Health Surveillance System and Disease Prediction with Social Media Data

P.Varalakshmi[1], *Hemanth N*[2], *Rubak Preyan G* [3], *Yogeeswar S*[4]
Department of Computer Technology
Anna University,MIT Campus
Chennai,Tamil Nadu
{*varanip*[1], *hemanthnov2001*[2], *rubakkeerthi123*[3], *yogeeswar9656vm*[4],}*@gmail.com*

***Abstract*—Public health surveillance is the systematic collection, analysis, and interpretation of data, closely integrated with the timely dissemination of these data to those responsible for preventing and controlling disease and injury. With a growth of population a proper public health surveillance system is a must, but large-scale collection of data related to mental health is traditionally difficult and carried out infrequently. In this paper we have proposed a health surveillance and disease prediction model with social media data mainly tweets from twitter and data from Reddit, using various natural language processing techniques. We propose a model that uses ARIMA and LSTM for forecasting a disease trend overtime, forming a correlation between available clinical data and the twitter data collected. The resulting insights are expected to help facilitate faster response to and preparation for epidemics and also be very useful for both patients and doctors to make more informed decisions.**

***Index Terms*—Twitter API, NLP ,BERT Model, DistilBERT, XLNet Model, Mental Health Analysis, Disease Prediction.**

I. Introduction

One of the most important lessons from the Covid 19 pandemic is early detection, which would significantly reduce the impact of an outbreak. Apart from this, in India, particularly Tamil Nadu, many people are prone to diseases like dengue. With the help of surveillance systems in the areas with an increasing number of cases, steps required to prevent the cause of disease, like for example mosquito breeding, can be implemented as a preventive measure in the early stages.

With the proliferation of the internet, a new potential for data sources has evolved. An increase in advent of technology, also means it is necessary for a mechanism to monitor the amount of useful and socially relevant data from the social media platforms, especially twitter. Messages propagating in real time is one of the unique features of analyzing twitter data. Researchers overtime have been using data from Twitter to predict a number of real world outcomes. Thus future prediction of disease levels could open up a new scope in the healthcare area since such predictions would provide useful and workable insights for the public health sector which is further used for better planning, allocation of resources , public health treatments and their prevention. As an identified improvement to other approaches, we propose a system that not only evaluates current disease activities with an accurate ratio, but also predicts future disease trends in advance. Apart from this, the proposed system also monitors the mental health of people in that particular area which also helps the government and healthcare professionals to take appropriate actions.

II. Objectives

With a dramatic increase in popularity and usage of social media, especially twitter, where people share ideas, events, as well as life stories, we can identify growing needs to extract useful information from an enormous amount of data, and increase the potential for valuable healthcare insights. So, the main objective of this research is to enhance the real time surveillance in the field of health informatics with the help of social media data, this surveillance system is mainly proposed for an early prediction of upsurge in diseases and also future trends of the disease in a particular geographical location using the various data from the social media users. Traditional surveillance systems might miss a rare event (like a new viral outbreak), and will definitely lack the real time monitoring abilities and demographic reach which the social media can provide. The proposed system's objective would be to estimate the magnitude and level of disease overtime and also estimate the overall mental health status of people in that particular area, and provide assistance to healthcare authorities to provide required medical services and organise awareness camps based on the possible extent of any disease in an area, and ensure safety of the citizens from communicable diseases.

1) To predict the impact of a particular disease in a particular area, using various NLP (Natural Language Processing) models, after data acquisition using twitter API.
2) To evaluate current trends, as well as predict future trends of a particular disease, combining real time streaming data and observed CDC (centres for disease control and prevention) data.
3) To predict the overall mental health of the people in a particular area using supervised and unsupervised learning, considering that users of social media do not constitute the whole of the population, with the speculation that few people might not post publicly about health status, and people in discomfort might not be active users.

III. Literature survey

SreeJagadeesh Malla et al[6]. examined the type of disease and interrelated with various other disease types and also concentrates about the identification of COVID - 19. This process is done with the help of Tweets which are uploaded in the social media Twitter. The work of this paper mainly concentrates on the Tweets which are informative in detecting the disease present in this pandemic situation and provide the necessary data to the government and all the other organisations. The data which is abstracted from the Tweets are categorised with the help of unique resources to help the sufferings. The process of identification is done with the help of concepts like Deep Learning , NLP and also do Sentimental analysis by using the mentioned algorithms. They use MVEDL (Majority Voting technique - based Ensemble Deep learning) which is mainly used to identify the informative tweets. Some of the states of MVEDL like CT-BERT and BERTweet are used. The core novelty of the paper was proposing a model with the help of deep learning technique which is mainly used to examine the informative tweets by eliminating the worthless tweets and to detect the relevant latent issue and emotional classification on the COVID - 19 tweets.

