IJCSN International Journal of Computer Science and Network, Volume 6, Issue 2, April 2017
ISSN (Online): 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Sentimental Analysis of Social Networks using MapReduce and Big Data Technologies

[1] Vikas Chauhan; [2]Anupam Shukla

[1] Indian Institute of Information Technology & Management, Gwalior, Madhya Pradesh India

[2] Indian Institute of Information Technology & Management, Gwalior, Madhya Pradesh India

Abstract -Social networks have become important medium of communication in presents days. This communication is regularly evolving towards the goal of making it as human and real as possible. Social network sites are used to express thoughts, emotions and opinion. There is a need of a framework that can recognize emotions present in communication or emotions of the involved users in order to enrich user in social networks. The data is generated by social network sites is very huge in amount and it is generated regularly, So in our study we have proposed a Big Data architecture to identify emotions present in text data generated by Social networks. The proposed architecture comprises following phases. 1. Collection of data. 2. Model building for data management. 3. classification of text according to emotions from text using mode. We have used Apache mahout, Hadoop, NaiveBayes, OpenNLP, LMClassifier to create model, and classification of text. The traditional classifiers has restriction that these can work only for subset of large set of data which has to be taken for analysis and the size of data generated by social networks grows in exponentially. Hadoop is a framework that is used for distributed storage and parallel data processing. Our proposed architecture is able to create a combined clustering and classification of data that run on Hadoop to classify text data generated by social networks in to positive and negative classes. We tried to optimize the performance of Big data analysis and used map reduce paradigm in hadoop architecture. It is very useful in the distributed environment for huge number data and performs classification faster than standalone system.

Keywords - Bigdata, Hadoop, MapReduce, NaiveBayes, LMClassifier, Social Networks.

1. Introduction

Social networking site (SNS) uses, emotions extraction to extract useful information from multimedia data and SNS text to evaluate real intention of their users. In particular, Emotion extraction or sentimental analysis has been used to understand the real meaning and intention of the users on social network sites. SNS users rapidly generates huge amounts of data so it is necessary to develop such technologies that can extract meaningful Information and predict the emotion from data.This information can be proved useful in different fields, social security, recommendation, politics, culture and opinion mining. However, there are a lot of text and multimedia data on SNSs, and they have anatypical nature, so those technologies are important which can efficiently executes and process on large amounts of unstructured SNS's data. Using Social media people interact each other and they share, create and exchange information in virtual world .Many forms of multimedia information are contained in Social media. For example, Twitter contains different types of information, text, image, and video. In this study, we have focused on textual data and we study methods to extract emotional information from textual data. This data is generated in huge amount and very frequently so We have used big data technologies with conventional text classifiers LMClassifer, NaiveBayesClassifier ,OpenNLPclassifer. In order to extract meaningful information from large amounts of unstructured data on social network , a structuring process is required for the unstructured data. Up to now, different technologies to process unstructured data have been analysed and studied. Thus, research techniques on text mining [1] [2] has been developed, in which some useful information is extracted from atypical or semistructured text data, and it is based on natural language processing(NLP).A statistical, periodic algorithm are used in these methods based on machine learning to extract meaningful and required information and to extract the useful information from the text data. Based on text mining technologies, some research work has also been carried out to determine the sentiment tendency of social network users, such as positive, negative preferences [3]. Recently, various open sources associated with the processing of big data have been provided. The Hadoop system [4] [5] is very famous big data processing system and it is most commonly used. More information on Hadoop system will be given in the next section. In this study we proposed a Hadoop Distributed File System (HDFS) [4] and MapReduce [6] approach based on Hadoop , which stably collect and store a variety of data generated by social

networks and analyze the sentiments of users on networks.Emotions have been studied in behavior sciences and psychology , They have also attracted the attention of researchers in computer science, especially in the human computer interaction field of research.Emotions in textual data is becoming increasingly important from an applicative point of view. Consider for example the tasks of market analysis ,opinion mining ,natural language interfaces,affectivecomputing.This paper describes experiments concerned with the emotion analysis of sentences.We have selected twitter social netwrok for our experiments and we have collected tweets of different dataset based on twitter network and we proposed a methodology on Map Reduce algorithm with OpenNLP, LMClassifier and NaiveBayes classifier on Hadoop to classify sentences into positive and negative classes.

2. Related works

Text analysis is a process extract and discover usefull information from the data with computational linguistics, and natural language processing.Social network is a powerful weapon to exhibit peoples ideas and views.[7] We can filter the text to permeate offensive messages using rule based and classification techniques of text. Social networks textual data has big diversity and size, This diversity and size create the need to find emotions contained in text data generated by social networks sites. It becomes a difficult classification task because users presents their emotions using emocations ,short word or sometimes idioms and There are some non usefull information like web link are attached in the text generated by social netwroks.Haddi and Liu [8] explored the role of text preprocessing in emotion finding in text, and reported that we can use feature selection and representation to preprocess the text data. Behavioral economics [9] shows us that emotions can strongly affect individual's behavior and decision making. Emotions affect the decision making collectively in society and mood of public is used to correlate or predict economic measures. Large scale twitter feeds are correlated [9] and daily emotions of users are analyzed into happy or sad and cross validated the resulting mood time series by comparing their ability to detect the publics response in presidential election and Thanksgiving day in 2008. Zhang and Li [10] have classified the reviews of their customer Support Vector machne and naiveBayes classifier. XueBai [11] proposed a Markov blanket model with heuristic search enhancement and it was able to capture the dependencies among words in text and provide a vocabulary that is helpful to find emotion in textual data. He has computed results on online movie reviews and collections of online text news showed that this method is able to identify a set of predictive features,and he suggested that sentiments can be captured by conditional dependencies among words in text as well as by high-frequency words or keywords . Some people express their emotions in different way likr idioms and phrases so role of idioms become important for emotion extraction in textual data. [12] Three measured precision, recall and F-measure can be used to improve the accuracy of results.There is another works based on text analysis on where Emotion extraction from online travel blog was done by using three supervised machine learning algorithms of Naive Bayes,SVM and the character based N-gram model and it shows [13] that machine learning techniques can be used to find positve and negative emotion from textual data. We can use simple methods tocatagorize the textual data [14] described by Peng ,which with autonomous and multiple sources. [15] As the Data increasing very drastically this is a major challange to organize and manage the data very efficiently. This emerged as the necessity of current machine learning techniques.So various libraries are available like mahout which are compatible with hadoop and mapreduce and these deal with this generated big data. Big data is a field that deals with this rapid growth of data by using storage techniques, dedicated infrastructures and development frameworks for the parallelization of defined tasks and its consequent reduction. Reduction techniques are efficiently used in big data online applications to improve classification [16] problems. Big Data concerns about [17] complex, large-volume, growing data sets with multiple and autonomous sources.

