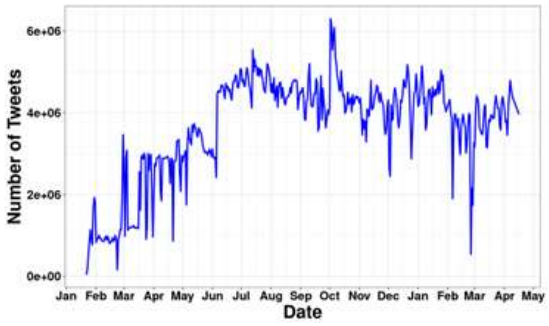


Fig 6. tweets vs date (PANACEA)

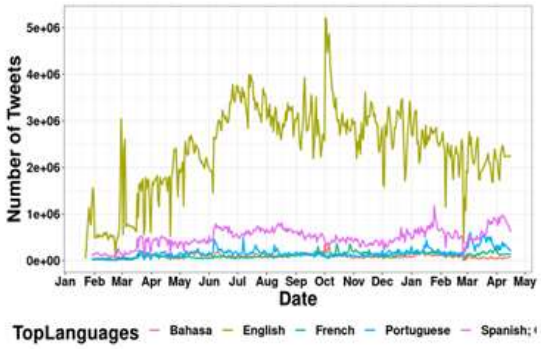


Fig 7 (a): Line plot of the language of the dataset. (b): Plotting the number of the tweets with respective month.

In line plot the number of tweets that have been plotted with respect to the time frame from which they were extracted and are shown in figure 7(b). The peak shown will correlate to the time during the pandemic when the majority of the dialogue in social networks was noticed.

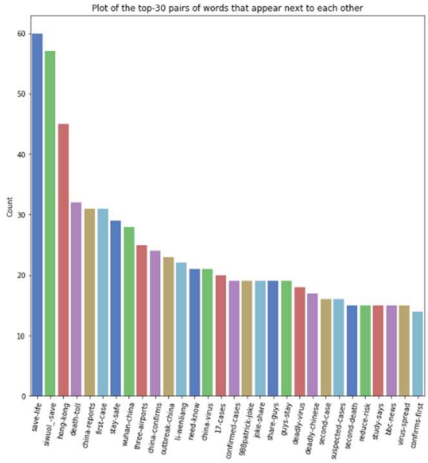
TABLE 1. PREPROCESSING STEPS

Steps	Processed Text	Data description
1	"we want to share the work we're doing to surface the right information, ... and the Tweets flowing is one of our top priorities in these difficult times"	Raw Data
2	we want share the work we're doing surface the right information, ... and the Tweets flowing is one our top priorities these difficult	Prepositions removed
3	share're doing surface information, ... Tweets flowing is top priorities	common words removed
4	share doing surface information Tweets flowing is top priorities	junk removed
5	share doing surface information Tweets flowing top priorities	processed data

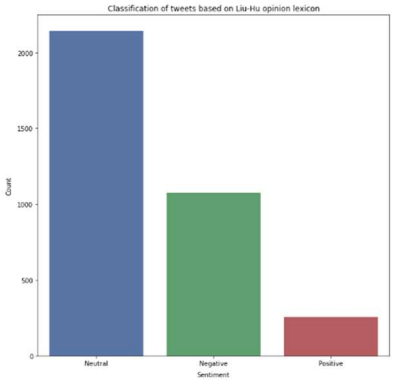
Table 1 has displayed the preprocessing, which demonstrates that the preprocessing procedures are being carried out in the manner depicted in [24]. The unprocessed text from the dataset is displayed here. A text is included for the reference, in which the in the prepositions, such as one of, and, etc., have been omitted because they contribute nothing useful to the sentimental analysis. afterwards having trash words deleted in order to receive the processed data and thereafter obtaining sentimental analysis. We can see the counts and weightage that were used to set the sentiment categorise by looking at table II, which shows the processing data that was used with hashtags and weightage that was added.