Son Doan and et al. in [9] have formulated an approach using natural language processing to find causality from tweets. They have collected about 24 million tweets from major cities for about a time span of a month, these tweets are first filtered with some target keywords relating to healthcare following this various NLP techniques such as sentence splitter, lemmatizer, parts of speech tagger, and a dependency parser. Finally based on dependency parser generated syntactic relations, causal-relations are identified. The cause-effect relation is extracted by creating rule set templates which include trigger nouns and verbs, which were found from WordNet. Some of the limitations in this paper are, one number of rules and patterns are small which may overlook cause-effect relations, two twitter has diverse ways of representing cause-effect relation such as hashtags but this paper is not able to extract those, third they have not considered synonymous expressions of target.

Tim Mackey et al. in [5] have used Machine learning techniques to detect the self reporting symptoms , finding out the testing access and recovery strategies of COVID - 19 with the help of Tweets which are posted in the social media “Twitter”. The main motive of the study is to find out and categorise the user generated conversations (Tweets) which are related to COVID - 19 indications or experiences or the recovery by using the Machine learning approach . The process is done by filtering the Tweets by using some of the general keyword terms which are related to COVID - 19 and they are furthermore filtered that are self - reported by the users. Here the Tweets are examined by the unsupervised machine learning approach which is called BTM (Biterm topic model) where this model is used to categorise the Tweets which are related and those categories are separated in clusters.

Oduwa Edo-Osagie, Ian Lake in [7] have summarised different approaches to diagnose each disease with the help of social media data available. For each article that they have reviewed, information like disease on focus, ML algorithms used, country in focus were all considered and results were well tabulated along with the methods like ARIMA, Binomial Regressions, Bayesian Inference, Temporal Ailment Topic Aspect Model for different diseases like Haemophilus, Food Borne Illness, Heat Related Illnesses, Gastrointestinal Illnesses, Cholera. They have actively reviewed many literature of Public health research that uses twitter api, but their main focus was on such literature which were limited to research that concerned with the monitoring, detection and forecasting of public health conditions.

Nelisetti Ashok et al. in [8] have used Deep Neural Network which was implemented using TenserFlow framework to model the probabilistic layer that intuitively identifies markings of schizophrenia discourse and performs classification algorithms on the same. Intuitively distinguishing patients with and without schizophrenia has been the singular goal of the proposed research work. They have proposed a dynamic and unique machine learning approach with a neat flow diagram to forecast the schizophrenia discourse using data from Twitter API.

Iwan Syarif and et al. in [10] have built a mental health detection model using publically available twitter data using rule based decision tree method with features that can be extracted from social media posts, For collecting data from twitter they have used twitter API and various preprocessing on the data has been done, such as filtering out the users with less than 25 tweets and also only selecting those whose tweets are at least 75 percent in english. They have considered two features, linguistic style feature and expression sentiment feature. In linguistic style features they have considered the use of absolutist words and self-referring words. For expression sentiment style they have considered emotional analysis, sentiment analysis and frequency of words that match the depression lexicon word count.

Kathy lee and et al. in [2] presented a system that will give more accurate, real time influenza activities notices and also mainly combines the data which is obtained from social media in real time and the dataset of CDC (centres for disease control and prevention) is used in order to obtain the precise predictions. This paper work is the improved version of the older paper which gives more advancement in the accuracy and also to predict the activities of the flu which may occur in future. This predictive model with the help of social data and the observed CDC data. The model is made up of multilayer perceptron with backpropagation where it uses large amounts of Twitter data which is mainly used to track the present and future flu prediction with large amounts

of accuracy.. The final aim of the project is to identify the influenza activities for future purposes which will help in arrangement , interference , resource allocation and avoidance.

