Sentiment classification on big data using Naïve bayes and logistic regression

Conference Paper · January 2017

DOI: 10.1109/ICCCI.2017.8117734

2 authors, including:

Vikas Khullar
Chitkara University Institute of Engineering and Technology
22 PUBLICATIONS **53** CITATIONS

SEE PROFILE

Sentiment classification on Big Data using Naïve Bayes and Logistic Regression

Anjuman Prabhat [1]
Department of Computer Sciences and Engineering
CT Institute of Engineering, Management and Technology
Shahpur, Jalandhar, India.
Email: anjuman.prabhat@gmail.com

Vikas Khullar [2]
Department of Computer Sciences and Engineering
CT Institute of Engineering, Management and Technology
Shahpur, Jalandhar, India.
Email: vikas.khullar@gmail.com

Abstract—**The huge expansion of world wide web has involved a contemporary fashion of conveying the attitude or viewpoint of human being. It is a channel where anybody any visualize opinion and sentiments of different customers. It is also possible to see opinion classified into different categories and ratings given on different products. This information plays a supreme role in sentiment classification task. The huge amount of data stored online can be mined effectively to extract valuable information and do decision based on extracted information. The real time twitter reviews are feed to different supervised machine learning classifier. After training the classification is carried out by variousclassifiers. The tweets as categorized as positive or, negative. In this paper we have used Naïve Bayes and Logistic Regression for the classification of twitters reviews. The performance of algorithmshas been evaluated on the basis of different parameter like accuracy, precision and throughput.**

Keywords—Big Data; Sentiment Classification; Naïve Bayes;Logistic Regression

I. INTRODUCTION

Sentiment analysis is a process in which we evaluate people idea, opinion, feeling, attitude, thought, and belief about a particular subject on specified topic or concept. The topic could be a business organization, a news forum, an enterprise or an online product. It is sometimes referred to as opinion mining [1][2]. In these growing scenarios social media plays a prime role in dealing with such an enormous amount of information. Traditional data mining techniques does not yield good results due to the increasing amount of data in web each second. So to overcome the problem of data mining different machine learning procedure is enforced. Machine learning classifier can easily deal with large amount of data which was not possible by traditional techniques. The information collected from social media can serve as an important parameter for online enterprises if the information is properly dealt with for knowledge discovery purposes [3][4].

Sentiment analysis is tracheotomy in nature. Most of the author uses document level classification because it does not involve complexity. Document level categorizes overall document sentiment to be as positive, negative or neutral. Second is sentence level which classifies each sentence to be positive, negative or neutral. Third one is aspect level classification which is considered to be most difficult level categorization study but it yields better result and accuracy. Aspect level touches all the aspects it touches in the context and the overall opinion it introduced in that particular context [4].

Text classification using machine learning classifier can be carried out using two approachesi.e supervised and the unsupervised learning. The supervised machine learning algorithm is characterized by labeled input whereas the later learning algorithm works on unlabeled resources. After labeling the dataset the training procedure is carried out to retrieve feasible output for further knowledge discovery [5][6]. The clustering algorithms deal with unlabeled dataset.This paper uses labeled data for classification purpose. The Twitter reviews are predominantly unstructured in nature. Therefore preprocessing step is carried out to convert unstructured data to structured form. During preprocessing stage undesirable information like special symbols, stopwords, URL etc is filtered out. After this stage Vectorization is handled in which text information are converted into unique numbers in matrix representation. After preprocessing of reviews, the algorithm needs to go through a process of Vectorization which converts the text data into matrix of numbers. This matrix also known as feature set which serve as an input for the classifier. The sentiment analysis is further carried out by different classifier.

There are three kinds of classifiers like generative, probabilistic and discriminative classifier. Naïve Bayes is a probabilistic classifier based on Bayes' Theorem with independence assumptions between its features. Naive Bayes classifiers finds applications in fields mainly in text classification, target tracking, clustering, fusion systems etc[7][8]. Information fusion algorithms might not scale up for large datasets. Logistics regression is a discriminative classifier. Logistics regression is a form of regression which allows prediction of outcome variables by combination of continuous and discrete predictors. In regressions the explanatory variables are independent to each other which in turns yield better result as compared to Naive Bayes where dependency exists between variables. The major augmentation of this paper is as following:

(i) The classification is carried out by both the classifier using unigram technique.

978-1-4673-8855-9/17/$31.00 ©2017 IEEE

(ii) The representation or implementation of classifier is showed by using attribute such as accuracy, precision and throughput.
(iii) Comparative study of results obtained from both learning techniques.

