TEST Engineering & Management

May - June 2020
ISSN: 0193-4120 Page No. 4950-4954

Machine Learning Based Sentiment Analysis of Distributed Customer Product Reviews Data on Amazon

[*1]Varsha V, [2]Akram Pasha

[1]PG Student, [2]Professor, [1,2]School of C and IT, REVA University, Bangalore, India
[1]varsha.vk777@gmail.com, [2]akrampasha@reva.edu.in

Article Info
Volume 83
Page Number: 4950-4954
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Abstract

Recognizing and Analysis of textual data generated from various social media platforms has become one of the most essential requirements in today's Big Data era. The result of such analysis helps in many crucial businesses to gain clear insights about their business models and to eventually take crucial business-oriented decisions to improve their businesses. In this paper, an attempt is made to perform sentiment analysis on the distributed computing framework using the many Machine Learning (ML) models and Hadoop-based Spark programming model. The existing approaches towards sentiment analysis are limited to only a few brands and their products. Therefore, to integrate the learning abilities with distributed computing models on large textual data, we developed the recommendation framework that recommends the product to users according to user's feature requirements collected as the huge textual data. The study implemented the Gaussian Naive Bayes (GNB) and Random Forest (RF) on the Spark Big Data analytics platform to process huge textual data. The experimental results have shown that the two algorithms produce superior efficiency over other methods while processing big sentiment datasets.

Keywords: *Big Data; Sentiment Analysis; Machine learning; Apache Spark; ML Pipeline; GNB; RF*

1. Introduction

The large data produced by social media consists of big data features like variety, velocity and volume hence they require machine learning and big data tools for sentiment or textual analysis. The data sets that are large and very complex in nature are known as big data [1]. Big Data is defined as a huge amount of data which is both structured and unstructured, the big data can be of any form/ anywhere like data from social media, videos, news, issues. This set of data is known as big data which can be further used for the sentiment analysis [2- 4]. Sentiment Analysis is defined as the task of finding the individual opinions on the particular structure [5]. Sentiment Analysis is basically the real tone of individuals towards a particular product, structure, events, issues and their attributes [4].

Companies mainly make use of sentiment analysis for decision making of their brands in business [6]. Machine Learning is a subset of artificial intelligence [7]. It helps system to automatically master and improve from past experience with human interference [8]. Machine learning computations aim to derive predictive models from models from the existing and experience data [9]. Machine Learning is broadly grouped into (Classification, Regression, Clustering, Anomaly detection, Recommenda -tion, and Dimensionality reduction) [10]. The sentiment approaches have been extensively employed in many applications with respect to big data and helps in different positive and negative predictions of movies, reviews [11]. Rapid growth in the amount of web-sentimental rich social media has resulted in increased interest among researchers in sentimental analysis and opinion mining.

May - June 2020
ISSN: 0193-4120 Page No. 4950-4954

Rapid growth in the amount of web- sentimental rich social media has resulted in increased interest among researchers in sentimental analysis and opinion mining. With so much social media on the internet, however, Sentiment Analysis is now considered a Big Data activity. The research's main goal was to find such a technique that could effectively perform Sentiment Analysis on Big Data sets. It offers very good efficiency in managing large data sets of sentiments [12]. The major aim of this study is the problem to identify large-scale sentiment data as Positive sentiment [True Positive, False Positive] or Negative sentiment [True Negative, False Negative]. To solve this issue, an application[Amazon] is used to collect and evaluate user feedback of a particular product. It will segregate the comments into both positive and negative assessments. The negative reviews would be useful to businesses in further developing their product based on input from the customer. The application also sets out the pros and cons of the product's individual function. The application would also include updates on the sentiment analysis carried out on the goods. The study's primary objective is to develop the recommendation framework that recommends the product to users according to the user's feature requirements. GNB and RF machine learning algorithms are used to classify feelings, a feature vector generation method is provided for sentiment polarity categorization, and two sentiment polarity categorization experiments are performed on the basis of sentence level and review level, and the output of three classification models is evaluated and compared on the basis of their experiment results.