Jia Xue and et al. in [13] have examined COVID-19–related discussions, concerns, and sentiments using tweets posted by Twitter users. Their study aims to examine the public discourse and emotions related to the COVID-19 pandemic by analysing more than 4 million tweets collected between March 7 and April 21, 2020. They ran a one-tailed z test to examine if each of the 8 emotions is statistically significantly different across topics. This research by the authors identified the need for a vaccine to stop the spread, quarantine and shelter-in-place orders, and protests against the lockdown. The main limitation of the paper comes in their dataset where they have sampled only 20 hashtags as the key search terms to collect Twitter data.

Lu He and et al. in [12] evaluated three general purpose sentiment analyzers popularly used in previous studies such as Stanford Core NLP sentiment analysis, TextBlob and VADER based on two online health datasets and a general purpose datasets as the baseline. The results show that these general purpose sentiment analyzers were unable to produce consistent results when applied to the same dataset and their performance varies when applied to different datasets. These findings suggest that general purpose sentiment analyzers developed in non healthcare domains may perform poorly on online health data.

Muhammad Riaz and et al. in [11] aims to identify the main potential prcusoors of the social media health information seeking intension (ISI) and examine their effects on heath information re-sharing behaviours and PHH during coronavirus (COVID - 19) pandemic. The data is collected through an online survey conducted in two different universities situated in highly COVID - 19 affected cities - Wuhan and Zhengzhous , China . The valid data consists of 230 useful responses from WeChat users and to analyse the final data set structure equation modelling (SEM) is used. The paper investigates the health information intentional behaviour precursors and their consequences via WeChat during COVID - 19 pandemic. Future studies may conduct research by examining online information behaviours on other social media platforms like Twitter , WhatsApp , Facebook etc.

IV. Proposed work

Some of the limitations in early research relating to similar idea and possible solution proposal are as follows,

1) C0 - Sentiments of the tweet are not considered, which plays an important role because the user may have put that in a sarcastic way when that is considered it will be a noise in the dataset. In such cases more training may be needed to modify the existing algorithms. To overcome this limitation, natural language processing, vectorization techniques could be used.
2) C1 - Most machine learning models were created for only a particular disease but the monitoring system should consider all or most of the diseases.

The main source of social media data that we have identified for the surveillance system is Twitter data, that would enable real time analysis to those involved in investigation of a potential epidemic outbreak, the main reason being that social media has that extra edge in covering topics that might not be covered in traditional data sources available for public health. Analyzing the posts and messages which are posted during the pandemic by the people will help the health care agencies and other volunteers to get a proper understanding about public's impertinence sentiments and necessities which in turn help to deliver the necessary information. The assistance to health authorities in devising practical public health services strategies would require structured social media information, hence ensuring a strict quality in the fetching of tweets is essential. Efficient and effective models might help in processing such data to evaluate the incidence of disease.

The model proposed would use various ML and deep learning models on a huge dataset from twitter, hence forecasting current and future extent of a disease, and positively with a high accuracy. Usually data would be available as self-reported syndrome and also clinically confirmed data. Twitter texts precisely reflect the real world, and the NLP techniques can extract the useful information from twitter streams, which also identifies user social circles with similar medical experiences.

An optimized model proposed is planned to achieve by applying various NLP components like sentence splitter, lemmatizer, Part-of-Speech (POS) tagger, and a dependency parser and finally, identifying the disease and features based on syntactic relations generated by the dependency parser. Continual collection of the twitter data would improve the performance of the model, for forecasting disease activity, and better resource planning.

The Fig 1. explains the architecture of the proposed system, first using the twitter API and the selected keywords tweets are in the form of JSON format, These data are then processed using natural language processing techniques such as tokenization which breaks down sentences into words, lemmatizer where different inflected forms of a word is groped together, these are then parsed and POS tagging is done which is then converted into a numerical encoding using TF-IDF algorithm, this is then fed into machine learning models to analyze the sentiment of the tweet this is done to ensure the correctness of the tweet as the user may have lied or used a sarcastic tone this is also used in analyzing mental health of the people in that particular area. These counts of the correct tweets of that particular disease are aligned with the CDC data to predict the future impact of that disease using the Time series model.