With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including biological ,physical and biomedical sciences.Reduction in big data usually falls in one of two main methods: (i) reduce the dimensionality by pruning or reformulating the feature set; (ii) reduce the sample size by choosing the most relevant examples. Both approaches have benefits, not only of time consumed to build a model, but eventually also performance-wise. We can use [7] the distributive approach to do the task with distributive approach.Gandomi and Haider [18] have given consolidated description of big data by integrating definitions and focused on the analytic methods used for big data they attempts to offer a broader definition of big data that captures its other unique and defining characteristics. Hopscotch hash scheme is provided to improve the performance of data storing and indexing of Naive-Bayes algorithm,[19]and presents a software implementation of Naive-Bayes text classification mapped in Topo-MapReduce model and it's results show that the improved hopscotch hash speeds up by 33% at maximum compared to the original hash, and the proposed MapReduce speeds up the Naive-Baye algorithm by 29% at maximum compared to the original MapReduce.Accuracy of a text classifiers for emotion

extractionin training proceess is difficult [20] in parallel processing and has slower computation so they have proposed Mapreduce approach for large scale textual data.Mapreduce model is appropriate [21] for classification model and it reduces the computation time in training process on large scale text data.The rise of different big data frameworks such as Apache Hadoop and, more recently, Spark, for massive data processing based on the MapReduce paradigm has allowed for the efficient utilisation of data mining methods and machine learningalgorithms in different domains. The combination of big data technologies and traditional machine learning algorithms has generated new and interesting [22] challenges in other areas as social media and social networks. They have shown that different frameworks are currently appearing under the umbrella of the social networks, social media and big data paradigms.We have used tweets of twitter social network for classification and used three classifiers naivebayes, LMClassifier and OpenNLP classifier for our experiment. First We have classified tweets using thsese three classifiers and after that we have used these classifiers using mapreduce approach on Hadoopcluster.Our work is described in following sections.

3. Methodology for Big Data Analytical Architecture for social networks sentiments

We have developed an architecture to extract emotions from text generated by Twitter social network using Hadoop and created a cluster of two and three system.Flow diagram of Metholodogy of our work is shown in figure 1. HDFS have all the text data into files and these files are splitted into block. these blocks are stored in different systems of cluster.We have created model for all three classifiers NaiveBayes, LMClassiferand OpenNLP. We have trained the data set and classified the text into two classes postive or negative acording to the emotion contained by text. We have used three classifer for classification and also used these classifer with Hadoop cluster by using MapReduce paradigm to classify this large amount tweets and analyzed the results related to accuracy and execution time.Used components and classifiers of our architecture are described in next subsections.

3.1. Hadoop

Hadoop is a framework and it allows to process and store large amount of data in a distributed environment in different computers in a cluster. It is designed in a way that we can scale up the cluster size from single servers to thousands of machines and each of these machine offers local computation and storage.

3.2. Hadoop distributed file system (HDFS)

The hadoop distributed file system is a distributed file system and it executes on computer system. The hadoopdistributed file system provides fault tolerance and it is duitable for those applications that works on large amount of dataset. The hadoop distributed file system provides high throughput. Streaming of data can also be enabled using hadoop distributed file system. It works on master slave architecture. A system called name node works as a master whose responsibilities are to manage file system and file access among system. Storage attached with each system is managed by datanode and this datanode is usually on one node per system. Hadoop distributed file system exposes file system name space and data is stored in files. This file is splitted in blocks and these blocks are stored in datanodes. Namenode executes on file system

3.3. MapReduce Algorithm

MapReduce algorithm is used to process lrge data sets.We can create a map function that processes a key and value pair and generates a set of intermediate key and value pairs. There is another function reduce function that merges all intermediate values associated with the same intermediate key. Lots of real world tasks can be done by this model and these map and reduce function can be implemented in this way that task can be automatically executed on a cluster of system . Run time system of map reduce takes care of detail of input data, handling machine and system failure, scheduling across machine and communication among machine. Map reduce model provides easy utilization of distributed system to user without any experience to it's deep detail. Our implementation of MapReduce algorithm on Hadoop runs on

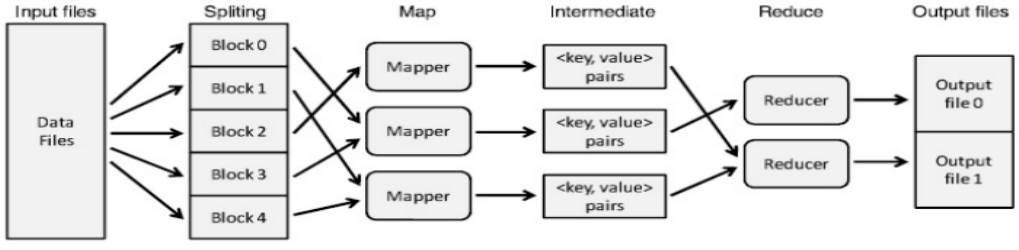


Figure 1: Map Reduce

a large cluster of machines and it is highly scalable. There are two map and reduce functions are described below.

