ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Scalable Sentiment Classification for Bigdata Analysis Using Logistic Regression Classifier

Anjuman Prabhat[1], Vikas Khullar[2]

M. Tech Student, Department of CSE, CT Institute of Engineering Management & Technology, Jalandhar, India[1]

Assistant Professor, Department of CSE, CT Institute of Engineering Management & Technology, Jalandhar, India[2]

ABSTRACT: Machine learning technologies are very much popular in text classification study. It can easily learn from training dataset which help companies in decision making with high efficiency. But there are algorithms which do not scale up with increasing volume of data. In this paper, we focused on the scalability of Logistics Regression in large datasets. We have implemented logistics regression on mahout which keeps track of the huge datasets while training process. The performance of logistics regression is enhanced with the increasing amount of data. The accuracy of the classifier reaches 80% when size of dataset increases.

KEYWORDS: Bigdata, Sentiment Classification, Logistics Regression, Term Frequency-Inverse Document Frequency.

I. INTRODUCTION

In this era internet accommodate an enormous amount of lexicon data. The textual data is fast developing frequently since it is easier to produce and report it. Twitter which is one of the most popular micro- blogging websites allows people to express their emotions, opinions and attitudes towards a product. Apart from these online sites allows users to interact, present and mutate the content. This facilitates users to explicit peculiar and personal sentiment about topics related to different domains. Twitter contents are short texts which provide real time data that express ideas and sentiments of different circumstances. Therefore tweets are good source of streaming data for sentiment classification.

Sentiment analysis is the operation of determining whether the sentiment conveyed in a subject or topic is expressed in a positive or negative way. Sentiment analysis is also known as Subjectivity detection or opinion mining. To find linguistic and statistical patterns in the text sentiment analysis uses machine learning methods and natural language processing [15][16]. In recent year's machine learning technologies and natural language processing is gaining popularity due to its numerous applications and usages. A lot of business organizations use machine learning algorithms to summarize feedback so that they can bring a change in their product. During election campaigns machine learning methods plays a key role in making a decision in different circumstances [17].

Supervised machine learning algorithms generally involves following steps to obtain better accuracy and performance. The steps included are:

(i) Collect a training dataset
(ii) Labeling the dataset
(iii) Feature extraction and selection
(iv) Training on classifier using training dataset
(v) Testing using test dataset

In this work, sentiment classification is accomplished in a real- time fashion. Therefore we develop a classifier like logistics regression that scale up with the increasing volume of data. This task is achieved by executing algorithms in a distributed fashion. To process vast amount of data we use Apache hadoop which is an open source framework for distributed storage and processing of large datasets. Apache Hadoop consists of two modules, a distributed storage

DOI: 10.15680/IJIRCCE.2016. 0411215

system known as Hadoop Distributed File System (HDFS) and a distributed processing system known as Hadoop Mapreduce[9][10]. The primary objective of HDFS is to store data efficiently even in the existence of a breakdown or failure including Namenode failures. Hadoop Mapreduce is a software framework which helps users to build distributed applications so as to process vast amount of data on a cluster of machines in a parallel manner [11][12][13]. We have also used Apache Mahout in our paper. Mahout which is an Apache project is a scalable machine learning library. It allows users to execute machine learning algorithms on Hadoop.

II. RELATED STUDY

A lot of work has been done on sentiment analysis using bag-of-words and rule based method. But very few research works is done on sentiment analysis using machine learning methods. A vast amount of valuable information about different products is available in web. This information can be used by vendors, companies etc to refine in areas they lack and make necessary changes according to it. But the major challenge is that data in web so extensive that to classify these data is a tough job.

