Assorted Model of Sentiment using Mapreduce Framework

Saurabh Dhyani, G. S. Thakur

Abstract: ***Social networking sites platforms, such as Facebook and Twitter, are being broadly used by community to share their feelings on different matters. Consequently, social networking site becomes an admirable and major open source for collecting public opinion. To perform sentiment analysis on such huge data, computational assorted models of single node are ineffective. Two ways to grip data that are big ,either by using super computers or by using parallel processing or by distributed processing. Where it is costly to use super computer, most models of parallel processing such as MPI, are difficult to implement and scaling, MapReduce is one of the parallel processing models that is highly scalable, tolerant to fault, and easy for using, in this research paper, we have proposed assorted model of sentiment analysis for twitter using MapReduce Framework. mapreduce based naïve bayes training algorithm was proposed for this purpose. Only single mapreduce job is executed for this algorithm which makes it different from earlier previous work. Training model is deployed to to classify million of tweets of twitter computers are costly, most parallel programming models, such as MPI, are difficult to use and scale. MapReduce is one of the parallel programming models that is highly scalable, fault tolerant and easy to use. This paper proposes a scalable framework for sentiment analysis of Twitter using MapReduce model. For this purpose a MapReduce based Na¨ive Bayes training algorithm is proposed, this algorithm uses only one MapReduce job which makes it different from previous works. The trained model is deployed to classify millions of tweets. Accuracy and Scalabilty of our propsed model is well compared to previous models.***

Keywords: Sentiment Analysis, Classification, MapReduce, Social Media, Big Data

I. INTRODUCTION

Now days, social media has become important part of daily routine in the life of people. Internet and social media sites can be used for various kinds of purposes including financial trends and advertisement, casting political opinion, extracting comments of user about products, spreading of news and spreading of spams[1]. Social networking sites generate virtual connection among users, in which users can express through feelings and develop connection through posts, likes and messages, comments. Social media permits users to expose their feelings , thoughts and opinion with other users easily and instantly.

Revised Manuscript Received on March 15, 2020.

Saurabh Dhyani, Research Scholar, Department of Computer Application MANIT Bhopal (MP), India.
(Email id:saurabhdhyani29@gmail.com)

G. S. Thakur, Assistant Professor, Department of Computer Application MANIT Bhopal (MP), India. (Email id:ghanshyamthakur@gmail.com)

While social networking sites are commonly employed for the purpose as information, social events and advertising, communication[2], they can also used for sharing poltical opinion on these kinds of platforms. Sentiment analysis of opinion of public is the most sought after information whether it is business related to marketing, political campaign and stock exchanges. Sentiment analysis is also known as opinion mining. It is a branch of data mining. It concentrates for fetching the polarity of sentiment from textual content. Textual content is usually in natural language. Sentiment analysis on the social media deals with categorization of the data that are available on social media sites such as facebook, twitter and IMDB. Sentiment analysis is the problem of classification and mostly solved with the help of supervised technique and unsupervised technique. Data collection, data preprocessing, feature extraction and creating feature vector, classification are Major components of opinion mining. Mostly lexicon based feature is used for unsupervised classification. Supervised classification include learning technique such as support vector machine(svm), neural networks, naïve bayes. For categorizing opinion of public automatically, lot of work has been done on preprocessing of data and classification technique such as supervised /unsupervised machine learning technique, where supervised technique beats to the unsupervised technique in terms of finding accuracy[3]. Now days use of social networking sites is increasing rapidly such as Facebook and Twitter. Social media sites is generating huge volume of data. This huge volume of data comes under the category of Big Data. Five hundered million tweets are generating per day according to the statics and 80% of these tweets are generated from a mobile device[4]. Furthermore number of internet users reduced 22% of the number of active users of the twitter in the world[5]. Twitter provides API for downloading tweets that can be used for finding sentiment analysis after giving training to a classifier. For sentiment analysis, processing large volume of data is very time taking process with the help of single node. One way for processing large volume of data is to take more powerful machine like supercomputers, these computers are not only costly but they are not also suitable for scalability. Parallel processing is the other way to handle such large volume of data. however, parallelization provides growth to the problem of following problems

- Distribution of resources
- Sharing of resources
- Synchronization
- Handling of failure

Mapreduce is the model of programming which hides above covered mentioned problems from users and provides large scalable cluster for processing large volume of data[6]. MapReduce technology has been designed by Google as a Proprietary techonology. Several open source implementation are available for MapReduce but among these open sources Apache Hadoop is most popular implementation. Mapreduce technology is inspired from functional programming. Map function and Reduce function are programmed by user. Recently, mapreduced technique combined with supervised technique and mapreduce technique combined with unsupervised technique have been proposed for finding sentiment from large volume of data. Classifying contents of tweets is generally different than classifying reviews of movie, as reviews are more formal and structured than tweets. to classify movie review of amazon with the help of naïve bayes classifier achieved 82% accuracy[7]. In[7], authors have designed algorithm for making the number of negative and positive training instances of dataset equally and providing priority probability of each classes is equal to 50 %. This cannot be considered for huge volume of tweets and needs to be calculated in the explicit way. In[8], MapReduce based naïve bayes model is used for finding sentiment from twitter dataset. But testing methods and accuracy were not reported. In [8] Two mapreduce job also is used to for training the classifier which would require more time.

In this research paper we have proposed cost effective and scalable framework for assorted sentiment analysis model using Mapreduce. We have used mapreduce algorithm with naïve bayes classifier for finding sentiment analysis from twitter dataset. To train naïve bayes model , we have used only one mapreduce job, That makes it different from previous work which has been done previously by many researchers. To be benefitted from the processing power of mapreduce data should be come in the category of big data. Some earlier work has been done related to sentiment analysis on big data but lacked in reporting higher accuracy for tweets from twitter .

