A Framework for Forecasting Outbreak of Infectious Diseases Based on Climate Variability and Social Media Content

Juliet Johny, Linda Sara Mathew

Abstract*: *The amount of data has risen significantly over the last few years, due to the popularity of some of the data generation sources like social media, electronic health records, sensors and online shopping sites. Analyzing, processing and storing this data is very prominent since it helps to uncover hidden patterns and unknown correlations. A big data analysis and prediction System is proposed in this context, which combines weather observations, health data and social media content in order to forecast the outbreaks of infectious diseases in a locality. Finding information about the determinants of disease outbreaks are required to reduce its effects on populations. An In-mapper combiner based MapReduce algorithm is used to calculate the mean of daily measurements of various climate parameters like temperature, atmospheric pressure, relative humidity, solar and wind. The climatic parameter that may leads to the outbreak of a disease is identified by finding the correlation between the parameters and disease incidence count. To evaluate how user's tweeting patterns and sentiments matched with the outbreak of diseases, all tweets containing keywords related to diseases are collected using twitter streaming APIs and are analyzed and processed using Spark framework. The performance of proposed model is improved due to the presence of tweet processing. This indicates that the real-time analysis of social media data can provide more effective result rather than working on the historical data.

Keywords* : *Apache Spark , Hadoop MapReduce, Kafka, Spark MLlib

I. INTRODUCTION

In recent years, the volume of data is increased enormously. This data is said to be big data when it is beyond the processing capabilities of a system. Huge volume of data is generated from various sources. One of the popular data generation sources is sensors. They are some electronic devices that continuously monitor various parameters such as heartbeat, body temperature, pressure, humidity etc. The data that is collected from them are continuously fed to some systems and further used for various applications such as, real time notifications in medical areas. Another popular data generation source is social media. Popular social medias such as twitter, Facebook, Instagram and whatsapp give us the permission to do various activities such as uploading pictures, like, share, comment etc. While doing all these activities, with or without our knowledge, we are actually generating huge volume of data. All these data will be stored in the corresponding databases. Various transactions in the bank sector also generate lots of data. Nowadays banks provide their web and mobile applications like GooglePay and FedMobile. While transfer-ring money to other accounts and recharging mobile phones through such applications, lots of data are generating in the background. We have some huge volume of data. So a proper method is always prominent for the efficient storage and processing of this data. Here comes the need of big data analytics. It is an area of studying about huge volume of data which helps to uncover useful information such as unknown correlations and useful patterns. Big data analytics [1] is a major area which has applications in various fields. In health sector, different medical data and electronic health records are generating day by day, which can be considered as a type of big data. By analyzing and studying about the health records of a particular persons, an efficient machine learning algorithm can predict, whether there exist any chances of occurrence of some for him in the near future. In an organization's point of view, big data analytics helps to collect product user's information and identify customer preferences and market trends. This helps them to make better business decisions. So big data analytics leads to efficient marketing. Climatology [2] is an area of studying about climate. Climate data is collected from various weather stations all over the world. Major application of climate data analytics is weather forecasting. Weather data is collected and analyzed for the better forecasting. There always exist a connection between climate of a region and the agriculture that is cultivated in that region. Similar to this, the disease that outbreak in a locality may have some connections with the climate of the region. As an example, consider the case of an infectious disease dengue. Aedes aegypti mosquito is the primary factor of dengue disease. High temperature condition is very suitable for its growth and that is the reason why this disease is very high in a locality during summer days. Similar to this there are various other diseases whose outbreak depends directly on the climatic condition of that locality. Our aim is to develop a framework for predicting outbreaks of such infectious diseases through climate data analytics and social media content. Twitter is a popular and widely used social media. People usually uses twitter to express their opinions [3] regarding various issues and socially relevant problems. By processing twitter data, the sentiments of the people during a climate change or during the outbreak of any infectious disease can be analyzed. The rest of the paper is as follows. Section 2 describes the related works in this area. Section 3 briefly explains the proposed work. Section 4 describes obtained results and the last section concludes the paper.