II. RELATED STUDY

2.1 Techniques and algorithm

A)MapReduce Paradigm

The MapReduce framework is in use since the original paper was published to process large datasets [9].Google's clusters process huge amount of data every day in the range of petabytes of data. All the data are executed in MapReduce fashion [10][11]. MapReduce paradigm helpsGoogle file system (GFS) store large amount of data. Google file system apprehends data in a distributed manner which makes its framework automatic parallelization. Apart from it also manages file replication, fault tolerance, data distribution and load balance [12][13][14]. Replicated files are stored in the form of blocks and chunks. It comes into use whenever system crashes or any unknown activity takes place.

A MapReduce task consists of a mapper and reducer functionspecifically. The function of mapper is to process a key-value pair which it receives from its input end and gives group of key-value pair as an output which serve as an input for the reducer phase. The key-value pair of the mapper coming from its output is sorted before sending it to various reducers. The pairs are distributed between reducer according to matched key. In this process the reducer receives a key and a set of associated values that has similar key.

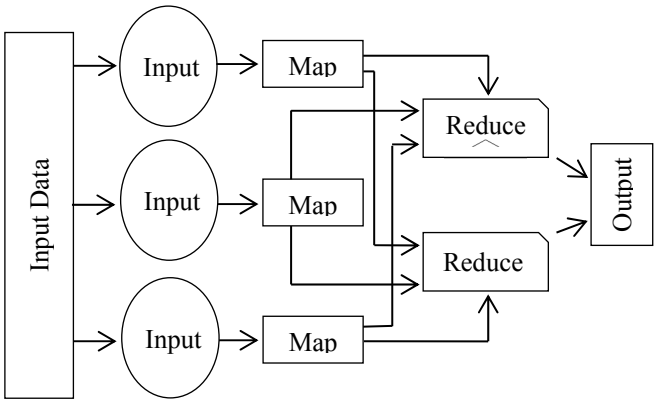

Fig 1. Execution of Map-Reduce Framework

Hadoop is an open source tool implemented on the MapReduce programming paradigm. Hadoop finds its applications in industry and academics for processing of bigdata. Hadoop framework consists of Hadoop ` distributed File System.

B) Mahout

Mahout is a popular tool for large scale Machine Learning. Mahout finds applications in project like Recommender systems and Text classification.Mahout has lot of usage in machine learning algorithm like text classification, clustering etc.

A) Naïve Bayes Classification(NBC)

Naïve Bayes classifier is a probabilisticclassifier which is refers to Bayes Theorem. NBC is based on conditional probabilities with independent assumptions between its features. The Naïve Bayes classifier assumes categorical classlabels and categorizes data based on training set and values in test data. Naïve Bayes finds applications in Spam Filtering, Text classification,Hybrid Recommender System [15][16][17]. Mathematically Naïve Bayes classifier is defined as below:

$$P(X|E_1,\dots,E_n) = \frac{P(E_1,\dots,E_n\ |X)P(X)}{P(E_1,\dots,E_n)} \dots\dots(i)$$

Where, X is the probability of an event
E is the given evidence
$P(E_1 \dots E_n|X)$ = Likelihood
$P(X)$ = Prior
$P(E_1 \dots E_n)$ = Normalization constant

Naïve Bayes has proved to be optimal and efficient in machine learning text classification [18][19].pang et.al categorizes reviews on the basis of n-gram technique. It groups or categorizes the polarity of sentence as either positive or negative. In this paper the author had used a supervised and a unsupervised learning algorithm. The supervised learning algorithm used is support vector machine and unsupervised learning method used is k-mean clustering. The evaluation is done using comparison between three labeling functionalities including K-mean labeling, Polarity labeling and SentiWordNet labeling. To implement machine learning they had used bag of words as a feature model in which they got best result from SVM algorithm [20]. Salvetti et.al.used overall opinion polarity feature for classification using machine learning algorithm such as naïve Bayes and markov model.The comparativeresult concludes that wordnet was comparatively less efficient as compared to POS filter or tag [21].

B) Logistics Regression

Logistics regression is extensively used in applications of machine learning. This model apprehends a vector of variables and evaluates coefficients or weights for each input variable and then predicts the class of stated tweet as a word vector [21]. Looking mathematically logistics regression function estimates a multiple linear function which is defined as:

$$logit(S) = b_0 + b_1M_1 + b_2M_2 + b_3M_3 \dots b_kM_k \dots\dots\dots (ii)$$

Where, S is the probability of presence of feature of interest.
$M_1, M_2 \dots\dots\dots M_k$is the Predictor value and
$b_0, b_1 \dots\dots\dots b_k$ is the intercept of the model
Assumptions of logistics Regression:

- Linear relationship between the dependent and independent variable does not exist in Logistics Regression.
- The dependent variable must be dichotomous.
- The independent variable must be linearly related, neither normally distributed, nor of equal variance within a group.
- The groups must be mutually exclusive.