II. LITERATURE REVIEW

in the subject Syrian civil war and following crisis of refugee, a series of analyses of sentiment using data belong to twitter were performed[9]. they collected tweets in two languages : English and Turkish. English tweets were carrying less positive sentiment about refugees and Syrian when compared to the Turkish tweets among 35 % of all tweets. For election monitoring and prediction, Sentiment analysis of twitter is inexpensive and quick tool[10]. They have proposed hybrid topic based sentiment analysis for prediction of electin using tweets. Their approach gave promising results by enhanching prediction of vote share compared to methods which are existed. An ensemble classifier proposed by combining classifiers related to base learning with the intention of improving accuracy and performance of classification technique of sentiment analysis[11].sentiment analysis technique beats learning based technique when training datset is not adequate[12]. Social networking sites permits users to generate and share diverse information anytime and anywhere by allowing users freely generate and shared contents. Since huge volume of data are generated and shared in real time, an effective method of processing data is necessary. For quick and effective processing of data of social networking sites uses mongodb[13] that comes under nosql technology which is extended technology of relational database. Since data of social networking sites are not in standardized form unlike existing traditional data, for analysis , pieces of information fetched after short writing[14]. There are lexicon related methods and machine learning methods for extracting sentiment from textual content. Two techniques is used for detailed sentiment analysis[15].lexicon technique is based on unsupervised learning. In lexicon technique textual data are classified into a set of predetermined classes of sentiment. Sentiment lexicon is a dictionary consisting of sentiment score associated to word Based on the sentiment lexicon, sentiment score of textual data are calculated[16]. Sentiment analysis extract people's opinion, feelings , behavior and thoughts from content of textual data using NLP(Natural Language Processing) Techniques[17]. Moreover, sentiment analysis is as same as opinion mining ,with concentration of the problem of classification of textual data. Extracting opinion or sentiment from textual data of social media can be very expensive and challenging due to large volume of data[18]. With the popularity of social media from year 2000, people started sharing their opinion and feelings through social media and it can be influencing significantly. Unlike traditional methods of data mining, sentiment analysis and text mining are used for dealing unstructured data[19].

Recently, twitter has become so popular that scientist has come forward for conducting research studies from various prospective. Behavior pattern of journalists was found after analyzing twitter data[20]. During tsunami incident, warning was given to the region of Indonesia and after that reactions among users of twitter have been analyzed[4]. Behavior pattern of cancer of cancer patients has been examined after using twitter[21]. Based on 36 million tweets was extracted from twitter, Real time Sentiment analysis model was proposed for classifying election tweets during election of president of US in year of 2012[22]. In 2012 presidential election of south korea, issue discussed in TV debate and communication pattern of conversion of twitter users on the topic of election were investigated [23]. a recent study analyzed the political campaign in india during general election and focused to find out the impact of first time voters on the internet[24]. Public data of twitter was gathered for the intention of extracting sentiment from it as well as tracking scenario of terrorism on the basis of graphical visualizations [25].

In[7], authors have proposed mapreduce based navie bayes classifier for extracting opinion from dataset. They tested opinion mining model using AMAZON movie review dataset and movie review dataset of Cornell university and obtained 82 % accuracy. [26] used unsupervised technique with dictionary based technique for finding sentiment from Korean language. They gathered huge volume of unstructured data in the form of tweets from twitter. Four map functions were used to analyze unstructured data after preprocessing it.

These map function performed analysis of preprocessing and detecting polarity, syntactically analysis of word, morpheme analysis and analysis of prohibited word in sequential manner. Positive score and negative score of contents of tweet from twitter was collected using reducing function. If positive score of content of tweet from twitter was greater than negative score of content of tweet, then that tweet was treated as positive and vice versa. A tweet was taken into next processing stage if positive score of tweet is equal to the negative score. Lexicon based feature was used for analysis and MongoDB was used for storing labeled tweets. 70% , 15% and 14 % were accuracy for positive, negative and neutral dataset respectively. In[27], DOM mobile application was proposed for sentiment analysis of Thai public. DOM is dependent on Hadoop. it utilizes MongoDB for storage and Mapreduce framework for analysis of unstructured data. Unstructured dataset was gathered from several social related websites in the server side of application . after gathering data it is stored in the MongoDB. For performing categorization of topic, Mapreduce job is used . Mapreduce job is also used for lexicon based sentiment analysis and categorizing data into positive,negative or neutral. A user can query particular topic from client side . results of experiment were calculated by gathering around 12 gb of dataset . the experimental result of DOM were compared to the annotated results of human. accuracy of long and short text was reported to be 76.32% after using DOM application. In[28], mapreduce Framework and HBase were used for sentiment analysis from dataset of twitter as well as these technologies used for building a lexicon database from dataset of twitter. HBase work with the hadoop framework. It is a distributed database system which provide storage as a backend in the hadoop framework. Lexicon database was used with the regression algorithm in the intention for classifying the tweets from twitter and it reported that accuracy was found around 72% -74%. Table 1 described summary of related work for finding sentiment after combined technique of machine learning with mapreduce framework. Table 2 described the summary of related work for sentiment analysis using machine learning technique without using mapreduce framework. Previously machine learning technique was rarely used with mapreduce framework. Mapreduce framework come into picture for big data technology. If data is so big then it cannot be handled in the efficient way with the help of single machine we need to distribute such big data into small chunks to the different nodes and write a program to process chunks of big data in the different nodes for the propose of finding out some new information from the big data. Hadoop framework provides the facility of distribute and process big data. Hadoop provides mapreduce framework for processing big data and HDFS(Hadoop Distributed File System) for storing data to the small chunks into different nodes.

in this research paper we have used Hadoop framework for implementing assorted system analysis model. We have implemented through Mapreduce Framework. Assorted model of sentiment would find output after covering different phases, these phases are namely collecting phase, filtering phase, preprocessing phase , and classification phase. For classification phase we have used naïve bayes classifier. We have implemented naïve bayes classifier into distributed environment. For checking accuracy of our proposed model we have taken two dataset namely predefined dataset and newly generated dataset based on hashtag. Our proposed model perform well for both dataset compare to previous model[7].