Different techniques have been used to classify this information. Tripathyet.al., have applied machine learning algorithms like naïve Bayes, maximum entropy, stochastic gradient descent and support vector machine to classify human sentiments on movie dataset. The performance of different method is accessed on the basis of parameter like accuracy, precision, recall and f-measure. Best result comes by using unigram and bigram as feature selection methods [1]. Barnaghi et.al. have used Bayesian Logistics Regression (BLR) and Naïve Bayes classification method to classify tweets related to FIFA world cup 2014. Firstly it looks for correlation between twitter sentiment and event that have occurred. To extract features from tweets the author has used unigram, bigram and TF-IDF (Term Frequency-Inverse Document Frequency) approaches to filter commonly used features. The results showed the positive and negative response of people regarding such events and how it alters based on occurrence during these events [2][14]. Kulcuet.al., have used naïve Bayes, logistics regression and complementary naïve Bayes for classification of Turkish tweets as positive or negative and link tweets to news items. Natural language processing techniques along with zemberek library have been used for morphological analysis and stemming. For mapping purpose the author has used bag-of-words method. The result showed naïve Bayes outperforms compared to other two algorithms [3][18][19][22]. Bhuvanet.al., have used Logistics Regression , Naïve Bayes and support vector machine to predict the sentiment polarity of movie reviews along with sentiment score for individual review. Features are extracted using matching semantic patterns of context specific grammar. Different domains have different sematic peculiar to that domain. Experimental results showed that Logistics regression with Stochastic Gradient Descent on Apache spark achieve superior result [4]. Qasemet.al., have compared the accuracy of machine learning techniques like neural network and logistic regression using stock related tweets. To compare both techniques weighting schemes like unigram term frequency-inverse document term frequency (TF-IDF) and bigram term frequency (BTF) has been used. Experimental result showed that unigram TF-IDF performs better than bigram term frequency [5][25]. Gautamet.al., have used naïve Bayes, maximum entropy and support vector machine for sentiment analysis of customer reviews. Semantic based wordnet is used to extract synonyms and similarity of different feature extracted from the original dataset. The experimental result showed that naïve Bayes using unigram approach gives better result compared to maximum entropy and support vector machine [6][20][21]. Liu et.al., have used naïve Bayes classifier to evaluate its scalability on large datasets instead of using mahout library. Result shows that naïve Bayes classifier can scale up easily. It does not need a database to store its contents when the size of dataset varies from one thousand to one million reviews [7][23][24]. Alec et.al., have proposed a novel approach to classify twitter messages using distant supervision. Machine learning algorithm like naïve Bayes, maximum entropy and support vector machine has been used. In this paper the author has used tweets with emoticons for distant supervision. Training machine learning algorithms with emoticons can help to increase high accuracy [8].

III. YSTEM DESCRIPTION

A. EXPERIMENTAL SETUP

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

Table 1. Configuration of single Node Machine

B. DATASET

In this research work, we have used real time twitter review. The tweets have been collected from Twitter streaming API using specific keywords which provides low latency access to the user.

IV. METHODOLOGY

To process twitter dataset we have used Hadoop environment using Eclipse IDE. To train dataset logistics regression is used which can easily scale up with increasing amount of data. Logistics regression prepares a model based on given input and based on the model predicts the outcome of test dataset.

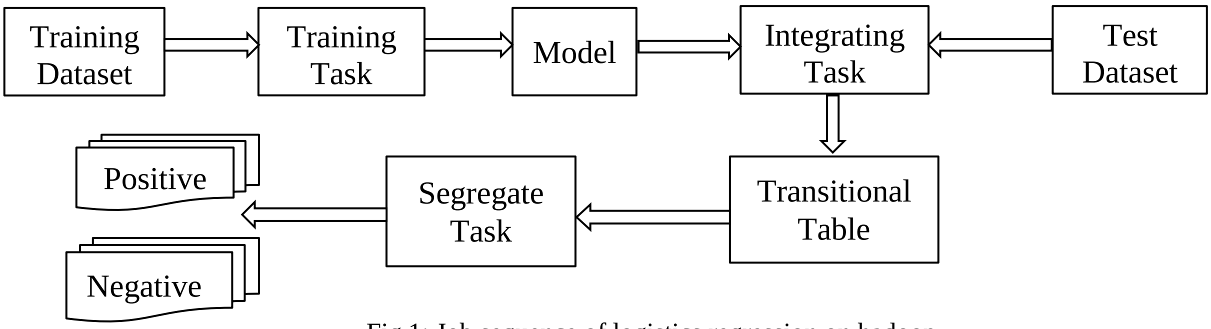

Fig 1: Job sequence of logistics regression on hadoop

A. GENERAL WORKFLOW

(i) **Pre-processing:** In this step the twitter text are converted into a common format. After preprocessing each review become one line in the dataset with document ID and sentiment prefixed either positive or negative. During preprocessing we deleted unwanted information like numbers, symbols, punctuation and meaningless words.

