Machine learning on text analytics and categorization through R-Language

[1]B.Rajesh, [2]G.Bhargavi, [3]L.Soundraya, [4]K.Ramadevi, [5]N.Anand Kumar

[1]Assistant.Professor, [2,3,4,5]B.tech Students
Computer Science and Engineering,
Mother Theresa Institute of Engineering and Technology, Palamaner, India

***Abstract*— Text mining and Text analytics is a process of extracting use full information from Text Documents. Due to the extreme growth in online textual information.eg: "Email messages and social medias online News" To organizing this data is one of the major problem. so we organize and handle this type of E-learning documents we introduce a approaches called topic modeling and sentiment analysis these two methods are classify textual data and documents in a systematic manner, where Topic modeling implemented by the use of LDA(Latent Dirchlet Allocation) it converts the text documents into sentences and words then DTM(Document Term Matrix) assigns the frequency for each and every Term Present in Text document based on number of occurrences then LDA groups relevant Topics. Sentiment analysis implemented by using SVM classifier, SVM (Support vector Machine) mainly concentrate on opinion mining to find out the emotion of the particular person on that particular Document like happy, sad, satisfied, unsatisfied etc. Thus E-learning documents can be simply retrieved and classified using these methods which is also proven by experimental verifications. From the experimental result real world data from various areas shows that our proposed system out performs more than a few other baseline methods.**

***Index Terms*— Topic modeling, LDA, classification, sentiment analysis, SVM, DTM, E-Learning, term frequency.**

I. INTRODUCTION

Data mining technology used to extract Relevant Information from various sources in that one of the major field called text mining, it is similar to Data mining concept but it is used to analyze only text documents. In 21st century e-learning documents are increased day by day so we organize those documents for an growth of organizations. in the form of statistics. i.e Each and every Business organization maintain his own data base like "Products sales and reviews, ratings " so at the end of the every year they can go for auditing to find the overall status. Where Text mining is it extracts use full information from text documents generated in various environments.

E-learning is also called as electronic-learning which is used to learn the user generated data from the internet, and here we can use the e-learning data from social media, micro blog and other sources. Here we can analyze the large number of text documents and finding the interesting key words and classifying by using topic modeling, In previous paper they can find top key words based on ranking of each tweet, but here we use word frequency to list the top keywords. High frequency words are said as interesting key words and the low frequency words are un-interesting contents. Basically the large amount of data is not an easy task to analyze because of this problem we adopt a statistical and popular unsupervised approach called Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). It is a one type of text clustering model used mostly for data mining problems. The functioning of LDA is, first it adopts a large amount of text data set and then predict that data, split into similar size of documents and, again split each document into a number of words after that we assign a frequency to each word and classify it based on frequency.

Topic modeling is a one type of classifier, it can classify the input data into a number of topics, and each topic containing similar words for each label. Labels are areas like politics, sports and movies, products, cars etc.., the LDA model contains two types: the word probability of words under the probability of topics, it's under the probability of documents, and test and train LDA data can be used in same labels.In this study, we mainly focus on statistical method to identify interesting and popular terms to a wide audience. Here we perform LDA-based topic model analysis and find latent topics and assign word frequency for each latent word. Our contributions are summarized as follows:

1. We proposed a novel unsupervised approach that identifies high frequent key terms in text documents and filters low frequency words said to be un-interesting key words from the twitter data set.
2. The text classification method called Latent Dirichlet Allocation (LDA) is used to analyze the data mining documents and propose an LDA based topic modeling to trending topic analysis.
3. Based on topic identification, we assign frequency to each word and topics as per its relative importance.

Sentiment analysis is also called as opinion mining, It is a process used to know the emotions behind the collection of words, which will be used to estimate the attitudes, opinions and emotions expressed within different online reviews or any online products. **Sentiment analysis uses:** Sentiment analysis is very useful in monitoring of social media as it allows us to collect an overview of the people opinions and emotions behind different topics. Social media monitoring tools like Brand watch Analyze and mobile phones analytics make that process faster and easier than previous methods, real-time monitoring is useful for the many capabilities. The uses of sentiment analysis are very high and powerful. The ability to extract deep understanding from social data is a practice that is being legally taken and brings it up as one's own by organizations across the world. For sentiment analysis we use the SVM (support vector machine) classifier. Support vector machine will represent a hyper plane or set of hyper planes in a low or high or infinite dimensional space.

R-LANGUAGE

R is a one type of statistical programming language. Which can provide various techniques for text mining like classification and clustering etc,.“R environment run based on CRAN mirror. it is a collection of packages and tools building to Run R-programs in a R-studio.”R is a combining process of facilities for data modification, calculation and display it mainly includes an effective data storage and data handling facility.R allows the user to add different operations by defining new operations.The packages involved in R are tm (textmining package),NLP(natural language processing) etc…

Then the remaining paper is organized as section 2, section 3, and section 4, in section 5 we examine related work, proposed work, and experimental result and finally conclude the paper.

