I.J. Intelligent Systems and Applications, **2016, 8, 66-70**
Published Online August 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2016.08.08

Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies

Mr. B. Narendra, Mr. K. Uday Sai, Mr. G. Rajesh, Mr. K. Hemanth,
Mr. M. V. Chaitanya Teja, Mr. K. Deva Kumar
Sree Vidyanikethan Engineering College,Tirupathi, Andhra Pradesh
E-mail: bnarendracse@gmail.com, udaysai.karnam@gmail.com, rajesh.gudi33@gmail.com,
hemanth2k16@gmail.com, mvchaitanyateja@gmail.com, kdevakumar@gmail.com

***Abstract*—**Social Networks such as Facebook, Twitter, Linked In etc… are rich in opinion data and thus Sentiment Analysis has gained a great attention due to the abundance of this ever growing opinion data. In this research paper our target set is movie reviews. There are diverge range of mechanisms to express their data which may be either subjective, objective or a mixture of both. Besides the data collected from World Wide Web consists of lot of noisy data. It is very much true that we are going to apply some pre-processing techniques and compare the accuracy using Machine Learning algorithm Naïve Bayes Classifier. With ever growing demand to mine the Big Data the open source software technologies such as Hadoop using map reducing paradigm has gained a lot of pragmatic importance. This paper illustrates a comparitive study of sentiment analysis of movie reviews using Naïve Bayes Classifier and Apache Hadoop in order to calculate the performance of the algorithms and show that Map Reduce paradigm of Apache Hadoop performed better than Naïve Bayes Classifier.

***Index Terms*—**Sentiment Analysis, Machine Learning, Naïve Bayes Algorithm, Apache Hadoop, Map Reduce paradigm.

I. INTRODUCTION

One of the key influencers of human behaviour are opinions. We often depend or take suggestions from others regarding almost every aspect of life including the filed of entertainment. As the usage of internet has gone viral it is very common that users express their feelings, opinions, perceptions in different social networking sites and many private blogs in the form of statuses and reviews. As these reviews or statuses are pretty much generic processing must be done to obtain precise results.Sentiment Analysis plays a vital role in obtaining the feedback from the people for both the newly released products and usability issues of existing products. As these type of feedback mechanism is growing viral it is extended to different fields such as mood prediction of song lyrics, news comments, sports and games, movie reviews etc… It is not only true for individuals but also for government sectors and many private enterprises. Sentiment Analysis helps these organisations immensly in measuring the performance of their products.

Movies are perhaps the best entertainment mankind has got and it is very usual that people watch the movies and express their views and opinions on them by going online either in social networking sites or their own blogs. These type of reviews have a materialistic impact on the movie makers and even on the other people who tend to got to the movie. So, rather than reading the enormous content posted by users we can analyze the textual records by sentiment Analysis and conclude what is the overall impact movie has created in the people.

There are three learning methodologies that are widely available. They are as follows

Supervised Learning: In this type we build the classifier based on the labelled data.

Unsupervised Learning: The learning mechanism is build with out using the unlabelled data.

Semi-Supervised Learning: The Classifier(or)Model is build based on the labelled and unlabelled data.

“Sentiment Analysis” or “Opinion Mining” is the specialized branch of Data Mining Stream which deals with the classification of statuses or textual reviews into positive, negative and neutral as well. In this paper we will be focussing on classification into positive or negative.

There are many ways to perform Sentiment Analysis which mainly boils down to two main methods which are elucidated as follows:

Bag of Words Model(BoW): This model emphasizes more on words rather than focussing on context in which the words are spoken. This method of study contains a sort of “Dictionary” which consists of words that add weight which is reffered to as sentiment in this context. The textual records consists of tokens which have a specific “weight” when mentioned in the context. The sentiment valuation is simply the result of the addition of the weights derived from all the textual records. This model offers no importance to grammar, language essentials and thus lacks in the filed of Human Computer

Interaction(HCI).

Natural Language Processing(NLP): This concept from the field of Artificial Intelligence gives us way to understand the context, string of words and sentence structures. The machine need to understand the grammar principles. Tagging parts of speech, named entities are some of the techniques to perform the task of Natural Language Processing.

Both the process has acquired fair results when applied to the different tasks. But the area of expertise may have an impact on the accuracy with which the task is done.