(ii) **Constructing Input Datasets:** Here input dataset is constructed for all tests trails for the convenience of ten-fold cross validation. The input dataset is sent to correct location in Hadoop distributed file system for all cases.

(iii) **Classification of sentiment using Hadoop:** This is the prime step in entire process. Once the training and test data is available in HDFS, the algorithm starts to build a model for training job. After training process the model is combined with test data to form an intermediate table. At last the organizer calculates the probabilities of each review in two classes subsequently and makes a decision about the sentiment of each review.

(iv) **Collection of results:** After classification process the results are stored in HDFS.

V. RESULTS AND DISCUSSIONS

In this set of experiment we discuss the results obtained by using logistics Regression on the basis of parameter like accuarcy, throughput , accuracy disruption and processing time.The performance of the classifier is evaluated on the basis of parameter like accuracy, throughput and computation time. Fig. 2 shows the accuracy of logistics regression classifier on different sizes of datasets. As the dataset size increases the values of accuracy becomes stable. But when the size of dataset is small the accuracy is unstable because the training data is large enough for the training model to learn knowledge about each class. As the size of dataset increases above 500K, the value of accuracy approaches 81% and finally climbs unto 85%. This proved that the average accuracy of logistics regression classifier becomes stable when the size of dataset increases. In fig. 3 we plot the classification accuracy into true positive, true negative followed by false positive and false negative. The experimental result showed that the values of true positive and true negative increases while the value of false positive and false negative decreases with the increase in size of data. Fig. 4 shows throughput of the classifier with the increase in size of dataset.