II. RELATED WORK

Text analytics and categorization be a use full methodology implemented by using some classifications techniques like topic modeling and sentiment analysis proposed by S.swathi and P.lalitha in 2017[1]. Text data mining is the analysis of data contained in natural language processing (NLP). The text data mining performs the transferring words in unstructured data into numeric values which can be linked with structured data in text database. Text databases are growing day by day due to the increasing amount of information available in e-documents-mails and www. Now a days the data stored in text databases are structured or unstructured data. it is more difficult is implementation of semi structured data in present databases. This is the major principle problem .to overcome this problem we use the survey of text data mining: such as retrieval, extraction and indexing techniques [2]. Knowledge Discovery in Database (KDD) is also known as data mining or text data mining. KDD defines the extraction of information from the large database. Data mining contains the unnecessary data and extracting the interesting useful information from the structured data.in the data mining they are different sources that convert structured data in very large amount.appications of data mining are increasing in digital word this is the drawback.to overcome this problem we use the applications such as educational data mining(EDM),life sciences and medical etc. this applications are determine how the KDD can be used in different fields and we can easily classify the different models adapted in the KDD[3]. In a micro-blogging area, topic modeling and semantic analysis are widely used to analyze textual data and help in many users related modelings. (Content filtering (Duan & Zeng, 2013; Pennacchiotti & Gurumurthy, 2011), sentiment analysis (Lin & He, 2009), Martinez-Romo & Araujo, 2013), to classify and analyze the topical differences between traditional media and twitter using twitter- LDA for investigating short messages. LDA based topic modeling regards the combination of trending topics and times to distribute them, the time distribution is based on the word frequencies. (Ramage, Dumais, & Liebling, 2010) deal with Labeled-LDA model here a tweet uses its labeled information, and then built the probability distribution for latent topics to represent the tweet’s content. Ramage et al. (2010) study the similarities between the topic modeling and interestingness, hear interesting contents are based on the word frequencies and trending topics also collect from those. In proposing system Min-Chul Yang, Hae-Chang Rim (2013) invent a topical analysis for identifying interesting twitter contents from the large amount of the micro - blog dataset.

III. PROPOSED SOLUATION

Due to the disadvantages described in the existing system, we propose a new method to overcome its drawbacks. The new system implements Latent Dirichlit Allocation (LDA) based topic modeling along with a text mining concept and Sentiment analysis to prove that the proposed system outperforms the existing system. The Process involved in the proposed system can be described as follows: A text dataset containing much number of messages in the form of unstructured format. In that dataset we can perform a pre-processing mechanism, where we can remove stop words, numbers, punctuations, mailing constants and finally stemming words this process is said to be a data cleaning. Then we can get a structured data and create Document term matrix (DTM), it can generate a number of Documents and containing terms as well as words. After creating a document matrix we must calculate the word frequency for all terms containing in the data set. There are two types of matrixes head frequency matrix represent an high frequent keywords and low frequency specifies low frequent key terms. Now, we must compute the term weighting factor by using the product of the frequency of words. This weight is plotted on the matrix which in turn becomes the word cloud or a topic model graph. The graph values can be processed by using LDA to represent high frequency words. Thus the proposed system outperforms the existing system in accuracy, performance and efficient retrieval and classification of documents.

LDA Algorithm

Latent Dirichlit Allocation is the most widely used classifier in topic modeling. LDA assumes “M” number of documents in that document containing “N” number of words and each word “W” contain assigned frequencies.

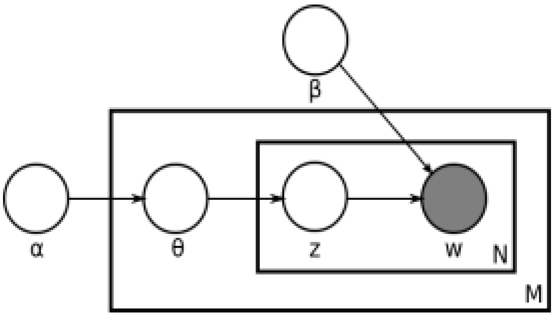


Fig.1. Plot notation for LDA algorithm

One of the main advantages of LDA is that it could easily split the documents in to words and cluster top key terms; LDA is a type of un-labeled topic modeling classifier. Working of LDA generative algorithm The working of LDA has the following steps

1. Collecting “M” twitter documents
2. Determine term distribution
3. Determine topic distribution for documents
4. Split documents in to number of topics
5. Split topics into number of words
6. Determine word frequency’s
7. Classify based on frequency

Above plot representation for LDA can represent a sequential flow of the model, here “M” represents a document (collection of tweets) N represents a number of topics present in the document and the W is a number of words in the each topic, then the parameters represent probability distributions.

SVM Algorithm

Sentiment analysis is a one type of classification technique where we use SVM classifier to predict the emotions present in particular document. The process of sentiment analysis is first it takes large textual documents as input and then we perform pre-processing, after create Document term matrix (DTM). DTM holds documents as rows and terms as columns and it assigns frequency for each word. Based on word frequency we creates labels and containers some text taken as input and some text for test label then we find precision and recall, based on these two things we find accuracy for SVM label.