• The Map: The map function's job is to process the input. Usually this input data is provided in the form of file it is stored in Hadoop file system (HDFS). This input file is passed to the map function line by line. The map processes the data and creates several small chunks of data and returns key value pair.
• The Reduce: The Reduce functions job is to take all value of specific key and process that data which comes from the mapper. After processing of these values and data ,it produces a new set of output, and this output is stored in the HDFS.

3.4. NaiveBayes Classifier

For text classification our first classifier is Naive ayes classifier. It worked well on classification of tweets. Naive Bayes classifier is based on Bayes' theorem and it is a probabilistic classifier also based on strong independence assumptions. All attributes in NaiveBayes are independent so it is also called conditional independent classifier. It assumes that each feature is conditional independent to other features in the the class. We have taken two classes 1 for positive and 0 for negative in our training set. K.Ming Leung
[23] has described the Bayes rule and these are shown in equations 1 and equation 2.

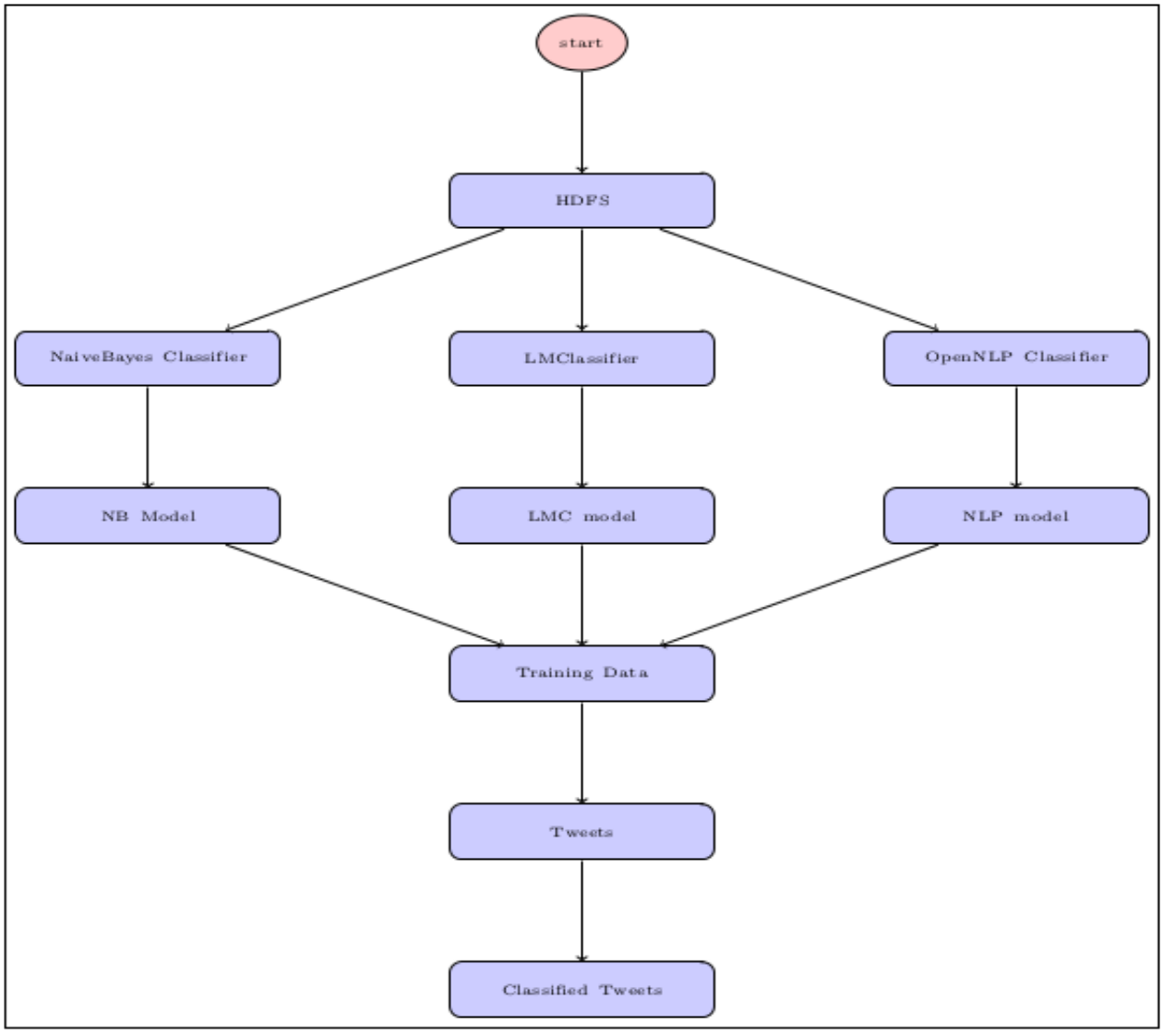


Figure 2: Distributed architecture of text classification

$$P(doc|V_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|V_j) \quad (1)$$

Where P (a i = w k |V j) is probability that word in position i is w k given v j is classified class positive or negative.

