*International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)*
*Volume. 1, Issue 5, September 2014, PP 1-5*
*ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)*
*www.arcjournals.org*

# Text Classification Using Mahout

**G.V. Ramana Reddy**
Assistant Processor, Dept. of CSE
BITS-Knl, JNTUA University
*gvrreddy@gmail.com*

**K. Mounika**
IV Year Student of ECE, Dept. of ECE
BITS-Knl, JNTUA University
*kattamounika.reddy93@gmail.com*

**A. Chinmayi**
IV Year Student of ECE, Dept. of ECE
BITS-Knl, JNTUA University
*chinu.chinmayi@gmail.com*

**S. Fareed Hussain**
IV Year Student of ECE, Dept. of ECE
BITS-Knl, JNTUA University
*sfareed143@gmail.com*

**Abstract:** *The storage, processing and analysis of BIGDATA present a plethora of new challenges to computer science researchers and IT professionals. Mahout is a set of distributed data mining libraries that interface with an underlying distributed system. The frame-work for the distributed system is Hadoop, which implements Mapreduce. Mahout provides a library of scalable machine learning algorithms useful for big data analysis based on Hadoop or other storage systems. Classification techniques decide how much a thing is or isn't part of some type or category, or how much it does or doesn't have some attribute. Classification, like clustering, is ubiquitous, but it's even more behind the scenes. This paper exhibits the classification technique by using Mahout. The sample data was taken from 20 Newsgroups and the resulting Confusion matrix is presented.*

**Keywords:** *Bigdata, Hadoop, Mapreduce, Classification, Mahout, Confusion Matrix*

## 1. INTRODUCTION

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, storage, search, sharing, transfer, analysis and visualization.[3]

### 1.1. Key Characteristics to Define Big Data

The main challenge is that as things stand today, there is no single technology that can cope with all the characteristics of big data – volume, velocity and variety – all at once. [1]

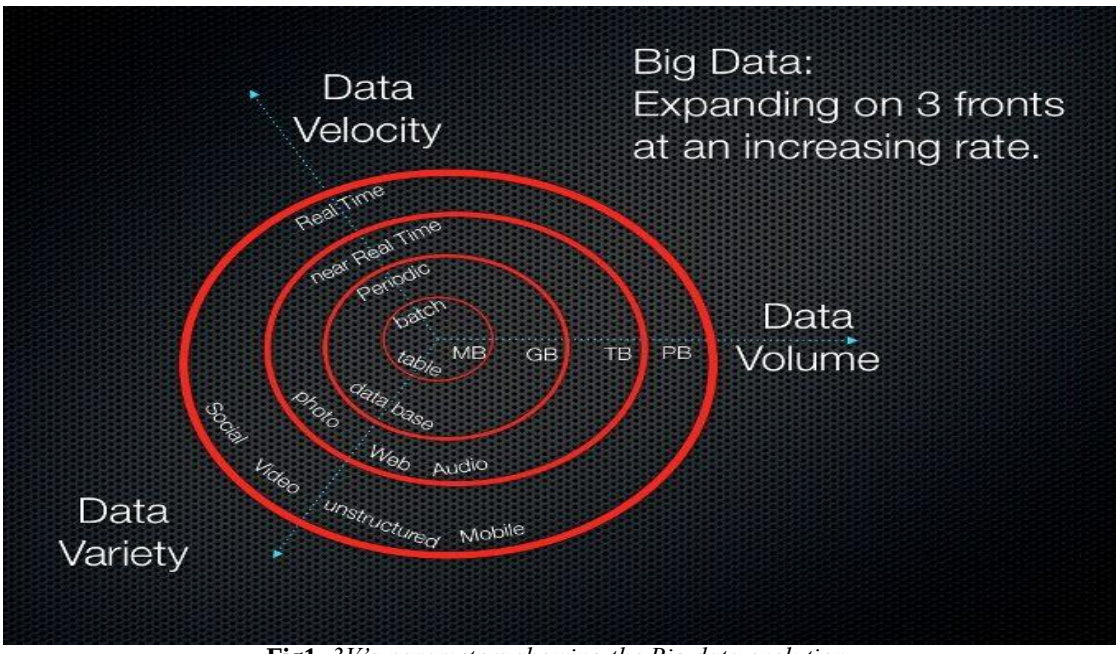


**Fig1.** *3V's parameters showing the Big data evolution*

## 2. MAHOUT

Mahout began life in 2008 as a subproject of Apache's Lucene project.Mahout is the tool used for analysis of **BIG DATA.**It is a Java library**.**It can be used with Hadoop to deal with large scale data.Currently Mahout supports mainly three use cases: Recommendation mining takes users' behaviour and from that tries to find items users might like. Clustering takes e.g. text documents and groups them into groups of topically related documents. Classification learns from existing categorized documents what documents of a specific category look like and is able to assign unlabelled documents to the (hopefully) correct category.[3]