C) Inspiration for proposed approach

(i) Most of the research work is carried out using bag-of-words as feature selection.But in this paper TF-IDF (Term Frequency-Inverse Document Frequency) feature selection method hasbeen used.

(ii) It is also seen that a large number of traditional research work is carried out using part-of-speech (POS) label for classification. But POS label does not give good accuracy since the grammar for a word changes in different context. In this paper each term is considered for sentiment analysis.

III. SYSTEM DESCRIPTION

We scaled up Naïve Bayes to Logistic Regression Classifiers using MapReduce and Hadoop Distributed File system for categorizing millions of movie reviews. The different components of the system are Work Controller, Parser, Collector, Terminal etc.

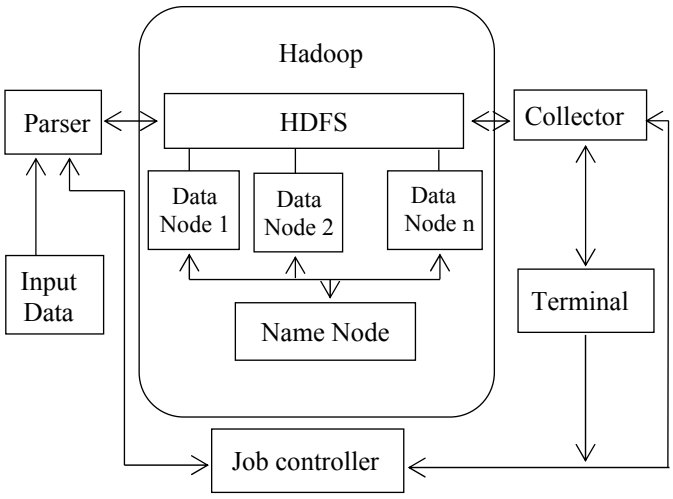

Fig 2. System to process data input using Logistic Regression or Naïve Bayes.

A) Constituent of system

The system consists of four components on Hadoop environment: job controller (JC), parser, terminal and collector. The objective of this system is to compare naïve Bayes and logistics regression on twitter reviews. The parser converts the input data to a format desired to process each review. The input job is submitted through terminal. The output of the experiment result is collected from collector after it completes its collection. The function of job controller is as follows:

(i) To control the work flow of entire system i.e to transfer source code to name node.
(ii) It informs parser the desired input format and the appearance of output data.
(iii) Prompt the collector to gather computing results from Hadoop Distributed File System (HDFS).

C) Dataset

In this research work, we have used real time twitter review.These datasets contain equal number of positive and negative reviews which make machine learning algorithm easy to classify the reviews.

D) Experimental setup

Configuration of single Node Machine is in Table 1

Table 1.Configuration of single Node Machine

Hadoop Version	Hadoop 2.7.1
Bench Program	Naïve Bayes& Logistics Regression
File System	Hadoop File System
Operating system	Linux Mint 17.2
Processor	Intel®Core(TM)i3 CPU
Clustered Node	Single Node
RAM	4 GB
System Type	64- bit operating system
Mahout	Mahout 0.9

IV. METHODOLOGY

We have used Hadoop, Mahout and Eclipse interface using java programming Language to train and classify the Naive Bayes and Logistics Regression. In total we haveused dataset of size 6 MB to categorize the tweets.

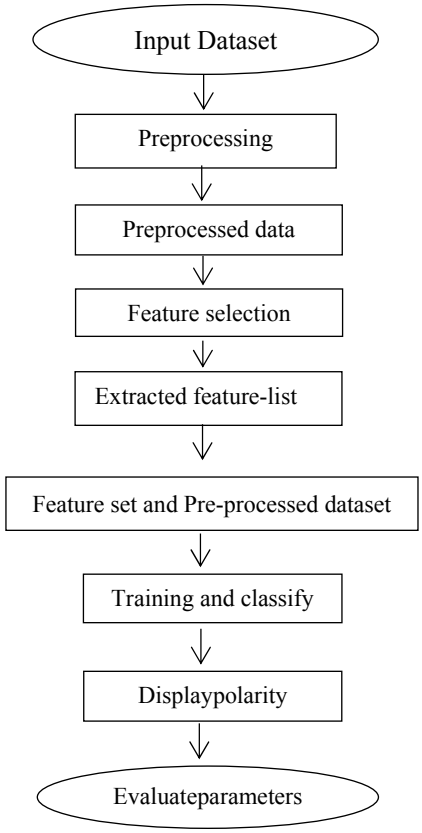

Fig 3: Flow Diagram of Proposed methodology

Explanation of the algorithm in pseudo format:

Input: Labeled Dataset.