$$P(w_k|+) = (n_k + 1)/(n + |Vocabulary|) \quad (2)$$

Where P (w k |+) is probability of being positive word at position k in tweet. n k is number of times word k occurs in class positive.

$$P(w_k|-) = (n_k + 1)/(n + |Vocabulary|) \quad (3)$$

n is number of words in positive class and P (w k |–) is probability of being negative word at position k in tweet. n k is number of times word k occurs in class negative n is number of words in negative class The multinomial model [18] based on NaiveBayes calculates word frequency information from the text in document. Maximum Likelihood Estimate is simply a relative frequency and it corresponds to most likely value of each parameter given in the training data.The problem with the Maximum Likelihood Estimate estimate is that it is zero for a term-class combination if it is not occurred in the training data.so to eliminate this zero probability problem [18], We add one one to each of the count.Adding one smoothing can be interpreted as uniform prior(each term occurs once for each class) that is then updated as evidence from the training data comes in. After creation of model We can classify tweet in positive or negative class using following equation.

$$V_{NB} = argmax_{v_j \in V} P(V_j) \prod_{w \in words} P(w|V_j) \quad (4)$$

Where V,N,B are the vaule returned by maximum probability of being positive or negative tweet.

3.5. LMClassifier

The utilized LingPipesLanguageModel (LM) classifier is based on probability and techniques of statistical language modeling, it classifies text into nonoverlapping categories. It is widely used clasifer in modern time because it is easily available .In this classifier probability of sequences of word s=c1,c2,c3...cn is calculated from document of text. Peng [14] the simple approach for language modelling which is based on n gram model. An n-gram is a sub-sequence of n items from a given sequence.If a sequence is given then then n gram is sub sequence of these n items. Characters or words can be items.

$$P(c_i|c_{i-n+1}^{i-1}) = \frac{\#(c_{i-n+1}^{i})}{\#(c_{i-n+1}^{i-1})}$$

If the language model is based on character sequences, it is called character language-model. The probability of any item sequence is calculated as frequency of the observed patterns According to language modeling classification, .where #( ) represents the total occurrences of a specified gram in the training data. We have used LingPipe'sLMClassifier and it creates a character based language-model for each category (positive or negative)during the training phase of training data; then, at classification time of text data, it calculates conditional and joint probabilities of each category for the classified object. It calculates also a score which represents the character cross-entropy rate normalization and allows between-document comparisons. This score is ordered in the same way as the joint probabilities. Finally, LingPipe classifier returns one best category (positive or negative acording to text) as result of classification process. In the classification process first category array and n-gram size is initialized then for each category a loop is continues. Training data is available in the directory by category name then training files are read using LMClassifier. After it resultant data is used to classify the text into two category positive or negative. LingPipe provides the best class acording to emotion represented by text. We have classified twitter network's tweet and categorized these into positive or negative classes.

3.6. OpenNLP Classifier

OpenNLP classifier uses maximum entropy classifier for classifing text. In maximum entropy we set constraints on training data on the conditional distribution. Characteristic of training data is expressed by each constraint and it should always present in a learned distribution.Let We use any function having real value of document[24] and a feature, f i (d, c) of class. Maximum entropy classifier provides us to restrict the distribution of model to have same expected value of this feature as seen in training dataset, D, So, we can say that the learned conditional distribution P (cld) follows the property: Usually , P(d) document distribution is

$$\frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} f_i(d, c(d)) = \sum_{d} \mathrm{P}(d) \sum_{c} \mathrm{P}(c|d) f_i(d, c)$$

unknown, and we are not interested to model it. Thus, we can use our training data of text , without class labels, as approximation to document distribution, and it enforce the constraint:

$$\frac{1}{|D|}\sum_{d\in D} f_i(d,c(d)) = \frac{1}{|D|}\sum_{d\in D}\sum_{c} P(c|d) f_i(d,c) \tag{5}$$

Thus in maximum entropy ,First of all we identify a set of feature functions which are usefull for classification of text. After it for each of these feature we measure it's expected value over training dataset and make it to constraint for model distribution. If we estimate the constraints in these manners then unique distribution will exist and it will chase maximum entropy and it has exponential form [25] as.

$$P(c|d) = \frac{1}{Z(d)} e^{\sum_i \lambda_i f_i(d,c)} \tag{6}$$

$$Z(d) = \sum_{c} e^{\sum_i \lambda_i f_i(d,c)} \tag{7}$$

where feature is f i (d, c) , λ i is a parameter for restimation and Z(d) is simply normalizing factor to ensure a proper probability:When labeled training data is used to estimate constraints then solution of the maximum entropy problem is also solution of a dual maximum likelihood problem for models of the same exponential form. Additionally, it is guaranteed that the surface of likelihood is convex, and it has a no local maxima and single global maximum.

$$\frac{\partial B}{\partial \delta_i} = \sum_{d\in D} f_i(d,c(d)) - \sum_{c} P_A(c|d) f_i(d,c) e^{(\delta_i f^{\#}(d,c))} \tag{8}$$

So we can guess any initial correct form exponential distribution as starting point then, perform hillclimbing in likelihood space,The maximum likelihood solution will be converged for exponential models by this, and it will also be a solution which is global maximum entropy solution because there are no local maxima. We can say that likelihood will increase if we can find a λ i such that B is positive. If we differentiate B with respect to change in each parameter δ i then we will get best λi and solve it for maxima. Afterstaring the hadoop ,hdfs cluster We have classified the tweets into two categories positive and negative in distributed environment. We have used NaiveBayes Classifier, OpenNLP tool for implementation of maximum entropy classifier and LMClassifier to classify the tweets into two categories.All the work is executed in non distributed and distributed environment. Then We have compared the execution time and acuuracy of the three classifiers in distributed and non distributed environment also.