Table 1: Summarization of related work of sentiment analysis Using MapReduce Framework

Paper	Dataset	Machine learning Technique	Classifier	Classes	Accuracy	Comments
Liu et al.[7]	Reviews of movie	Supervised learning	Naïve Bayes	Binary	82%	Equal probability assumed
I.Ha. et al [26]	Topsy, Twitter	Unsupervised Learning	Lexicons	Binary	33%	Accuracy may be improved
S. Prom-on et al[27]	Social websites	Unsupervised Learning	Lexicons	Multiclass	76.32%	Accuracy can be enhanced
V. N. Khuc et al[28]	Twitter	Hybrid learning	Regression and lexicon	multiclass	73.70%	Accuracy can be further enhanced

Table 2: Summarization of related work of sentiment analysis using machine learning technique

Title of Research article	Year	Technique/Concept	Dataset	Results
Analyzing the emotions of crowd for improving Emergency Response Services[29]	2019	Change point detection technique and emotion analysis	LasVegas Shooting tweets from twitter	
Social Media's impact on the consumer mindset: when to use which sentiment extraction tool?[30]	2019	SVM(Support Vector Machine)	Product and service provider page of industries. Page is related to Facebook and YouGovt	
An efficient MapReduce Based assorted and Hybrid NBC-TFIDF algorithm to mine the public sentiment on diabetics mellitus[31]	2018	Naïve Bayes Classifier and TFIDF(Term Frequency and Inverse document frequency)	Diabetic Hashtag Related data from twitter	High accuracy for Bigram feature among 1-3 grams
Sentiment analysis model of tweets with non-language Features[32]	2017	Scoring methodology for finding sentiment polarity of tweets	6 datasets related to query IPL cricket, commonwealth game, election, Amazon, flipkart, Snapdeal	Highest accuracy (84%) for dataset related to query of Snapdeal

III. METHODOLOGY

this section provide detail of all modules of sentiment analysis model which we have proposed as shown in the fig 1. first we have collected training dataset from Twitter and filtered it to discard redundancies . after preprocessing of dataset classifier was trained. The trained model is saved in distributed file system and implement for testing data using classification technique. All the module of proposed model, we are explaining in the section 3.1, 3.2 and 3.3 . in section 3.1, we have explained how we have collected trained dataset using API . We have explained in details about dataset in the result and discussion section, In Section 3.2, after collecting dataset we have performed preprocessing on trained dataset. In section 3.3, we have explained naïve bayes classifier on distributed environment using MapReduce Framework. We have applied naive bayes classifier on training and testing dataset. 2 algorithms were generated for this propose one for training dataset and another for testing dataset

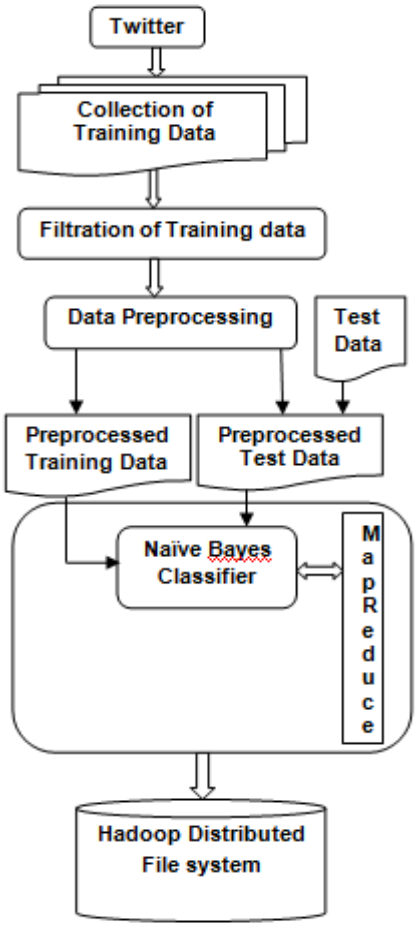

Fig1: Proposed Methodology

3.1 Training data collection and filtering

We have used API of Social media to gather training dataset by using query to the API of twitter. we have gathered tweets for finding emotional related hashtags that were synonyms for "Profit" and "Loss". extracted tweets are directly dumped in NoSQL database . We have taken HBase tool for NoSQL database . We have filtered tweets for removing redundancies before storing to Hadoop Distributed File System. Main aim of gathering this training dataset is two folded. First to analyze that what speed up has been achieved by proposed model after using large dataset second to analyze the impact of volume of training dataset on accuracy of classification algorithm.

3.2 Data Preprocessing

We have used regular preprocessing method of text including converting text in the lower case, discarding stop words and words that retaining characters which are non alphabetical. Hashtags and URLs were also discarded contents of tweets were changed to feature vector using unigram approach and term count as weight of feature.

