Review on Text Mining

Aarushi Rai[#1], Aarush Gupta[*2], Jabanjalin Hilda J.[#3]

[#1] School of Computer Science and Engineering, VIT University, Tamil Nadu - India
[#2] School of Computer Science and Engineering, VIT University, Tamil Nadu - India
.[#3]Associate Professor, School of Computer Science and Engineering, VIT University, Tamil Nadu - India

ABSTRACT:
Text data mining is the process of deriving high quality information from text, it is a relatively novel area of computer science and the use of it has grown as the unstructured data available continues to increase exponentially. In the present day there has been a surge in the amount of textual information available on the internet. Text mining has applications in several areas such as customer care service, fraud detection, contextual advertisement, healthcare etc.

KEYWORDS: data, text mining, classification, clustering

INTRODUCTION:
Data mining is the method to find meaningful answers or patterns from a large dataset.The primary goal of data mining is to fetch information from a database and transform it into a data set which is meaningful to the user. The user can then decide how he/she wishes to use the data to find answers to various questions. Data mining is still a very evolving technology. Using the techniques and concepts of data mining on a database consisting of text refers as text mining. Data mining can be used to solve a lot of business problems. Some of those business problems could be like answering questions such as, "How should one shop place the items in order to increase sales" or "How should business advertise in order to target areas better". Using text mining in order to solve such problems is referred as text analytics. Text mining can allow organizations to obtain valuable business insights of the market from text-based sources such as word documents, emails, pdfs and various social media platforms.

A recent study showed that almost 80% of a company's information is stored in the form of text. Since the amount of information on text is so high, text mining is believed to have a huge scope. Text mining is a complex task since most of the text data is unstructured data, which requires cleaning in order to get good results. Text mining involves a lot of steps within it such as information retrieval, clustering, visualization, machine learning, data mining etc.

TEXT DATA
The text data can be stored in various formats such as word documents, pdfs, social media etc. It can mainly be categorized into two types:

(i) Structured Data
(ii) Unstructured Data

STRUCTURED DATA
Structured data is the format of data that can be loaded spreadsheet. Structured text data can be stored in the form of rows and columns. Structured text data that is present in a spreadsheet may or may not have each cell filled. Such kind of data is consistent, uniform and data mining friendly i.e. after applying various text mining techniques on it, it is expected that more accurate and desired results will be generated.

UNSTRUCTURED DATA
Unstructured data is generally present in the form of word documents, html web pages, pdf documents. Unstructured data is inconsistent and isn't data mining friendly i.e. it won't produce expected results. Structured data is often converted into a semi-structured data first to perform data mining techniques. Semi-structure data is mostly in XML format.

STAGES OF TEXT MINING

INFORMATION RETRIEVAL
Information retrieval (IR) is the primary step in the process of text mining. IR is the process of extracting relevant information from all the sources. It involves the use of a query written by the user to collect all the documents that have relevant information. One of the example of IR would be search engines such as Google. Search engines run the query to match all the relevant documents present on World Wide Web to the keywords entered by the user.

NATURAL LANGUAGE PROCESSING
Human language for a computer can be difficult to process due to the subtle nuances. For example, synonyms are words which are spelled same but have different meaning. The meaning of the sentence can only be inferred through the context in which it is used. Natural Language Processing (NLP) is a complex task and is a part of the artificial intelligence domain. It refers

to the inference of the human language by computers. Shallow parsers categorize only the major grammatical components in a sentence like nouns, adjectives and verbs. On the other end, deep parsers build an extensive characterization of the grammatical structure of a sentence. NLP is the domain of text mining is responsible for transferring data in the form of linguistic data to the information extraction phase.

INFORMATION EXTRACTION

After linguistic data is produced by natural language processing it is transferred to the information extraction(IE) stage. IE stage is responsible for structuring the linguistic data. The data after being structured in this stage is transferred to the data mining stage where the text analysis occurs.

DATA MINING

Data mining is a method that helps to make sense of a huge amount of information. It is used to find meaningful patters and answers from large data sets. When used in text mining, the techniques or algorithms of data mining are applied on the structured data processed from the information extraction stage. The results obtained from the data mining stage are stored into another database which can be accessed by end-user queries.

