Large-scale Sentiment Analysis Using Hadoop

Dasari Prasad[1], G.N.V.G. Sirisha[2], G. Mahesh[3], G.V. Padma Raju[4]
P.G Student, CSE Department, SRKR Engineering College, Bhimavaram, India [1]
Assistant Professor, CSE Department, SRKR Engineering College, Bhimavaram, India [2]
Associate Professor, CSE Department, SRKR Engineering College, Bhimavaram, India [3]
Professor, CSE Department, SRKR Engineering College, Bhimavaram, India [4]
Email: {prasad.dasari1126[1], sirishagadiraju[2], gadirajumahesh[3], gvpadmaraju[4]}@gmail.com

Abstract: Sentiment analysis involves the usage of text analytics to identify and categorize the polarity of opinions expressed in a piece of text. Sentiment analysis analyzes the intension of a customer from a given feedback text. Supervised machine learning techniques are one of the popular methods for sentiment analysis. The accuracy of the algorithms increases with the increase in size of training data. Large volumes of user reviews are available online, to leverage them Hadoop-based Sentiment Analysis system is proposed in this paper. The proposed system applies Naïve Bayesian Classifier for detecting the polarity of users' opinions. The system achieved 94% accuracy and 86% accuracy when tested on two datasets namely product review dataset and movie review dataset. These accuracies even without applying pre-processing steps like Parts of Speech tagging.

Keywords: sentiment analysis, Hadoop, supervised machine learning, opinions, text analytics, naïve Bayesian

I. INTRODUCTION

Sentiment analysis helps in identifying the writer's attitude towards an individual, organization, event, product or topic is positive, negative or neutral. World Wide Web has become a part of everyone's life. More and more people are using www for a number of tasks like information retrieval, e-commerce, social networking etc. it has enabled companies, organizations, political parties to easily reach their customers and citizens through online advertising, online campaigning. People are also using blogs, social networking sites to share their opinions with friends, family and society at large.

Compared to traditional media, broadcasting and narrowcasting are much easier and cost effective with World Wide Web. This is only one side of the coin; the other side of coin is that the fake news, negative opinions are also spreading at a faster rate through www. People are sharing their dislikes, dissatisfaction, and anger about a product, event, individual or party through blog posts, reviews and social media. If actions like immediate dialogue with dissatisfied customers,

Compensation to product errors are taken the spread of negative sentiment can be reduced. Organizations and governments are resorting to find the success of product/scheme by analysing customers or citizen's response in social media networks, reviews, and tweets etc. so, Analysing the sentiment polarity of a piece of text has become the need of the day. From data granularity point of View, there are three levels of sentiment analysis; document level, sentence level and aspect (feature) level. There is not much difference between document level and sentence level sentiment analysis as sentences are nothing but short documents as in [1]. Sometimes, the writer expresses different views about different aspects (features) of same product.

Human beings are social beings and it is natural for us to ask the opinion of others before we make any choice. Earlier generally people used to take the opinions of friends, family but with the availability large volumes of opinion data online in the form of reviews, ratings etc. we now rely on them. So, it is also necessary to automatically identify the polarity of large review datasets and consolidate them so that it becomes easy for the users to use them.

A number of tools and techniques were developed in the area of sentiment analysis over the past decade. Some of the examples for tools are face book insights, red opal etc. as in [3]. Sentiment analysis methods are broadly classified into machine learning methods, lexicon based methods and hybrid methods as in [2]. Machine learning methods in turn classified as supervised and unsupervised method. Lexicon based approaches are classified as dictionary based approach and corpus based approaches. Hybrid approaches uses both machine learning and lexicon based method. Among these supervised machine learning methods are most frequently used owing to their natural inclination in solving such problems. Machine learning approaches are more accurate compared to lexicon-based approaches but they are time consuming as in [3]. Lexicon based approaches are said work well when there is clear difference between positive and negative opinion texts.

Supervised classification algorithms work in two steps. The first step is training and validation phase. The second step is usage of the model for classifying unseen document/sentence.

The set of opinion texts whose polarity is already is known is required to build the classifier. This set is in turn divided into training set and validation set. During training phase, training set is given as input to the algorithm and a classifier is built. During the validation phase the accuracy of the classifier is tested using validation set. If the accuracy of the classifier is good, it is used to classify unseen opinion texts. Many classification algorithms like decision trees, support vector machine, neural networks, and rule based classifier, naïve Bayesian classifier, Bayesian networks etc. are used. Among these Naïve Bayesian Classifier is efficient and simple to use.