Manuscript received on January 06, 2021.
Revised Manuscript received on January 15, 2021.
Manuscript published on January 30, 2021.
* Correspondence Author
Juliet Johny, Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India, julietjohny5@gmail.com
Linda Sara Mathew, Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India, lindasaramathew@gmail.com

II. RELATED WORKS

Many previous works were done to predict the occurrence of diseases prior and many works are conducted using big data analytics tools like Hadoop MapReduce and apache spark. Similarly many methods are already implemented for sentiment analysis. But this work focuses on incorporating all these concepts under a single umbrella. This is a big data processing and prediction framework that integrates climate data, health data and twitter data in order to forecast the occurrence of infectious diseases based on climate variability and social media content. Artificial neural networks have the potential to predict various infectious diseases based on the electronic medical records of the patient. It can model some relationships between the inputs and outputs. This is done by training the model with large number of input data such as health records of several patients. When a new data which is not part of the training data, is loaded into the model, it will predict a new outcome. It helps to identify whether the new patient has the chances of occurrence of a disease. According to Arifianto [4], a polynomial neural network can be used for malaria incidence forecasting by using the incidence record and weather pattern as inputs. Long training time and bad input selection of neural networks often make it difficult to use. Diseases can be predicted using fuzzy expert system [5] which is a knowledge based system. Expert systems have been successfully applied in various domains for the prediction purposes. It is a rule based system which includes several IF-THEN rules. One can design an expert system for the early diagnosis of various diseases using fuzzy inference system. It uses medical records and symptoms of people as input variables and convert them to fuzzy IF-THEN rules. Most common symptoms like fever, headache, body pain etc are considered as input variables and generate rules like, IF headache then probable dengue. It classifies a person's record as no disease, probable disease and disease confirm. Accuracy of the prediction depends on the number of rules. One can expect more accuracy if there exist more rules. But writing each and every rule is very time consuming.

Apache hadoop [6] is a popular and widely used framework for dealing with huge volume of data. HDFS (Hadoop Distributed File System) is the storage module of Hadoop which has the potential to storage large amount of data. Efficient processing of the data can be done with the help of Hadoop MapReduce, which is the programming module. Hadoop is best when dealing with historical data which is stored in HDFS wherein which it is inefficient for streaming data processing. According to Yao Q [7], a medical big data processing system can be developed based on Hadoop. Extraction, transformation and loading of the data is done through sqoop where in which sqoop is a module in hadoop that loads data from external sources. This system is helpful to the medical professionals in providing proper individual or patient based recommendations. This is achieved using the mahout based recommendation engine.

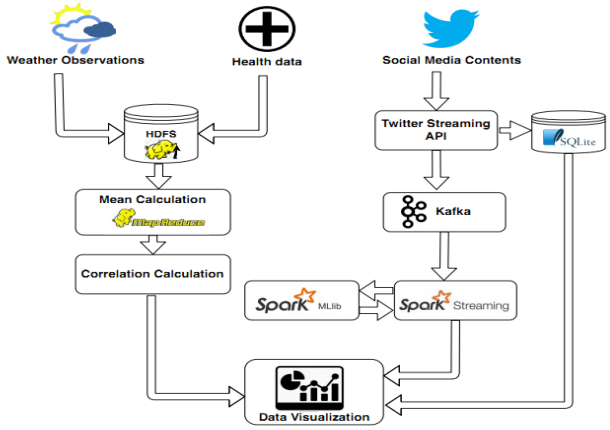

Fig. 1.Proposed model

Sentiment analysis [8] is the process of identifying, analyzing and categorizing the opinion of the people which is expressed in social media such as twitter. In 2015, a new method is proposed to analyze the sentiments retrieved from tweets. Since social medias like twitter has become an efficient platform for expressing public opinions, the sentiments of the people during a natural disaster or climate change is analyzed. Tweets containing the keyword " climate" are collected between September 2008 and July 2014 and are analyzed through a sentiment measurement tool called Hedonometer. This research [9] provides an overall analysis of opinion of the people during a natural disaster through the popular social media twitter and it shows that there exist less happiness among the people during such circumstances.

III. PROPOSED WORK

Proposed model is a data processing and prediction framework. It integrates climate data, health data and twitter data in order to forecast diseases arising as a result of climate change. Fig.1 shows the architecture of overall proposed model.