FRAMEWORK OF TEXT MINING

Text mining can broadly be visualized in two phases:

(i) Text Refining
In the text refining stage, the input is data in the form of text and the output is an intermediate form. Intermediate form can be either semi-structured/structured.

(ii) Knowledge Distillation
In the knowledge distillation phase the input is the intermediate form generated in the text refining phase and the output are the different patterns mined from the data.

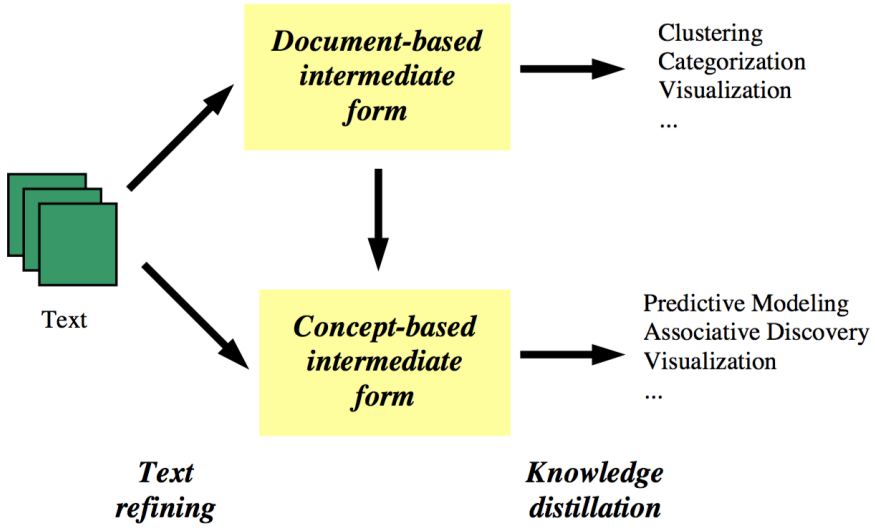

Figure i Framework of text mining

TEXT MINING ALGORITHMS

The text mining algorithms can be broadly classified into three categories:

(i) classification
(ii) categorization
(iii) clustering

K-MEANS CLUSTERING ALGORITHM

Clustering is the method of dividing a group of data points into a small number of clusters. K-Means clustering is an unsupervised learning algorithm that is used to partition data set into k groups. Each group or cluster has a cluster centroid. The main goal of the k-means algorithm is to reduce the total sum of the squared distance of every point in the data set to its corresponding cluster centroid. Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq$ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares where μi is the mean of points in Si.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Advantages to Using this Technique
K-Means clustering algorithm has an advantage over other clustering methods such as hierarchical clustering when there is a large number of variables. K-means will be faster in computation. Also, K-Means produce tighter clusters if the cluster are global.

Disadvantages to Using this Technique
The main disadvantage of using K-Means clustering is being unable to compare the quality of the clusters generated. K-means algorithm doesn't work well when it comes to non-globular clusters. Also, different iterations of initial partition will eventually end up in an output of different results.

APPLICATIONS OF K-MEANS CLUSTERING

- Unsupervised learning of neural networks
- Machine Vision
- Pattern recognition
- Artificial intelligence
- Classification analysis
- Image Processing