IV. SYSTEM DESIGN

The structural evaluation of the project may be described as, the text dataset is a collection of Documents in the form of unstructured format or said to be raw data. Every document contains valuable data and also useless data; from here we can gather useful and interesting or top key terms.

A. DATASET

Dataset is a collection of text documents collected from various sources, dataset may be in txt, csv ,.. etc formats.

B. CREATING THE CORPUS

A corpus might be a collection of user generated data in the form of E-Learning; each corpus has a number of topics and tweets, where we create a Data frame based on input dataset.

C. PRE-PROCESSING THE DATA

Preprocessing is a type of data analysis technique it can do mainly in two types Data Cleaning and Stemming, in Data cleaning we can perform some transformations including converting the text to lower case, removing stop words and punctuation, , removing re-tweet entities, removing numbers, removing html links, removing unwanted spaces. In Data Stemming we can remove the same words and identifying synonyms. Stemming use an algorithm that removes frequent word endings for English words, such as \es", \ed" and \'s". All this transformation is done under the text mining.

D. DOCUMENT TERM MATRIX

Document term matrix (DTM) is a one type of matrix having rows as terms and columns as topics, term document matrix is an inverse of DTM and it contains frequency of each word in the dataset. Head frequency denotes top key words in twitter dataset. In classification algorithm also use this to classify topics based on word frequency.

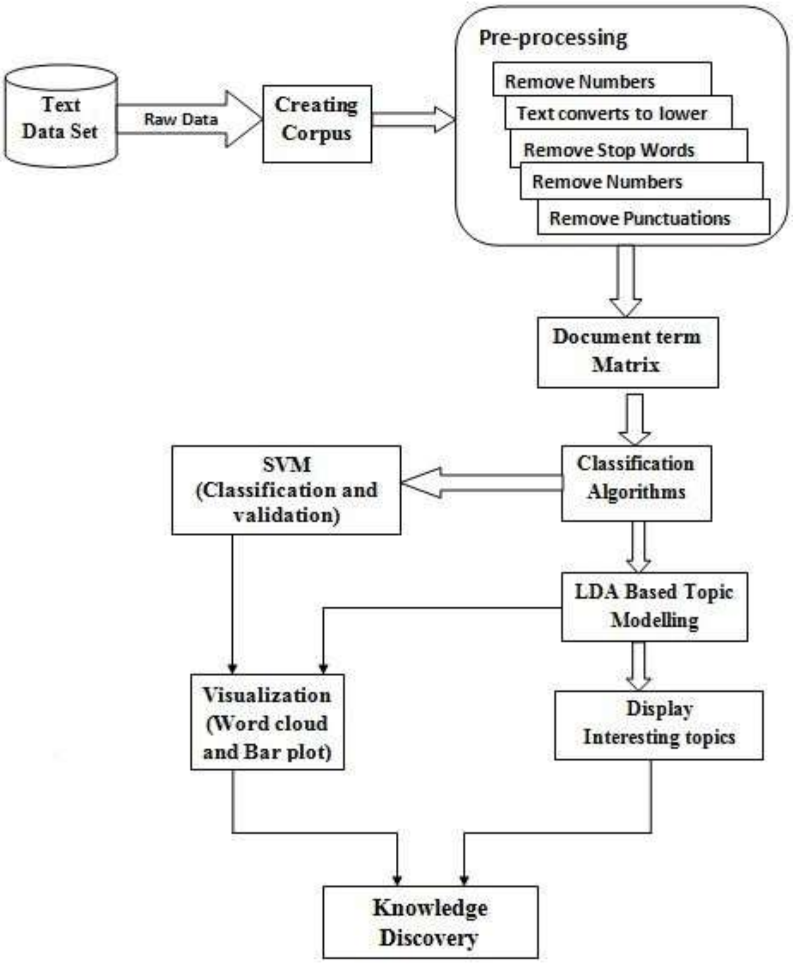


Fig.2. System Architecture

E. CLISSIFICATION ALGORITHM

Classification is a process of supporting the area of topics, Here we can use an LDA based topic modeling classifier, used mostly in large data mining areas where the trending and interesting key words are found to classify the similar words and topics. SVM classifier used to find emotion of the writer on that particular document.

F. WORD CLOUD GENERATION

A word cloud is a type of plot, it can represent the number of words in cloud forms the high frequency word can be displayed in bold and big size. The words mostly follow colors in random order the use of word cloud is that we easy to find the most popular and interesting topic in the dataset.