V. Implementation

From the Fig. 1 we can divide the system into 3 main modules, one to collect data and process it another to analyse

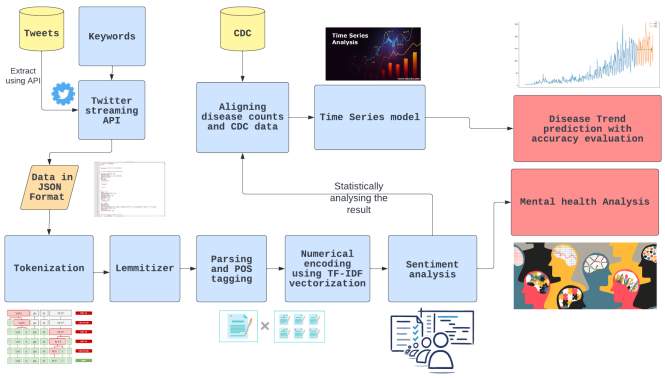

Fig. 1. Architecture Diagram

the overall mental health of a particular area and the last module if takes care of predicting the impact of a particular disease based on the data collected, all these modules are explained in detail below.

A. Data collection

Approval of elevated access from the Twitter Developer platform, allowed extraction of tweets from twitter using a set of keywords relating to a disease. Raw tweets had several discrepancies, hence preprocessing techniques like CountVectorization and TF-IDF Vectorization were usedto remove the discrepancies. These tweets were then used to create a word-cloud, using packages in python, in order to extract the most important words out from the tweets. Fig. 2 shows a word cloud formed with the collected tweets. These are then used as keywords and tweets are again fetched using twitter's application programming interface (API), following which these tweets are sent to Mental health analysis module.

B. Mental Health analysis

Word embeddings, Document embeddings, TF-IDF analysis, Topic models, Dimensionality reduction, Clustering Language models like BOW were used that allowed exploring how different embedding approaches provide good ways to separate the labels and vectorize full sentences and plot them in a 2-dimensional space. The foremost aim was to distinguish between different mental health conditions - between 3 labels: anxiety, depression and normal. The improved model by hyper-parameter tuning and reduction of dimensionality perform better than the dumb classifier, but not better than the standard model without hyper parameter tuning in a statistical significant way. In this method emotion detection on related keywords are used. Keywords are matched with predefined emotional keywords and mental health have been classified into 7 categories - anxiety, suicidewatch, casual, BPD, bipolar, depression, and happy.

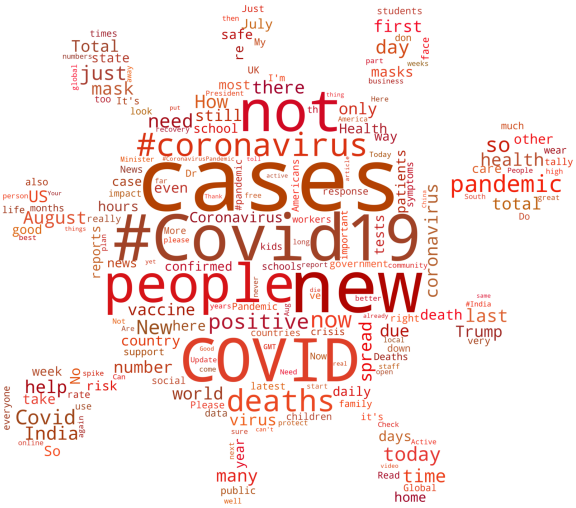

Fig. 2. Word Cloud

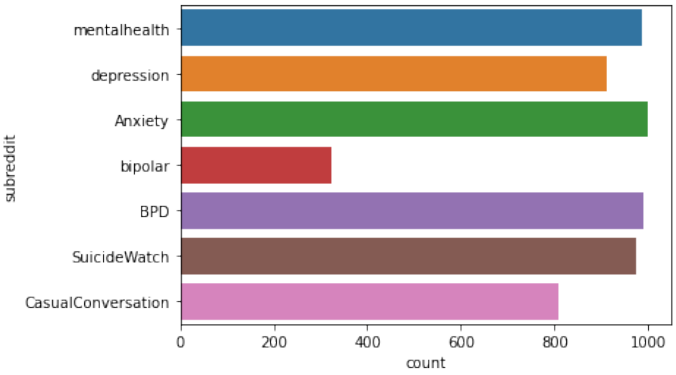

Fig. 3. Reddit data analysis

C. Disease prediction

Disease forecasting has been done using ARIMA models and LSTM models which have been used extensively in other fields, but have not been used wisely in public health. Our findings indicate that the ARIMA model is valid for identifying periods of heightened activity on Twitter related to health and disease outbreaks. The model offers an objective and empirically based measure to identify periods of greater interest for timing the dissemination of credible information related to disease forecasting. We identified and corrected for two types of bias present in Twitter data: (1) demographic variance between US Twitter users and (2) natural language ambiguity, which creates the possibility that mention of a disease name may not actually refer to the disease. The correlation between disease prevalence and Twitter disease mentions both with and without bias correction were measured

that allowed us to quantify each disease's over representation or under representation on Reddit, relative to its prevalence.