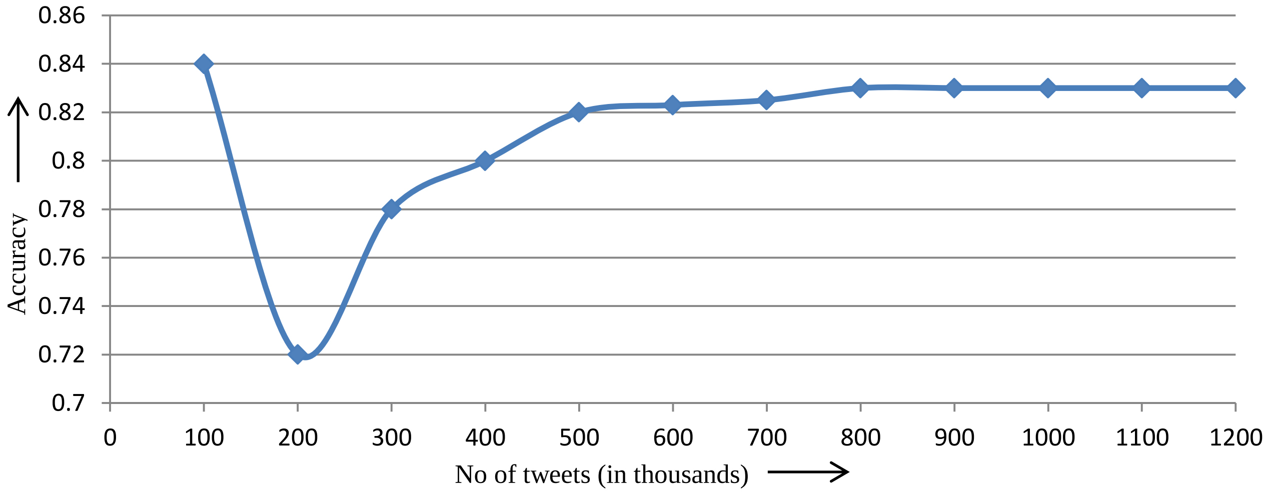

Fig 2. Variation of accuracy with dataset size

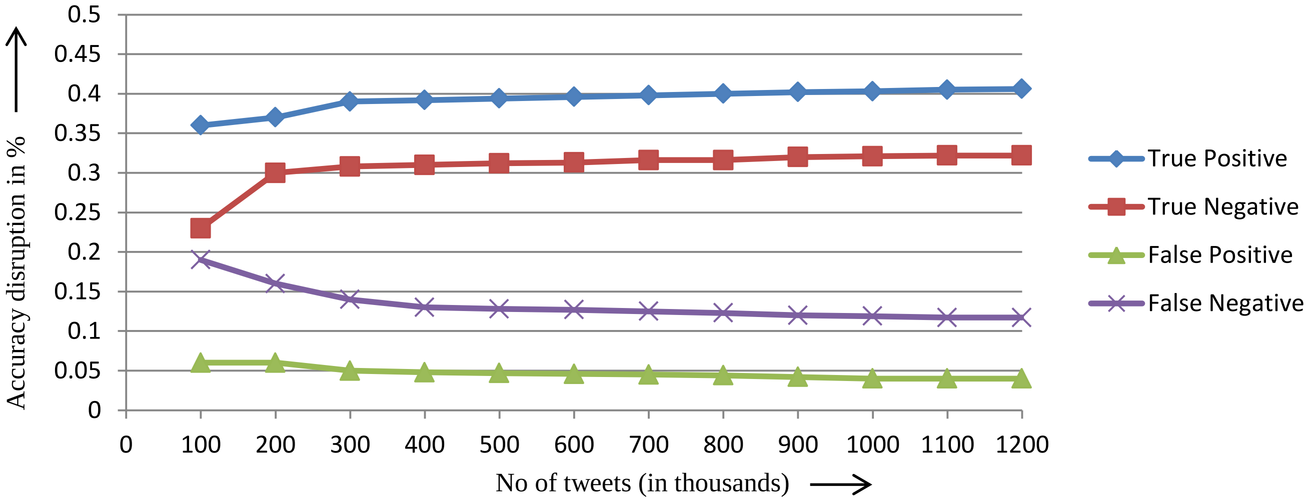

Fig 3. Variation of accuracy with dataset size

Copyright to IJIRCCE DOI: 10.15680/IJIRCCE.2016. 0411215

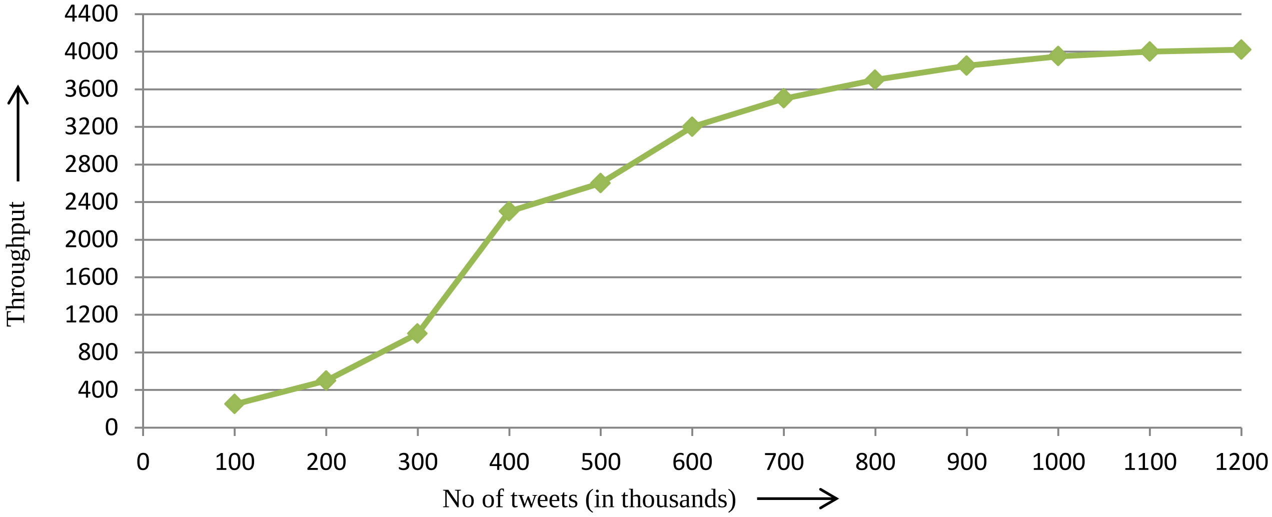

Fig 4. Variation of throughput with dataset size

Table 1 shows the processing time of logistics regression classifier with the increasing size of dataset. At initial stage when the size of dataset is small enough then Hadoop does not benefit from its parallelization concept because the input data is small than the size of one block in Hadoop distributed file system. Once the size of dataset increases the benefit of Hadoop starts to appear in that the processing time for same amount of data is severely reduced compare to small size of data.

VI. CONCLUSION

In this paper we have implemented logistics regression for sentiment mining on large datasets on Hadoop ecosystem. The experimental results show that logistics regression can easily scale up when the size of dataset increases.

REFERENCES

1. 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. P. Barnaghi, J. breslin, P. Ghaffari, “Opinion Mining and Sentiment polarity on Twitter and Correlation Between Events and Sentiment”, Journal of second International Conference on Big data Computing Services and Applications(IEEE), pp.52-57, 2016.
3. S. Kulcu, E.Dogdu, “A Scalable Approach for Sentiment Analysis of Turkish Tweets and Linking Tweets To News”, Journal of Tenth International Conference on Semantic Computing (IEEE), pp.471-476, 2016.
4. M. Bhuvan, V. Rao, S. Jain, R. Guddeti, “Semantic Sentiment Analysis Using Context Specific Grammar”, International Conference on Computing Communication and Automation (IEEE), pp.28-35, 2015.
5. M. Qasem, R. Thulasiram, P. Thulasiram, “Twitter Sentiment Classification Using Machine Learning Techniques for Stock Markets”, Journal of International Conference on Advances in Computing, Communications and Informatics (IEEE), pp.834-840, 2015.
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. 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.
8. A. Go, R. Bhayani, L. Huang, “Twitter Sentiment Classification using Distant Supervision”, Journal of Cloud Computing pp.82-90, 2012**.**