## 3. CLASSIFICATION

Classification techniques decide how much a thing is or isn't part of some type or category, or how much it does or doesn't have some attribute. Often these systems learn by reviewing many instances of items in the categories in order to deduce classification rules. Classification helps decide whether a new input or thing matches a previously observed pattern or not, and it's often used to classify behaviour or patterns as unusual. It could be used to detect suspicious network activity or fraud. It might be used to figure out when a user's message indicates frustration or satisfaction.[3]

The famous classifier engines are:

- Spam emails
- Google's Picasa
- Optical character recognition software
- Apple's Genius feature in iTunes [4]

### 3.1. Using Mahout for Classification

Mahout is most useful with extremely large or rapidly growing data sets where other solutions are least feasible. Up to about 1, 00,000 examples, other classification systems can be efficient and accurate. But generally, as the input exceeds 1 to 10 million training examples, something scalable like 'Mahout' is needed. [3]

**Table1.** *Choosing the Classification approach based on size of examples*

| **System size in number of examples** | **Choice of classification approach** |
|---|---|
| <1,00,000 | Traditional non-Mahout approaches should work very well. Mahout may even be slower for training. |
| 1,00,000 to 1million | Mahout begins to be a good choice. The flexible API may make Mahout a preferred choice, even though there is no performance advantage. |
| 1million to 10 million | Mahout is an excellent choice in this stage |
| >10 million | Mahout excels where others fail |

The reason Mahout has an advantage with larger data sets is that as input data increases; the time or memory requirements for training may not increase linearly in a non-scalable system. In general, the classification algorithms in Mahout require resources that increase no faster than the number of training or test examples, and in most cases the computing resources required can be parallelized. This allows you to trade off the number of computers used against the time the problem takes to solve. [3]

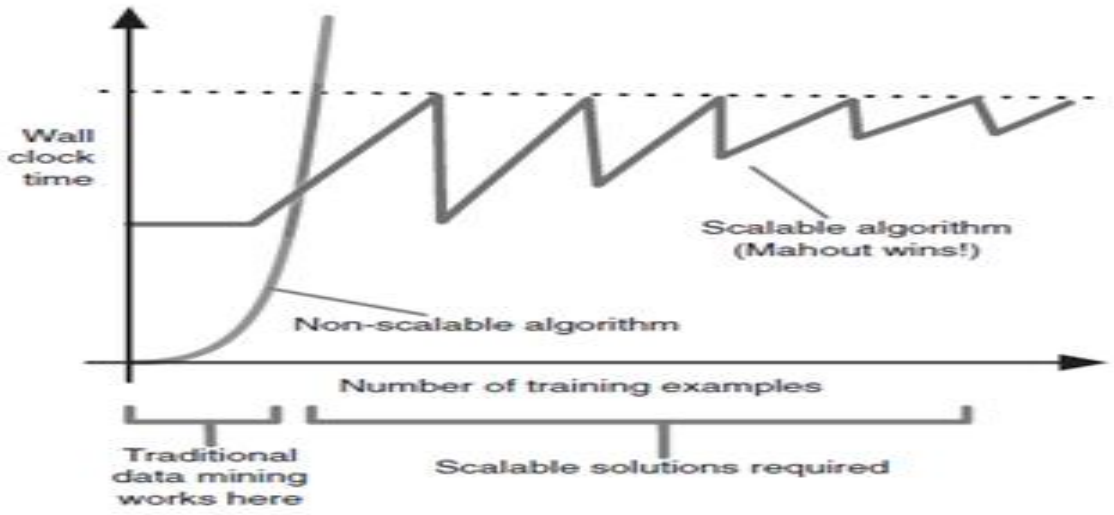


**Fig2.** *Advantage of using Mahout for large scale systems*

### 3.2. Working of Classification

There are two main phases involved in building a classification system:

- ✓ The creation of a model which is produced by a learning algorithm, and the use of that model are to assign new data to categories.
- ✓ The selection of training datas, output categories (the targets), the algorithm through which the system will learn, and the variables used as input are key choices in the first phase of building the classification system.[3]