TABLE II: SAMPLE TWEET DATA THAT HAS BEEN PREPROCESSED

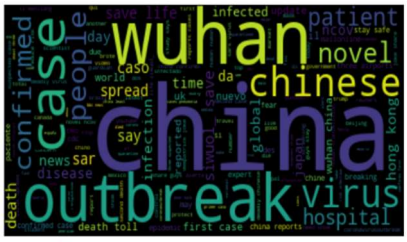
Sl No	Processed Data	Tweet Counts	Hashtags Used	Weightage
1	trying times share	4249	coronavirus	5600
2	italy trying times	4229	nan	4573
3	stand italy trying	4228	covid	4439
4	times share support	4223	19	4334
5	share support italian	4217	covid19	4321
6	support italian friends	4215	people	4254
7	italian friends colleagues	4196	trump	4252
8	tests positive corona virus	1966	via	4246
9	Hre frelno follow	1959	us	4231
10	true href rel	1913	virus	4225



(8a)



(8b)



(8c)

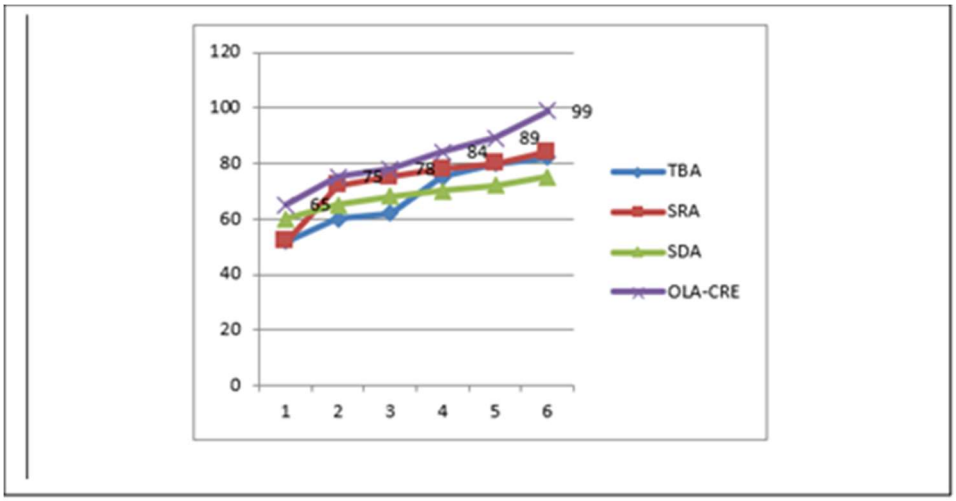
Fig 8 (a): Results of feature extraction effected on the input (b) SA of input tweet (c) Word cloud

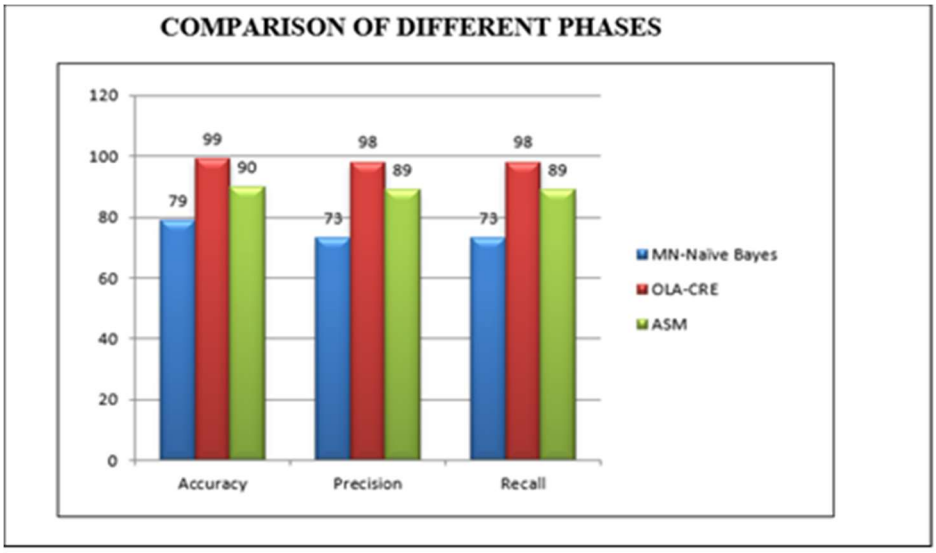
The outcomes of the preprocessing step is given in figures 8(a), (b), and (c), respectively. This wordpair reveals the word that indicates the feeling as pointed out by the pair of words, which can be explained as the similar field of the words having closer position in a token. Similarly, this near word pair reveals the words that indicate the feeling as pointed out by the pair of words. An explanation has been provided for the histogram of the word pairings that are located next to each other. However, we also see negative emotions, particularly risk and death, which have been noted showing a significant amount of emotion among people who were preoccupied with anxiety regarding the pandemic. While the pair of words that show the neutral are significantly present in the conversations that have been

noted, we also see negative emotions, particularly risk and death, which have been noted showing a significant amount of emotion. It has come to our attention that there are word combinations that are analogous, and despite this, we have identified the distribution of words that communicate emotion in 8 (b).