Success of any classifier depends on the size and quality of training data. The larger and better the training data the better the accuracy of classifier. Large volumes of review data, tweets, and comments are easily available from e-commerce and social networking sites. But processing such large volumes of data for sentiment analysis is either time consuming or sometimes may not possible on a single machine. Since Naïve Bayesian Classifier can be applied in a distributed manner, the Hadoop ecosystem which can handle large volumes of data using distributed file system is used to analyze large volumes of data in this paper. Map reduce programming model was used to implement the Naïve Bayesian Classification algorithm.

The rest of the paper is organized as follows. Section 2 discusses the literature survey on sentiment analysis. Section three discusses the architecture and detailed methodology. Section four discusses the datasets. Section five presents the results. Section six concludes the paper.

II LITERATURE REVIEW

This section gives a brief review of different approaches and algorithms for sentiment analysis.

Ankur Goel et al. has proposed a framework which uses Naïve Bayes algorithm along with SentiWordNet to classify tweets as in [11]. The positivity, negativity and objectivity of scores of words are found. The dataset used is Sentiment140 dataset which contains 16 million tweets. The accuracy of the system is found to be 58% when only Naïve Bayesian classifier is used. SentiWordNet score of each preprocessed word in the tweet is added to the prior probability and posterior probability when calculating class conditional probabilities. The authors claim that usage of Naïve Bayesian Classifier together with SentiWordNet improves the accuracy of the system.

Deebha Mumtaz et al. have proposed a senti-lexical algorithm to find polarity of movie reviews as in [3]. Sentiment keywords, emoticons and negation words along with preprocessed review data are given as input to the algorithm. They have applied the algorithm on 300 tweets and the accuracy of the algorithm is found to be 70%.

Xing Fang et al., have proposed a general process for sentiment polarity categorization as in [4]. The proposed algorithm is applied on 5.1 million Amazon product review dataset. The sentences that contain at least one positive or negative word are extracted and tokenized. An algorithm was proposed for negative phrase identification, sentiment score computation. Based on the words and phrases identified, a feature vector is constructed. Sentence level and review-level polarity categorization methods are proposed.

Pallavi Sharma et al. have proposed feature level sentiment analysis for movie review dataset as in [5]. The sentiment scores of the reviews are calculated using SentiWordNet. When finding sentiment score at feature level certain problems like negation, intensifier, synonyms occur. The authors have proposed methods to handle them. The sentiment score is calculated using SentiWordNet 3.0. The accuracy obtained is 81%.

Mouthami et al. have proposed fuzzy classification algorithm with parts of speech tags to classify Movie Review data as in [6]. They applied the algorithm for document level sentiment classification.

Chaitali et al. have proposed an algorithm for aspect level sentiment mining based on aggregate score of opinion words and aspect table as in [7]. The authors emphasize on using identification and usage of implicit reviews along with explicit reviews to improve accuracy of the system. The proposed is tested on restaurant dataset of a single restaurant from Zomato website. Using the system the authors are able to identify among the aspects food, staff, service, ambience and rate which aspects have positive sentiment and which aspects have neutral or negative sentiment.

Minara Panto et al., have proposed a framework for obtaining automatic rating of products from Twitter through opinion mining as in [8]. Data is collected using Twitter4J API, it is preprocessed, POS tagged and then classified in to opinion sentiment words using SVM and Stanford dictionary. Opinion sentiment words frequency is found using unigram approach. These frequencies are used to find product ratings for products.

Khaled Ahmed et al., has presented a survey about sentiment analysis methods, available APIs and tools as in [9].

Purtata Bhoir et al., have proposed a method to aspect level sentiment analysis as in [10]. The proposed framework is applied on movie review data. The data is initially preprocessed and POS tagged. The sentences in the review are classified as objective and subjective sentences using Naïve Bayesian Classifier and SentiWordNet. Subjective sentences are sentences that contains opinions expressed in them. Initially the classifier is trained with a set of 5000 objective and 5000 subjective sentences. Then from the extracted subjective sentences, feature-opinion word mapping is done. These feature-opinion pairs are then used to find sentiment at aspect level. As a final phase, for the different aspects like songs, story etc. summarized polarity is generated.

Further, an extensive discussion about sentiment analysis algorithms and applications could be found in [2]

III ARCHITECTURE AND METHODOLOGY

The proposed system uses Naïve Bayesian classifier to classify the polarity of reviews. The system consists of three phases which are preprocessing phase, training phase and testing phase. As large and large review datasets are available online and as the size of the user reviews generated in a site is growing from time to time, to handle them efficiently, the proposed system is implemented using Map reduce framework and Hadoop. The three phases of the system working is described as follows

1). Preprocessing Phase

The reviews are tokenized. Special symbols and stop words are removed.

