MINING COLLECTION OF DOCUMENTS: CLUSTERING AND CATEGORIZATION

Ashok Suragala   
Department of Computer Science & Engineering  
JNTUK- University College of EngineeringVizianagaram, India  
ashok.cse@jntukucev.ac.in

Srikanth Kolli  
Department of Information Technology  
JNTUK- University College of EngineeringVizianagaram, India  
srikanth.kolli969@gmail.com

*Abstract*—Technical Documents from Web pages and News Articles, Transcripts of phone calls with customers, Customer complaint letters through E-mails, Insurance Claims is Corporate Knowledge “Ore”. Information is highly unstructured form not readily useful for statistical analysis; huge collection of documents (Big Data) is challenges in Text Mining.

[Sentiment  Analysis](https://www.sciencedirect.com/topics/engineering/sentiment-analysis) (SA) is an ongoing field of research in text mining field. SA is the computational treatment of opinions, sentiments and subjectivity of text.

The main contributions of this paper include the sophisticated Clustering and categorizations of a large number of recent articles and the illustration of the recent trend of research in the sentiment analysis and its related areas.

The corpus-based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical, two phase approach for text mining- Feature extraction and Information Distillation.

 In Text Categorization, document collection are processed and grouped into categories that are predetermined based on a user-provided taxonomy.

In Clustering, document collection are processed and grouped into clusters that are dynamically generated by algorithm.

Information from micro-blogs, blogs and forums as well as news source, is widely used in Text mining. This media information plays a great role in expressing people’s feelings, or opinions about a certain topic or product. Using social network sites and micro blogging sites as a source of data still needs deeper analysis. There are some real world data sets movie-pang02, chicago-affnia especially in reviews which are used for Machine Learning Classification and Clustering Algorithms evaluation using R tool.

Keywords—Text Mining, corpus, tf-idf, Partition based clustering. Naive Bayes classification, movie-pang02, chicago-affnia

# Introduction

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people’s opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. The two expressions SA or OM are interchangeable. They express a mutual meaning. However, some researcher stated that OM and SA have slightly different notions [1]. Opinion Mining extracts and analyzes people’s opinion about an entity while Sentiment Analysis identifies the sentiment expressed in a text then analyzes it.

Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity as shown in Figure 1. Sentiment Analysis can be considered a classification process as illustrated in Figure 1. There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions. Wilson et al. have pointed out that sentiment expressions are not necessarily subjective in nature. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents. Classifying text at the document level or at the sentence level does not provide the necessary detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity like this sentence ‘‘Good but not without problems’’. This survey tackles the first two kinds of SA. These fields include Emotion Detection (ED), Building Resources (BR) and Transfer Learning (TL). Emotion detection aims to extract and analyze emotions, while the emotions could be explicit or implicit in the sentences. Transfer learning or Cross-Domain classification is concerned with analyzing data from one domain and then using the results in a target domain. Building Resources aims at creating lexica, corpora in which opinion expressions are annotated according to their polarity, and sometimes dictionaries.

# I.I. need for text mining:

# I.II. HOW TEXT MINING IS DIFFERS FROM DATA MINING:

In Data Mining process we first identify data set , apply various pre-processing techniques , select features and analyze distribution.

where as in Text Mining, identify documents data ,apply various pre-processing techniques, extract features and construct structured data , select features and analyze distributions.

The key difference is extraction of features and construct structured data from un-structured document data.

# I.III. SENTIMENT ANALYSIS PROCESS:

Overall process of Sentiment Analysis process is depicted in the below figure.

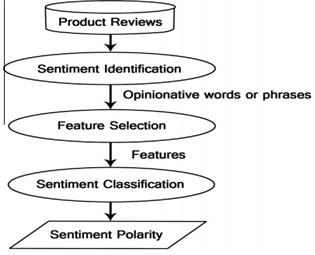


Figure 1. Sentiment analysis process on product reviews

These fields include Emotion Detection (ED), Building Resources (BR) and Transfer Learning (TL). Emotion detection aims to extract and analyze emotions, while the emotions could be explicit or implicit in the sentences. Transfer learning or Cross-Domain classification is concerned with analyzing data from one domain and then using the results in a target domain. Building Resources aims at creating lexica, corpora in which opinion expressions are annotated according to their polarity, and sometimes dictionaries.

# I.IV. SENTIMENT CLASSIFICATION TECHNIQUES:

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach(ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. The text classification methods using ML approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is difficult to find these labeled training documents.

The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach.

The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods. There is a brief explanation of both approaches’ algorithms and related articles in the next subsections.

I.IV.I. MACHINE LEARNING APPROACH:

Machine learning approach relies on the famous ML algorithms to solve the SA as a regular text classification problem that makes use of syntactic and/or linguistic features. Text Classification Problem Definition: We have a set of training records D={X1,X2,...,Xn} where each record is labeled to a class. The classification model is related to the features in the underlying record to one of the class labels. Then for a given instance of unknown class, the model is used to predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.

I.IV.II.NAIVE BAYES CLASSIFIER:

The Naı¨ve Bayes classifier is the simplest and most commonly used classifier. Naı¨ve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

I.IV.III. LEXICON BASED APPROACH:

Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together are called opinion lexicon. There are three main approaches in order to compile or collect the opinion word list. Manual approach is very time consuming and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The two automated approaches are presented in the following subsections.

A) DICTIONARY-BASED APPROACH:

presented the main strategy of the dictionary-based approach. A small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well known corpora Word Net or thesaurus for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the pro-cess is completed, manual inspection can be carried out to remove or correct errors. The dictionary based approach has a major disadvantage which is the inability to find opinion words with domain and context specific orientations.

B) CORPUS-BASED APPROACH:

The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus. One of these methods were represented by Hatzivassiloglou and McKeown. They started with a list of seed opinion adjectives, and used them along with a set of linguistic constraints to identify additional adjective opinion words and their orientations. The constraints are for connectives like AND, OR, BUT, EITHER-OR. The conjunction AND for example says that conjoined adjectives usually have the same orientation. This idea is called sentiment consistency, which is not always consistent practically. There are also adversative expressions such as but, However which are indicated as opinion changes. In order to determine if two conjoined adjectives are of the same or different orientations, learning is applied to a large corpus. Then, the links between adjectives form a graph and clustering is per-formed on the graph to produce two sets of words: positive and negative.

I.V. FEATURE SELECTION IN SENTIMENT ANALYSIS:

Sentiment Analysis task is considered a sentiment classification problem. The first step in the SC problem is to extract and select text features. Some of the current features are **Terms presence and frequency:** These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears, or one if otherwise) or uses term frequency weights to indicate the relative importance of features. **Parts of speech(POS):** finding adjectives, as they are important indicators of opinions. **Opinion words and phrases:** these are words commonly used to express opinions including good or bad, like or hate. On the other hand, some phrases express opinions without using opinion words. For example: cost me an arm and a leg. **Negations:** the appearance of negative words may change the opinion orientation like not good is equivalent to bad.