G. KNOWLEDGE DISCOVERY

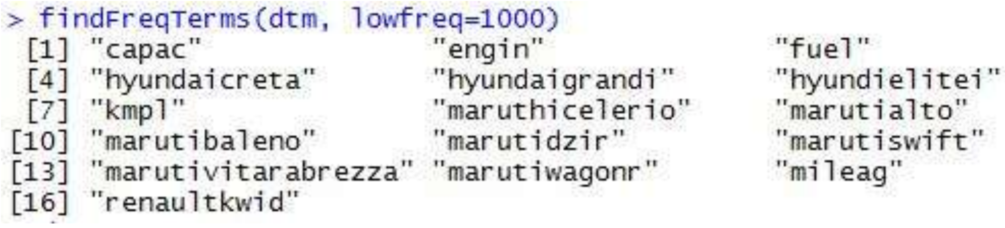
Knowledge means the final results and performance of algorithms used in this paper, where we use some classification techniques like LDA and SVM for these we acquire knowledge.

V. EXPERIMANTAL RESULTS

A primary thing is to find high frequent terms and emotions of the text data. Generally every document must have some message and that can deal with the particular area. In this paper, we can deal with the large amount of data grouped together to form a corpus or data set, we can use the LDA based topic modeling to apply classification and SVM classifier for sentiment analysis on the input data. Before we apply the LDA algorithm do the data cleaning process, it contains two main techniques Data pre-processing and stemming. Pre-processing is a type of data cleaning process here we can remove unwanted data from input data like removing numbers, punctuation, spaces and http links. Data stemming is a process of removing synonyms.

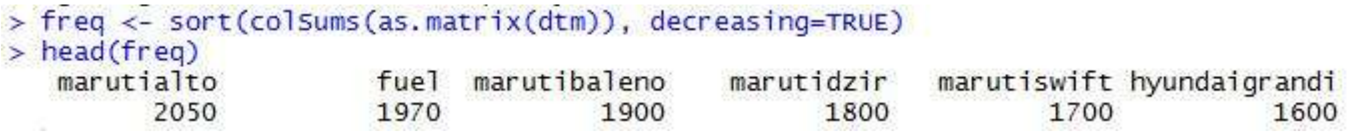
After that data cleaning process creates a Term Document Matrix and Document Term Matrix. These two matrices contain words as columns and rows containing frequency’s of all key words in dataset. Then we find the head frequencies of key terms, using topic modeling with an LDA algorithm to classify the interesting topics based on frequency. Here I am taking 20 cars dataset in a year.

Topic modeling:



```
> findFreqTerms(dtm, lowfreq=1000)
 [1] "capac"              "engin"              "fuel"
 [4] "hyundaicreta"       "hyundaigrandi"      "hyundielitei"
 [7] "kmpl"               "maruthicelerio"     "marutialto"
[10] "marutibaleno"       "marutidzir"         "marutiswift"
[13] "marutivitarabrezza" "marutiwagonr"       "mileag"
[16] "renaultkwid"
```

Fig 3: the list of terms greater than 1000 frequency in a document



```
> freq <- sort(colSums(as.matrix(dtm)), decreasing=TRUE)
> head(freq)
   marutialto          fuel  marutibaleno    marutidzir   marutiswift hyundaigrandi
         2050          1970          1900          1800          1700          1600
```