A. Data Block

This is the data layer of the proposed model. Generally three types of data are collected and utilized for the implementation of proposed model. It includes climate data, health data and twitter data. Climate data indicates the climatic information collected from various weather stations across Karnataka from January 2010 to December 2018. Daily measurements of various climate parameters such as temperature, relative humidity, atmospheric pressure, wind, precipitation and solar are included in the climate data. Health data includes the number of monthly occurrence of diseases in each latitude and longitude. Numerous diseases can be taken

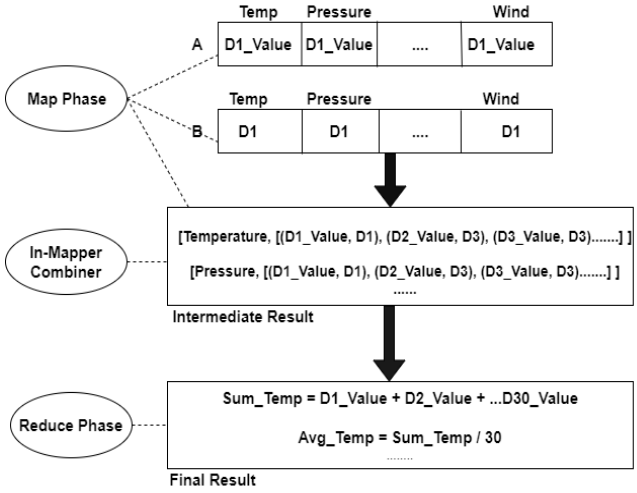

Fig. 2.MapReduce Algorithm

into consideration. Social media content represents the tweets collected from various localities of Karnataka.

B. Mean Calculation

Daily climate data of various localities of Karnataka is taken for the study. This data is pre-computed in such a way that, the monthly mean of various climate parameters is calculated from the daily measures. Rather than dealing with the data as a whole, the average of the data is taken as a monthly basis. The data is shortened when we adopted this strategy. This mean calculation is done by using In-Mapper combiner based MapReduce algorithm. MapReduce is a programming model, one of the components in Apache Hadoop framework. Hadoop is a data processing and storage framework which is used for the analysis of huge volume of data. It mainly include two components. MapReduce and Hadoop Distributed File System (HDFS). MapReduce is the programming model and HDFS is the storage file system.

- *In-mapper combiner MapReduce algorithm*:

MapReduce model generally includes two phases. Map phase and reduce phase. MapReduce always deals with data in form of key-value pairs and data is taken as input from HDFS. Map phase, which is the primary phase, read every line from the input and process the key-value pairs. The output of this phase is just an intermediate one and it will be the one that is forwarded as input to the reducer phase. Reducer joins all the values of a specific key and gives the ultimate output. In the proposed model, to calculate monthly mean of various climate parameters, MapReduce paradigm is adopted which make use of the concept of In-mapper combiner [10]. Combiner is an optional phase in between map and reduce phase. It is used to aggregates all key-value pairs having the same key. In case of an in-mapper combiner, the combiner phase exist within the mapper phase itself, which means that, rather than waiting for the mapper to finish all of its task, combiner combines the key-value pairs continuously as soon as it receives two pairs with the same key.

These are the basic steps in this MapReduce algorithm.

- Begin the Map phase
- Create two associative arrays for storing sum and number of days
- Add values to first associative array on daily basis
- Add respective days to the second associative array
- Return key-value pairs having the same key aggregated together as intermediate result
- Begin Reduce phase
- Calculate the sum and average of values from intermediate result
- Return the average

Fig. 2 shows the internal working of the algorithm. It takes daily wise climate data from HDFS as the input. Apache Hadoop can be started in our system from command line using the command $HADOOP_HOME/bin/start-all.sh. It will start the name node, data nodes, secondary name node etc. HADOOP_HOME variable need to be set in bashrc file. In the end we have to make ensure that all these nodes stops running using $HADOOP_HOME/bin/stop-all.sh. Data can be loaded to HDFS from our local file system with the help of CopyFromLocal command. $HADOOP_HOME/hadoop –CopyFromLocal path/to/localfile path/to/HDFS directory. In the mapper phase two associative arrays are used. They can be considered as data structures that are used to create hash tables where in which each values are stored as key-value pairs. Since they have the property of memorization, which help them to remember the answers of computation of previous stages, they are best fit in our scenario. In mapper phase, the primary associative array stores daily measurements in association with the keys (climate parameters). Each key holds key-value pairs having same key, and passing them to the reducer phase as an intermediary input. It is in the reducer phase the actual mean measurements takes place.