Last but not least, the word cloud for the aforementioned dataset may be seen in 8. (c). To create the tag clouds, first the phrases or tags are arranged, then they are placed in the cloud with varying sizes and positions that are allotted to them based on the frequency with which they appear.

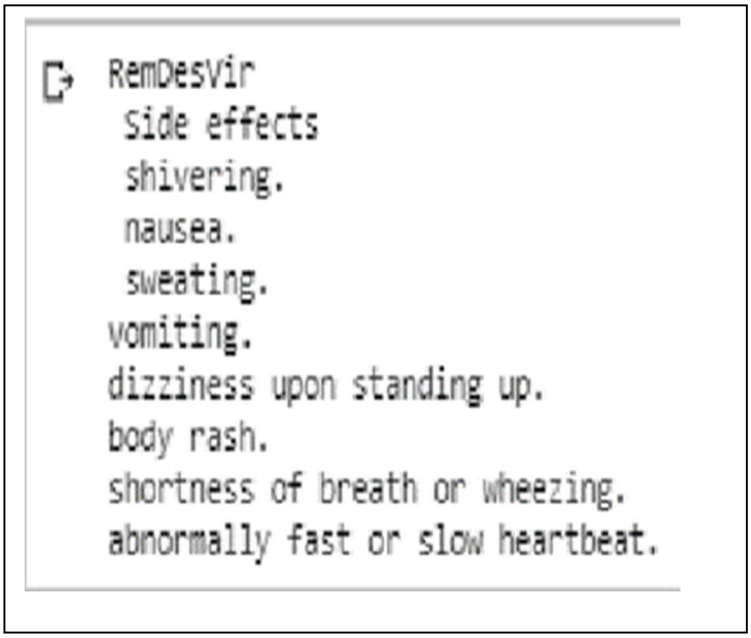
COMPARISON PLOTS ON
EXISTING METHODS VS OLA-CRE APPROACH





MEDICINE SUGGESTION

- The evaluation of the system being presented faces a number of challenges, the most significant of which is the management of the enormous dataset along with the prescribed sample analysis using the software environment.
- In order to work around this issue, the python integrated collab platform makes use of Google Collaboratory in order to improve the visualisation of the large dataset. The necessary library functions that are required for the Python code are automatically retrieved by the Collab environment on their own.



- The dataset containing global languages is then analysed using the NLK toolkit with python as the primary programming language. In light of the fact that the garbage data are influenced more by the live dataset, it is necessary to derive the common principle. Therefore, in order to construct an accurate model, it is necessary to choose a benchmark dataset in the beginning.

IV. CONCLUSIONS

This paper presents a suggestion for enabling sentiment analysis for medical purposes by making use of pre-processing techniques and constructing natural language processing (NLP) systems. Twitter was one of the online media platforms that we explored for our inquiry. Tweets were compiled in order to get a better feel for how Indians are feeling about the lockdown. From March 1st to July 10th, 2020, tweets were separated using the following prominent hashtags to be more specific: #COVID, #Coronavirus, #Lockdown, #Pandemic, and #PMCares. The total number of tweets that were taken into consideration for the analysis was 5,74,108. Python was used to complete the analysis, and a number of diagrams were drawn up to show how the many tweet-based hypotheses related to the topic fit together. The emphasis of the suggested approach is on a lightweight and adaptable model, and it makes use of an opinion Lexicon process in conjunction with a Bags of Words model. The proposed approach achieves a better accuracy of 99% and, once the analysis result is obtained, formulates the

medicine that is most usually suggested. The medication recommendation was trained with the CTI dataset (Clinical trials government). The unique methodology that has been proposed can be expanded upon as a result of the vast amounts of data that have been collected all around the world. More datasets need to be gathered around the world in order to improve the accuracy of the evaluation process.
It is recommended that you use the expanded lexicon method through the deep learning methodology.