The succeeding sections of this work shall be defined as follows. Section 3 provides a description of the relevant research. Section 4 sets out the methods used in this work. The experimental setup and results are discussed in Section 5. Section 6 sets out the conclusion and possible development of this study.

2. Related Work

The literature review is focused on collecting the most important research in the area of Big Data Sentiment Analysis. Sentiment analysis has become an evolving field of technology and study in which several scholars have carried out work to offer such innovations or valuable methods that can be further applied in order to provide full assistance in emergency medical situations. A lot of work has already been performed by a scientific researcher who can identify or forecast real-time phenomena using both supervised and unsupervised learning algorithms.

A. Mining big data in real time

The real-time analysis of big data is very relevant. The k-means clustering process for sentiment analysis is carried out in the published work [1]. For such a work, the same volume of data will be generated every two days. The approach used in this research was low cost, but as the data produced in real time would begin to expand, it presented several challenges.

B. Big data sentiment analysis using Hadoop

Hadoop is an open source program for storing data and running applications on commodity hardware clusters. Hadoop map reduce methodology has been used for emotional research in this published work [13]. The research was carried out using Hadoop on large data sets of tweets and the efficiency of the procedure was calculated in speed and precision. Through sentimental research, the development of Hadoop maps provided very good productivity through managing massive data sets of sentiments. Nevertheless, it was less than sixty percentage accuracy.

C. Sentiment analysis using SVM

SVM is the most widely used algorithm for sentiment analysis among machine learning algorithms. In the published work [14] it explains the experimental findings used by SVM on the benchmark datasets to train the sentiment classifier. The most classical characteristics were derived using N-grams and various measuring schemes. It also used the chi-weight feature in order to pick the informative features for classification. Chi-square attribute collection greatly increased the precision of the classification. The only demerit of this method was that it took a lot of time to train the device.

D. Sentiment analysis using product review data

Analyzing client preferences results in market analysis and assists in taking choices on their products. In the work published [15] the methodology known as natural language processing (NLP) has been implemented. The key goal was to tackle the question of emotion polarity categorization, which was the basic problem of sentiment analysis. The precision was good, up to eighty per cent compared to Hadoop sentiment analysis. Yet the difficulty of time was greater than other methods.

Compared to all the above methods the proposed system experimental results states that it can identify large-scale sentiment data as Positive sentiment [True Positive, False Positive] or Negative sentiment [True Negative, False Negative].The study's primary objective is to develop the recommendation framework that recommends the product to users according to the user's feature requirements.

3. Methodology

The model used in this analysis is seen in the most general context in Fig.1. The major components implemented in this system are discussed in the following part of this section.

May - June 2020
ISSN: 0193-4120 Page No. 4950-4954

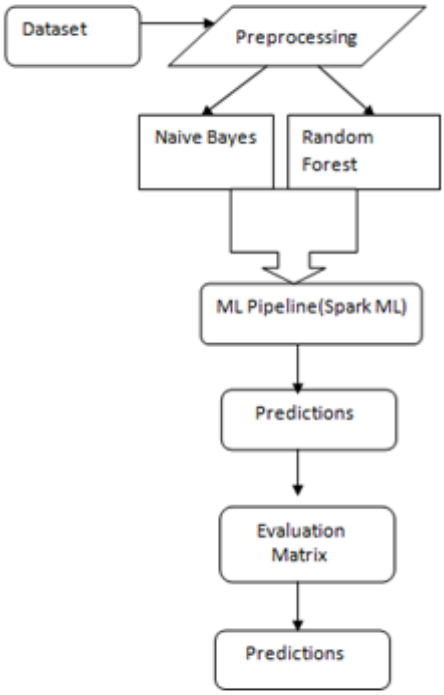

Figure 1: Proposed model for distributed big data sentiment analysis using ML

A. Dataset

The data set is already usable and ready for use by supervised machine learning algorithms in the field of sentiment analysis. The data set contains reviews from three different areas, namely cinema reviews, amazon product reviews and hotel reviews. A favorable review is identified by 1.0 while a negative review is identified by 0.0. This method is only useful for amazon product reviews through amazon.com.