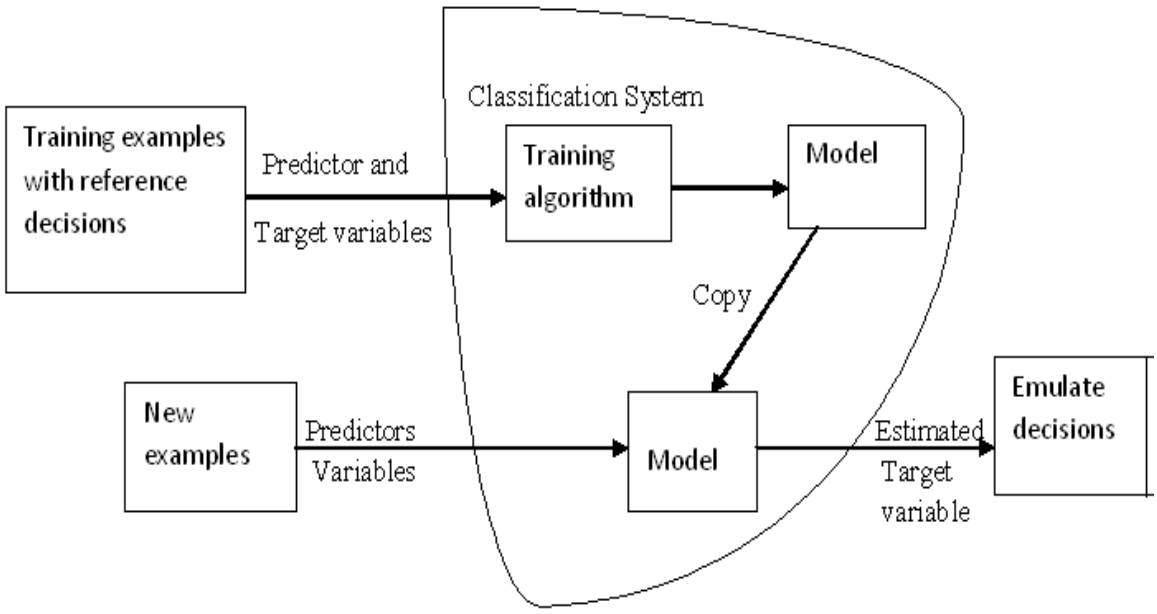


**Fig3.** *Working of a Classification System*

A computer program that makes decisions; in classification, the output of the training algorithm is a model. A classification algorithm learns from examples in a process known as training. The output of this training process is called a model. This model is a function that can then be applied to new examples in order to produce outputs that emulate the decisions that were made in the original examples.

### 3.3. Confusion Matrix

**Confusion matrix** is a cross-tabulation of the output of the model versus the known correct target values in the data. The matrix has rows for each actual value of the target variable and columns for the output of the model. Let us consider an example,

If a classification system has been trained to distinguish between cats, dogs and rabbits, a confusion matrix will summarize the results of testing the algorithm for further inspiration. Assuming a sample of 27 animals- 8 cats, 6 dogs and 13 rabbits and the resulting confusion matrix could look as below

**Table2.** *Example on Confusion Matrix*

| | **Predict class** | | | |
|---|---|---|---|---|
| **Actual class** | | Cat | Dog | Rabbit |
| | **Cat** | 5 | 2 | 1 |
| | **Dog** | 2 | 3 | 1 |
| | **Rabbit** | 1 | 1 | 11 |

From the above table we have the sample data(cat,dog,rabbit) and this data is given as the input the training algotithm. Here the algorithm is trained and a model is produced as the output. Now this model is given as the input to the test data and based on the training which is present and test data the model produces the output which is the target variable. The accuracy is judged. Actually we have taken 8cats and at the output the model identified 5 as cats, 2 as dogs and 1 as rabbit. During the training we give both inputs(predictor variables) and target variables as the input. During testing only inputs(predictor variables) are given based and the model generates the output(target variable) as required. If output is not appropriate the model is again trained for the good result.

## 4. RESULTS AND DISCUSSIONS

By following the above procedure we have generated a model which includes the following steps:

- ✓ Creating a working directory and copying the data
- ✓ Creating a sequence files from 20newsgroups data
- ✓ Converting the sequence files into vectors
- ✓ Creating the training and test data with 60-40 split
- ✓ Training the data using Naive Bayes model
- ✓ Self testing the training data [4]

By following the above steps as a result a Confusion Matrix is generated which gives the overview on the performance of the model.