Fig 4: head frequent words present in text document

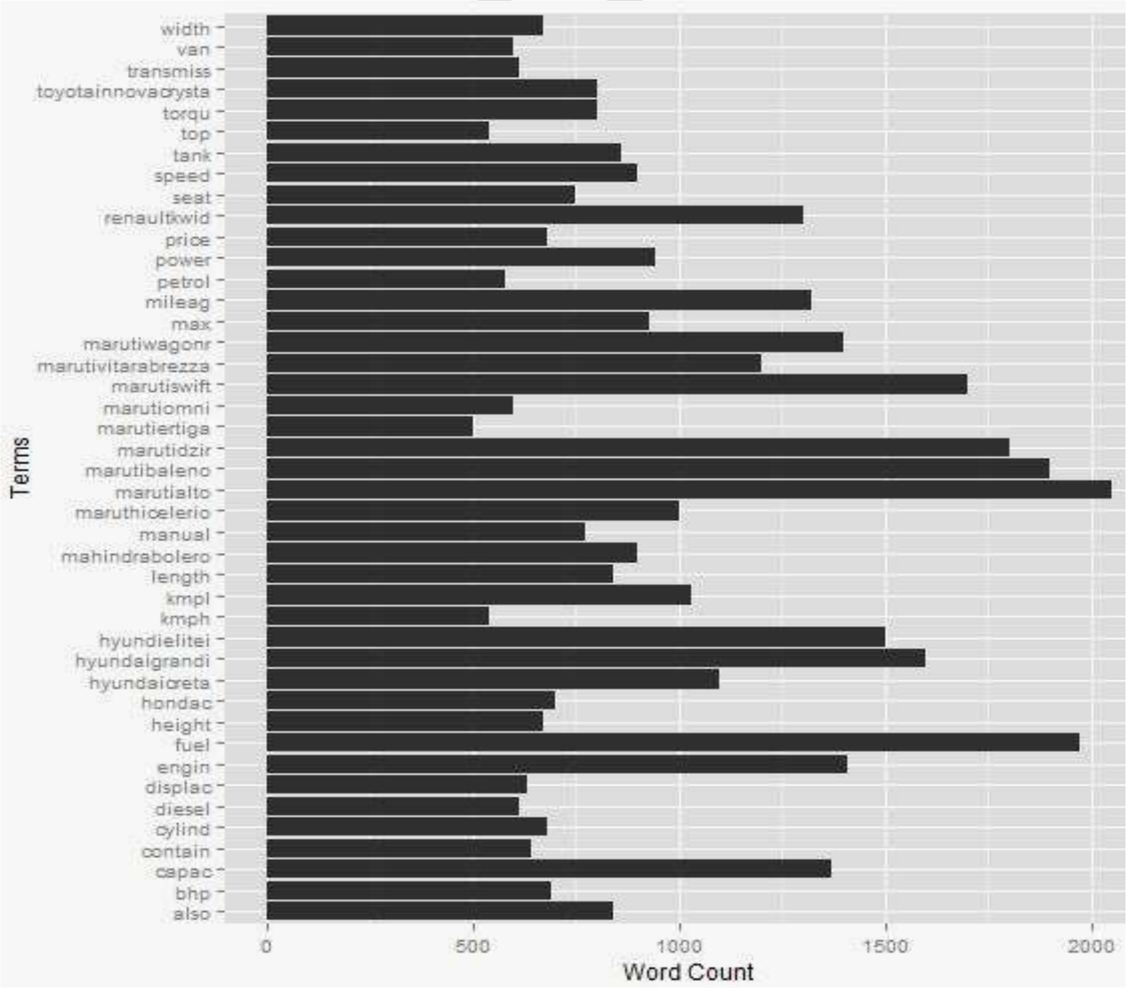
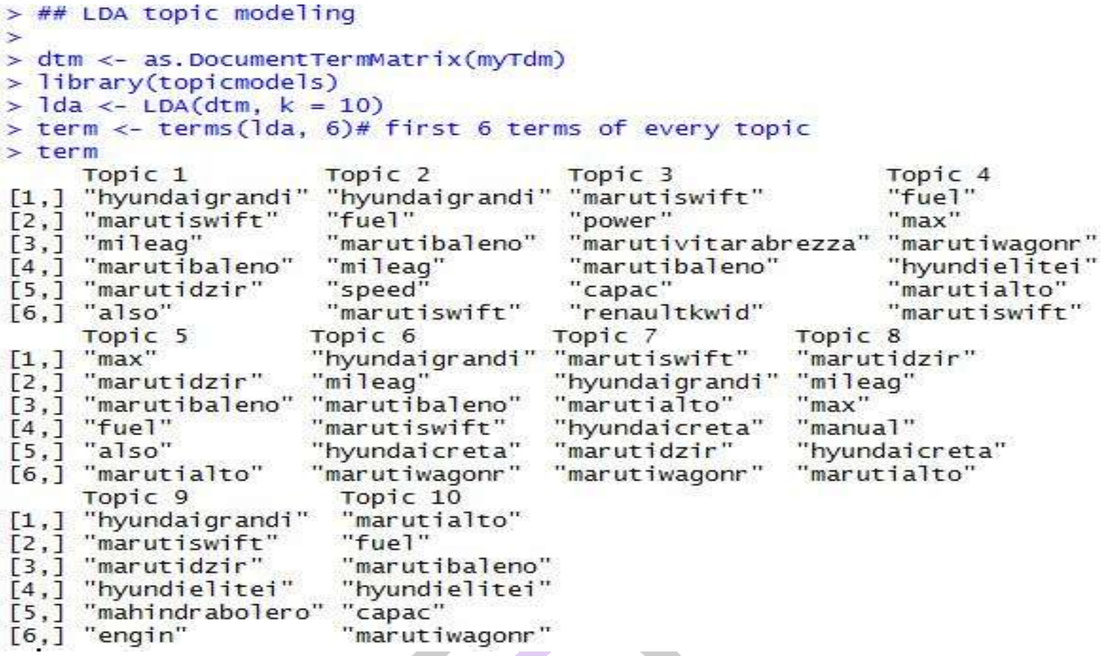