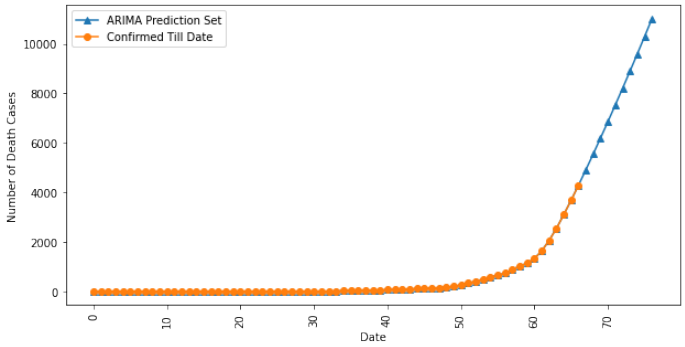

Fig. 4. ARIMA prediction

VI. Results and Discussion

Our sample included more than 710,000 tweets. Adjusting disease prevalence to correct for Twitter demographics more than doubles the correlation between Twitter disease mentions and disease prevalence in the general population - from almost .113 to nearly .258. Ambiguity correction was applied to our Twitter corpus that achieved a correlation between disease mentions and prevalence of .208. Simultaneously applying correction for both demographics and ambiguity more than triples the baseline correlation to .366. Compared with prevalence rates, cancer appeared most over represented in Twitter, whereas high cholesterol appeared most under represented in twitter and Reddit collectively.

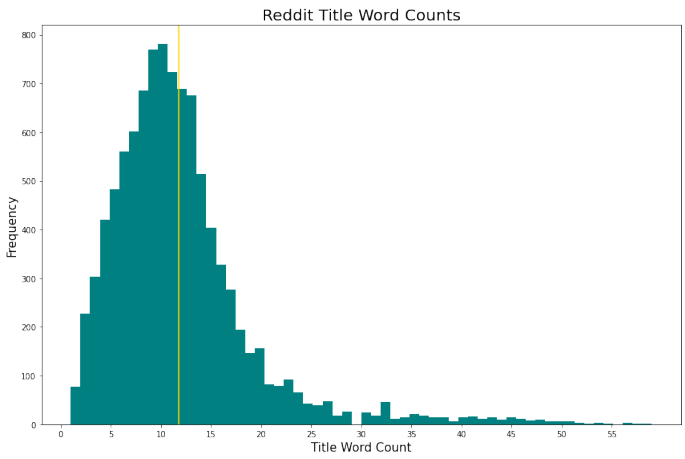

Fig. 5. Word Count Frequency

Bidirectional Encoder Representation of Transformers (BERT) is a self-supervised technique for NLP tasks. The model is usually pre-trained on large datasets like Wiki and then fine tuned for various different tasks in natural language processing. In order to classify the tweets as real or fake, we have added a classification layer on top of the pre-trained BERT architecture. The BERT model performs well in most of the cases, than XLNET architecture, that performs well with few parameters but not better than BERT when many parameters and larger dataset of tweets were considered.