3.3 Naïve bayes classifier using mapreduce:

in this section we have used multinomial based naïve bayes classifier using MapReduce programming model. We have performed binary classification for classifying the data in either positive category or negative category. We have used multinomial naïve bayes classifier as it provides good accuracy for classification of textual kind of data. There are two type of classifier using navie bayes, Bernoulli and Multinomial. For our proposed model we are providing MapReduce based multinomial naïve bayes classifier as it provides good accuracy for classification of textual data. Naïve bayes classifier is based on probabilistic model.

$$\text{argmax}_{d_k \in C} P(d_k) \times P(D/d_k) \quad (1)$$

$$P(d_k) = \frac{N_d k}{\sum_{k_1 \in C} N_{dk_1}} \quad (2)$$

$$P(D/d_k) = \prod_{i=1}^{n} P(t_i/d_k) \quad (3)$$

$$P\left(\frac{t}{d_k}\right) = \frac{Td_k t}{NTd_k + V} \quad (4)$$

In equation (1), d_k is the probability of class k contains document D. $P(d_k)$ is also called the priori kind of probability. If training dataset contains equal instances of all classes then $P(d_k)$=1/total number of classes and can be discarded from equation (1). In equation (2) and (3), Collection of classes is C and $N_d k$ indicates class k contains number of training instances. In equation(3), $P(t_i/d_k)$ is the probability that documents of class d_k contain i^{th} term in document D. where n is the total number of terms in D. from equation (4), $P(t/d_k)$ is calculated Where V is the vocabulary size. To train classifier from above equations we need vocabulary size V , $Td_k t$ for all terms t in V, For all classes requirement of Nd_k.
To train the classifier following parameter need to be calculated

- P, N: - Number of Training instances of positive and negative respectively.
- T_P,T_N:- Number of terms in the positive and negative instances of training dataset respectively.
- V: - Size of Vocabulary. After preprocessing total number of distinct terms remain in training dataset
- $T_P N_t$,$T_N N_t$: - Positive and negative instances of training dataset contains number of times term t occurs respectively

All the above parameters have to be calculated in distributed environment. Fig 2 shows small example of flow of the phase of training with the help of mapreduce framework. MapReduce Framework [2] is inspired by programming which is functional. Map and Reduce are functions that are programmed by user. Key and value pair are generated from dataset. First input dataset is distributed across Mappers and mappers execute this dataset in parallel and after that mappers produce intermediate key value pairs. Reducers are assigned to these key value pairs. Reducers execute these key value pairs in parallel and also emit output in the form of key value pairs. There is no global memory shared for mappers and reducer in the MapReduce Framework. Counters are only global objects that are provided by MapReduce Framework. Mappers and reducers are used to read these counters. Counter is a type of long and among many operations counter support only increment operation mappers and reducers can also used to update counter using increment operation.map and reduce function can be used by programmers for using the global counter. Counters can also be used by Mapreduce framework for keeping track of status of job such as number of bytes generated by mappers, number of bytes written to file by reducers etc. $T_P N_t$ and $T_N N_t$ are collected using key value pairs. Using counters of MapReduce, T_P and T_N are counted. The training model which contains all the parameters or statics, is stored in the distributed file system.