The "Bag of Words" model requires massive amounts of machine learning concepts to be built in. Algorithms such as Support Vector Machine(SVM), Naïve Bayes Classifier, Maximum Entropy(MaxEnt) recognise patterns and add the weights. It is very common practice to use Bag of Words representation in Natural Language Processing for obtaining better results. The Bag of Words representation is obtained by using Naïve Bayes Classifier and the processing is done in Natural Language Tool Kit(NLTK). This method of processing help us to obtain promising results.

Naïve Bayes Classifier is a machine learning algorithm which belongs to the family of probability classifiers which depends on independence between the tokens or words in the textual records. It is a highly scalable algorithm which reqires many of the parameters to be linear in the problem to be solved. In this paper we explore the performance obtained using Naïve Bayes Classifier.

In section II we present the related work done regarding this field and in Section III we proceed to Experimental Setup. In section IV we interpret the results and conclusion and Future Scope is mentioned in section V.

II. Related Work

Sentiment Analysis has been adressed at different levels of granularity like at the document level, sentence level and many others. In this section we furnish you with the some of the relevant work which has already made its mark in this field of study.

In [1], the authors implemented Naïve Bayes, Perceptron, Rocchio Classifiers on the corpus which consists of 1131 positive and negative instances for both the training and testing data which is collected from the Facebook relating to Languge Learning and showed that the precision of Naïve Bayes Classifier has an edge over the remaining classifiers.In [2], the authors designed a TeachingSenti-Lexicon for teaching evaluation by an automated sentiment orientation polarity definition and applied Naïve Bayes, Support Vector Machine(SVM), ID3 on 175 instances and SVM classifier outperformed remaining classifiers with an accuracy measured over 97%. In [3], the author performed a comparitive study on Tunsian users statuses on facebook during "Arabic Spring" and compared the results of SVM classifier and Naïve Bayes Classifier.In [4], the authors implemented Sentiment Analysis at multiple levels namely as local sentiment which adds up to global sentiment and this technique performs better than the orthodox SVM classifier..In [5], the authors implemented this task on mood classification of lyrics of 185 songs and used SentiWordNet to find out the words which have sentiment and applied the Naïve Bayes Classifier, SVM, K-Nearest Neighbour(KNN) and Naïve Bayes Classifier has an edge over remaining algorithms.

In [6], Rawan T. Khasawneh performed Sentiment Analysis using SentiStrength and SocialMention tool and compared performance.In [7], Jalel Akaichi and his co-scholars compared the accuracy of SVM and Naïve Bayes Classifier using unigram, bigram and Trigram collocation features.In [8], Addlight Mukwazvure and K.P Supreethi performed Opinion Mining using SVM and KNN classifier and showed that SVM performs better than KNN using various metrics.In [9], Ms.K.Mouthami and his team used a Fuzzy Classification Algorithm for Natural Language Processing and parts of speech tags and Bags of Words model using SVM classifier.

In [10]. Sudipto Shankar Dasgupta and his colleagues made the study of different Brand watch outs using Hadoop Map Reducer using R Sentiment package from CRAN and calculated the Accuracy for five different brands. In[11], the authors has proposed a speedy data uploading mechanism using HDFS structure in Apache Hadoop and used a hybrid structure which uses open source software as well as commodity hardware which increases the profits of IT industries.

III. Experimental Setup

In order to perform this task we choose a tool namely Natural Language Tool Kit(NLTK)[12] in which is an exclusive tool provided for the Natural Language Processing in pyhton. We firstly install appropriate version of python and download all the data regarding NLTK. The method of study is represented in the form of following figure.

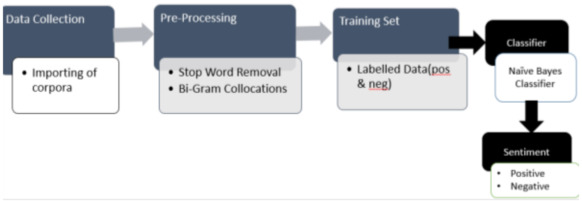

Fig.1. Steps used in Sentiment Analysis

The mathematical formula used in the Naïve Bayes Classifier are as follows:

$$P(\frac{tag}{traits}) = P(tag) * \frac{P\left(\frac{traits}{tag}\right)}{P(traits)}) \qquad (1)$$

$$P(\frac{tag}{traits} = P(tag) * P\left(\frac{t1}{tag}\right) \ldots * P(tn/tag)/P(tags) \qquad (2)$$

Where P(tag) represents the probability whether the instance is assigned a positive or negative tag and P(traits) represents the features present our data set and the rest of the formulae are based on the Bayes rule of conditional probability provided the independence between the traits of our instances are preserved.