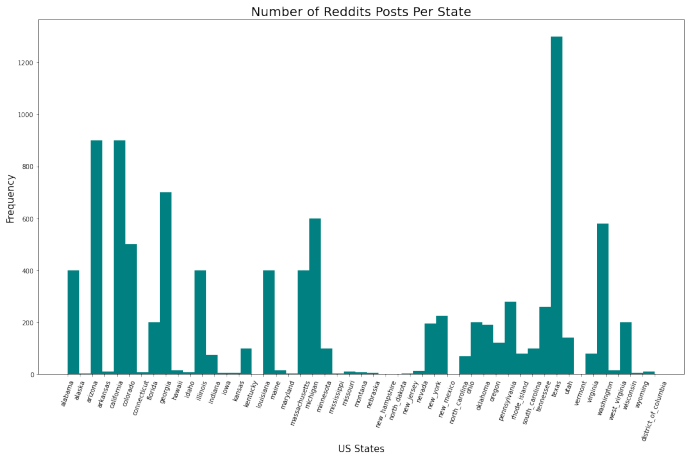

Fig. 6. Number of reddit users

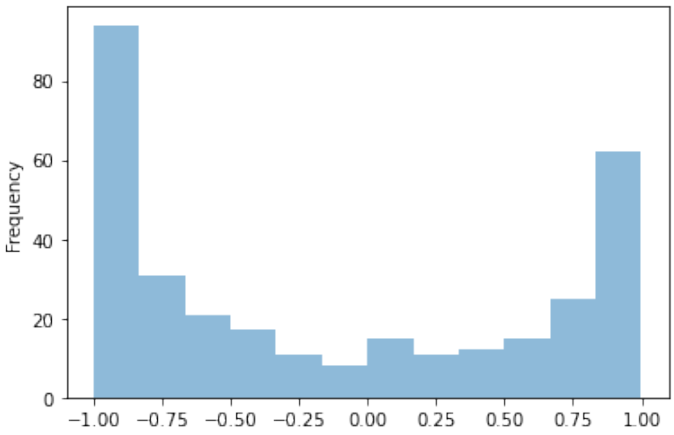

Fig. 7. Average vs Frequency

The plot is able to distinguish anxiety from the others main conditions, but overlaps depression and suicide posts. The goal should be to spot those suicide posts which are more extreme that the average depression post, indicating a worsening of the condition. The sentiment of Casual is skewed towards the positive spectrum. SuicideWatch has the distribution most skewed toward the negative spectrum. Interesting observation is that bipolar and BPD show more even distributions. This is in line with the specific conditions, which tend to have both highs and lows.

Comparing the two types of lemmatizations, the one keeping only Nouns and Adjectives seemed better after severe analysis. Even though it is more aggressive, meaning that we lose more content, it reduces the variability drastically, yet still conveying the important information and the pre- and post- processing analysis is depicted in the below graph that includes number of characters and words in the text before and after cleaning.