C. Correlation Calculation

After obtaining the monthly mean of weather observations, this data is aggregated with the number of monthly occurrence of diseases. The health data used in proposed model is the

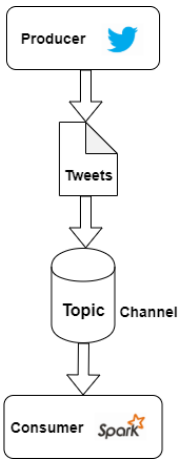

Fig. 3.Kafka Architecture

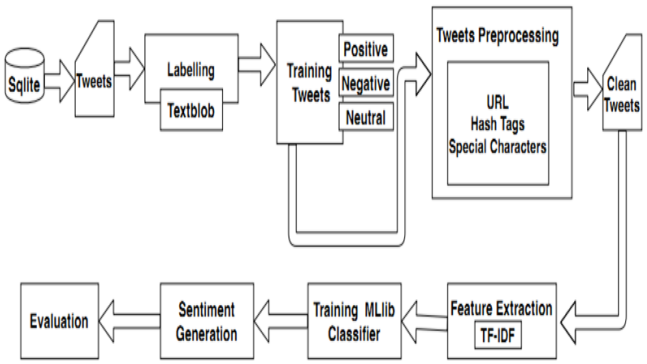

Fig. 4.Analysing Tweets

number of incidence of diseases in a specific latitude and longitude during a monthly span. In order to analyze relationship between each of the climate parameters and disease incidence count, Pearson correlation coefficient is calculated. This measurement helps to identify the relationships between two columns. For example, temperature and disease incidence count. Values falls between -1 to +1. Values closer to positive number indicates that, the two columns have the highest correlation.

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}} \quad (1)$$

PCC denotes the correlation value. x and y denote the mean of x and mean of y respectively. x shows the disease incidence and y indicates the climate parameter. For a specific latitude and longitude, we calculated the Pearson correlation coefficient [11] of each of the climate parameter and disease incidence. For the locality which having higher correlation value (for example, temperature), it can be estimated that, the chances of occurrence of that disease in this area is very high during the respective climate.

D. Tweet Collection Phase

Twitter is a popular and widely used social media platform. Since it provides more access than any other social medias, it can be easily used for analyzing peoples opinion about a specific topic. Twitter streaming APIs are required to fetch tweets from Twitter. To obtain such APIs, we need to create a new application in the twitter developer account. After that we need to create our access tokens. By clicking on 'Create my access token', four types of keys are generated.

- API Key (Consumer Key)
- API Secret Key (Consumer Secret)
- Access Token
- Access Token Secret

To access these twitter API, a python library called 'tweepy' is used. In order to collect tweets from a specific location like Karnataka, respective longitudes and latitudes are required. This is obtained using bounding box tool which gives two lon-gitudes and two latitudes of the specific area. In case of Bengaluru it gives, westlimit=77.460102; southlimit= 12.834012; eastlimit=77.784051; northlimit=13.143665. To collect tweets in English language only, in the filter area, languages can be set as English. Tweets are collected based of specific keyword such as 'flu'. It can be specified using 'track' keyword.

E. Twitter Data Storage

After accessing tweets using twitter streaming API via tweepy library, the next step is to store them in a proper manner. In proposed model, two different kinds of storage mechanisms are proposed. Apache kafka and Sqlite. Sqlite is a relational database which can be used for the effective storage of data. It is much different from the traditional databases like MySQL and SQL Server since it allows loose scheme and follows a file like database.

Apache Kafka is a module used for transferring messages from one application to other and it follows a publish subscribe messaging system. Kafka has mainly three functionalities. Publish, subscribe and store data. Producer is the one who will be publishing the messages and in proposed model twitter is acting as the producer. Consumer, in proposed scenario, spark, will be retrieving messages from the producer. The tweets from producer, will be loaded into a topic that we created. Topic acts as the channel between producer and consumer. Fig.3 shows the architecture of kafka. Following are the steps to be followed for working with kafka.