B. Natural Language Processing(NLP)

In this work they have used 3 techniques from natural language processing and they are

1) Feature: A feature indicates an observation property or is also known as a variable. Given the tabular set, the row indicates the observation and the column indicates the feature. For example, consider a tabular set containing the student details age, gender, class, srn is called the features, while each student detail is known as observation.

2) Feature Extraction: Term frequency-inverse document frequency (TF-IDF): it is a function vectorization tool that is used mainly in text mining to know the meaning of the word to a document in quantity. In the MLlib, TF and IDF are divided to make more flexible.

Bag-of-words: Text recognition that describes the presence of words within a document is known as bag-of-words. Every word is described as a function since machine learning algorithms cannot work with raw text data.

3) Feature Transformers: Tokenization: is a converter that automatically transforms the given input string to a lowercase, and by using whitespaces, splashes the string into characters, i.e. splashes the input string into a sequence of characters. Stop Words Remover: This function is very helpful because stop words are terms that are not helpful and frequently used in a sentence, and certain kinds of terms can be omitted. The stop word remover uses the tokenizer output string as an input string and eliminates all stop words from the input string.

C. ML pipelines

Machine learning is used to perform a pseudo code sequence to generate and learn from the results. For example, a processing of input text data may require several stages within it. Using the feature vector and symbols, translating the input text to terms, terms into a graphical feature vector and eventually learning the prediction model. MLlib describes such a workflow as a pipeline composed of a series of pipeline stages within .ML Pipelines offer a standard range of high-level APIs built on top of the Data Frames that help users build and balance realistic machine learning pipelines[13].

D. Gaussian Naive Bayes Classification

It is a subclass of a supervised learning system. GNB is primarily derived from the Bayes Theorem built on the use of n number of steps and the estimation of a solution using previous experience. The specification of the big data platform spark. ml follows both the multinon Naïve Bayes and Bernoulli Naive Bayes. This GNB is often used to distinguish positive and negative sentiment in emotion analysis. This GNB is very simple and fast to train, and they are used in real time and are often used in various recommendation systems to provide valuable input on material to individuals.

E. Random Forest Classification

One of the supervised learning algorithms is RF. It is used for both classification and regression which consists of multiple decision trees in it. They can manage the different missing values in the data and retain the vast proportion of data. The RF works primarily on the bagging process.

F. Implementation using Spark

This section will explain the implementation of sentiment analysis on the named text data set using ML Algorithms and Libraries such as Classification Algorithms (GNB and RF) and the Spark ML package, the code will be written in the Python Programming Language. First, the data set will be loaded as input to the device and the data will be saved. Upon loading the data set into the program, before implementing and testing the templates, the data set is arbitrarily divided into 80 percent train data set and 20 percent check data set. Then the pre processed steps

May - June 2020
ISSN: 0193-4120 Page No. 4950-4954

begin on the training data set, using certain NLP principles for sentiment analysis such as Feature Extractors and Feature Transformers. Through here, approximately the code used by both algorithms will remain the same till some step and one of the case model used will vary between the GNB and RF. Then all of the next steps for both algorithms remain exactly the same until the conclusion of this process. Therefore only one written method for both algorithms will be provided here and implemented using the ML Pipeline (SparkML).

G. Experimental steps

1) Data set: The data set is already usable and ready for use by supervised machine learning algorithms in the field of sentiment analysis. The data set contains reviews from three different areas, namely cinema reviews, Amazon product reviews and hotel reviews. A favorable review is identified by 1.0 while a negative review is identified by 0.0. This method is only useful for Amazon product reviews through amazon.com. The sample reviews consists of 407 positive and 401 negative reviews. Overall the system consists of 808 trained data sets. Table 1 demonstrates the analysis of the Amazon sample reviews with its label.