2). Training Phase

The system is trained using product reviews and movie review datasets separately where 2/3 data is used for training and 1/3 is used for testing.

3). Testing

The accuracy of the proposed system is tested for dataset of different sizes like 415MB, 2.6GB and 6GB and the test results are shown in Section V.

The architecture of the proposed system is shown in figure2

All the phases like preprocessing, training and testing are implemented using Hadoop Ecosystem. Hadoop 2.7.2 is used for the implementation. Hadoop ecosystem is as follows

A. Hadoop Framework

Hadoop is an open source software framework which provides distributed computing and distributed storage. It uses Hadoop distributed file system which is a distributed memory formed by combining the available memory of all data nodes with in the cluster. It runs on the large clusters of commodity hardware. Commodity hardware is the system having 2GB RAM. Hadoop cluster contains one master and any number of slaves. In our Project we have implemented 8 node cluster with hadoop. Hadoop cluster is as shown in figure 1

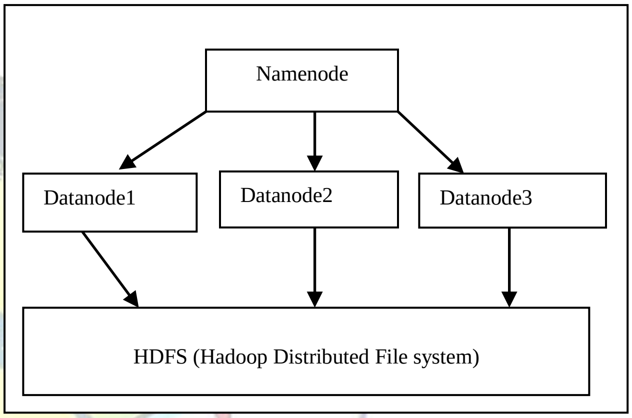


Fig 1: Hadoop Cluster

5 daemons run on master and slaves as

1). Namenode- It is a master daemon which keeps track of all data nodes with in the cluster by getting heart beat signal from every data node. Namenode maintains metadata corresponding to previous log record and file system image.

2). Datanode- It is a slave daemon which process the task given by the master in mapreduce manner and it contributes the available memory for formation of HDFS for distributed storage.

3). Secondary Namenode- This daemon can be reside in the master or as a separate node based on how we configure. It maintains metadata from Namenode corresponding to recent log record. If namenode fails, using the metadata secondary namenode can serve in place of master.

4). Node manager- It is a slave daemon which runs on the datanode and which keeps track of space utilization and available space with in the data node.

5). Resource manager- It is a master daemon which keeps track of amount of memory used and available memory by asking node manager of a datanode with in the cluster of all data nodes.

This paper uses Naïve Bayesian Classifier which is a machine learning algorithm for sentiment classification. Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem.

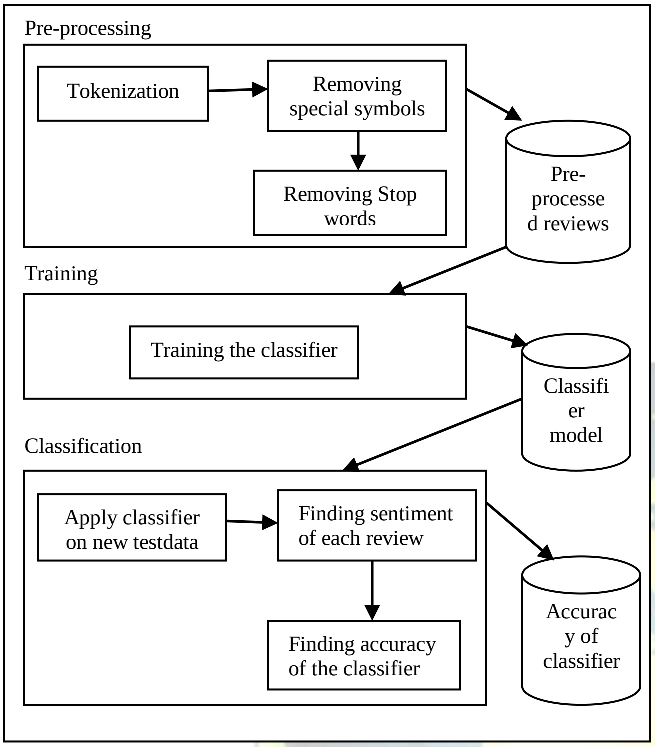
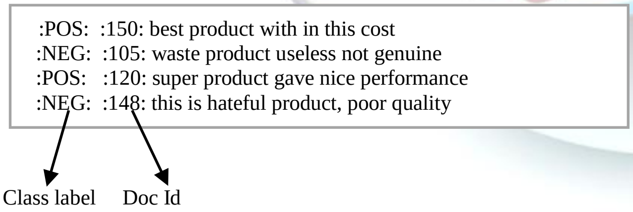


Fig 2: Architecture of Large-scale sentiment analysis using Hadoop

IV DATASETS DESCRIPTION

Two datasets namely product review dataset and movie review dataset are used in this paper. Product review data originally had 1300 reviews. This dataset is generated by us which has the format shown in figure 3.