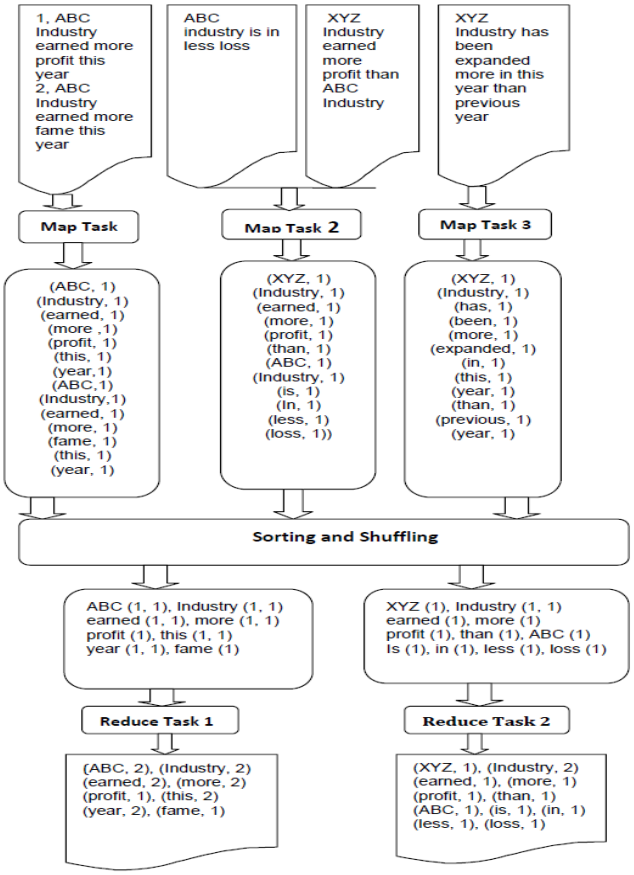

Fig 2: An exmple of mapreduce based naïve bayes classifier on training dataset

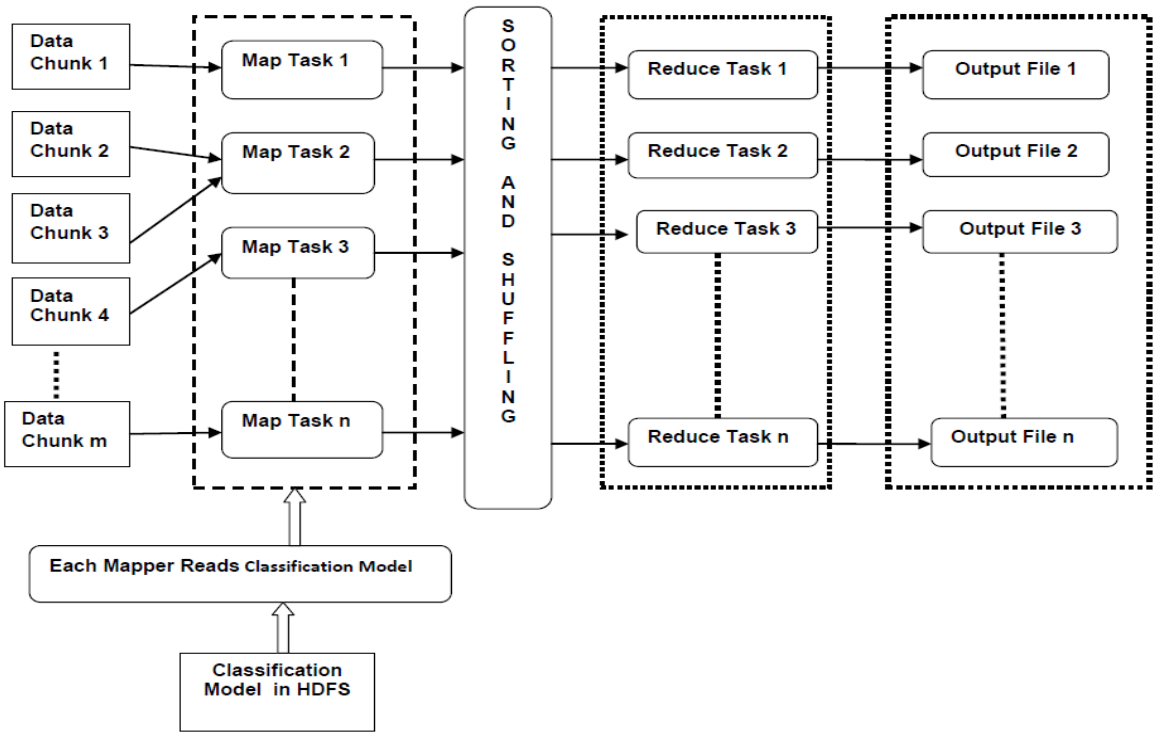

Fig 3: Mapreduce based naïve bayes classifier on testing dataset

Fig 3 shows the mapreduce based naïve bayes classifier on testing dataset. Equation 2.0 is used for calculating classification of test dataset. Trained classifier should be present in all the nodes in the hadoop framework. Each mapper stores the trained classifier model in memory. driver program provides path to the model as an input. For each test instances of test dataset to be classified, Mappers predict class corresponding to the test instance. reducer used as an identity function and dump the resultant output to the hadoop distributed file system(HDFS). The counters like True Positive, False Positive, True Negative, False Negative are also managed to find accuracy algorithm 2 wil also be used for classification of unannotated data and counters which calculate accuracy can be ignored

we have used dataset from twitter therefore the word tweet is used for training instances in the algorithm but it can be any kind of text. Binary classification would be implemented using algorithm 1 using more counters, given algorithm can be modified for implementing multiclass classification.

Algorithm 2 shows the pseudo code of naïve bayes classifier in the distributed environment using MapReduce framework. Intention of running naïve bayes classifier in the distributed environment is to find out classification of big data. Mappers will map naïve bayes classifier to different nodes of hadoop framework. And reducer would find the result of the classification of the data in the some node. Resultant node would be subset of whole nodes available in the framework. Result set contains less data compare to the previous data set. algorithm 1 and algorithm 2 are checked for both training and test dataset respectively first we have trained model and then applied test data for testing propose of our classifier. we have one intention to design two algorithms so that accuracy can be improved of classifier. if accuracy is improved of classifier then timing for finding sentiment polarity from dataset would be improved .

Algorithm 1

```
Class Map
counters {P, N} //Map Counters
function Mapper(Key k, Value v)
    // Key k : Tweet ID
    //Value v : instance of training (class, Tweet)
    Class = v.getclass()
    Tweet = v.gettweets()
    if Class = 'Positive' then
        Counter.P = Counter.P + 1
    end if
    if Class = 'Negative' then
        Counter.N = Counter.N + 1
    end if
    Tokens[] = preprocess(Tweet);
    for each token t in Tokens[] do
        Emit(t, Class)
    end for
end function
Class Reduce
C{V, T_p, T_N) //C : Counters of Reduce
function Reducer(Key k, List_of_Values < V >)
    //key k : tokens
    //List_of_Values : List of classes
    C.V = C.V + 1
    int NegCount = 0
    int PosCount = 0
    for each value v in < V > do
        if Class = 'Negative' then
            NegCount = NegCount + 1
        end if
        if Class = 'Positive' then
            PosCount = PosCount + 1
        end if
    end for
    Emit(k, (p : PosCount, n : NegCount))
    C.T_p = C.T_p + PosCount :
    C.T_N = C.T_N + NegCount :
end function
```

Aaa

Algorithm1 is used for building trained model for classifier in distributed environment and algorithm 2 is used for building testing model in the distributed environment. Algorithm 2 would perform well if algorithm 1 is performing well. it means algorithm 2 is dependent on the algorithm 1. Algor 1 is designed so that it would train to the system in the efficient way and also algo2 is designed so that testing performed well