9. J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters", In 6th symposium on operating system design and implementation, Vol- 6, pp.137-149, 2004.
10. 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.
11. 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.
12. Q. Zhou, C. Zhang, "Sentiment Classification of Chinese Reviews in Different Domain: A Comparative Study", Springer, pp.13-23, 2014.
13. J. Dean and S. Ghemawat, "Mapreduce: Simplified Data Processing On Large Clusters", Journal of Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.
14. S. Aravindan, A. Ekbal, "Feature Extraction and Opinion Mining in Online Product Reviews", Journal of Computer Society (IEEE), pp.94-99, 2014.
15. B. Liu, "Sentiment Analysis and Opinion Mining," , A Review Article on Synthesis Lectures on Human Language Technologies, Vol. 5, No-1, pp. 1-167, April- 2012.
16. R. Feldman, "Techniques and Applications for Sentiment Analysis", Journal of Communications of the ACM, Vol. 56, No-4, pp. 82-89, 2013.
17. T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning (Second Edition Springer), 2009.
18. E. Blasch, Y. Chen, G. Chen, D. Shen, and R. Kohler, "Information fusion in of cloud-enabled environment", Journal of High Performance Semantic Cloud Auditing (Springer), 2014.
19. B. Liu, Y. Chen, E. Blasch, K. Pham, D. Shen, and G. Chen, "A holistic cloud-enabled robotics system for real-time video tracking application," International Workshop on Enhanced Cloud Fusion, Sept. 2013.
20. 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.
21. 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.
22. D. Lewis, "Naıve (Bayes) at forty: The independence assumption in information retrieval", Journal of Machine Learning, pp. 4–15, 1998.
23. A. Balahur, M. Turchi, "Comparative experiment Using Supervised Learning and machine Translation for Multilingual Sentiment Analysis", Journal of Computer speech and Language (Elsevier), pp.56-75, 2014.
24. P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss, "Machine learning, Journal of Transactions on Pattern Analysis and Machine Intelligence(IEEE), vol. 29, No. 2-3, pp.103–130, 1997.
25. 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.