- Start Zookeeper which acts as the managing component with the command, $KAFKA HOME/bin/ zookeeper -server-start.sh config/zookeeper.properties
- Start kafka server using, $KAFKA HOME/bin/kafka-server-start.sh config/server.properties
- A topic is created named 'tweetz' using the command, $KAFKA_HOME/bin/kafka-topic.sh –create –zookeeper localhost:2158 –replication-factor 1 –partitions 1 –topic tweetz
- When running our program, the data will be loaded into this topic in JSON format. To view the tweets in this topic, the following command can be used . $KAFKA_ HOME/bin/kafka-console-consumer.sh – bootstrap -server localhost:9092 –topic tweetz –from-beginning

F. Analysis Phase

Although Hadoop can be used for storing and processing huge volume of data effectively, when coming to the speed of processing data, Spark always outperforms. This is because, spark utilizes the specific data structure RDD (Resilient Distributed Dataset) and performs in-memory computations. All operations in spark are carried out via RDD transformations and actions and they take place in memory rather than in disk. In proposed model spark is used for processing and analyzing the twitter data. The data that is loaded into the kafka is analyzed using spark streaming library of spark and the sentiment of the people is extracted from the tweets. After downloading spark in our system, we need to set path in .bashrc file such as export SPARK_HOME = /usr/local/spark and export PATH = $PATH:/usr/local/spark/bin. Fig.4 shows the steps in analysis phase. By analysing the tweets, the sentiment of the people can be easily identified regarding an epidemic outbreak [12]. The tweets are classified as positive, negative and neutral. This helps

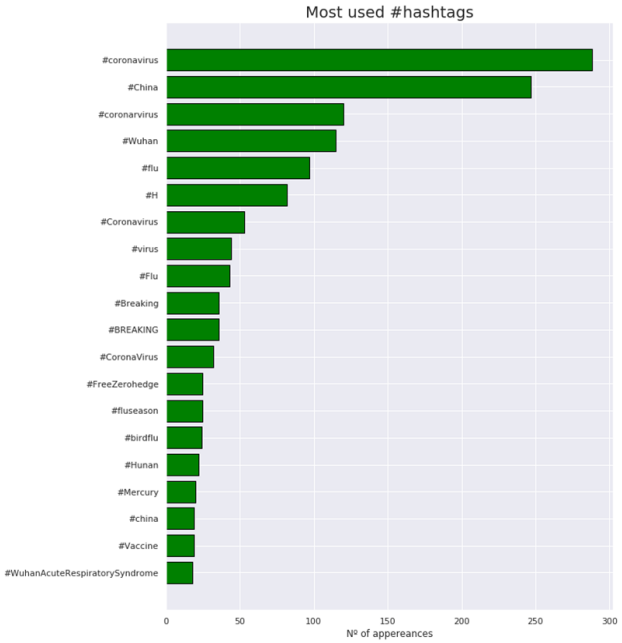

Fig. 5.Most Used Hashtags

to predict the areas wherein which chances of occurrences of diseases is very high. To achieve this Machine Learning methodology is taken into consideration.

- *Preprocessing:* For training the classifier tweets are labeled into three category using textblob library [13] and the results are stored in sqlite database. Before analyzing data, the tweets needs to be pre-processed in order to remove unwanted characters, urls etc. This is achieved using Stanford Core NLP Library. For that we need to import libraries such as nltk, re etc. TweetTokenizer method of nltk is used for tokenizing the tweets.
- *Feature Extraction:* Before passing the tweets directly to the classifier, they needs to be translated to some feature vectors. The mlllib.features package contain several feature generation methods such as Bag of Words, TF-IDF etc. In text classification [14] approaches, the number of times each word occur is considered as a feature Term Frequency, TF(t,d) is the number of times term t appears in document d and Document Frequency DF(t,D) is the

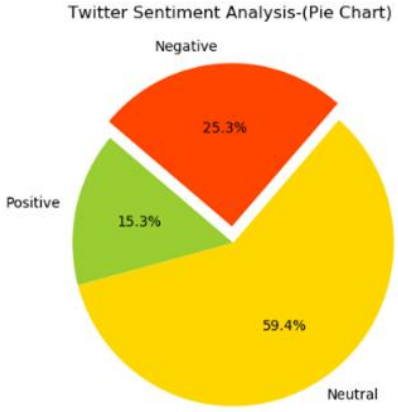

Fig. 6.Sentiment Analysis Pie Chart

Fig. 7.Live Twitter Sentiment

number of documents that contains term t. Since stop words like 'a','the' etc has no special meaning in the whole set of documents IDF is used. TF-IDF scheme identifies how important a word in a document of the entire data set. TF-IDF of a term t is the product of TF and IDF of the word. The entire text is transformed into a feature vector, which is then passed to a ML model for classification. Both HashingTF and CountVectorizer can be used to identify the TF. To import HashingTF and IDF use, pyspark.ml.feature import HashingTF, IDF.