Algorithm 2

```
Class Map
Naive Bayes Model NB
//CTP : Counters True Positive, TP : True Positive
//TN : True Negative, FP : False Positive, FN : False Negative
function MAPSETUP()
    NB = Load_Naive_Bayes_Model(Path_of_Model)
end function
function Mapper(Key k, Value v)
//Key k : Tweet ID, Value v : (Class, Tweet)
    Tweet = v.gettweet()
```

```
Trueclass = v.getclass()
Pr_pos = Pr_pos * (NB * P)/(NB * P + NB * N)
Pr_neg = Pr_neg * (NB * N)/(NB * P + NB * N)
Tokens[]= Preprocess(Tweet)
for each token t in Tokens[]do
  Pr_pos = Pr_pos * ((NB * T_P * t) + 1)/(NB * T_P + NB * V)
  Pr_neg = Pr_neg * ((NB * T_N * t) + 1)/(NB * T_N + NB * V)
end for
Assignedclass = (Pr_pos ≥ Pr_neg ? pos : neg)
switch(Trueclass + Assignedclass)do
   case pos_pos: TP+1
   case neg_neg : TN+1
   case pos_neg: FN+1
   case neg_pos: FP+1
end switch
emit(k,{Assignedclass, Trueclass})
end function
Class Reduce
function Reducer(Key k, Values < V >)
  for each value v in < V > do
    emit(k,v)
  end for
end function
```

IV. RESULT AND DISCUSSION

In this section we have discussed description of dataset, sentiment extraction from dataset using proposed model, accuracy and scalability of our proposed model. and we have also discussed what would be computation time of proposed model without using MapReduce framework and after using mapreduce framework.

4.1 Description of Dataset:-

mostly training dataset which are available for sentiment analysis is not of big size. We have collected training dataset using twitter API. We have given query to twitter for extracting tweets which Hashtags of synonyms of "Profit" and "Loss". Regarding the technique of this is described in [33] and [34]. Data was collected for the month of October 2019. Synonyms of "Profit" and "Loss" were chosen from thesaurus of Roget . tweets related to synonyms of profit were considered as positive tweets and tweets related to synonyms of loss were considered as negative tweets. Table shows specific Hashtags and emotion associated with it. In table we have collected synonyms of profit and loss from thesaurus dictionary. All synonyms related to profit and loss express as query terms for twitter.

With all the query term we have collected tweets into two buckets one bucket is associated to positive sentiment and another bucket is associated to negative sentiment. Before using these dataset for training all these hashtags from tweets are removed to avoid biasing. Our dataset contained tweets only in English language. We have analyzed one case for negation for tweets for example not #profit, not such cases were found. User uses #notprofit to express negative sentiment instead of not #profit. Query term associated to emotion is shown in Table 3.

Table 3: query term associated to emotion

sentiment	Query Terms
Positive	#benefit, #earning, #interest, #receipt, #revenue, #return ,#saving, #turnout, #surplus, #yield, #value, #advantage, #production, #income, #output, #advancement, #return
Negative	#catastrophe, #casualty, #damage, #defeat, #destruction, #disaster, #ruin #failure, #trouble, #calamity, #debt #cataclysm, #deficiency, #deprivation #harm

The size of data was collected 2 GB, containing 30781532 positive tweets and 10245670 negative tweets. Both classes do not contain same number of tweets, therefore priori probability cannot be assumed to be equal. We have also taken predefined dataset of Sentiment140 which contains 800,000 positive and 800,000 negative tweets [35]. We have implemented our proposed model using hadoop framework and python programming. 2node, 4 node and 6 node cluster were created for implementations propose.

4.2 Sentiment extraction from dataset using proposed model

we have extracted sentiment from proposed model using naïve bayes classifier in the distributed environment. Fig 4 shows the percentage of positive, negative and neutral sentiment on the sentiment140 training dataset after applying navie bayes classifier in distributed environment. Fig 5 shows the percentage of sentiment polarity on the dataset which are generated from hashtag of twitter. we have generated dataset on the synonym of profit and loss using hashtag after using the concept of twitter API. after finding sentiment on the both training dataset we have found accuracy of the proposed model based on the testing instances of same dataset.

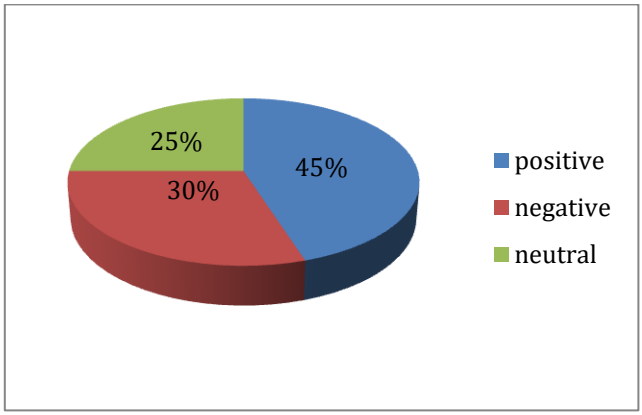

Fig 4: percentage of sentiment polarity on the predefined dataset after applying proposed model

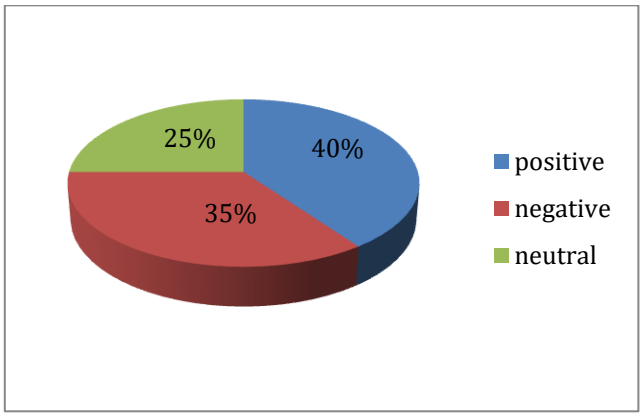

Fig 5: Percentage of sentiment polarity on the hashtag based generated dataset from twitter