Fig 5: Bar plot for frequent terms and frequency

The histogram represents the high frequency terms and their frequencies, in that the high frequency word is data and it presents more than 180 times in the twitter data set.



```
> ## LDA topic modeling
>
> dtm <- as.DocumentTermMatrix(myTdm)
> library(topicmodels)
> lda <- LDA(dtm, k = 10)
> term <- terms(lda, 6)# first 6 terms of every topic
> term
     Topic 1          Topic 2          Topic 3              Topic 4
[1,] "hyundaigrandi"  "hyundaigrandi"  "marutiswift"        "fuel"
[2,] "marutiswift"    "fuel"           "power"              "max"
[3,] "mileag"         "marutibaleno"   "marutivitarabrezza" "marutiwagonr"
[4,] "marutibaleno"   "mileag"         "marutibaleno"       "hyundielitei"
[5,] "marutidzir"     "speed"          "capac"              "marutialto"
[6,] "also"           "marutiswift"    "renaultkwid"        "marutiswift"
     Topic 5        Topic 6         Topic 7         Topic 8
[1,] "max"          "hyundaigrandi" "marutiswift"   "marutidzir"
[2,] "marutidzir"   "mileag"        "hyundaigrandi" "mileag"
[3,] "marutibaleno" "marutibaleno"  "marutialto"    "max"
[4,] "fuel"         "marutiswift"   "hyundaicreta"  "manual"
[5,] "also"         "hyundaicreta"  "marutidzir"    "hyundaicreta"
[6,] "marutialto"   "marutiwagonr"  "marutiwagonr"  "marutialto"
     Topic 9          Topic 10
[1,] "hyundaigrandi"  "marutialto"
[2,] "marutiswift"    "fuel"
[3,] "marutidzir"     "marutibaleno"
[4,] "hyundielitei"   "hyundielitei"
[5,] "mahindrabolero" "capac"
[6,] "engin"          "marutiwagonr"
```

Fig 6: LDA Topic modeling

Above screenshot represents topic modeling of 6 words of each topic, contain related top key words, the words are linked from the term document matrix. DTM contains the words and their frequent rights, every topic containing the high frequent key words the top word on every topic are the most interesting word in data set.

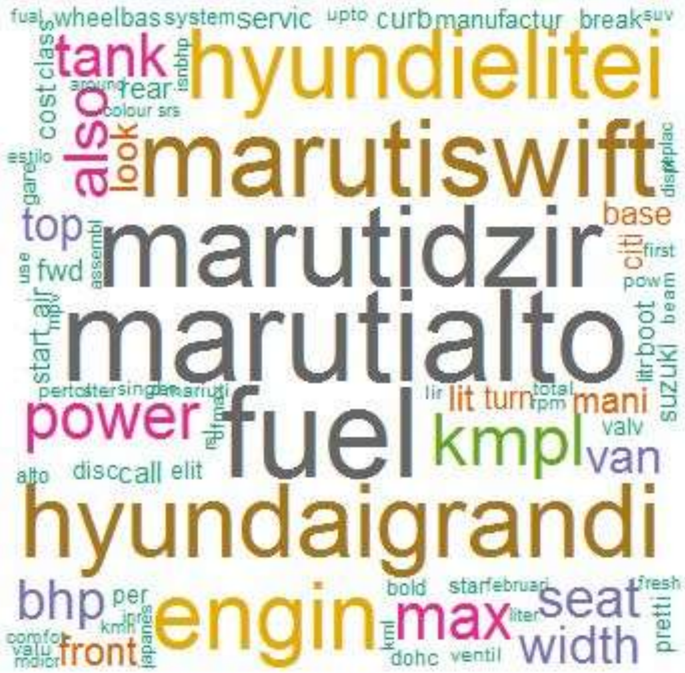
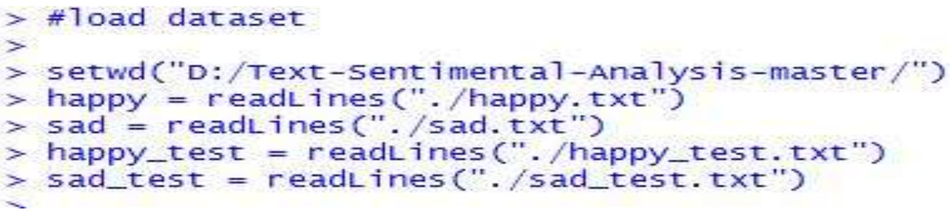


Fig 7: word cloud generation for dataset

Word cloud represents an easy understanding and visualizing of the top key words, this may be useful in understanding of topics. In that the bold word is top key word in giving twitter data set.