I.V.I. FEATURESELECTION METHODS:

Feature Selection methods can be divided into lexicon-based methods that need human annotation, and statistical methods which are automatic methods that are more frequently used. Lexicon-based approaches usually begin with a small set of ‘seed’ words. Then they bootstrap this set through synonym detection or on-line resources to obtain a larger lexicon. This proved to have many difficulties as reported by Whitelaw et al. Statistical approaches, on the other hand, are fully automatic. The feature selection techniques treat the documents either as group of words (Bag of Words (BOWs)), or as a string which retains the sequence of words in the document. BOW is used more often because of its simplicity for the classification process. The most common feature selection step is the removal of stop-words and stemming (returning the word to its stem or root i.e. flies ->fly). In the next subsections, we present three of the most frequently used statistical methods in FS and their related articles. There are other methods used in FS like information gain and Gini index.

# Literature Review

In this study, numerous sources in literature were reviewed. Reviews were focused on Sentiment analysis using machine learning approach.

The approaches used in this sentiment analysis include lexicon based approach naïve bayes classification. Currently, Sentiment Analysis concentrates for subjective statements or on subjectivity and overlook objective statements which carry sentiment(s).

[1] Pang and Lee and Liu. They focused on the applications and challenges in SA. They mentioned the techniques used to solve each problem in SA.

[2]Cambria and Schuller et al., Feldman and Montoyo and Martı´nez-Barco. have given short surveys illustrating the new trends in SA. And different approaches mentioned.

[3]Tsytsarau and Palpanas. have presented a survey which discussed the main topics of SA in details. For each topic they have illustrated its definition, problems and development and categorized the articles with the aid of tables and graphs. The analysis of the articles presented in this survey is similar to what was given but with another perspective and different categorization of the articles.

[4]Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings.

[5]Jalaj et al.[2] discussed about exiting methods, approaches to do sentimental analysis for unstructured data which reside on web here, they proposed new approach to classify and handle subjective as well as objective statements for sentimental.

[6]Popowich et al.[8] developed an application. The application makes use of a natural language processing (NLP) engine, together with application-specific knowledge, written in a concept specification language. Using NLP techniques, the entities and relationships that act as indicators of recoverable claims are mined from management notes, call centre logs and patient records to identify medical claims that require further investigation.

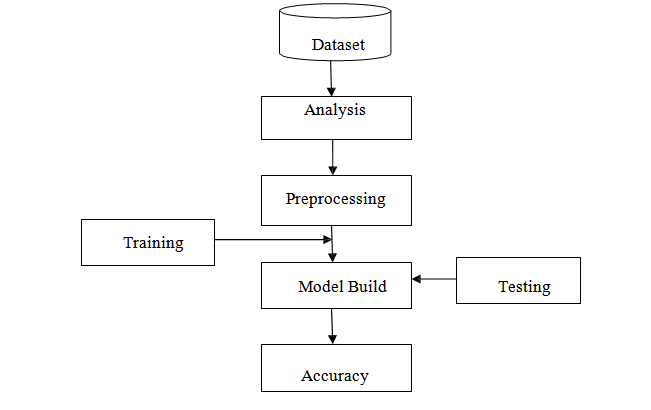
# methodology

Methodology is a framework that is used to structure, plan, and control the process of developing an information system.

III.I. PROBLEM STATEMENT:

The objective of this project is to develop naïve bayes classifier and clustering model for prediction of Chicago affnia and movie pang data. This project shows how data mined Chicago affnia and movie pang can be used to develop an objective model for prediction of different reviews and opinions. Various classification algorithms like naïve bayes classification and clustering algorithm like k-means clustering applied on these data sets compare with these techniques.

III.II. system architecture:



III.III.I.TEXT DATA ANALYSIS:

Data analysis is also known as analysis of data or data analytics is a process of inspecting, cleaning, transforming and modeling data with goals of discovering useful information, suggesting conclusions and support data making.

III.III.II.TEXT PRE PROCESSING:

Text pre-processing is the initial step of text mining which reads one text document at time and processes it. This step divides into following main three subtasks.

PRE PROCESSING OF TEXT DOCUMENTS:

1. The first step when faced with a raw text document is to stem the words.
2. This means that one cuts words to their root: for example, "tax" from taxing, taxes, and taxation.
3. The next step is to search the text documents for a list of stop words containing irrelevant words marked for removal.
4. if, and, but, who, what, the, they, their, a, or, and so on are examples of stop words that need to be removed.
5. Also, one usually removes words that are extremely rare using "tf-idf " measure.
6. For example, in a corpus of n documents one can base the screening of words on the tf-idf (term frequency/inverse document frequency) score.

tf-idf = fij \*log(n/dj)

fij = the (relative) frequency of word j in document i

**n** is the number of documents

dj the number of documents containing word j.

III.III.III.TOKENIZATION:

Generally text document contains multiple sentences. So this process divides whole sentence into words by removing comma, spaces, punctuations etc.

III.III.IV. STOP WORDS REMOVING:

This process removes stop words such as “the”, “are”, “a” or any tags like HTML tag etc.

III.III.V. STEMMING:

Stemming is applied after stop word removal by reducing the word to its root word. E.g. “playing”, “played” are stemmed to “play”.

III.III.VI. TEXT TRANSFORMATION:

Text transformation has the role of conversion of text document into words so that it will useful for further processing.

FEATURE EXTRACTION: Extraction of features and construct structured data.

Before construction Bag of words representation we must pre-process the document data.

The Analysis of Phrase counts from text documents is the current state of the art.

The "bag of words" representation of text assigns frequencies to words or combinations of words.

Divide the number of occurrences of each word in a document by the total number of words in the documents.

However, considerable preprocessing of text is needed before one can obtain frequency information on words and before one can start the statistical analysis.

III.III.VII. FEATURE SELECTION:

It performs removing features that are considered unrelated for mining purpose.

III.III.VIII. PATTERN DISCOVERY OR ANALYZE DISTRIBUTIONS:

Pattern discovery is one of the important processes that use methods for discovering patterns. Methods include clustering, classification, summarization, information retrieval, topic extraction etc.

III.III.IX. MODEL BUILDING:

Model build means based on data set feature selection missing value removal outlier detection we build the model.

III.III.X. TRAINING:

Dataset can be divided into training and testing. Training data and testing data not be equal in my project training data is 70% in my data set. Training data is given to the model as input and result will be taken as output of the training data.

III.III.XI. TESTING:

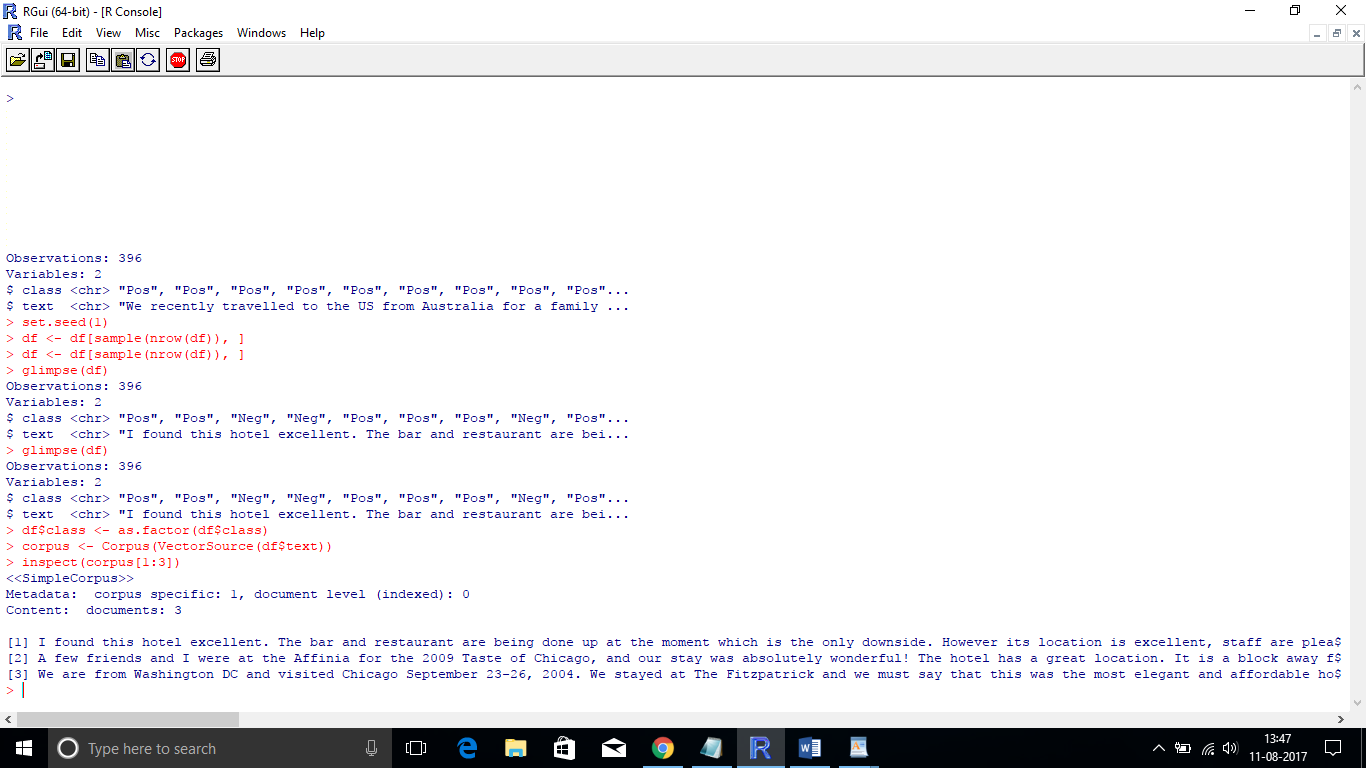
In data set remaining 30% are considered as testing data. Testing data is given to the model after training data output will be obtained. Both training and testing data result will be approximately equal.

III.III.XII. RESULT:

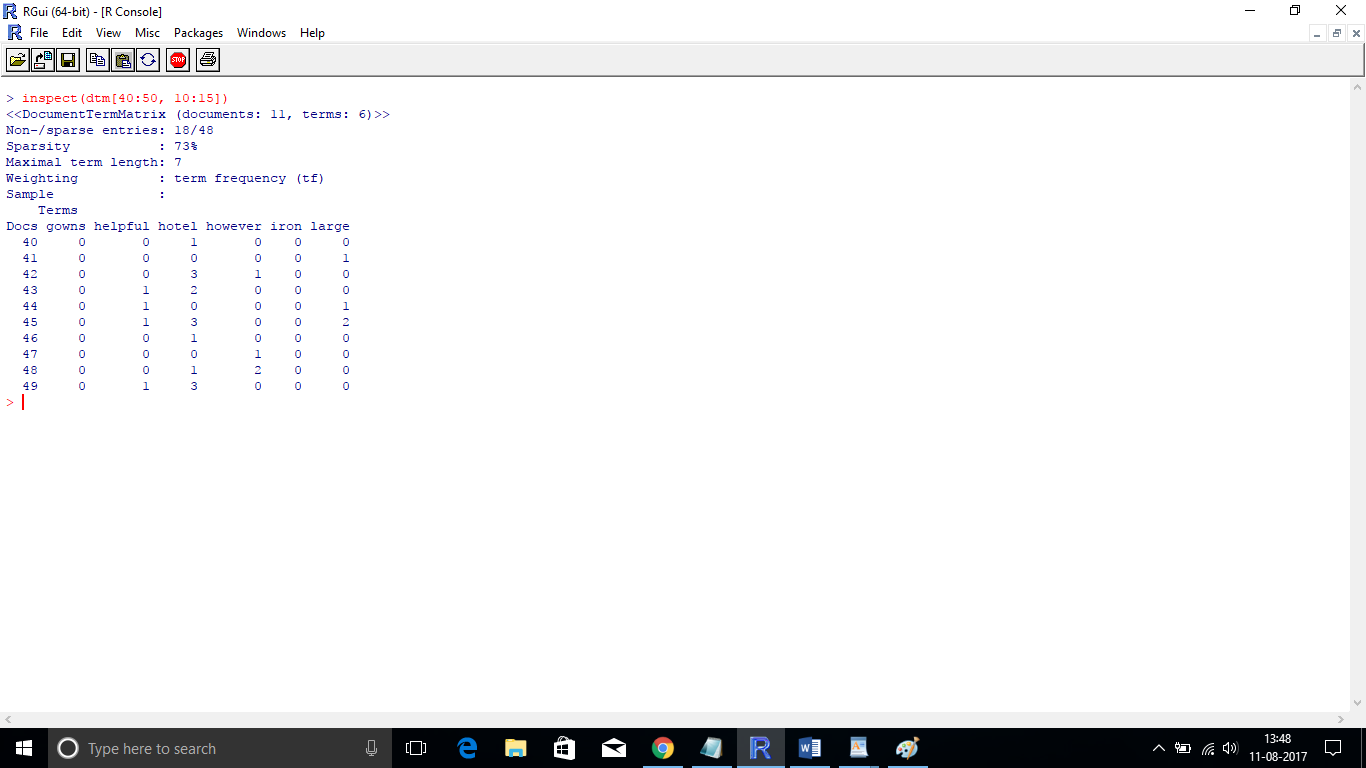
Result for the project is both training and testing data outputs. Both will be approximately equal then data set is fitted to the model.