Table 1: Amazon Sample Reviews
(FIRST 10 ROWS)

Label	Text
0.0	So there is no way for me to plug it in here in the US unless I go by a converter
1.0	Good case Excellent value
1.0	Great for the Jawbone
0.0	Tied to charger for the conversations lasting more than 45 minutes MAJOR PROBLEMS!!
1.0	The mic is great
0.0	I have to jiggle the plug to get it to line up right to get decent volume
0.0	If you are Razr owner you must have this
1.0	Needless to say I wasted my money
1.0	The sound quality is great
0.0	Very good quality though

2) Train and Test Data: The data set is primarily split into two sections, which are train data and testing data to prevent error creation. The provided input data set is initially divided into 80 percent train data set and 20 percent randomly tested dataset after loading the datasets into the program before progressing to the further stages of implementing and evaluating models. The test data is often independent of the learned data so identifying measures such as precision is useful for testing the consistency of the training models and the cross validation is performed to split the given dataset.

3) Pre-processed Data: The input data set will be loaded into the device by an 80 percent train and a 20 percent random check. Later, using the natural language processing methods described above, the attributes are extracted into the lowercase and the words are split into white spaces. Stop words will be removed with the use of the stop word remover and the data set will still not be in the form. Therefore the TF-IDF is used to get a bag of words. Then the Sentimental analysis algorithms were implemented using a bag of words. While, the ML pipeline (SparkML) has been optimized.

4. Experimental Setup and Discussion of Results

The experimental setup has been constructed into 2 different minimum requirements - software and hardware requirements. The software requirements consists of Windows 7 Operating System (OS), python 3.0 coding language, Spark 2.0 big data tool and Anaconda environment. The hardware requirements consists of 15”Led monitor with Intel i3 processor, 120 Giga Byte (GB) Hard Disk and 8GB RAM. Figure 2 and Table 2are the findings of both the GNB and RF algorithms for the textual analysis conducted on the Amazon product review from amazon.com. These are the final results produced from the 80 percent train data and the 20 percent test data. In the figure 2 the x-axis represents the classifiers whereas the y-axis represents the accuracy in percentage.

Table 2: Average accuracy of the classifiers

Techniques with evaluation Metrics	Naive Bayes (GNB)	Random Forest (RF)
Accuracy of classifier 1	0.7425	0.6906
Accuracy of classifier 2	0.7430	0.7412
Accuracy of classifier 3	0.7447	0.7375
Accuracy of classifier 4	0.7537	0.7164
Accuracy of classifier 5	0.7449	0.6730
Accuracy of classifier 6	0.7478	0.6319

A. System Evaluation Results

The machine assessment can be carried out with any regression evaluator. In this analysis to initially know the estimation they provide the estimates in the train and evaluate results on each of the measurements. Second, the models are tested using the precision to check the data sets. The findings are seen in the Figure3 where accuracy is indicated via x-axis and classifiers are indicated via y-axis. It concludes that the precision is higher in GNB than RF.