Sample Product Reviews Format:

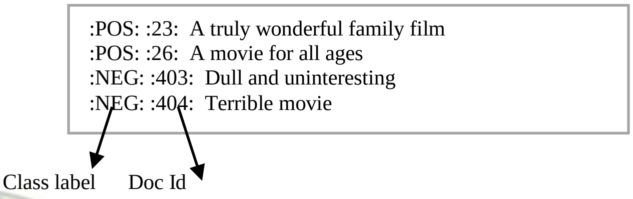


:POS: :150: best product with in this cost
:NEG: :105: waste product useless not genuine
:POS: :120: super product gave nice performance
:NEG: :148: this is hateful product, poor quality

Fig 3: Sample product reviews

From the data shown above, each review clearly has words that state whether the review has positive or negative polarity. So, the accuracy of the classifier will be definitely high for this dataset. This data is replicated to generate a huge dataset. Initially it is replicated to generate 415MB, 2.6GB and 6GB data. The dataset having 415MB size contains 6.5 million positive reviews and 6.6 million negative reviews for a total of 13.2 million reviews. The dataset having 2.6GB contains 20.9 million positive reviews and 36.2 million negative reviews for a total of 57.2 million reviews. The dataset having 6GB contains 48.2 million positive reviews and 83.7 million negative reviews for a total of 132 million reviews. The reason for data replication is to test the efficiency of the algorithm.

Sample Movie Reviews Format**:**



:POS: :23: A truly wonderful family film
:POS: :26: A movie for all ages
:NEG: :403: Dull and uninteresting
:NEG: :404: Terrible movie

Fig 4: Sample Movie Review Data

Movie review data is collected from Amazon Movie Review dataset [12]. It contained 8 Million reviews. For each review, it contains details like product ID, user Id, profile Name, helpfulness, score, review time, review summary, review text. Out of these details review summary is extracted using python program. The data size after transformation is 415MB. As it is a standard dataset, accuracy of our system on this dataset is reliable. This data is later replicated to generate 2.6GB and 6GB data sizes. The dataset having 415MB contains 1.3 million positive reviews and 8 million negative reviews for a total of 9.4 million reviews. The dataset having 2.6GB contains 8.4 million positive reviews and 49.3 million negative reviews for a total of 57.7 million reviews. The dataset having 6GB contains 19.5 million positive

V. RESULTS

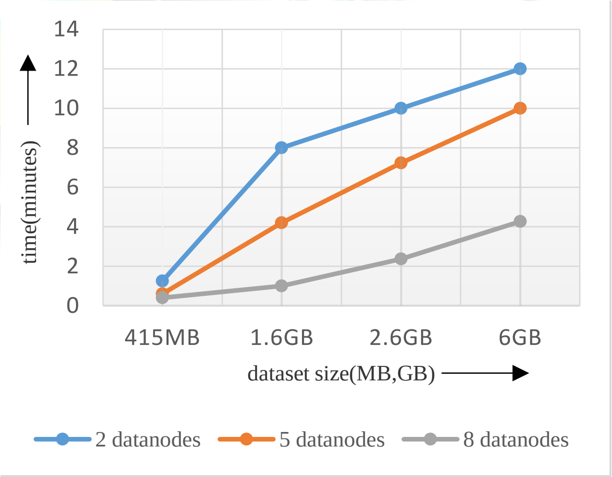


Fig 5: Time taken for different data sizes on computing cluster with 2 nodes, 5 nodes and 8 nodes respectively.

Figure 5 shows that by increasing the number of nodes from two to eight the time required to detect the polarity of sentiments has reduced from 12 minutes to 4 minutes. For datasets of terabytes size, which are impossible to be processed on a single computing node, Hadoop implementation is an attractive alternative.