$$IDF(t,D) = \frac{|D|+1}{DF(t,D)+1} \quad (2)$$

$$TF-IDF_{(t,d,D)} = TF_{(t,d)} * IDF\_\{t,D\} \quad (3)$$

- *MLlib Classifier:* Spark MLlib [15] machine learning library support various classifiers such as Naive Bayes, SVM, Logistic Regression etc. These three classifiers are adopted for predicting the tweets' sentiment. The class labels and feature vectors are required for classification purpose. To carry out the classification process, we need to import the classifier from pyspark.ml.classification package. The trained model is loaded to memory and it process the incoming tweets to identify the sentiments from the text. Tweets are then classified [16] as either positive, negative or neutral. These three models are compared and finally it is obtained that, for tweet classification purpose SVM gives the higher accuracy and TPR rate.

G. Visualization Phase

Visualization is mainly done using Matplotlib, basemap and dash frameworks. Raw tweets are visualized to identify the most used hash tags and most active users. It is shown in fig.5.This visualization is done with the help of Python libraries pandas, numpy, matplotlib, json, seaborn, wordcloud etc. Visualization of sentiments of such tweets can be done using matplotlib pie chart and bar graphs. The total number of

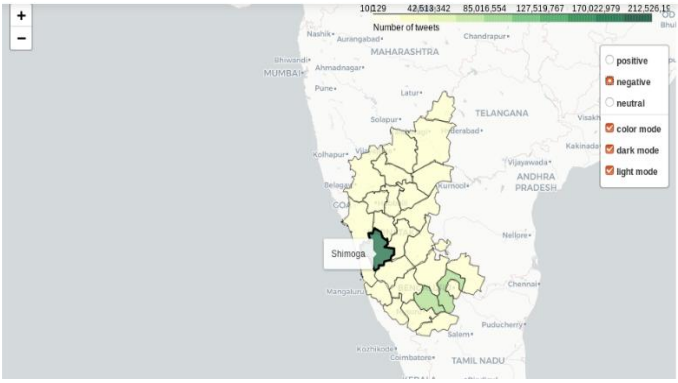

Fig. 8.Epidemic hit areas

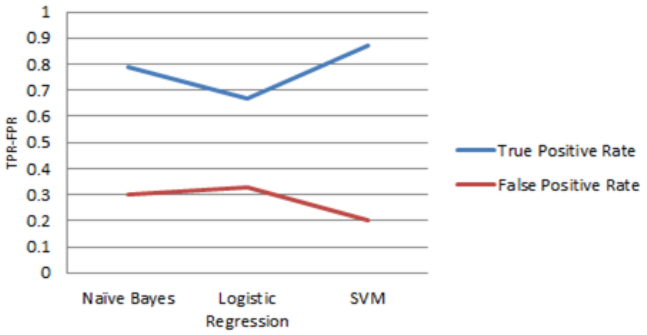

Fig. 10.TPR-FPR

tweets and the count of positive, negative, neutral tweets can be displayed. It is shown in fig.6. Live sentiment analysis can be calculated real time and it can be visualized using dash framework. It gives a polarity score to each of the tweet and it is visualized real time in a web browser [17]. The epidemic hit areas can be predicted based on the number of tweets collected. To visualize the results in a base map, the geographical measurements of the Karnataka state needs to be stored as a geojson file. Also the latitudes and longitudes of each of the districts is also used.

IV. RESULTS AND DISCUSSION

The proposed model can be used to analyze the sentiments of the people during a time and it helps to identify their overall concern regarding a specific matter. Tweets can be categorized into positive, negative and neutral and a score is given to them. This score can be visualized real time in a browser. This is shown in Fig.7 and it is observed that most of the sentiment coming from the real-time tweets are negative. So there can be chances of outbreak of a specific disease in the near future. Since tweets are location filtered, the sentiment results can be visualized in a cholropath map of Karnataka. Epidemic hit areas of Karnataka is visualized by considering the number of positive, negative and neutral tweets in each of the district. Fig.8 shows the map of Karnataka based on order