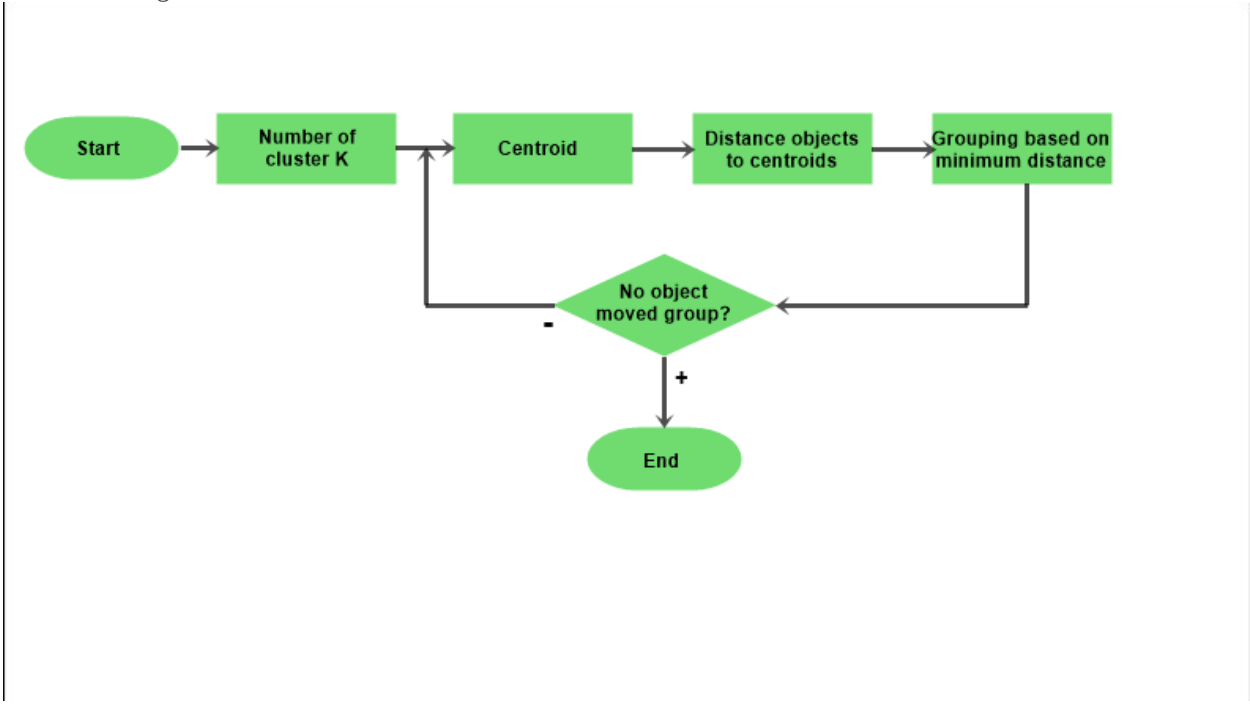

Figure ii Flow of K-Means

NAIVES BAYES
Naïve Bayes classifiers are linear classifiers. The term naïve refers to the assumption that the features in the database are mutually independent and Bayes comes from the theoretical use of Bayes theorem. In real time scenarios, the assumption of two independences are disobeyed. However, Naïve Bayes classifiers works better than most of the alternatives for a small sample size. The Naiver Bayes classifiers seem to find a lot of lot of application in various fields since they are relatively robust, easy to implements, fast and accurate.

$$\text{posterior probability} = \frac{\text{conditional probability} \cdot \text{prior probability}}{\text{evidence}}$$

LINEAR SUPPORT VECTOR MACHINE LEARNING CLASSIFICATION
Support vector machine learning (SVM) is a very modern algorithm used for regression and classification.

SVM has a procedure for optimizing predictive accuracy. SVM forwards the input data into a kernel space and later builds a linear model within this kernel space that is automatically chosen by SVM. SVM executes well with real world applications such as text classification, recognition of hand –written symbols, image classification etc. For the purpose of machine learning and data mining SVM is the most widely used tool.

SVM works best with sparse data and produces wonderful results with high-dimensional challenging data. SVM makes use of active learning. Active learning refers to the fact that as the size of the SVM model increases so does the size of the training sets.