The processing steps are elucidated as follows:

Movie reviews posted by users in all the private blogs are collected from the amazon site. In NLTK it comes as in built data in the corpora folder and we can just import it. The corpus consists of 2000 instances which consists of exactly 1000 positive and 1000 negative instances respectively. In this about 75% of the instances are used to built the classifier. The remaining instances are supplied under the name of testing set.

The pre-processing is done inorder to remove the noisy data. The pre-processing techniques used are Stop word Removal[13] which means that removing of words which does not add meaningful content to the dataset. The other technique is feature extraction in which n-gram feature extraction techniques are used. We use bigram feature extraction[14] in which the words that appear consecutively are extracted using this feature extraction.

Nextly we train the classifier using the 1500 positive and negative instances using the Naïve Bayes Classifier algorithm. Later the testing set which consists of 500 instances is implemented on the classifier to predict the sentiment which is either positive or negative. Our Naïve Bayes Classifier takes into account the independence between the words that are present in the dataset. The classification is done into two variants namely positive and negative. The mathematical formulae required to perform this task are done in the NLTK package through python[15]. We explore the results through different attributes and calculate accuracy at different stages of our implementioned which are implemented in the later section.

A Naïve Bayes Classifier that is a probabilistic classifier reflexes the independent assumptions among the features of the instances. The simplicity of the Naïve Bayes Classifier and the accurate results in the previous studies[3]make us to choose this as respective algorithm of our study machine learning algorithm of the widely available classifiers in the field of machine learning.Theoratically Naïve Bayes Classifier have the minimum error rate in comparision to all other classifiers. However, in practice this is not always the case owing to inaccuracies in the assumption made for its use, such as class conditional independence, and the lack of available probability data. So we propose a new insight in to data analytics usingApache Hadoop.

We now examine the process of sentiment analysis done in Apache Hadoop[16]. The Apache Hadoop is an open source frame work written in java which is used to process the large data sets using the mechanism of Map Reduce[17]. The Map Reduce is a parallel, disributed way of processing the results in a cluster. Apache Pig[18] is a data flow language which is return on top of the Map Reduce in a high level language known as Pig Latin. Every Pig operator inturn comes down to a individual Map Reduce job. It in turn supports the functions written in Java, Python and other programming languages. We now perform the task of sentiment Analysis using this Apache Pig.

The following steps delineate our prescribed work:

1. Start
2. Mapping the unstructured data and generate key-value pairs
3. Shuffle and Sort the data
4. Searching the data
5. Reducing the data
6. Stop

The task starts with mapping the unstructured data and there by generating the key-value pairs. Next we perform the task of shuffling the data and later on we sort it followed by searching the positive or negative instances based on the user input and later the Reducer class returns the number of positive or negative instances.

The task starts with mapping the unstructured data and there by generating the key-value pairs. Next we perform the task of shuffling the data and later on we sort it followed by searching the positive or negative instances based on the user input and later the Reducer class returns the number of positive or negative instances.

The simple programming models subsumed in the Apache Hadoop allows distributed processing of data sets across clusters of computers. It offers a flexible and scalable architecture for parallel processing. The library itself has the ability to detect and resolve the problems at the application layer and thus delivering highly-available service on top of a cluster of computers each of which may prone to the failure rate. This method of study help the enterprises in identifying the areas of customer satisfaction or grievance and improved feedback mechanism but the hectic procedure in setting up the environment make it a bit difficult in carrying this type of data analytics.

IV. Results & Analysis

There are many attributes to interpret the results in many forms such as Accuracy, F-Score, Precision, Recall of them which we focus on this metrics which can be defined by following mathematical formula[19].

The resultant class labels in the Sentiment Analysis are Positive and Negative and the following table represents the true positive and negative instances.

Table.1. Contengency Table

		Correct Labels	
		Positive	Negative
Classified Labels	Positive	True Positive(TP)	False Positive(FP)
	Negative	False Negative(FN)	True Negative(TN)

Accuracy:The efficiency of the used framework (Or) model in finding out the positive and negative instances is called accuracy and is represented as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precsion:The exactness of a classifier is measured using precision. The higher precision means less false positives(FP), while a precision means more false positives.