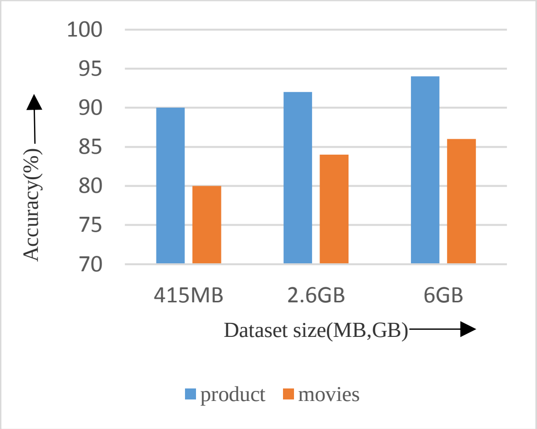


Fig 6: Accuracy of the proposed system for product and movie review Datasets of different sizes

Figure 6 shows the accuracy of the proposed system for product review dataset and movie review dataset. Product review dataset is self-generated data which contained words/tokens which clearly define the polarity hence naturally the accuracy of proposed system is 90% or above for this dataset. For movie review dataset the accuracy is 80% for size 415MB. It is 84% for 2.6GB data and 86% for 6GB data. These results are in par with the results discussed in state-of-the-art papers on sentiment analysis.

VI. CONCLUSION

Sentiment analysis is automatically analyzing and classifying large volumes of users' opinions. By finding the polarity of user's reviews, sentiment analysis helps in consolidating the reviews at document level, sentence level and feature level. The accuracy of sentiment analysis increases with the availability of large volumes of training data. Implementations like Hadoop based sentiment analysis as proposed in this paper have become the need of the day with the availability of large volumes of online reviews.

The proposed system uses Naïve Bayesian Classifier for identifying the polarity of user reviews and Hadoop for handling large volumes of user reviews. The system is tested on two datasets namely product review data set and movie review dataset. The accuracy of the system is 94% for product reviews and 86% for movie review datasets of 6GB size. The time taken by the system is 4 minutes for 6GB size data when the number of computing nodes in the cluster is 8. By increasing the computing nodes to few tens of nodes the time required by the system could be drastically reduced.

REFERENCES

[1] Liu B., "Sentiment Analysis and Opinion Mining", Synth Lect HumaLang Technol, 2012.

[2] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, 2014, doi:10.1016/j.asej.2014.04.011.

[3] Deebha Mumtaz, Bindiya Ahuja, "Sentiment Analysis of Movie Review Data Using Senti-Lexicon Algorithm", Proc. IEEE International Conference on Applied and Theoretical Computing and Communication Technology, 2016, doi:10.1109/ICATCCT.2016.7912069.

[4] Xing Fang and Justin Zhan, "Sentiment Analysis Using Product Review Data", Journal of Big Data 2015, doi: 10.1186/s40537-015-0015-2.

[5] Pallavi sharma, Nidhi Mishra, "Feature Level Senti ment Analysis on Movie Reviews", Proc. IEEE Intl. Conf. Next Generation Computing Technologies, 2016, doi:10.1109/NGCT.2016.7877432.

[6] K.Mouthami, K.NirmalaDevi, V.Murali Bhaskaran, "Sentiment Analysis and Classification Based on Textual Reviews", Proc. IEEE International Conference on Information Communication and Embedded Systems, 2013, doi:10.1109/ICICES.2013.6508366.

[7] Chaitali Chandankhede, Pratik Devle, Abhijit Waskar et al., "ISAR:Implicit Sentiment Analysis of User Reviews", Proc. IEEE International Conference on Computing, Analytics and Security Trends, 2016, doi:10.1109/CAST.2016.7914994.

[8] Minara Panto, Nivya Johny, Mejo Antony et al., "Product Rating Using Sentiment Analysis", Proc. IEEE International Conference on Electrical, Electronics and Optimization Techniques, 2016, doi:10.1109/ICEEOT.2016.7755346

[9] Khaled Ahmed, Neamat El Tazi, Ahmad Hany Hossny, "Sentiment Analysis Over Social Networks: An Overview", Proc. IEEE International Conference on Systems, Man, and Cybernetics 2015, doi:10.1109/SMC.2015.380

[10] Purtata Bhoir, Shilpa Kolte, "Sentiment Analysis of Movie Reviews Using Lexicon Approach", Proc. IEEE International Conference on Computational Intelligence and Computing Research, 2015, doi:10.1109/ICCIC.2015.7435796

[11] Ankur Goel, Jyoti Gautam, Sitesh Kumar, "Real Time Sentiment Analysis of Tweets Using Naïve Bayes", Proc. IEEE Intl. Conf. Next Generation Computing Technologies, 2016, doi:10.1109/NGCT.2016.7877424

[12] Webdata: Amazon movie reviews, https://snap.stanford.edu/data/web-Movies.html