Algorithm 1: Maximum entropy algorithm for OpenNLP classifier

1 Input: Collection of Labelled documents D and a feature set of functions f_i
2 Set the constraints (Equation 3.5) For each feature f_i , it estimates it's expected value on the training set of documents
3 Initialize all λ_i =0;
4 Iterate this untill it convergence
5 Calculate the expected class labels for each document with current parameters $P(c|d)$ (Equation 3.6)
6 For each parameter λ_i
7 Set $\frac{\partial B}{\partial \delta_i}$ =0 and solve for δ_i(Equation 3.8)
8 set λ_i=λ_i+δ_i
9 Output: classified text with class label positive or negative

Algorithm 2: Map Function

1 Get input from dataset ;
2 **if** *This is the first iteration* **then**
3 | Train the traing data into two classes with associated probabilities 0 for negative and 1 for positive class;
4 **else**
5 | take tweet from test data and classifiy it with OpenNLPclassifier whether it is positive or negative.
6 **end**
7 collect all tweets from classified classes positve and negative;
8 Assign output key-value as $< key, sum >$;

4. Experiment and Result

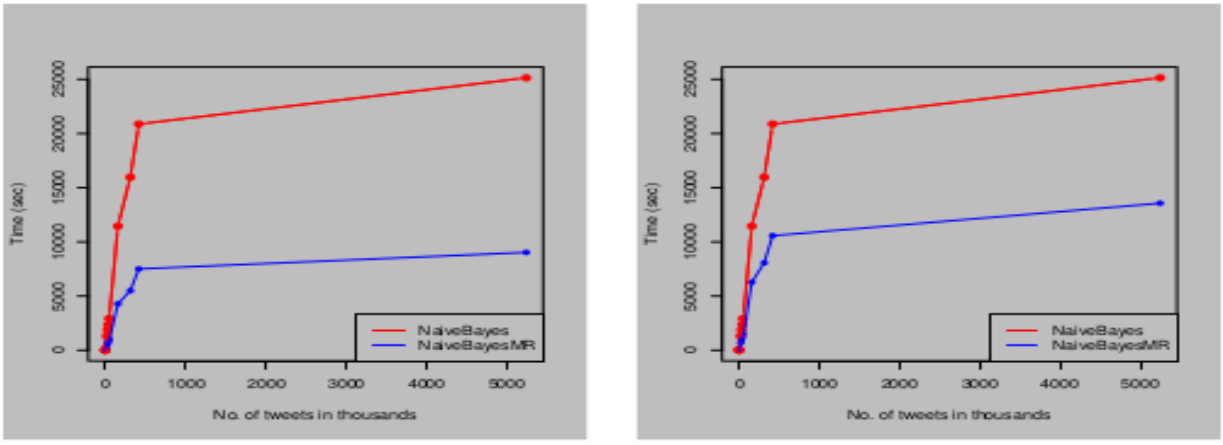
4.1. Experimental Environment

The experimental environment is shown inTable1. In this three system is used and each one has 4 GB memory, and 500 GB disk. To store data and to implement Mapreduce algorithm we use Hadoop version 2.7.0.We have used three classifier OpenNLP,LMClassifier and NaiveBayes Classifier to detect the sentiment of the text. We create one namenode and three datanodes. All of the machines use Ubuntu 15.10. as their operating system.

• **Stanford Twitter Sentiment:** Test Set (STS-Test) : We have used this dataset in two forms first one is manually

Dataset			OpenNLP					LMClassifier					NaiveBayes				
	% of test tweets	total tweets	TP	TN	FP	FN	accuracy	TP	TN	FP	FN	accuracy	TP	TN	FP	FN	accuracy
STS large Dataset	20	209715	98457	54291	34301	22666	72.83	94456	51910	36682	26667	69.79	107615	46400	37344	18356	73.44
	30	314572	147406	81733	49788	35645	66.50	142619	76272	55249	40432	66.47	15767	217331	77599	3875	74.10
	40	419430	197724	106484	67994	47228	72.52	193141	96516	77962	51811	69.05	210576	103494	67886	37474	74.88
	50	524287	244933	132613	85100	61641	72.01	243841	114746	102967	62738	68.39	281178	115498	88560	39051	75.66
STS small Dataset	20	19997	7989	5314	5389	1305	66.50	5704	7590	3113	3590	66.47	6234	7851	2852	3060	70.43
	30	29996	12343	7864	6972	2817	67.36	9064	10429	3113	3590	64.90	10215	10807	3999	4945	70.08
	40	39994	16722	10570	8476	4226	68.24	12475	13288	5758	8473	64.41	14118	13826	5220	6830	69.87
	50	49974	21265	12645	10431	5653	67.85	19205	14233	8843	7713	66.91	18567	17483	6767	7157	72.13
Sender Twitter Dataset	20	292	87	101	45	59	64.38	120	68	78	26	64.38	107	80	66	39	64.04
	30	438	119	152	67	100	61.87	180	103	116	39	64.61	167	114	105	52	64.15
	40	584	165	191	101	127	61	242	120	172	50	62	229	145	147	63	64.04
	50	730	222	248	117	143	64.38	286	180	185	79	63.83	264	208	157	101	64.65

Table 2: Accuracy table for three classifier using mapreduce with the percentage of training data



(a) cluster size 2 (b) cluster size 3

Figure 3: Execution time of NaiveBayesClassifier

4.2. Collection of data

A wide range of data is available which can be helpful to classify emotions in tweet as positive and negative.We have collected millions of tweets from various sources and description of those dataset is given below.