Recall: The sensitivity of a classifier is measured using recall. The false negatives means higher recall where as recall means more false negatives(FN).

F-Measure: The weighted harmonic mean of precision and recall and thus combined to produce single metric and equivalent to accuracy.

In this section we compare the accuracies for the stopword removal and bigram features and also for the raw data and later on we compare the acuracies with the accuracy obtained in the mapreducer in the hadoop using the pig latin code. The results are depicted as follows in the graphical way in Fig. 2.

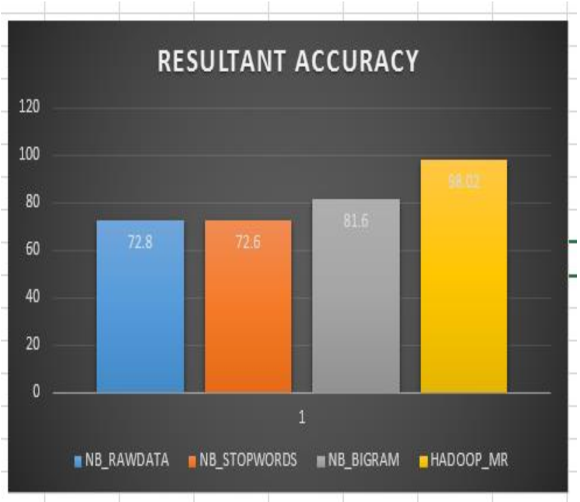

Fig.2. Graphical Method for Comparision of Results

The above graph illustrates the accuracies of four different cases in which our work was carried out. The first block depicts result of Naïve Bayes Classifier performed on raw and unstructured data followed by same classifier applied on the data using stopword removal technique. The third block illustrates the Naïve Bayes Classifier applied on the data using Bigram word features and finally the results obtained by performing the experiment in Hadoop using MapReduce framework are obtained.

The table 2 in which each row represents metrics and columns represents classification technique with preprocessing or algorithm technique used represents the comparision of above metrics for our experimental study.

The accuracy with which the positive and negative instances are classified for the rawdata is 72.8% and by using the stopword removal technique the accuracy obtained is 72.6% and from this we can conclude that stopwords play a materialistic role in the processing of movie reviews and later on moving to the accuracy obtained by the bigram features the accuracy obtained is 81.6%. Finally when we process the data using map and reducer classes in Apache Hadoop the accuracy is slightly higher than 98% which out performs every other method of processing the data. This is even reflexed when we compare it using other metrics such as precision, recall and F-score.

Table 2. Comparision of Metrics for different techniques used in our work.

	NB_ Rawdata	NB_ Stopwords	NB_ Bigram	Hadoop_ MR
Pos Precision	0.65	0.64	0.75	0.82
Pos Recall	0.97	0.98	0.94	0.95
Pos F-measure	0.77	0.77	0.83	0.87
Neg Precision	0.96	0.95	0.92	0.89
Neg Recall	0.49	0.47	0.69	0.76
Neg F-score	0.64	0.62	0.76	0.81

V. Conclusion

Thus we conclude that the best processing of movie reviews data can be obtained when we process it using the MapReducer framework in Apache Hadoop by using Apache Pig rather than using Naïve Bayes Classifier using both stopword removal and Bigram Collocation feature technique. Thus we propose to use the MapReduce framework implemented using Apache Hadoop as a better method of study when compared with machine learning algorithm(which we refer as Naïve Bayes Claasifier in this paper) and we like to address the Neutral class of sentiment Analysis in future.

References

[1] Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro, "Sentiment Analysis of Facebook statuses using Naïve Bayes classifier for language learning", Proceedings ofInformation, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on 10-12 july 2013

[2] Charit Pong-inwong, Wararat Songpan ,"TeachingSenti-Lexicon for Automated Sentiment Polarity Definition in Teaching Evaluation", Proceedings of Semantics, Knowledge and Grids (SKG), 2014 10th International Conference on 27-29 August 2014.

[3] Jalel Akaichi, "Social Networks ' Facebook' Statues Updates Mining for Sentiment Classification", proceedings of SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013

[4] Na Fan, Wandong Cai, Yu Zhao,"Research on the Model of Multiple Levels for Determining Sentiment of Text", Proceedings of 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application.