Due to the insufficient amount of data we have taken the same data for both training and testing and hence the confusion matrix is generated as below:

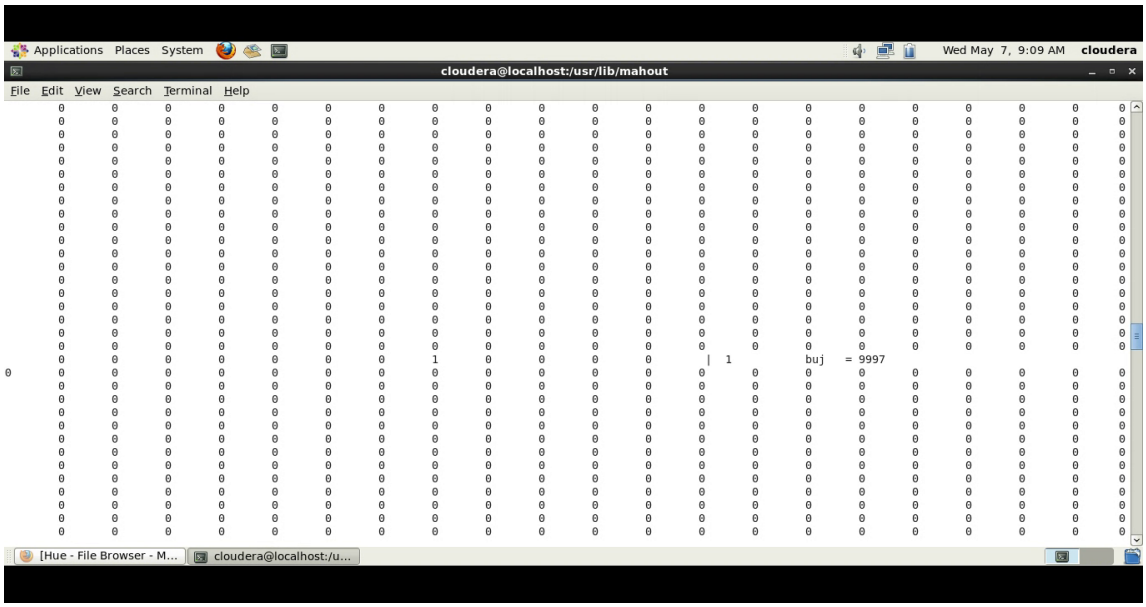


**Fig4.** *Confusion matrix*

The ingredients that are used for generating the model are

- ✓ Cloudera- Hue tool
- ✓ LINUX commands
- ✓ Data from 20newsgroups
- ✓ Naive Bayes algorithm to train the model [3]

## 5. CONCLUSIONS

In this paper, we have presented the statistical Naive Bayes classification algorithm that determines the probability of a set of words falling under different categories based on observations from the training data. This algorithm was applied on the data downloaded from 20news-groups. Linux commands were used to apply the algorithm and the related confusion matrix was generated. And it was found that the model was working well when the training data itself was fed as test data since the test data obtained was very small in size. However, it has to be tested with different data sets to establish the accuracy of the model.

## REFERENCES

[1] https://mahout.apache.org/

[2] https://mahout.apache.org/users/classification/twenty-newsgroups.html

[3] Sean Owen,Robin Anil et al, "Mahout in Action"Manning Publications Co.©2012.

[4] Apache Mahout cookbook- Book by Piero Giacomelli published Dec 2013 by Packtpub.

## AUTHORS' BIOGRAPHY

**GVRamana Reddy**, completed his post graduation from University of Hyderabad in specialization of Artificial Intelligence and presently working as an Assistant Professor in Brindavan Institute of Technology & Science, Kurnool A.P. Areas of interest are Natural Language Processing(NLP), Data Mining and Artificial Intelligence

**K.Mounika** pursuing Bachelor`s Degree in Electronics & Communication Engineering from Brindavan Institute of Tech. & Science, Kurnool A.P. Area of Interest is DBMS and Big Data Technology

**A.Chinmayi** pursuing Bachelor`s Degree in Electronics & Communication Engineering from Brindavan Institute of Tech. & Science, Kurnool A.P. Area of Interest is BigData and Advanced Technologies

**S. Fareed Hussain** pursuing Bachelor`s Degree in Electronics & Communication Engineering from Brindavan Institute of Tech. & Science, Kurnool A.P. Area of Interest is Big Data and Hadoop Technologies