• **Sanders Twitter Dataset:** The Sanders dataset consists of about 1400 . Each of these tweet was manually labelled as either positive or negative acording to the emotion contained in it.

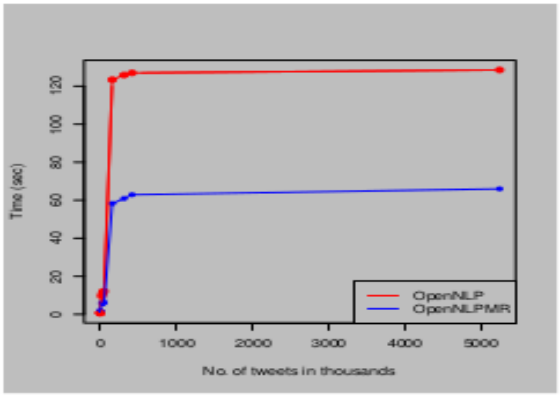
NaiveBayes Classifier with and without mapreduce approach respectively. We have calculated true positive, true negative , false positive and false positive values and calculated the accuracy of these different classifiers with and without mapreduce. Accuracy was almost similar and these results are shown in Table 2. For senders writers dataset OpenNLPclassifer performed less efficient than LMClassifer and NaiveBayesClassifier. NaiveBayesclassifer performed better than rest classifier. For STS small dataset OpenNLP performed better than LMClassifer and provided more accuracy than LMClassifer but NaiveBayesclassifer provided

more accuracy than other two classifer.For STS large datatsetOpenNLPprovied more accuracy than LMClassifer

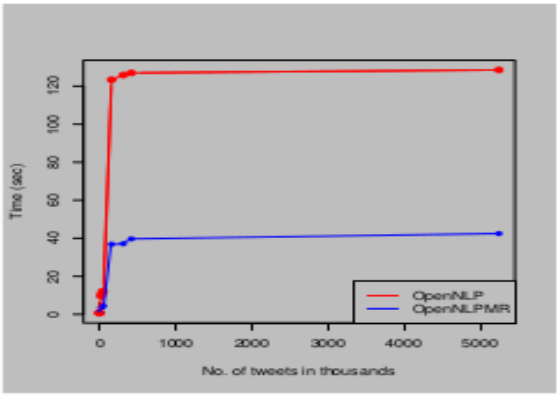
but NaiveBayesclassifier again provided more accuracy than LMClassifier and OpenNLPclassifier. Accuracy of all three classiferOpenNLP, LMClassifer and NaiveBayesClassifer is increasing with the size of test dataset and no. of tweets. We have calculated True positive ,True negative, False positive and False negative values from our 3 DataSets and calculated accuracy from the following formulaes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

where
T P = true positive
T N = true negative
F P = false positive
F N = false negative

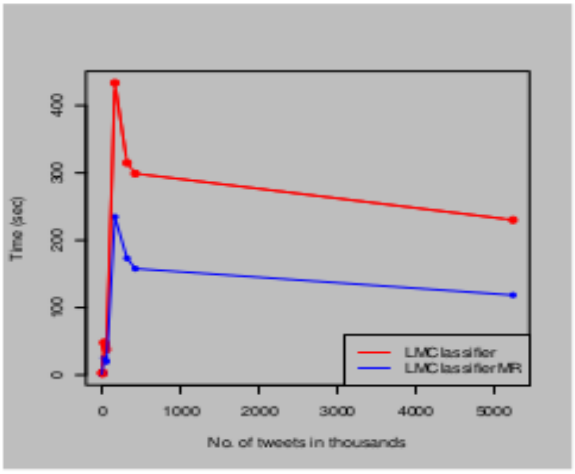


(a) cluster size 2

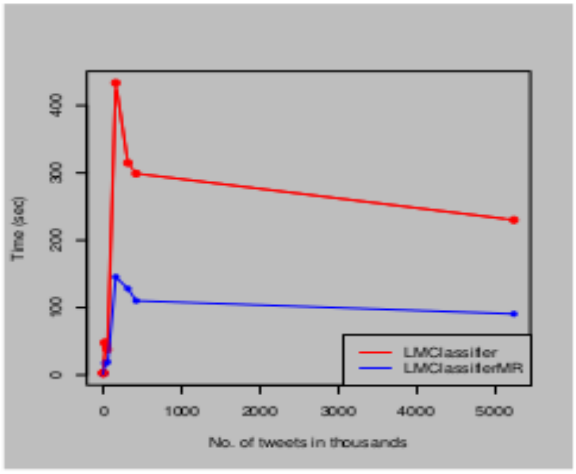


(b) cluster size 3

Figure 4: Execution time of OpenNLP classifier



(a) cluster size 2



(b) cluster size 3

Figure 5: Execution time of LMClassifier

We have found execution time is increasing propotional to no. of tweets in all three classifiers OpenNLP ,LMClassifier and NaiveBayes. We have reduced execution time by using these three classifers with MapReduce algorithm on cluster of 2 and 3 system respectively and results are shown in Figure 3,4,5respesctively for OpenNLP, LMClassifier and NaiveBayes classifier. We can manipulate no. of system in a cluster as per our requirement. Execution time is better for LMClassifer as per our experiment because it took less time to execute than OpenNLP and NaiveBayes Classifier. Execution time to LMClassifer is decreasing when size of text is increasing.NaiveBayes took highest execution time because it computes probability for each word for each sentence in test data so It is better to use Mapreduce algorithm with these classifiers and We can reduce the execution time if we use Mapreduce algorithm.NaiveBayes Algorithm provided highest accuracy and we can use it with mapreduce algorithm to reduce execution time. Figure 3,4,5 shows the graph between execution time and size of dataset with cluster

size 2 amd cluster size 3 and it shows that Execution time is decreasing when size is increasing in MapReduce approach. In conventional use of classifiers execution time increases with increment of size of dataset So applying classification techniques in distributed environment provides better execution time.