[5] Vipin Kumar, Sonajharia Minz,"Mood Classification of Lyrics using SentiWordNet" ,Proceedings of 2013 International Conference on Computer Communication

and Informatics(ICCCI-2013), Jan. 04-06, 2013, Coimbatore, INDIA

[6] Rawan T. Khasawneh, Heider A. Wahsheh, Mohammed N. Al-Kabi, Izzat M. Alsmadi, "Sentiment Analysis of Arabic Social Media Content: A Comparative Study, 2013", Proceedings of the 8th International Conference for Internet Technology and Secured Transactions(ICITST-2013)

[7] Jalel Akaichi, zeineb Dhouioui, Maria Jose Lopez-Huertas Perez,"Text Mining Facebook Status Updates for Sentiment Classification, 2013", Proceedings of System Theory, Control and Computing(ICSTCC), 2013 17th International Conference

[8] Addlight Mukwazvure, K.P Supreethi,"A Hybrid Approach to Sentiment Analysis of News Comments, 2015" ,Reliability, Infocom Technologies and Optimization(ICRITO), 2-4 Sept. 2015.

[9] K. Mouthami, K. Nirmala Devi, V. Murali Bhaskaran," Sentiment Analysis and Classification Based On Textual Reviews, 2014", Proceedings of Information Communication and Embedded Systems (ICICIES), 21-23 Feb 2013.

[10] Sudipto Shankar Dasgupta, Swaminathan Natarajan, Kiran Kumar Kaipa, Sujay Kumar Bhattacherjee, Arun Viswanathan,"Sentiment Analysis of Facebook Data using Hadoop based Open Source Technologies, 2015" Proceedings of Data Science and Advanced Analytics(DSAA),2015 19-21Oct.2015.

[11] Gaurav D Rajurkar, Rajeshwari M Goudar, "A speedy data uploading approach for Twitter Trend And Sentiment Analysis using Hadoop", Proceedings 2015 International Conference on Computing Communication Control and Automation.

[12] Mykhailo Lobur, AnDriy Romanyuk, Mariana Romanyshyn, "Using NLTK for educational and scientific purposes", Proceedings of CAD Systems in Microelectronics(CADSM), 2011 11th International Conference.

[13] http://www.nltk.org/book/ch02.html

[14] http://www.nltk.org/api/nltk.html

[15] http://www.nltk.org/book/ch01.html

[16] Shankar Gznesh Manikandan, Siddharth Ravi, "Big Data Analysis Using Apache Hadoop", Proceedings of IT Convergence and Security (ICITCS), 2014 International Conference on 28-30 Oct. 2014

[17] Milind Bhandarkar, "MapReduce Programming with apache Hadoop" ,Proceedings of Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on 19-23 April 2010.

[18] https://pig.apache.org/docs/r0.7.0/tutorial.html

[19] http://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/

[20] P. Dhana Lakshmi, K. Ramani, B. Eswara Reddy, "Feature Relevance Analysis in Online Marketing To Improve Productivity", proceedings journal of Software Engineering, Volume 9.No.3, jan-mar 15, ISSNO 973-5151.

Authors' Profiles

Mr. B. Narendra, working as Assistant Professor ,in Department of CSSE, at Sree Vidyanikethan Engineering College, Tirupathi, Andhra Pradesh, India. His interested areas include data bases, Data Mining, big data and pattern recognition. He published research papers in these areas..

K. Uday Sai, studying IV B. Tech in Department of CSSE, Sree Vidyanikethan Engineering College and Team Lead for the project entitled as above.

G. Rajesh, Studying in IV B.Tech in same department and interested in the field of Data Mining and Data Warehousing.

K. Hemanth, Studying in IV B.Tech in Department of CSSE, SVEC. His interested areas include data mining and data warehousing.

M.V.Chaitanya Teja, Studying IV B.Tech and have a penchant to solve the data storage problems in real world.

K. Deva Kumar, Studying IV year of B.Tech and very interested in the topic of Sentiment Analysis.

How to cite this paper: B. Narendra, K. Uday Sai, G. Rajesh, K. Hemanth, M. V. Chaitanya Teja, K. Deva Kumar, "Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies", International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.8, pp.66-70, 2016. DOI: 10.5815/ijisa.2016.08.08