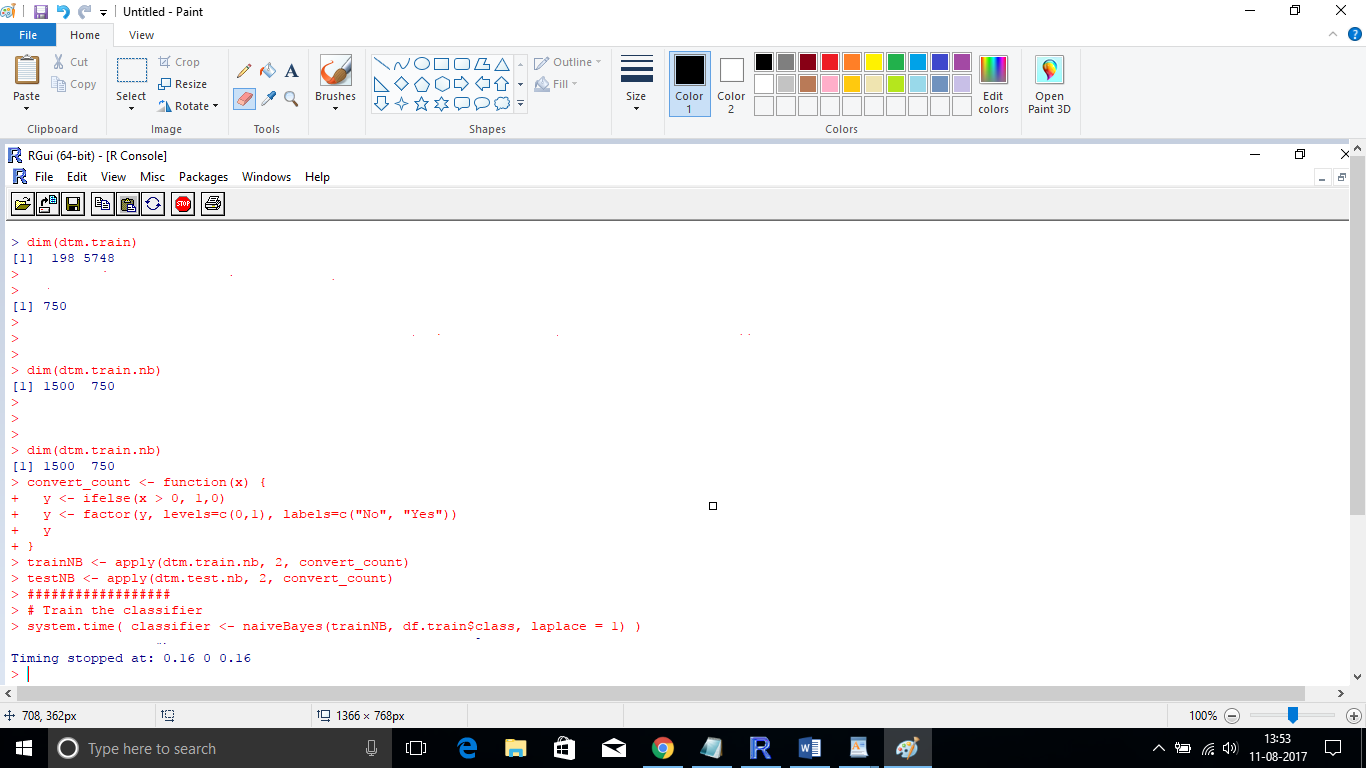
# FINDINGS

IV.I. INFORMATION EXTRACTION:

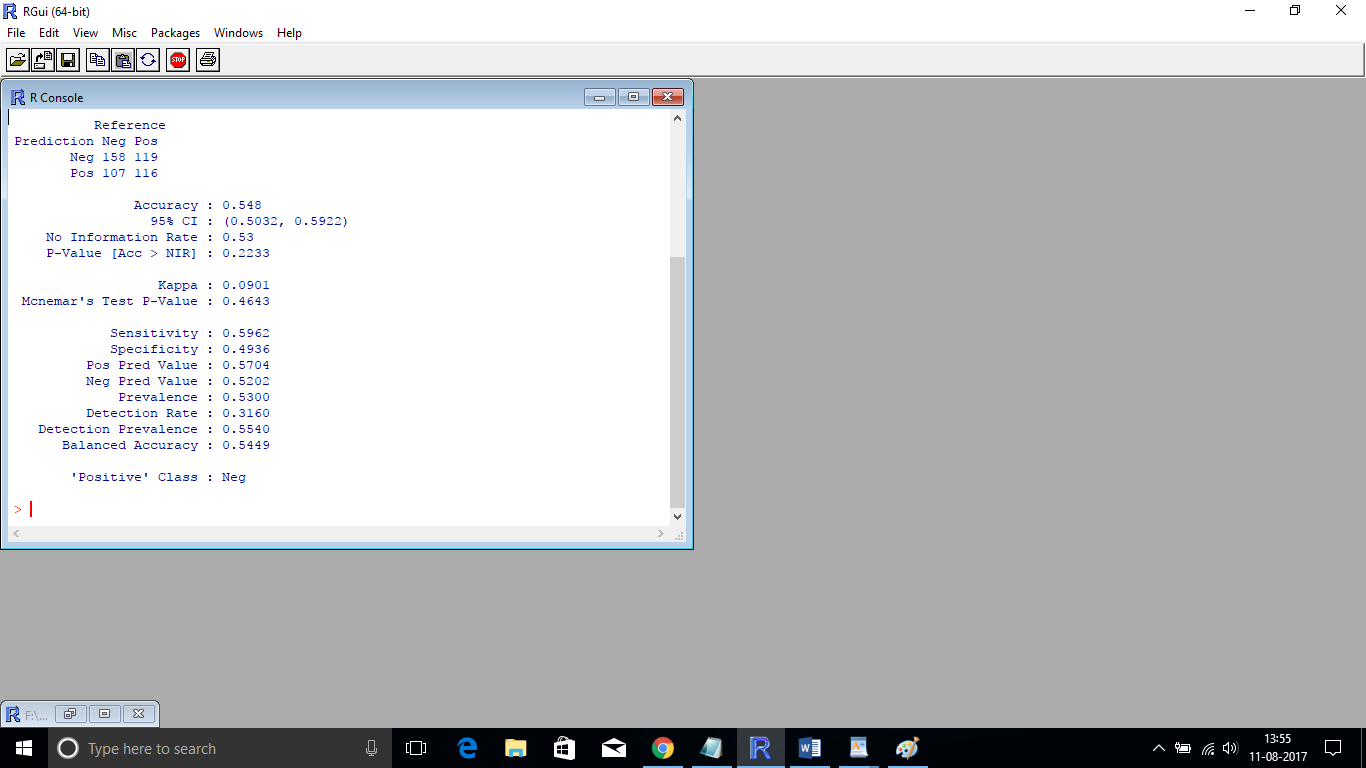


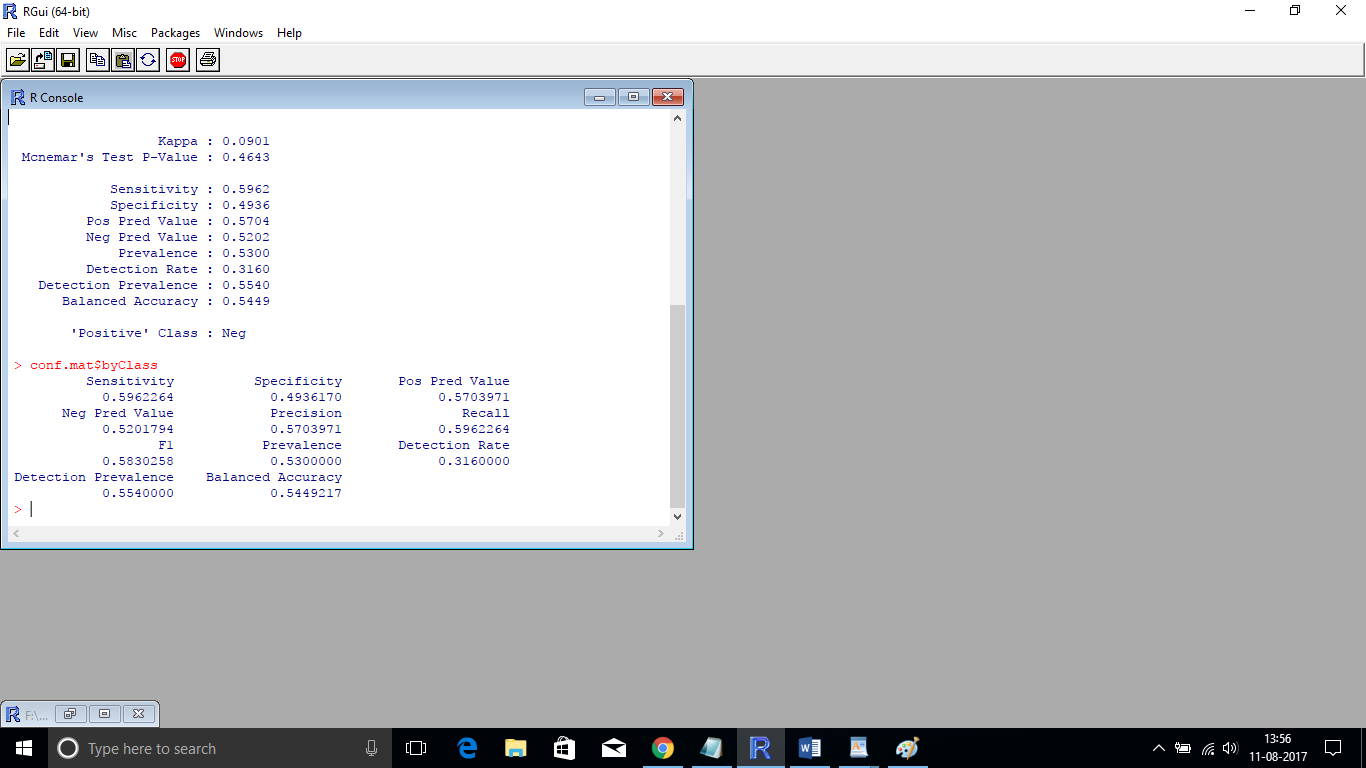
IV.II.NLP:



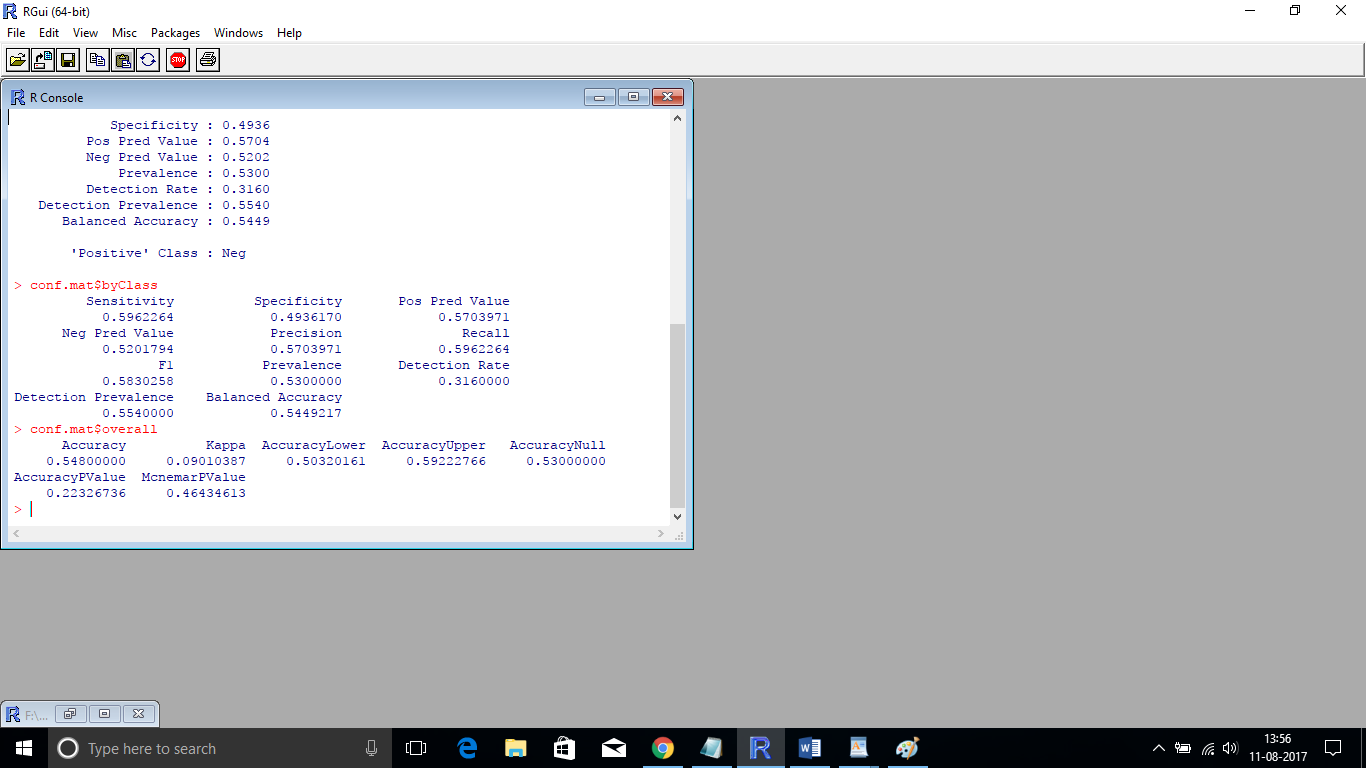


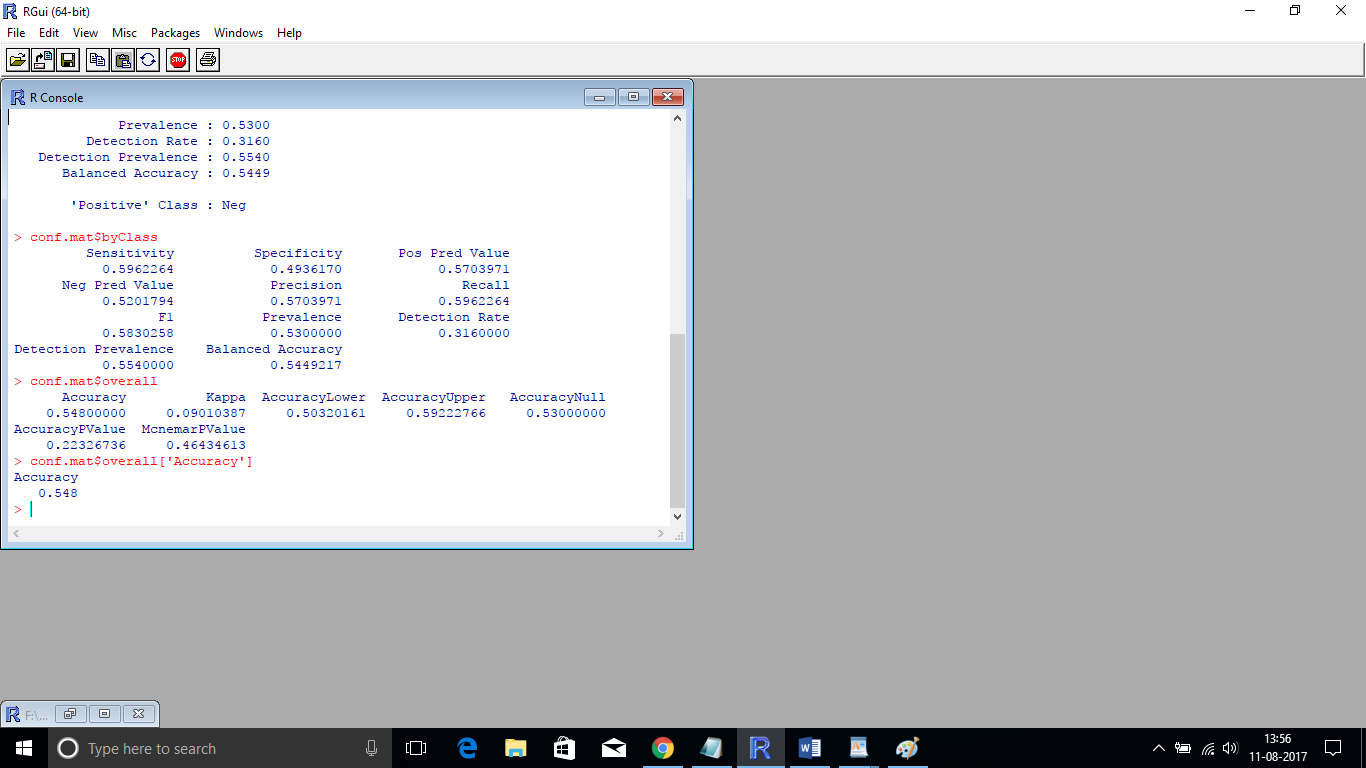
CLASSIFIER:

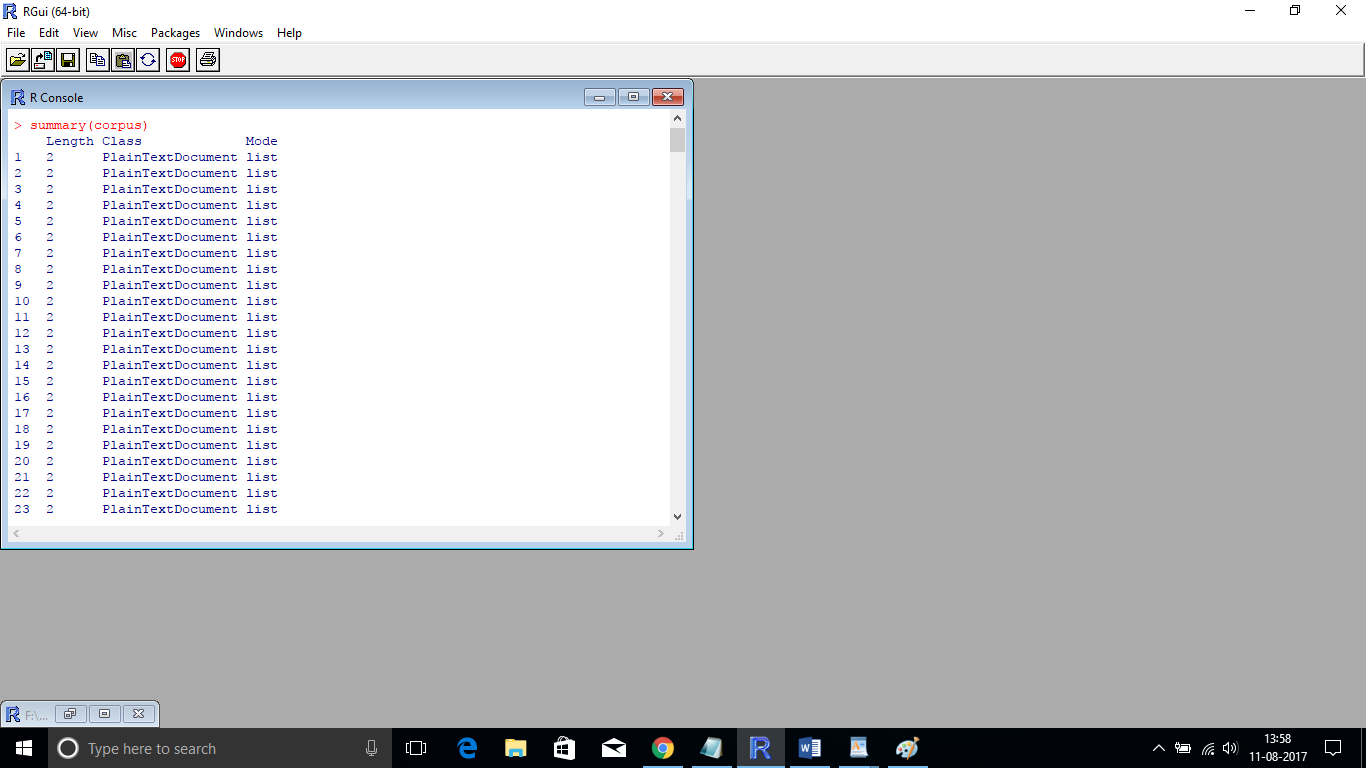


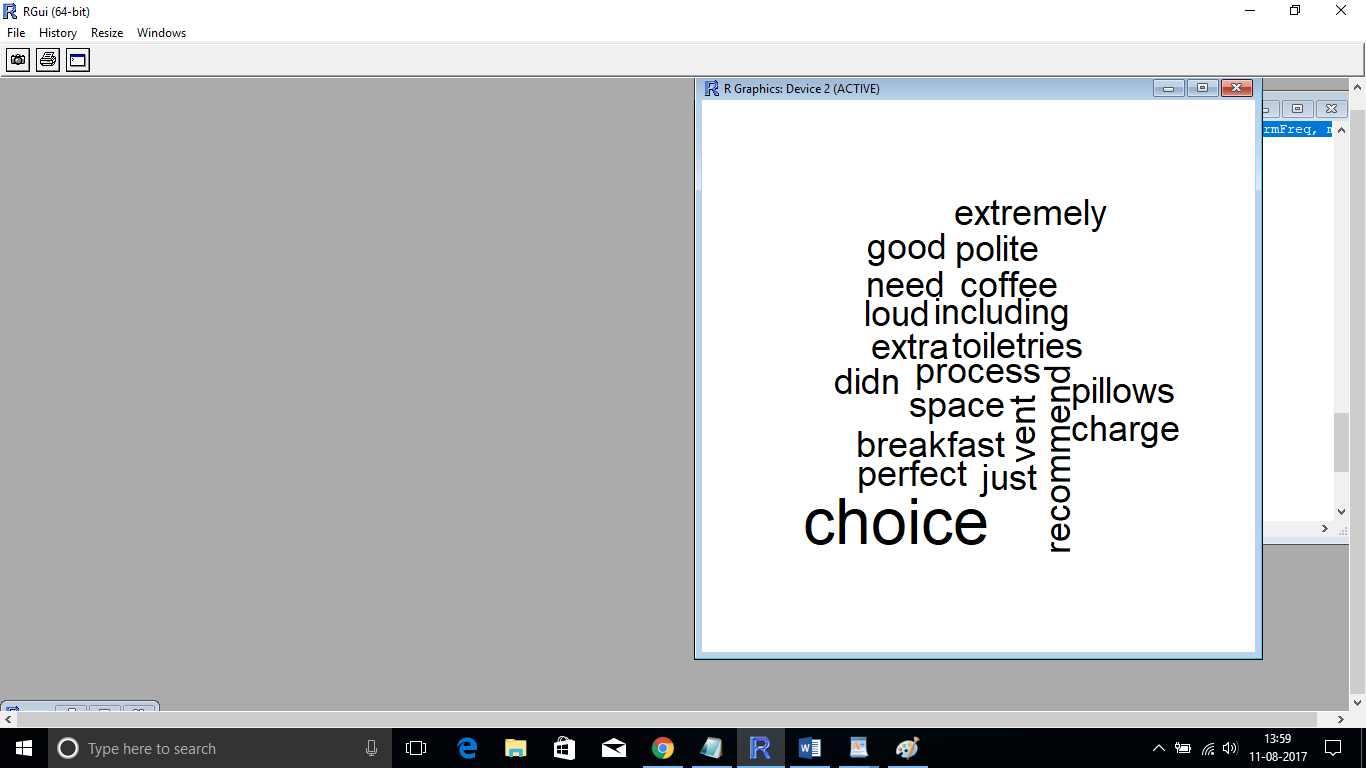


NAÏVE BAYES CLASSIFICATION:

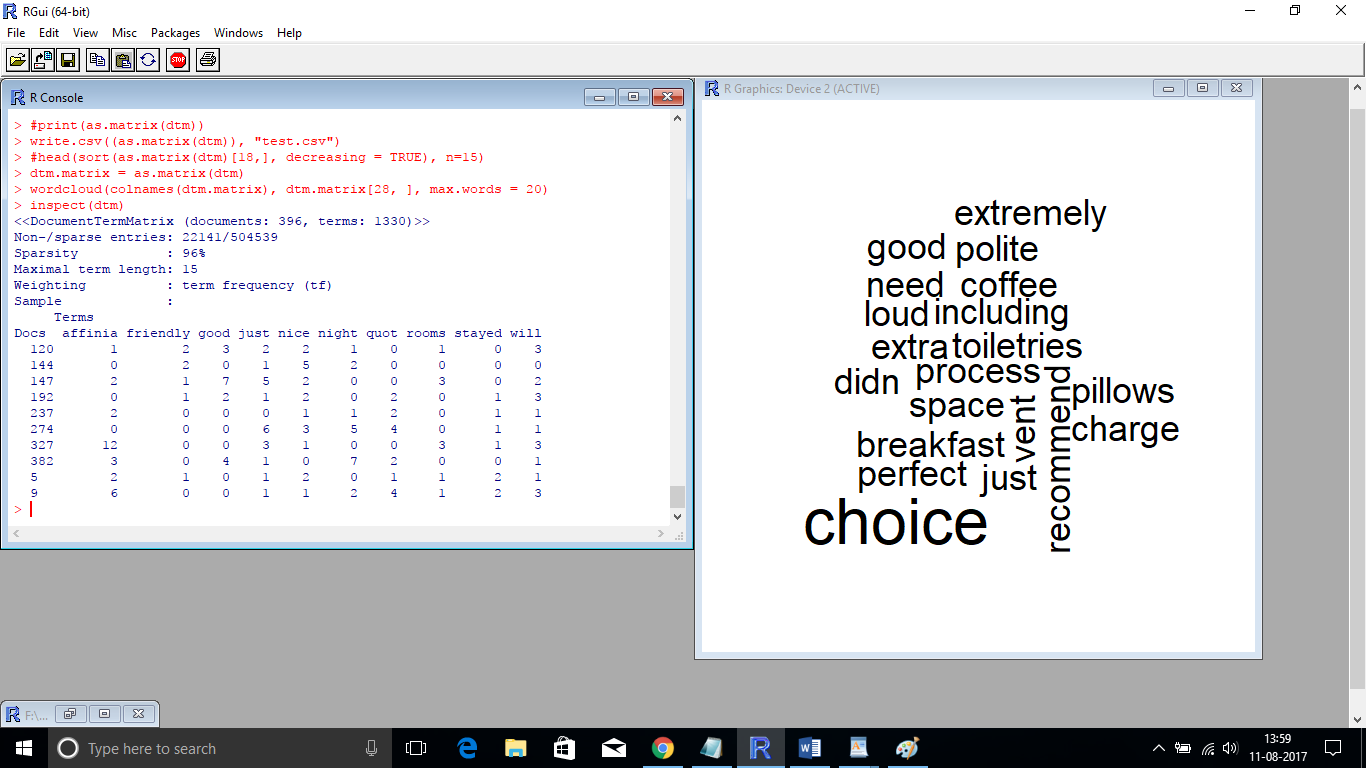




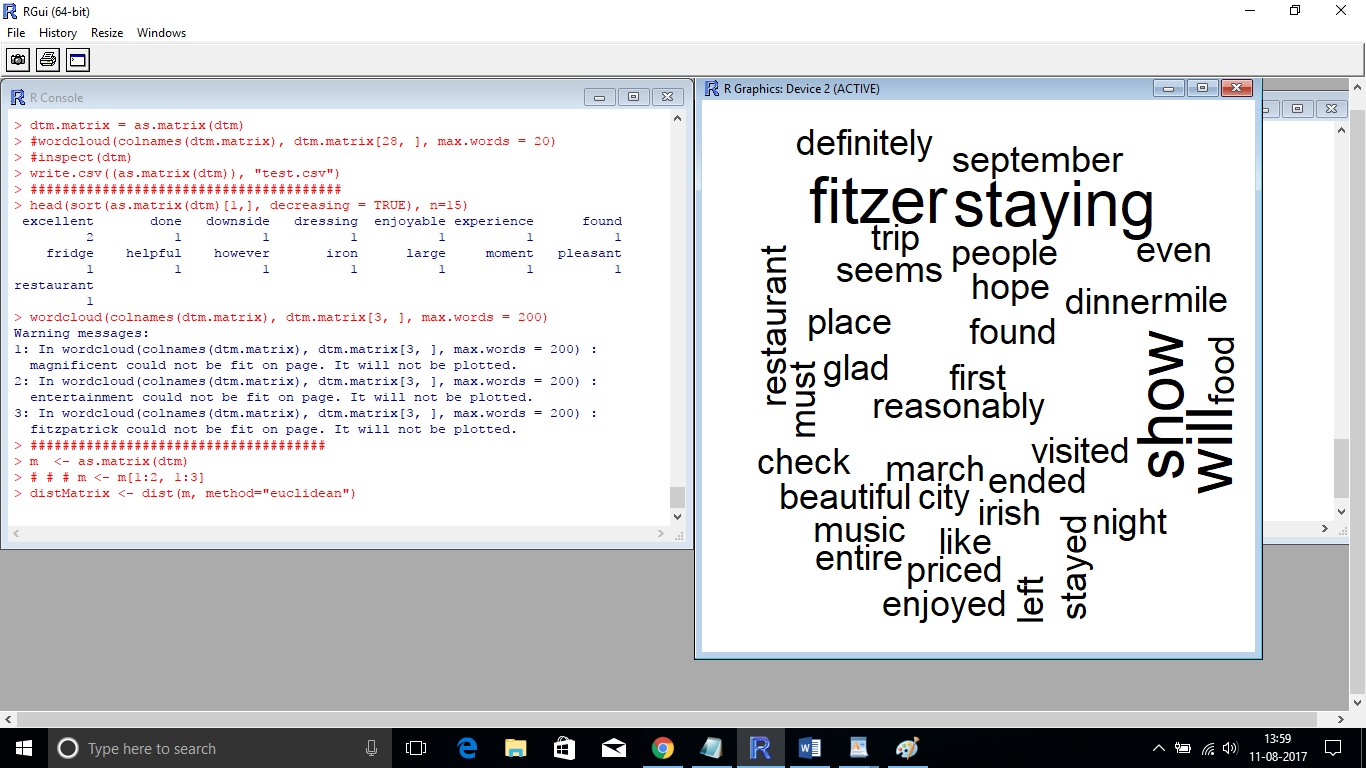




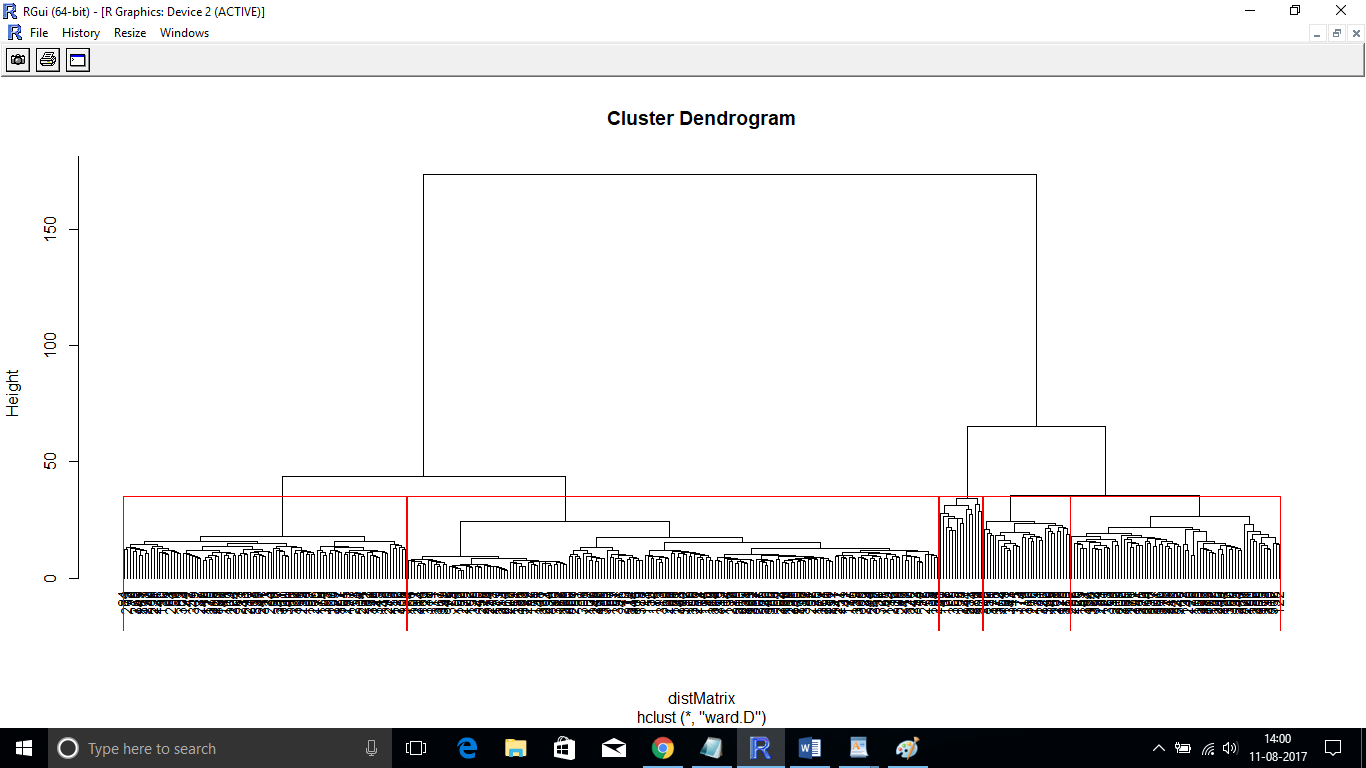
NLP:



WORD CLOUD:



CLUSTERING



V. CONCLUSIONS / IMPLECATIONS

This study examined the performance sentiment analysis two different approaches machine learning approach and lexicon based approach. the enhancements of SC and FS algorithms are still an open field for research. Naı¨ve Bayes and Support Vector Machines are the most frequently used ML algorithms for solving SC problem. The most common lexicon source used is Word Net which exists in languages other than English. Building resources, used in SA tasks, is still needed for many natural languages. Information from micro-blogs, blogs and forums as well as news source, is widely used in SA recently. This media information plays a great role in expressing people’s feelings, or opinions about a certain topic or product. Using social net-work sites and micro blogging sites as a source of data still needs deeper analysis. There are some benchmark data sets especially in reviews which are used for algorithms evaluation.

V.I. FUTURE WORK:

In future this methodology is applied for remaining Techniques Support Vector Machine(SVM), corpus based technique and different deep learning techniques for better results.

VI.REFERENCES

[1] Tsytsarau Mikalai, Palpanas Themis. Survey on mining subjective data on the web. Data Min Knowl Discov 2012;24:478–514.

[2] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/ EMNLP; 2005.

[3] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.

[4] Yu Liang-Chih, Wu Jheng-Long, Chang Pei-Chann, Chu Hsu-an-Shou. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl-Based System 2013;41:89–97.

[5] Michael Hagenau, Michael Liebmann, Dirk Neumann. Auto-mated news reading: stock price prediction based on financial news using context-capturing features. Decis Supp Syst; 2013.