4.3 Accuracy of Proposed model:

Accuracy is given to training dataset on accumulate test dataset. After implementing proposed model to the hashtag dataset and Sentiment140 dataset, both dataset provide the best accuracy compared to the previous model [7] as shown in Table 4. Naïve bayes is dependent on statistical model. Accuracy of proposed model does not differ from its sequential version. For each class, accuracy is given on the basis of its prior probabilities as 50% . hashtag related dataset which we have generated performed well for our proposed model. It gives highest accuracy 82.8%. Predefined dataset which we have fetched directly from website, provide 75.5% accuracy. Accuracy of our proposed model has improved by 8.6 % for hashtag related data and 2.6 % improved for sentiment140 dataset. in further section 4.4 we have discussed about scalability and we have checked scalability to our proposed model.

Table 4 : Accuracy of proposed model

Dataset	Accuracy of previous model	Accuracy of proposed model
Sentiment140	72.9%	75.5%
Hashtag Dataset	74.2%	82.8%

4.4 Scalability:

number of nodes in the hadoop cluster is inversely proportional to the computation time of a given size of problem. Increasing number of nodes reduce the computation time of given size of problem. We have checked computation time of proposed model for different instances of 2GB dataset on different cluster of nodes of hadoop as shown in Table 4. We have taken 2 node cluster, 4 node cluster and 6 node cluster for the propose of finding sentiment from instances of dataset through proposed model. For 2 node cluster we have taken size of instances of dataset namely, 0.50 GB, 1.00 GB, 1.50 GB and 2.00GB. Similarly for 4node and 6 node cluster we have taken same size of instances of dataset. 6 nodes cluster perform well for all size of instances because computation time is reduced compared to the two node cluster and 4 node cluster. It means our proposed model also provide good scalability Table 5 shows the relationship of time and size for proposed sentiment analysis model. We have checked scalability for 2, 4 and 6 node cluster in hadoop cluster and found that it reduce computational time as we add more nodes to cluster.

Table 5: Training time of proposed model for different size of training instances for 2, 4 and 6 nodes cluster

Size of cluster	Size of instances	computation time(sec)
2 nodes	0.50 GB	400
2 nodes	1.00 GB	600
2 nodes	1.50 GB	1100
2 nodes	2.00 GB	2000
4 nodes	0.50 GB	200
4 nodes	1.00 GB	320
4 nodes	1.50 GB	590
4 nodes	2.00 GB	1000
6 nodes	0.50 GB	90
6 nodes	1.00 GB	100
6 nodes	1.50 GB	190
6 nodes	2.00 GB	390

Scalabilty can be associated to classification time. If we add more node to system time of classification would be reduced.

Table 6: Classification time and of Dataset by using naïve bayes algo on distributed environment

Dataset	Size of Dataset	Size of Cluster	Classification time(sec)
Sentiment140	2GB	1 node	5800
Sentiment140	2GB	2 node	3120
Sentiment140	2GB	4node	1640
Sentiment140	2GB	6 node	890
Hashtag Data	2GB	1 node	5700
Hashtag Data	2GB	2 node	3000
Hashtag Data	2GB	4 node	1600
Hashtag Data	2GB	6 node	800

Table 6 shows the classification time of classifier in the distributed environment for same size dataset namely sentiment140 and hashtag data which we have generated through API of twitter. From table 6, it is indicated that sentiment140 dataset perform well for 6 node cluster because classification time is reduced most for 2GB dataset. From table 6 it is also indicated that hashtag based data perform well for 6 node cluster because it reduce classification time most for 2 GB hashtag related dataset.

Table 7 shows comparison of proposed model with the traditional model or one machine system with respect to time with different size of data. And we have found also speedup in the table after dividing training time of proposed model by training time of traditional model.

Table 7: comparison of Training time

Size of Data	Proposed model training time(Sec)	One machine training time(Sec)	Speedup
0.5	440	2150	4.88
1.0	860	3080	3.58
1.5	1400	4400	3.14
2.0	1900	5880	3.09

V. CONCLUSION:

This Research Paper Proposed scalable and effiecent naïve bayes classifier based on mapreduce framework for sentiment analysis. Navie bayes classifier classify tweets according to the polarity of sentiment on the distributed environment. The propsed model provides high scalabity and also provides more accuracy on the new large size training dataset based on hashtags which we have collected from twitter. On the large dataset, our model provides 8% more accuracy than small dataset.