TEXT MINING PROCESS FLOW

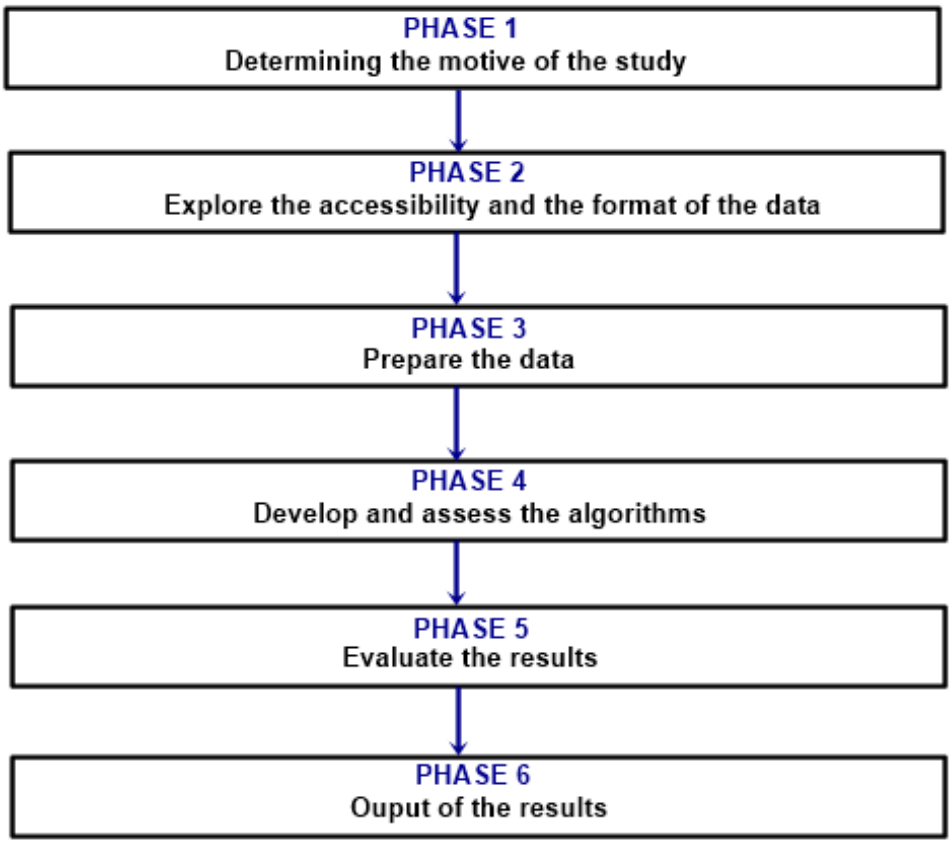

Figure iii Process flow of Text Mining

LIMITATIONS OF TEXT MINING

The main difficulties faced during text mining is that the text data consists of language which is ambiguous, Misspellings, abbreviations and spelling variants can also cause difficulties and error in finding the correct result.

Text mining is the process of finding patterns and answers from data. These data are already present and have these patterns hidden in them. We use the process of data mining or text mining to uncover these hidden patters. All in all, text mining doesn't generate any new facts. The limitations in text mining occurs due to unavailability of structured clean data. Apart from that since natural language processing is still an evolving domain, text mining is unable to generate more accurate results. This happens because the human language is difficult. The text data generated from it is ambiguous and contains some words can have the same spelling but different meaning(homographs) or different words that can have the same meaning (synonyms).

ACKNOWLEDGEMENT:
The authors are grateful to the authorities of School of Computer Science and Engineering of VIT University, Vellore.

REFERENCES:

1. A Comparison Of Document Clustering Techniques: Michael Steinback, George Karypis and Vipin Kumar
2. A tutorial review on Text Mining Algorithms: Mrs. Sayantani Ghosh, Mr. Sudipta Roy and Prof. Samir K. Bandyopadhyay
3. A Survey of Text Mining Techniques and Applications: Vishal Gupta and Gurpreet S. Lehal
4. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, LS-8 Report 23, Thorsten Joachims
5. A review on various text mining techniques and algorithms: R.Balamurugan, Dr. S.Pushpa
6. Text Mining Process, Techniques and Tools: an Overview, Vidhya. K.A , G.Aghila
7. Text Mining: The state of the art and the challenges
8. https://www.experfy.com/blog/k-means-clustering-in-text-data
9. https://web.stanford.edu/class/cs124/lec/naivebayes.pdf
10. https://www.slideshare.net/treparel/support-vector-machines-svm-text-analytics-algorithm-introduction-2012