[6] Tao Xu, Peng Qinke, Cheng Yinzhao. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. Knowl-Based Syst 2012;35:279–89.

[7] Maks Isa, Vossen Piek. A lexicon model for deep sentiment analysis and opinion mining applications. Decis Support Syst 2012;53:680–8.

[8] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.

[9] Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013;28:15–21.

[10] Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013;56:82–9.

[11] Montoyo Andre´s, Martı´nez-Barco Patricio, Balahur Alexandra. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 2012;53:675–9.

[12] Lu Cheng-Yu, Lin Shian-Hua, Liu Jen-Chang, Cruz-Lara Samuel, Hong Jen-Shin. Automatic event-level textual emotion sensing using mutual action histogram between entities. Expert Syst Appl 2010;37:1643–53.

[13] Cao Qing, Duan Wenjing, Gan Qiwei. Exploring determinants of voting for the ‘‘helpfulness’’ of online user reviews: a text mining approach. Decis Support Syst 2011;50:511–21.

[14] Fan Teng-Kai, Chang Chia-Hui. Blogger-centric contextual advertising. Expert Syst Appl 2011;38:1777–88.

[15] Zhou L, Li B, Gao W, Wei Z, Wong K. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Presented at the 2001 conference on Empirical Methods in Natural Language Processing (EMNLP’11); 2011.

[16] Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F. Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM’11); 2011.

[17] Zirn C, Niepert M, Stuckenschmidt H, Strube M. Fine-grained sentiment analysis with structural features. In: Presented at the 5th International Joint Conference on Natural Language Processing (IJCNLP’11); 2011.

[18] Hu Nan, Bose Indranil, Koh Noi Sian, Liu Ling. ‘‘Manipulation of online reviews: an analysis of ratings, readability, and sentiments’’. Decis Support Syst 2012;52:674–84.

[19] Gupta Sunil Kumar, Phung Dinh, Adams Brett, Venkatesh Svetha. Regularized nonnegative shared subspace learning. Data Min Knowl Discov 2012;26:57–97.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.