TEST Engineering & Management

May - June 2020
ISSN: 0193-4120 Page No. 4950-4954

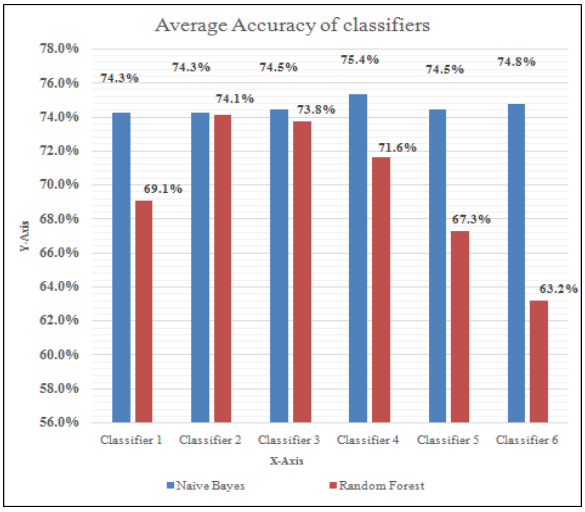

Figure 2: Average accuracy of the classifiers

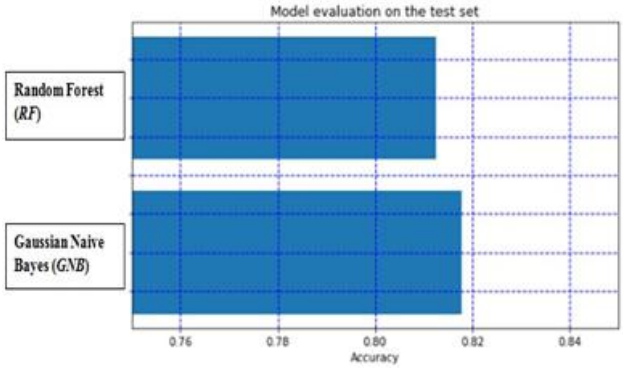

Figure 3: Model Evaluation on the test data

5. Conclusion and Future work

Here, the proposed framework was able to conduct a big data analysis of emotion over an immense volume at high speed in real time owing to the large data platform Spark that was used for implementation. The GNB and RF approaches are mainly used to assess the positive and negative feeling still the proposed system has been through the preprocessing stage, characteristics generation stage, classifier learning stage, and pipeline level. Through the experimental tests the system output is improved by the average measurement accuracy metric and compared to the RF the GNB has more accuracy. The experiments in this study are performed on the Amazon data set (reviews) only. In the future outlook, this study can be used on different data sets to predict accuracy using the big data tools.

References

[1] A Bifet, "Mining big data in real time, "Inform., vol. 37, no.1, pp.15–20, 2013.

[2] S. Lenka Venkata, "A Survey on Challenges and Advantages in Big Data," vol. 8491, p115–119, 2015.

[3] J. P. Verma., A. Smita, B. Patel., and P. Atul, "Big Data Analytics: Challenges and Applications for Text, Audio, Video, and Social Media Data," Int. J. Soft Comput. Artif. Intell. Appl., vol. 5, no. 1, pp. 41–51, 2016.

[4] Ramesh R, Divya G, Divya D, Merin K Kurian, and Vishnuprabha V, "Big Data Sentiment Analysis using Hadoop", IJIRST, Volume 1, Issue 11, pp. 92-98, 2015.

[5] Nurulhuda Zainuddin, Ali Selamat," Sentiment Analysis Using Support Vector Machine", IEEE International Conference on Computer, Communication, and Control Technology (I4CT 2014), Kedah, Malaysia,pp.333-337, 2014.

[6] Kamal Al-Barznji, Atanas Atanassov, "A Framework for Cloud Based Hybrid Recommender System for Big Data Mining", a journal of "Science, Engineering & Education", Volume 2, Issue 1, UCTM, Sofia, Bulgaria, pp. 58-65, 2017.

[7] Jason Bell, "Machine Learning: Hands-On for Developers and Technical Professionals", Published by John Wiley & Sons, Inc., Indianapolis, Indiana, 2015.

[8] Boštjan Kaluža, "Machine Learning in Java", first published: Published by Packt Publishing Ltd, UK, 2016.

[9] Benjamin Bengfort and Jenny Kim, "Data Analytics with Hadoop", Published by O'Reilly Media, Inc., First Edition. USA, 2016.

[10] Mohammed Guller, "Big Data Analytics with Spark", ISBN13 (pbk): 978-1-4842-0965-3, 2015.

[11] Rajat Mehta,"Big Data Analytics with Java", Published by Packt Publishing Ltd, ISBN 978-78728-898-0, UK, 2017.

[12] Kamal Al-Barznji, Atanas Atanassov, "Big data sentiment analysis using machine learning algorithms" 26th international conference on "Control of energy, industrial and ecological system",2019

[13] Ramesh R, Divya G, Divya D, Merin K Kurian, and Vishnuprabha V, "Big Data Sentiment Analysis using Hadoop", IJIRST, Volume 1, Issue 11, pp. 92-98, 2015.

[14] Nurulhuda Zainuddin, Ali Selamat," Sentiment Analysis Using Support Vector Machine", IEEE International Conference on Computer, Communication, and Control Technology (I4CT 2014), Kedah, Malaysia,pp.333-337, 2014

[15] X. Fang and J. Zhan, "Sentiment analysis using product review data," J. Big Data, pp. 1–14, 2015.