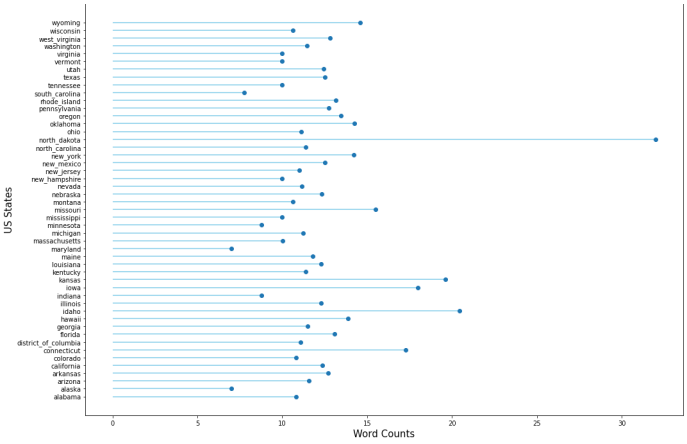

Fig. 8. Word Count by State

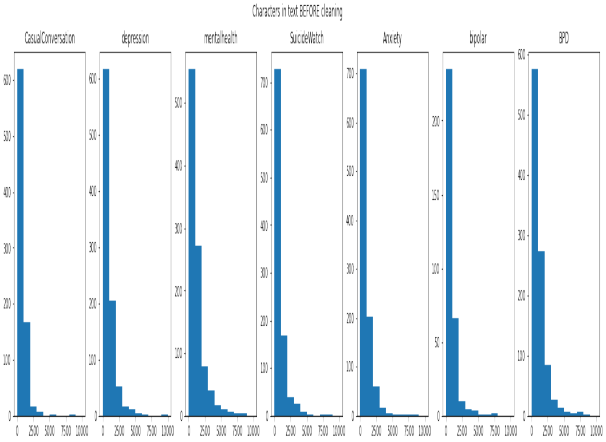

Fig. 9. Characters in text Before processing

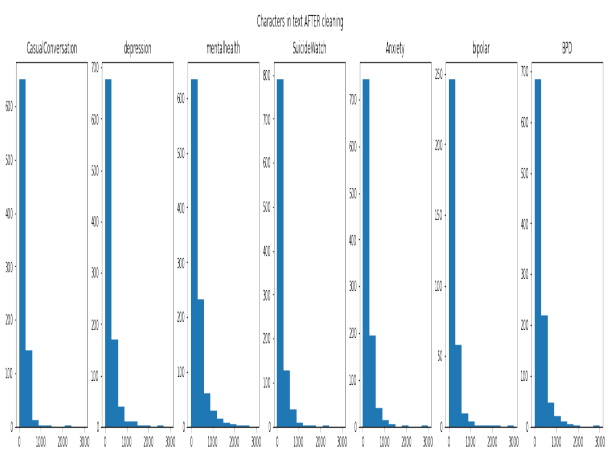

Fig. 10. Characters in text After processing

VII. Conclusion

Twitter is a potentially useful tool to measure public interest in and concerns about different diseases, but when comparing diseases, improvements can be made by adjusting for population demographics and word ambiguity. It is learnt that exploring and analyzing data can help to weed out unnecessary information. Considering that traditional survey methods are time-consuming and expensive,timely and proactive data sources to respond to the rapidly evolving effects of health policy on our population's mental health is necessary, and our model performs the function to an expected level.

The traditional way of public health surveillance does have its own limitations and known biases, hence if this epidemiology is digitalised, it does offer a new scope and vision to extract signals and data from social Media that might be complementary to official statistics. The complete clinical data is only available to a limited group, that too with an enormous amount of pay, and hence with an unprecedented growth of usage of social media, it is wise to extract relevant data to achieve better results with respect to disease-trend prediction in any geographic location.

References

[1] Akansha Jain, Sreejith Cherikkallil(2018), Twitter based Platform for Health Care , Analytics, Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2018)

[2] Kathy Lee, Ankit Agrawal, Alok Choudhary(2017) , Forecasting Influenza Levels using Real-Time Social Media Streams, 2017 IEEE International Conference on Healthcare Informatics

[3] Aakansha Gupta, Rahul Katarya, Social media based surveillance systems for healthcare using machine learning: A systematic review July 2020 Journal of Biomedical Informatics

[4] Calix, Ricardo A.; Gupta, Ravish; Gupta, Matrika; Jiang, Keyuan (2017). IEEE 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) - Deep Gramulator: Improving Precision in the Classification of Personal Health-Experience Tweets with Deep Learning

[5] Marcelo Aparecido Carlos, Marcelo Nogueira , Analysis of Dengue Outbreaks Using Big Data Analytics and Social Networks , November 2017, IEEE Conference Record of Annual Pulp and Paper Industry Technical Conference

[6] Malla, SreeJagadeesh, P.J.A, Alphonse. Applied Soft Computing, May 2021 ; 107:107495, 2021. COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets

[7] OduwaEdo, OsagieaBeatriz, IainLake, Obaghe Edegherec, July 2020 , A scoping review of the use of Twitter for public health research

[8] Nelisetti Ashok, Usha Nandhini, (2021) Advances in Systems, Control and Automations, Select Proceedings of ETAEERE - Unique and Dynamic Approach to Predict Schizophrenia Disease Using Machine Learning,

[9] Doan, S., Yang, E.W., Tilak, S.S. et al. Extracting health-related causality from twitter messages using natural language processing. BMC Med Inform Decis Mak 19, 79 (2019). https://doi.org/10.1186/s12911-019-0785-0

[10] Syarif, I., Ningtias, N., Badriyah, T. (2019). Study on Mental Disorder Detection via Social Media Mining. 2019 4th International Conference on Computing, Communications and Security (ICCCS).

[11] Riaz, M., Wang, X., , S. and Guo, Y. (2021), "An empirical investigation of precursors influencing social media health information behaviors and personal healthcare habits during coronavirus (COVID-19) pandemic", Information Discovery and Delivery, Vol. 49 No. 3, pp. 225-239. https://doi.org/10.1108/IDD-06-2020-0070

[12] He, Lu, and Kai Zheng. "How Do General-Purpose Sentiment Analyzers Perform when Applied to Health-Related Online Social Media Data?." Studies in health technology and informatics vol. 264 (2019): 1208-1212. doi:10.3233/SHTI190418

[13] Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, Zhu T Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach J Med Internet Res 2020;22(11):e20550