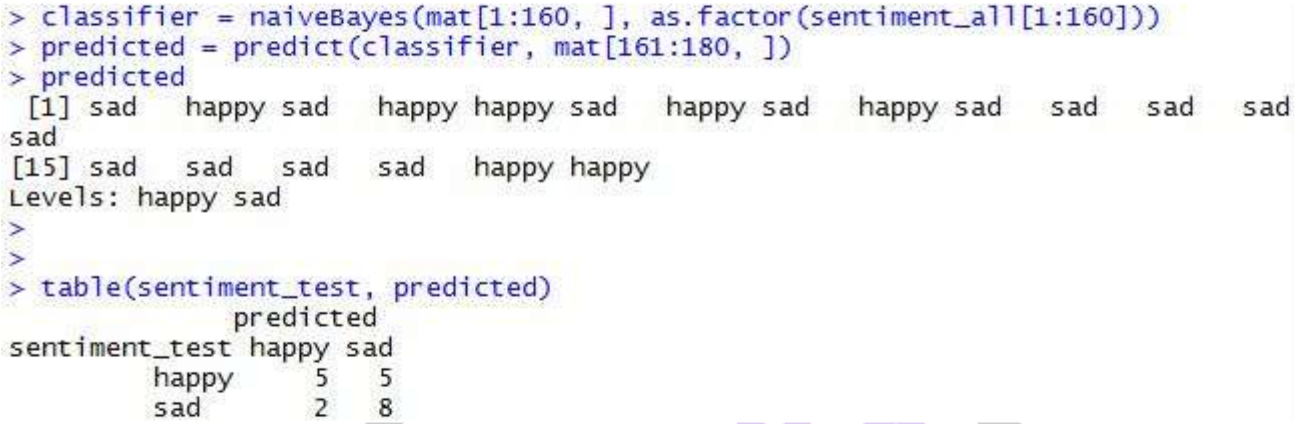
Sentiment Analysis:



```
> #load dataset
>
> setwd("D:/Text-Sentimental-Analysis-master/")
> happy = readLines("./happy.txt")
> sad = readLines("./sad.txt")
> happy_test = readLines("./happy_test.txt")
> sad_test = readLines("./sad_test.txt")
>
```

Fig 8: dataset loading for SVM

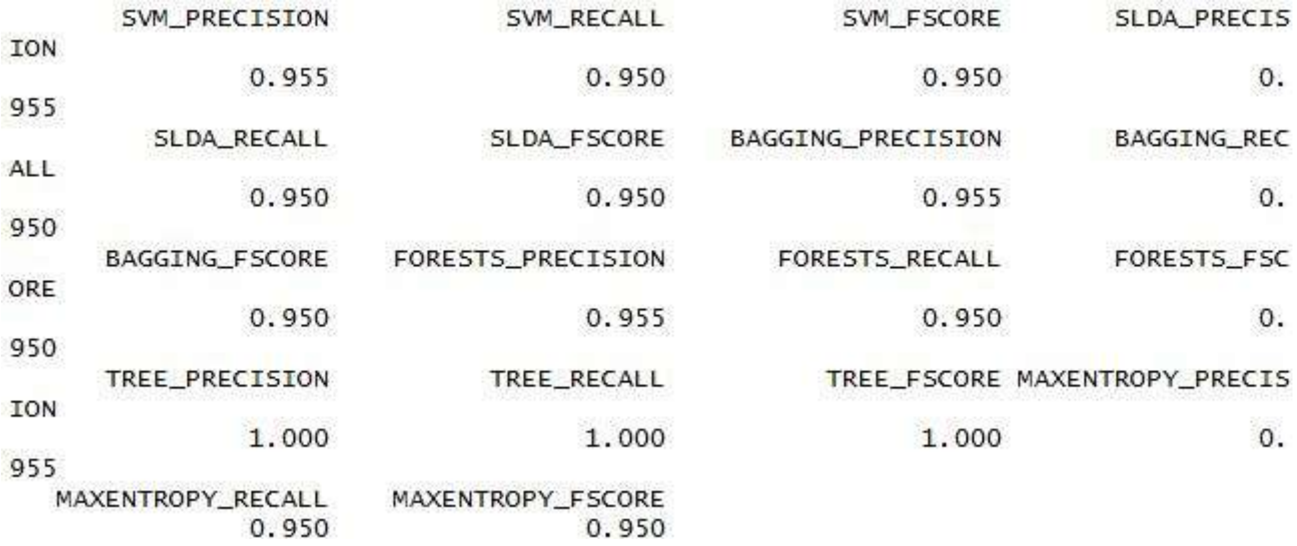
Above screen short shows loading of data from text sentiment analysis master, it gives data for prediction and testing.



```
> classifier = naiveBayes(mat[1:160, ], as.factor(sentiment_all[1:160]))
> predicted = predict(classifier, mat[161:180, ])
> predicted
 [1] sad   happy sad   happy happy sad   happy sad   happy sad   sad   sad   sad
sad
[15] sad   sad   sad   sad   happy happy
Levels: happy sad
>
>
> table(sentiment_test, predicted)
              predicted
sentiment_test happy sad
         happy     5   5
         sad       2   8
```