REFERENCES

[1] Official COVID-19 WHO tracking website https://covid19.who.int/

[2] Jensen, L., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7, 119–129 (2006). https://doi.org/10.1038/nrg1768

[3] Andrey Rzhetsky, Michael Seringhaus, Mark Gerstein Seeking a New Biology through Text Mining, VOLUME 134, ISSUE 1, P9-13, JULY 11, 2008, https://doi.org/10.1016/j.cell.2008.06.029

[4] Altman, R.B., Bergman, C.M., Blake, J. et al. Text mining for biology - the way forward: opinions from leading scientists. Genome Biol 9, S7 (2008). https://doi.org/10.1186/gb-2008-9-s2-s7

[5] Fiorini N, Canese K, Starchenko G, Kireev E, Kim W,

Miller V, et al. (2018) Best Match: New relevance search for PubMed. PLoS Biol 16(8): e2005343. https://doi.org/10.1371/journal.pbio.2005343

[6] Lever, J., Zhao, E.Y., Grewal, J. et al. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods 16, 505–507 (2019). https://doi.org/10.1038/s41592-019-0422-y

[7] Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. (2018) Best Match: New relevance search for PubMed. PLoS Biol 16(8): e2005343. https://doi.org/10.1371/journal.pbio.2005343

[8] Yu H, Yang C-C, Yu P, Liu K (2022) Emotion diffusion effect: Negative sentiment COVID-19 tweets of public organizations attract more responses from followers. PLoS ONE 17(3): e0264794. https://doi.org/10.1371/journal.pone.0264794

[9] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja, The Impact of Features Extraction on the Sentiment Analysis, Procedia Computer Science, Volume 152,2019, Pages 341-348, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.05.008.

[10] Imran Ahmed, Misbah Ahmad, Joel J.P.C. Rodrigues, Gwanggil Jeon, Sadia Din, A deep learning-based social distance monitoring framework for COVID-19, Sustainable Cities and Society, Volume 65, 2021, 102571, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2020.102571.

[11] Mohammad Shorfuzzaman, M. Shamim Hossain, Mohammed F. Alhamid, Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic, Sustainable Cities and Society, Volume 64, 2021, 102582, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2020.102582.

[12] International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 8, August 2014

[13] Sohan, Md. Fahimuzzman. “So You Need Datasets for Your COVID-19 Detection Research Using Machine Learning?” *ArXiv* abs/2008.05906 (2020): n. pag.

[14] Calabrese, B., Cannataro, M., Ielpo, N. (2015). Using Social Networks Data for Behavior and Sentiment Analysis. In: Di Fatta, G., Fortino, G., Li, W., Pathan, M., Stahl, F., Guerrieri, A. (eds) Internet and Distributed Computing Systems. IDCS 2015. Lecture Notes in Computer Science(), vol 9258. Springer, Cham. https://doi.org/10.1007/978-3-319-23237-9_25

[15] C. Zucco, H. Liang, G. D. Fatta and M. Cannataro, "Explainable Sentiment Analysis with Applications in Medicine," *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 1740-1747, doi: 10.1109/BIBM.2018.8621359.

[16] A. Yadav and D. K. Vishwakarma, "A Weighted Text Representation framework for Sentiment Analysis of Medical Drug Reviews," *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 326-332, doi: 10.1109/BigMM50055.2020.00057.

[17] Y. Han, M. Liu and W. Jing, "Aspect-Level Drug Reviews Sentiment Analysis Based on Double BiGRU and Knowledge Transfer," in *IEEE Access*, vol. 8, pp. 21314-21325, 2020, doi: 10.1109/ACCESS.2020.2969473.

[18] Hassan, Noha Y. et al. “Credibility Detection in Twitter Using Word N-gram Analysis and Supervised Machine Learning Techniques.” International Journal of Intelligent Engineering and Systems 13 (2020): 291-300.

[19] L. Li *et al*., "Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo," in *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 556-562, April 2020, doi: 10.1109/TCSS.2020.2980007.

[20] Twitter daily user growth rises as Musk readies to take control, 28 Apr 2022, https://www.aljazeera.com/economy/2022/4/28/twitter-daily-user-growth-rises-as-musk-readies-to-take-control#:~:text=%2C%E2%80%9D%20without%20elaborating.-,Twitter%20reported%20an%20average%20of%20229%20million%20daily%20active%20users,users%20in%20the%20previous%20quarter

[21] Introduction to Information Retrieval, By Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, Website: http://informationretrieval.org/, Cambridge University Press, © 2008 Cambridge University Press

[22] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja, The Impact of Features Extraction on the Sentiment Analysis, Procedia Computer Science, Volume 152, 2019, Pages 341-348, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.05.008.

[23] Nasser, Nidal et al. "*n*-Gram based language processing using Twitter dataset to identify COVID-19 patients." *Sustainable cities and society* vol. 72 (2021): 103048. doi:10.1016/j.scs.2021.103048

[24] PSYCHOLOGY AND EDUCATION (2021) ISSN: 0033-3077 Volume: 58(3): Pages: 1227-1232