5. Conclusion

In this paper, we have proposed Real-Time Big Data Analytical Architecture for emotion extraction from social networks tweets. The proposed architecture efficiently process and analyze real time and offline data and classify tweets in positive and negativetweets.We can use the big data technologies to classify text in distributed manner and can decrease the execution time. LMClassifier, OpenNLPClassifer and NaiveBayesclassifer these classifers can be used with mapreduce approach to reduce execution time for large amount of text data.If we use NaiveBayesclassifer then we can achieve highest accuracy with less execution time because in normal Naivebayesapprocah it tooks more time to execute than OpenNLPClassifer and LMClassifer. We can use Mapreduce with all three classifiers to reduce the execution time. Text data can arrives with high velocity, veracity, Volume, variety so in these conditions We can use Mapreduce approach because it handles all these scenerios in execution.

References

[1] R. J. Mooney, R. Bunescu, Mining knowledge from text using information extraction, SIGKDD Explor. Newsl.7 (1) (2005) 3–10. doi:10.1145/1089815.1089817.URL http://doi.acm.org/10.1145/1089815.1089817

[2] W.Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5 (4) (2014) 1093 – 1113. doi:http://dx.doi.org/10.1016/j.asej.2014.04.011.URL http://www.sciencedirect.com/science/article/pii/S2090447914000550

[3] J. Leskovec, D. Huttenlocher, J. Kleinberg, Predictingpositive and negative links in online social networks,in: Proceedings of the 19th International Conferenceon World Wide Web, WWW '10, ACM, New York,NY, USA, 2010, pp. 641–650. doi:10.1145/1772690.1772756.URLhttp://doi.acm.org/10.1145/1772690.1772756

[4] Hadoop.URL http://hadoop.apache.org/

[5] Apache.URL http://httpd.apache.org/

[6] Mapreduce:simplified data processing on large clusters.URLhttp://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

[7] V. Subramaniyaswamy, R. Logesh, V. Vijayakumar, V. Indragandhi, Automated message filteringsystem in online social network, Procedia Computer Science 50 (2015) 466 – 475, big Data,Cloud and Computing Challenges.doi:http://dx.doi.org/10.1016/j.procs.2015.04.016.URLhttp://www.sciencedirect.com/science/article/pii/S1877050915005177

[8] E. Haddi, X. Liu, Y. Shi, The role of textpre-processing in sentiment analysis,ProcediaComputer Science 17 (2013) 26 – 32, first International Conference on Information Technology and Quantitative Management.doi:http://dx.doi.org/10.1016/j.procs.2013.05.005.URLhttp://www.sciencedirect.com/science/article/pii/S1877050913001385

[9] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science 2 (1)(2011) 1–8.

[10] Z. Zhang, Q. Ye, Z. Zhang, Y. Li, Sentiment classification of internet restaurant reviews written in cantonese,ExpertSystems with Applications 38 (6) (2011) 7674–7682, cited By32.doi:10.1016/j.eswa.2010.12.147.URLhttp://www.scopus.com/inward/record.urleid=2s2.079951576641&partnerID=40&md5=3a4b2b3bbb81ed9982c089a374b0da2a

[11] X. Bai, Predicting consumer sentiments from online text, Decis. Support Syst. 50 (4) (2011) 732–742. doi:10.1016/j.dss.2010.08.024.URLhttp://dx.doi.org/10.1016/j.dss.2010.08.024

[12] L. Williams, C. Bannister, M. Arribas-Ayllon, A. Preece, I. Spasi, The role of idioms in sentiment analysis, Expert Systems with Applications 42 (21) (2015) 7375 – 7385.doi:http://dx.doi.org/10.1016/j.eswa.2015.05.039.URL http://www.sciencedirect.com/science/article/pii/S0957417415003759

[13] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Systems with Applications 36 (3, Part 2) (2009) 6527 – 6535. doi: http://dx.doi.org/10.1016/j.eswa.2008.07.035.URLhttp://ww

w.sciencedirect.com/science/article/pii/S0957417408005022

[14] F. Peng, D. Schuurmans, S. Wang, Language and task independent text categorization with simple language models, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Volume 1, NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 110–117.doi:10.3115/1073445.1073470.URLhttp://dx.doi.org/10.3115/1073445.1073470

[15] A. Srinivasulu, C. D. V. SubbaRao, K. Y. Jeevan, High dimensional datasets using hadoop mahout machine learning algorithms, in: Computer and Communications Technologies (ICCCT), 2014 International Conference on, 2014, pp. 1–1. doi:10.1109/ICCCT2.2014.7066727.

[16] C. Silva, M. Antunes, J. Costa, B. Ribeiro, Active manifold learning with twitter big data, ProcediaComputer Science 53 (2015)208215.doi:http://dx.doi.org/10.1016/j.procs.2015.07.296.URLhttp://www.sciencedirect.com/science/article/pii/S1877050915017998

[17] X. Wu, X. Zhu, GongQing, W. Ding, Data mining with big data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

[18] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management.