REFERENCES

1. Alarifi, A., Alsaleh, M., Al-Salman, A., 2016. Twitter turing test: identifying social machines. Inf. Sci. 372, 332–346
2. Chianese, A., Piccialli, F., 2016. International workshop on Data Mining of Iot Systems (DaMIS): a service oriented framework for analysing social network activities.Procedia Comput. Sci. 98 (2016), 509–514.
3. L. Augustyniak, T. Kajdanowicz, P. , M. Kulisiewiczand W. Tuliglowicz, An approach to sentiment analysisof movie reviews: Lexicon based vs. classification, inHybrid Artificial Intelligence Systems, Springer, vol. 8480, Jun. 2014, pp. 168-17
4. Carley, K.M., Malik, M., Landwehr, P.M., Pfeffer, J., Kowalchuck, M., 2016. Crowd sourcing disaster management: the complex nature of Twitter usage in Padang Indonesia. Saf. Sci. 90, 48–61.
5. Kayser, V., Bierwisch, A., 2016. Using Twitter for foresight: an opportunity? Futures 84 (A), 50–63.
6. J. Dean and S.Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, in Communications of the ACM 51.1, Jan. 2008, pp. 107-113
7. B. Liu, E. Blasch, Y. Chen, D. Shen and G. Chen, Scalable Sentiment Classification for Big Data Analysis Using Na¨ive Bayes Classifier, Proc. IEEE International Conference on Big Data. IEEE, Oct. 2016, pp. 99-104
8. Z. Li, Na¨ive Bayes Algorithm For Twitter SentimentAnalysis And Its Implementation In MapReduce, Diss. University of Missouri {Columbia, Dec. 2014
9. Nazan Öztürk, Serkan Ayva, Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis,Telematics and Informatics,pp. 136-147, 2018
10. Barkha bansal,Sangeet Srivastava,On Predicting elections with hybrid topic based sentiment analysis of tweets, 3rd Int. Conf. on Comp. Sci. and Comput. Intelli. vol. 135. pp. 346-353
11. Ankit,Nabizath S., An Ensemble Classification System for Twitter Sentiment Analysis, Int. Conf. on Comput. Intelli. and Data Science(ICCIDS), pp.937-946, 2018
12. Atanu Dey, Jitesh, J. Thakkar, Senti-N-Gram:An n-gram lexicon for sentiment analysis, Expert Systems and Applications, pp.92-105, 2018.
13. Wei-Ping, Z., Ming-Xin, L., Using MangoDB to implement text book management system instead of MySQL. Communication Software and Networks 3rd ICCSN conference, pp 110-115. 2017
14. Tang,J.,Chang,Y.,& Liu, H.(2013). Mining Social media with social theories:a survey, ACM SIGKDD Explorations Newsletter, 15(2), pp 20-29
15. Medhat,W., Hassan, A., & Korashy,H.(2014), Sentiment analysis algorithms and applications: a survey. Ain Shams Engineering Journal 5(4), pp. 1093-1113
16. Sun, S., Luo, C., Chen, J., 2017. A review of natural language processing techniques for opinion mining systems. Inf. Fusion 36 (2017), 10–25
17. Danneman, N., Heimann, R., 2014. Social Media Mining with R: Deploy Cutting-Edge Sentiment Analysis Techniques to Real-World Social Media Data Using R
18. Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., González-Castaño, F.J., 2016. Unsupervised method for sentiment analysis in online texts. Expert Syst. Appl. 58, 57–75.
19. Oza, K.S., Naik, P.G., 2016. Prediction of online lectures popularity: a text mining approach. Procedia Comput. Sci. 92 (2016), 468–474
20. [20]. Lee, N.Y., Kim, Y., Sang, Y., 2017. How do journalists leverage Twitter? Expressive and consumptive use of Twitter. Social Sci. J. 54 (2), 139–14
21. Crannell, W.C., Clark, E., Jones, C., James, T.A., Moore, J., 2016. A pattern-matched twitter analysis of US cancer-patient sentiments. J. Surg. Res. 206 (2), 536–542
22. Wang, Hao, et al., 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics.
23. Park, S.J., Park, J.Y., Lim, Y.S., Park, H.W., 2016. Expanding the presidential debate by tweeting: the 2012 presidential election debate in South Korea. Telematics Inform. 33 (2), 557–569
24. Ahmed, S., Jaidka, K., Cho, J., 2016. The 2014 Indian elections on Twitter: a comparison of campaign strategies of political parties. Telematics Inform. 33 (4), 1071–1087.
25. Cheong, Marc, Lee, Vincent C.S., 2011. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. Inf. Syst. Front. 13 (1), 45–59.
26. I. Ha, B. Back and B. Ahn, MapReduce Functions to Analyze Sentiment Information from Social Big Data,International Journal of Distributed Sensor Networks, vol. 11, Jun. 2015, pp. 5
27. S. Prom-on, S. Ranong, P. Jenviriyakul, T. Wongkaew, N. Saetiew and T. Achalakul, DOM: A big data analytics framework for mining Thai public opinions, Proc. International Conference on Computer, Control, Informatics and Its Applications, Oct. 2014, pp .1-6
28. V. N. Khuc, C. Shivade, R. Ramnath, J. Ramanathan,Towards Building Large -Scale Distributed System forTwitter Sentiment Analysis, Proc. 27th annual ACM symposium on applied computing, ACM, Mar. 2012, pp. 459-464
29. Neha Singh, Nirmalya roy, "Analyzing the emotions of crowd for improving the emergency response services", Pervasive and mobile computing 2019
30. Raoul. V. Kubler, Anatoli Colicev and Koen H.(2019). " Social Media's impact on the consumer mindset: when to use which sentiment extraction tools?", Journal of Interactive Marketing,

31. J .Ramsingh, V. Bhuvaneswari(2018). "An efficient Map Reduce based Hybrid NBC-TFIDF algorithm to mine the public sentiment on diabetes mellitus- A Big Data Approach", Journal of King Saud University - Computer and Information Sciences
32. Akilandeswari J.,Jothi G.(2018) "Sentiment Classification of Tweets with Non-Language Features", 8th int. Conf. on Adv in comput. and communication, pp 426-433.
33. S. M. Mohammad, #Emotional Tweets, Proc. First Joint Conference on Lexical and Computational Semantics, vol 1, Proc. Main conference and the shared task, vol 2, Proc. Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics, Jun. 2012, pp. 246-255.
34. S. M. Mohammad, S. Kiritchenko and X. Zhu, NRCCanada: Building the State-of-the-Art in Sentiment Analysis of Tweets, Proc. Seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), 2013
35. Sentiment140.URL: http://help.sentiment140.com/forstudents/