*Output:*Positive and Negative Polarity along with values of parameters like Accuracy, Precision and Computation time.

Step-1 Pre-processing of movie reviews:
(a) Remove special symbols
(b) Remove URL
(c) Convert into lower case
Step-2 Fetch feature vector list using feature selection.
Step-4 Combine Feature vector List + preprocessed dataset.
Stopwords=file path name
Preprocessed file=path name of file
Feature Vector List=path name of Feature Vector List
Step-5 Training of prepossessed dataset in step -4 by using both classifiers.
Step-6 Classify and display polarity.
Step-7 Show accuracy,precision and computation time**.**

A) Ratingparameters:

A confusion matrix is created to visualize the performance of text classification using machine learning algorithm. These parameters are used to evaluate the performance of machine learning algorithm. To compare tag of sets terms such as true positive, true negative, false negative and false positive have been used.True positive are positive analysis and are also categorized as positive by the classifier, whereas on the other hand false positive shows positive analysis but classifier does not group it as positive case. Likewise, True Negative are negative analysisand are also sorted as negative by the classifier, whereas False Negative represents the reviews which are negative but classifier does not group or categorize it as negative. To estimate the performance of classifier, parameters such as accuracy, Precision and computation time is used.

- **Accuracy**: It is defined as the number of correctly classified reviews to the total number of reviews. It is measured in percentage.

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

- **Precision**: It is the ratio of number of reviews correctly identified as positive to the total number of reviews identified positive by the classification algorithm. It is measured in percentage.

$$Precision = \frac{Total\ TP}{Total(TP + FP)}$$

V RESULTS AND DISCUSSIONS

In this set of experiment we discuss the results obtained by using Naïve Bayes and logistics Regression and their performance is compared on the basis of parameters namely, accuracy, precision and computation time.
Table 2 shows the performance measurement of naïve Bayes and logistics regression classifier in terms of accuracy, precision and computation time. Fig 4 present a complete view of accuracy of both the classifier. A relative computation on the basis of precision is shown in Fig 5. Similarly equivalent calculation on the basis of computation time is shown in Fig 6

Table 2:Comparable result of two algorihtms

Parameters	**Naïve Bayes**	**Logistic Regression**
Dataset size	6 MB	6 MB
Accuracy %	66.667	76.767
Precision%	69.23	73.575
Computation time (Mili-sec)	15732	3689