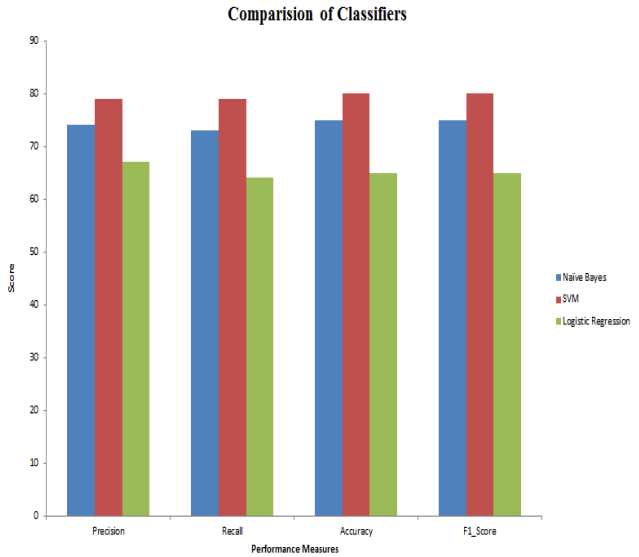

Fig. 9.Comparison of classifiers

of negative tweets achieved. This can be analyzed to identify the most probable epidemic hit areas in Karnataka. Since the most number of negative tweets are collected from Shimoga and Bangalore, it can be considered as the most probable epidemic hit area in the near future.

To determine the accuracy of the proposed model, we need to identify True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). TN are tweets that are negative and classified as negative. FP are tweets that are negative and classified as positive. FN are tweets that are positive and classified as negative and TP are the tweets that are positive and also classified as positive. A better model have precision, recall, accuracy and F1 score higher and FPR lower. Fig.9 shows the comparison of various MLlib classifiers. SVM works well for the sentiment classification.

Fig.10 shows the true positive and false positive rate of these classifiers where TPR is the percentage of tweets classified correctly. FPR is the percentage of negative tweets incorrectly classified as negative. Since it is an incorrect classification, its value must be lower for a better model. Fig.11 shows the ROC curve of these classifiers and it is obtained that Area Under the Curve (AUC) is higher for SVM classifier.

V. CONCLUSION

Proposed model is a big data processing framework to store and process the big climate data and social media content. The proposed framework is capable of monitoring the correlation between the climate parameters and disease incidence in a continuous manner. It is demonstrated in a HDFS environment with seven layers, namely, data block, mean calculation layer, correlation calculation layer, tweets collection block, tweets storage block, data analysis block and visualization blocks. Hadoop MapReduce and Spark are used for implementation of the analysis phase. Spark MLlib machine learning library is adopted for the implementation of the proposed model and this is used for feature extraction and model classification of tweets. In order to analyze whether the tweeting habits of users matched with the outbreak of diseases, all tweets containing keywords related to diseases are collected using twitter streaming API and using Apache Kafka they are analyzed and processed using Spark streaming. The performance of proposed model was improved when these social media data is included. The epidemic hit areas are identified using twitter sentiment analysis

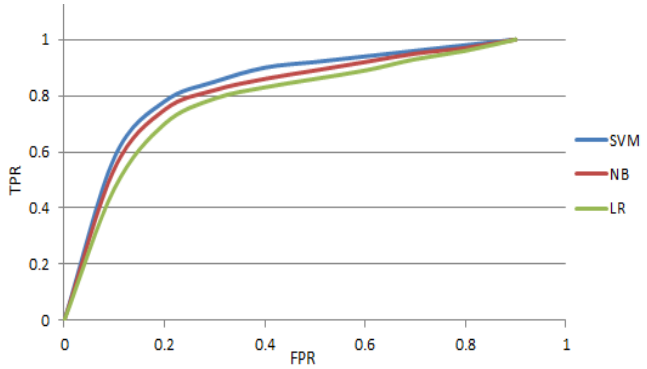

Fig. 11.AUC in ROC diagram

and the climate parameter that leads to the disease outbreak in those epidemic hit areas are identified with the help of correlation calculation.. This indicates that including real-time analysis of social media data can provide more effective result when working on the historical climate data.