[19] L. Zhou, Z. Yu, J. Lin, S. Zhu, W. Shi, H. Zhou, K. Song, X. Zeng, Acceleration of naive-bayes algorithm on multicore processor for massive text classification, in: Integrated Circuits (ISIC), 2014 14th International Symposium on, 2014, pp. 344–347. doi:10.1109/ISICIR.2014.7029490.

[20] X. Fei, X. Li, C. Shen, Parallelized text classification algorithm for processing large scale tcm clinical data with mapreduce, in: Information and Automation, 2015 IEEE International Conference on, 2015, pp. 1983–1986. doi:10.1109/ICInfA.2015.7279613.

[21] X. Chen, K. Wu, C. Wu, Constructing classification model with mapreduce, in: Multimedia Information Networking and Security (MINES), 2010 InternationalConference on, 2010, pp. 611–615. doi:10.1109/MINES.2010.134.

[22] G. Bello-Orgaz, J. J. Jung, D. Camacho, Socialbig data: Recent achievements and new challenges,Information Fusion 28 (2016) 45 – 59. doi:http://dx.doi.org/10.1016/j.inffus.2015.08.005.URLhttp://www.sciencedirect.com/science/article/pii/S1566253515000780

[23] Naivebayes.URLhttps://tom.host.cs.standrews.ac.uk/ID5059/L15-LeungSlides.pdf/

[24] K. Nigam, Using maximum entropy for text classification, in: In IJCAI-99 Workshop on Machine Learningfor Information Filtering, 1999, pp. 61–67.

[25] S. Della Pietra, V. Della Pietra, J. Lafferty, Inducingfeatures of random fields, IEEE Trans. Pattern Anal.Mach. Intell. 19 (4) (1997) 380–393. doi:10.1109/34.588021.
URL http://dx.doi.org/10.1109/34.588021

[26] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, Journal of Informetrics.

[27] Apache mahout.URL http://mahout.apache.org/

[28] D. Kangin, P. Angelov, J. A. Iglesias, A. Sanchis,Evolving classifier {TEDAClass} for big data, Procedia Computer Science 53 (2015) 9 – 18, {INNS}Conference on Big Data 2015 Program San Francisco, CA, {USA} 8-10 August 2015.doi:http://dx.doi.org/10.1016/j.procs.2015.07.274.URLhttp://www.sciencedirect.com/science/article/pii/S1877050915017779

[29] A. Nugumanova, A. Novosselov, Y. Baiburin, A. Karimov, Automatic keywords extraction from the domain texts: Implementation of the algorithm based on the mapreduce model, in: Current Trends in Information Technology (CTIT), 2013 International Conference on, 2013, pp. 186–189. doi:10.1109/CTIT.2013.6749500.

[30] X. Wu, X. Zhu, G. Q. Wu, W. Ding, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering 26 (1) (2014) 97–107. doi:10.1109/TKDE.2013.109.

[31] S. Sagiroglu, D. Sinanc, Big data: A review, in: Collaboration Technologies and Systems (CTS), 2013 International Conference on, 2013, pp. 42–47. doi:10.1109/CTS.2013.6567202.

[32] J. Dittrich, J.-A. Quiané-Ruiz, Efficient big data processing in hadoopmapreduce, Proc. VLDB Endow.5 (12) (2012) 2014–2015.doi:10.14778/2367502.2367562.URL http://dx.doi.org/10.14778/2367502.2367562

[33] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, A. Rasin, Hadoopdb: An architectural hybrid of mapreduce and dbms technologies for analytical workloads,
Proc.VLDBEndow.2(1)2009)922933.doi:10.14778/1687627.1687731.URL
http://dx.doi.org/10.14778/1687627.1687731

[34] H. Chen, R. H. L. Chiang, V. C. Storey, Special issue:Business intelligence research business intelligence and analytics: From big data to big impact.

[35] J. Santoso, E. M. Yuniarno, M. Hariadi, Large scaletext classification using map reduce and naive bayes algorithm for domain specified ontology building, in:Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on, Vol. 1,2015, pp. 428–432. doi:10.1109/IHMSC.2015.24.

First Author
Vikas Chauhan is an Assistance professor in Computer science and Engineering department in Madhav Institute of Technology and Science Gwalior,India. He is also connected with a professor, Dr. Anupam Shukla, who had served as his research supervisor for completion of his Master in Technology programme at Indian Institute of Information Technology and Management Gwalior in data mining and Big Data analytics. He was also associated with Infosys Limited for 2.5 years before joining of his masters in Indian Institute of Information Technology and Management Gwalior.

Second Author Professor Anupam Shukla is currently associated with ABV-Indian Institute of Information Technology and Management working as a Professor in the Department of Information Technology and has total 25 years of experience in teaching and research. He was

the member of Board of Governors from 2008 to 2013 and is presently a member of Academic Senate. He established the 'Research and Consultancy Cell' in the institute and was nominated as Professor-in charge for Sponsored Research Project and Consultancy Cell. He received Young Scientist Award from Madhya Pradesh Council of Science & Technology, Bhopal in year 1995 and Gold Medal from Jadavpur University, Kolkata in the year 1998 for his postgraduate studies. His main research area is Artificial Intelligence and he is currently focusing on Neural Networks and 'Evolutionary and Nature Inspired Computations' that have incalculable applications in Bioinformatics, Medical Expert System and Robotics. He has supervised 8 PhD students and 67 M Tech thesis in this area. He has published 161 research papers in various national and international journals/conferences and 7 book chapters.