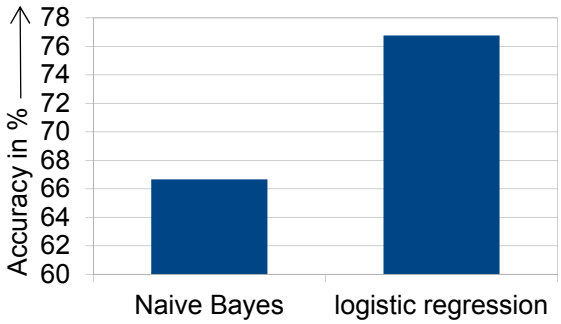

Fig 4. Comparision for accuracy

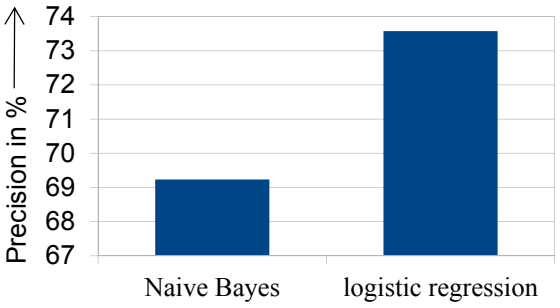

Fig 5. Comparison of precision

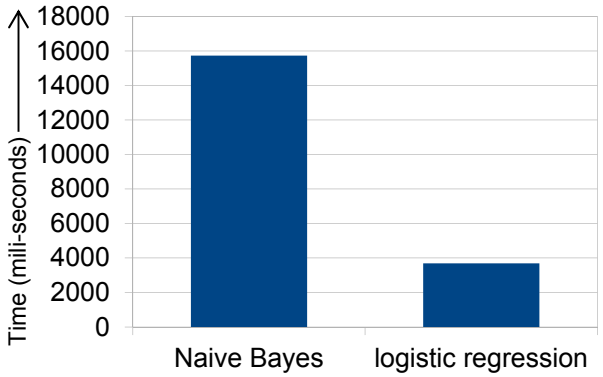

Fig 6. Comparision for computation time

VI. CONCLUSION

In this paper,twitter reviews have been used using machine learning algorithm like Naïve Bayes and logistics regression for classification task. We have implemented both the classifier on top of the Hadoop along with Mahout. Additional module like work controller is added to automate the experiment. The analysis with logistics regression gives 10.1% more accurate and 4.34% more precise results with almost one fifth implementation time for same size of dataset. As in this paper only text tweets have been used, further experiments may be organized on imaging tweets as a future scope of work. Hence sentiment classification with text as well as images will be more effective. Higher gram i.e bigram, trigram etc can also be used which may be proved more effective.

REFERENCES

[1] A.Tripathy, A.Agrawal, S.K Rath, "Classification of sentiment reviews using n-gram machine learning approach,"Journal of Expert Systems With Applications(Springer), Vol. 57, pp. 117-126, 2016.

[2] S. Aravindan, A. Ekbal, "Feature Extraction and Opinion Mining in Online Product Reviews", Journal of Computer Society(IEEE), pp.94-99, 2014.

[3] B.liu, E.Blash,Y.chen, G.chen, D.shen, "Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier", Journal of International Conference on Big Data (IEEE), Oct. 2013.

[4] B. Liu, "Sentiment Analysis and Opinion Mining,", A Review Article on SynthesisLectures on Human Language Technologies, Vol. 5, No-1, pp. 1-167, April-2012.

[5] R. Feldman, "Techniques and Applications for Sentiment Analysis",Journal of Communications of the ACM, Vol. 56, No-4, pp. 82-89, 2013.

[6] G. Gautam, D.Yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and semantic analysis", Seventh International Conference on contemporary computing(IEEE), pp. 437-442,2014.

[7] T. Hastie, R. Tibshirani, "The Elements of Statistical Learning (Second Edition Springer), 2009.

[8] E. Blasch, Y. Chen, G. Chen, D. Shen, and R. Kohler, "Informationfusion in a of cloud-enabled environment" ,Journal ofHigh Performance SemanticCloud Auditing(Springer), 2014.

[9] B. Liu, Y. Chen, E. Blasch, K. Pham, D. Shen, and G. Chen, "A holisticcloud-enabled robotics system for real-time video tracking application,"International Workshop on Enhanced Cloud Fusion, Sept. 2013.

[10] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing onlarge clusters" ,In 6th symposium on operating system design and implementation,Vol- 6, pp.137-149, 2004.

[11] S. Goraya, V. Khullar, "Map-Reduce Synchronized and Comparative Queue Capacity Scheduler in Hadoop for Extensive Data", Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 6, Ver. 5, pp. 64-75, Nov-Dec 2015.

[12] S. Goraya, V. Khullar, " Enhancing Dynamic Capacity Scheduler for Data Intensive Jobs", International Journal of Computer Applications, Volume 121 – No.12, pp. 21-24, July 2015.

[13] P. Barnaghi, J. breslin, P. Ghaffari, "Opinion Mining and Sentiment polarity on Twitter and Correlation Between Events and Sentiment", Second International Conference on Big data Computing Services and Applications(IEEE), pp.52-57, 2016.

[14] J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing OnLarge Clusters", Journal of Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[15] J. Li, M. Sun, "Experimental study on sentiment Classification of Chinese Review Using Machine Learning Techniques", Journal of computer society(IEEE), pp.393-400, 2007.

[16] J. Smailovic, M. Gracana, N.Lavrac, M.Znidarsic, "Stream-Based Active learning for Sentiment Analysis in the financial domain" Journal of Information Sciences (Elsevier),pp.181-203,2014.

[17] A.Balahur, M.Turchi,"Comparative experiment Using Supervised Learning and machine Translation for Multilingual Sentiment Analysis",Computer speech and Language(Elsevier), pp.56-75, 2014.

[18] D. Lewis, "Naıve (Bayes) at forty: The independence assumption ininformation retrieval", Journal of Machine Learning, pp. 4–15, 1998.

[19] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss,"Machine learning,Journal ofTransactions on Pattern Analysis and Machine Intelligence(IEEE), vol. 29, No. 2-3, pp.103–130, 1997.

[20] S. Liu, I. Lee, "A hybrid sentiment Analysis framework For Large Email Data, International conference on Intelligent System and Knowledge Engineering,. Conference Publishing Services, 2015.

[21] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up?: Sentiment Classification Using Machine Learning Techniques." In conference on Empircal methods in natural language processing in association for computational Linguistics, Vol.10, pp. 79–86, 2002.