REFERENCES

1. Lopez, D., Gunasekaran, M., Murugan, B. S., Kaur, H., and Abbas, K. M. (2014, Octo-ber). "Spatial BigData analytics of influenza epidemic in Vellore, India," in Proc. 2014 IEEE International Conference on Big Data (pp. 19–24). IEEE.
2. Lopez, D., Manogaran, G. (2016). "Big Data Architecture for Climate Change and Disease Dynamics" Eds. Geetam S. Tomar et al. The Human Element of Big Data: Issues, Analytics, and Performance, CRC Press.
3. Mart´ın, A., Juli´an, A. B. A., and Cos-Gay´on, F. (2019),"Analysis of Twitter messages using big data tools to evaluate and locate the activity in the city of Valencia (Spain)", Cities: 86, 37–50. Miller, H. J., and Goodchild, M. F. (2015). "Data-driven geography." Geo Journal, Elsevier 80(4), 449–461.
4. F. Ibrahim, M.N. Taib, W.A.B.W. Abas, C.C. Guan, S. Sulaiman, "A novel Dengue Fever (DF) and Dengue Haemorrhagic Fever (DHF) analysis using artificial neural network ", Computer Methods Programs in Biomedicine, Elsevier, 79, 273–281 (2015)
5. Szmidt, E., Kacprzyk, J., In: Abraham, A., Jain, L., Kacprzyk, J. (eds.), "An Intuitionistic Fuzzy Set Based Approach to Intelligent Data Analysis: An application to medical diagnosis", Recent Advances in Intelligent Paradigms and Applications, pp. 57–70. Springer, Heidelberg (2017)
6. Chinmayee Mohapatra, Siddharth Swarup Rautray , Manjusha Pandey,"Prevention of infectious disease based on big data analytics and MapReduce modeling", IEEE Journal of Biomedical and Health Informatics, 22-24, Feb. 2017
7. Yao, Q., Tian, Y., Li, P.-F., Tian, L.-L., Qian, Y.-M. and Li, J.-S. (2015), 'Design and development of a medical big data processing system based on hadoop', Journal of medical systems 39(3), 23.
8. O. Serban, N. Thapen, B. Maginnis, C. Hankin and V. Foot, "Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification", Information Processing and Management, Elsevier, June 2018.
9. Ozt urk, N., Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. Telematics and Informatics, 35(1), 136 - 147. DOI : https://doi.org/10.1016/j.tele.2017.10.006
10. Daphne Lopez,Gunasekaran Manogaran and Naveen Chilamkurti,"In-Mapper combiner based Map-Reduce algorithm for big data processing of IoT based climate data", Future Generation Computer Systems, April 2018.
11. Abderr ahmane Eddaoudy and Khalil Maalmi,"A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment", Journal of Big Data ,2019,6:104.
12. Abirami, M.A.M., Gayathri, M. V, 2016. "a Survey on Sentiment Analysis Methods and Approach". 2016 Eighth Int. Conf. Adv.Comput.72–76. https://doi.org/10.1109/ICoAC.2017. 7951748
13. O. Serban, N. Thapen, B. Maginnis, C. Hankin and V. Foot, "Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification", Information Processing and Management, Elsevier, June 2018
14. A. Neviarouskaya, H. Prendinger, M. Ishizuka, SentiFul., "A lexicon model for deep sentiment analysis and opinion mining applications", IEEE Transactions on Affective Computing, Vol. 2, No. 1, January-March 2011.
15. A. Kanavos, N. Nodarakis, S. Sioutas, A. Tsakalidis, D. Tsolis and G. Tzimas, "Large Scale Implementations for Twitter Sentiment Classification," Algorithms, vol. 10, no. 1, p. 33, 2017.
16. Sindhujaa N, Vanitha CN, Subaira AS (2016) An improved version of big data classification and clustering using graph search technique. Int J Comput Sci Mob Comput 5(2):224–229
17. Kucher, K., Paradis, C., and Kerren, A. (2018). The state of the art in sentiment visualization. Computer Graphics Forum, 37(1):71–96.

AUTHORS PROFILE

Juliet Johny received Bachelor of Technology in Computer Science and Engineering from Amal Jyothi Engineering College, Kanjirappally in 2017 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Big Data Analytics and Data Mining.

Linda Sara Mathew is currently working as assistant professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. She received her B-Tech Degree in 2002 and M-Tech in 2011 in Computer Science and Engineering from Mahathma Gandhi university and Anna University. She has around 15 years of teaching and industrial experience. She is interested in the areas of Data Mining, Neural Network, Image Processing, Soft Computing