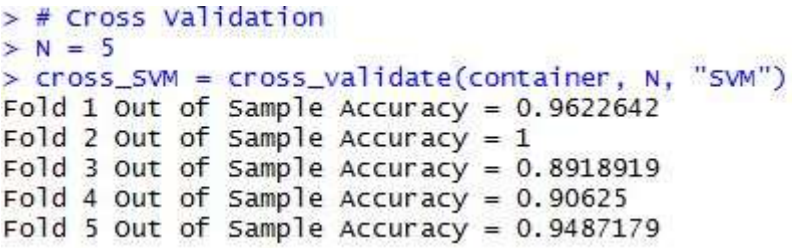
Fig 9: prediction table for sentiment

Prediction table gives emotion of documents in a tabular manner.



```
        SVM_PRECISION              SVM_RECALL              SVM_FSCORE         SLDA_PRECIS
ION
                0.955                   0.950                   0.950                  0.
955
          SLDA_RECALL             SLDA_FSCORE       BAGGING_PRECISION         BAGGING_REC
ALL
                0.950                   0.950                   0.955                  0.
950
       BAGGING_FSCORE      FORESTS_PRECISION          FORESTS_RECALL         FORESTS_FSC
ORE
                0.950                   0.955                   0.950                  0.
950
       TREE_PRECISION             TREE_RECALL             TREE_FSCORE MAXENTROPY_PRECIS
ION
                1.000                   1.000                   1.000                  0.
955
    MAXENTROPY_RECALL       MAXENTROPY_FSCORE
                0.950                   0.950
```

Fig 10: precision and recall, accuracy of SVM



```
> # Cross Validation
> N = 5
> cross_SVM = cross_validate(container, N, "SVM")
Fold 1 Out of Sample Accuracy = 0.9622642
Fold 2 Out of Sample Accuracy = 1
Fold 3 Out of Sample Accuracy = 0.8918919
Fold 4 Out of Sample Accuracy = 0.90625
Fold 5 Out of Sample Accuracy = 0.9487179
```

Fig 11: cross validation for SVM container

VI. Conclusion

We proposed a novel un-supervised approach to discover interesting key terms or top key terms from the user generated data. To analyze textual data more effectively, we developed a new LDA based topic modeling technique which is used to determine the interestingness of an individual tweet, we first extracted micro-blog topics based on our proposed functions. From the latent variables of LDA, we calculate word frequency of a topic by utilizing its representative words. Topic modeling is a one type of classifier; it can use the Term Document Matrix (TDM) in text mining to construct a number of topics. Also, for each topic we calculated two types of frequencies. One is high and another is low; here high frequency denotes interesting topics and low frequency denotes un-interesting key terms in text data. Compared to Gibbs sampler and other sentiment algorithms SVM gives better accuracy and performance. From our observations, each latent topic with a high weight value covers a specific topical theme. We then investigated how important topics spread with a target text. In a series of experiments, we demonstrated the ways in which our model can be naturally applied to recommend, filter, and understand textual posts in E-learning documents. So finally these two classification algorithms classify the topics in a effective manner.

References

[1] S. Swathi, P. Lalitha on 2017 A Study on Text Analytics and Categorization Techniques for Text Documents

[2] R. Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012.

[3] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, “Data mining techniques and applications–a decade review from 2000 to 2011,” Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[4] Nalini, K. and Dr. Jaba Sheela, L. “Survey on Text Classification”, July 2014.

[5] Martinez-Romo, J., & Araujo, L. (2013). Detecting Mali cious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications,40, 2992–3000.

[6] Armentano, M., Godoy, D., & Amandi, A. (2013). “Follo Wee recommendation in twitter based on text analysis of micro-blogging activity”. Information systems, 1116-1127.

[7] Lee, C.-H. (2012). “Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams”. Expert Systems with Applications, 39, 13338–13356.

[8] Hong, L., Dan, O., & Davison, B. D. (2011). “Predicting popular messages in twitter” In Proceedings of the 20th inter national conference companion on world wide web WWW ’11 (pp. 57–58).

[9] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or news media? In Proceedings of the 19th international conference on World Wide Web WWW ’10 (pp. 591–600). New York, NY, USA: ACM.

[10] Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: Experiments on recommending Content from information streams. In Proceedings of the SIGCHI conference on human factors in computing systems CHI 10 (pp1185–1194). New York, NY, USA: ACM.

[11] Lauw, H. W., Ntoulas, A., & Kenthapadi, K. (2010).Estimating the quality of postings in the real-time web. In Proce -edings of the WSDM 2010 workshop on search in social media SSM ’10.

[12] Alonso, O., Carson, C., Gerster, D., Ji, X., & Nabar, S. U. (2010). Detecting uninteresting content in text streams. In Proceedings of the SIGIR 2010 workshop on crowdsourcing for search evaluation CSE ’10 (pp. 39–42).

[13] Ramage, D., Dumais, S., & Liebling, D. (2010). Cairo -cterizing microblogs with topic models. In Proceedings of the Fourth international AAAI conference on weblogs and Social Media. AAAI.

[14] Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit Attribution in multi-labeled corpora. In Proceedings of the 2009 conference on empirical methods in natural language Processing: Vol. 1 – Vol. 1 EMNLP ’09 (pp. 248–256). Stroudsburg, PA, USA: Association for Computational Linguistics.

[15] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

[16] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., et al. (2011). Comparing Twitter and traditional media Using topic models. In Proceedings of the 33rd European Conference on advances in information retrieval ECIR’11 (pp. 338–349). Berlin, Heidelberg: Springer-Verlag.

[17] Duan, J., & Zeng, J. (2013). Web objectionable text Content detection using topic modeling technique. Expert Systems with Applications, 40, 6094–6104.