IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Special Issue 2, December 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Text Mining: A Comprehensive Review

M.Janaki M.C.A,M.Phil(Ph.D)

Asst. Professor, Department of Computer Science, Sacred Heart Degree College for Women, Bengaluru, India

ABSTRACT: Text mining is a technique which extracts information from unstructured data and find pattern which is novel and unknown earlier. It is also known as knowledge discovery from text (KDT), deals with the machine supported analysis of text documents are in semi-structured or unstructured format datasets such as emails, full-text documents, HTML files pdf.In this research paper we can analyse pre-processing, techniques, performance measure and algorithmsand future directions in text mining.

KEYWORDS: Text mining, NLP, Tools, Support Vector Machine

I. INTRODUCTION

Text mining is the process of mining the useful information from the text documents. It's also called knowledge discovery in text (KDT) or knowledge of intelligent text analysis. It extracts information from both structured and unstructured data and also finding patterns. Text mining techniques are utilized in various types of research domains like NLP, information retrieval, text classification and text clustering.

Text Mining is the tool which helps in getting the information cleaned up. It's themethod of converting unstructured text data into meaningful and actionable information. It automatically process all the data and generate valuable insights, helping companies make data-driven decision. It is used to retrieve hidden high quality information from document. Text Mining is to analyse large quantities of natural language text and it detects lexical patterns to extract useful information. Text Mining is beneficial for organization because most of the data is in text format. The following steps can be included in text mining[7]

- It converts the unstructured text into structured data.
- Identify the patterns from structured data.
- Analyse the patterns using Text Mining techniques.
- Extract the useful information from the text.

II.TEXT PREPROCESSING TECHNIQUES

Pre-processing mainly involves:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Discretization

First phase converts the data by filling in missing values, smoothens noisy data, identifies errors and removes outliers, and resolves the inconsistencies. Second phase combines data from several multiple sources into one coherent data store, using multiple databases, data-cubes or files. Third phase involves two steps: Normalization fits the information within a range, & Aggregation combines the normalized data. Fourth phase reduces the quantity of the information producing the outcome. Fifth phase is a part of data reduction that replaces the numerical attributes with nominal ones.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

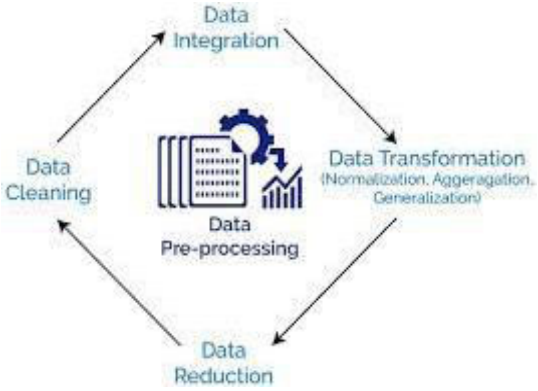


Fig.1. **Pre Processing Process**

2.1. Extraction (Tokenization)

This method is used to tokenize the file content into individual word.[1]

2.2. Stop Words Elimination

Stop words are a division in NLP. Removing stop words reduces the dimensionality of term space. The common words in text documents are articles, prepositions, and pro-nouns, etc. that does not give the meaning to the documents. These words are treated as stop words. Example: the, in, a, an, with, etc.

2.3. Stemming

This method is used to spot the root/stem of a word. For example, the words work,worked, working, connections all can be stemmed to the word "**connect**". The purpose of this method is to get rid of various suffixes, to reduce the number of words, to have accurately matching stems, to save time and memory space.[2]

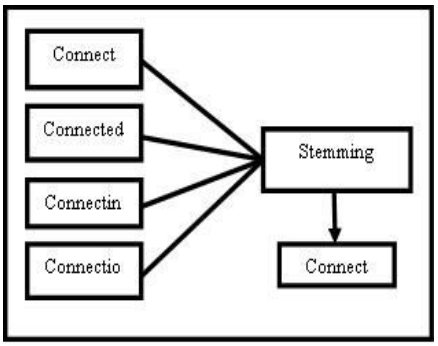


Fig.2.Stemming

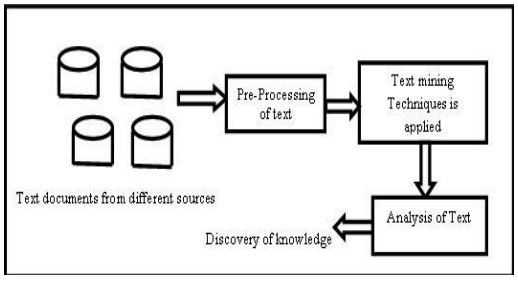


Fig 3.Structure of text mining

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

III. TEXT MINING TECHNIQUES

There are several techniques used for text mining, but most popular are Natural Language Processing (NLP) and Information Extraction (IE). Research is going on to analyse other techniques for text mining such as knowledge based, statistical, rule-based and machine learning-based approaches. NLP focuses on text processing where IE focuses on extracting information from actual text. To analyse, understand and generate text, technologies are produced by NLP. The technologies like knowledge extraction, summarization, categorization, clustering and information visualization, are used in the text mining process. In the followingsections we will discuss each of these technologies and the role that they play in text mining. The types ofsituations where each technology may be useful in order to help users are also mentioned.

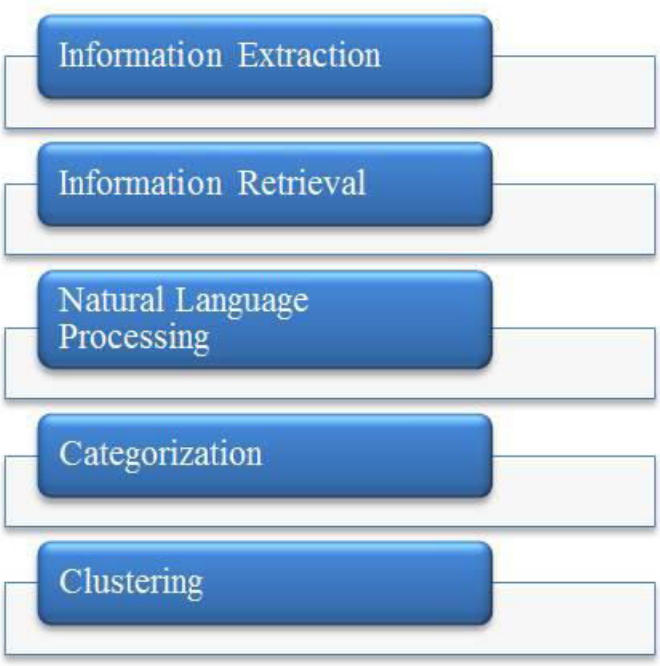


Fig:4 Techniques

1. **Natural Language Processing**
2. Information Retrieval
3. Information Extraction
4. Summarization
5. Topic Tracking
6. Categorization
7. Clustering
8. Concept linkage
9. Information Visualization
10. Association Rule Mining

3.1. Natural Language Processing

It is involved with interactions between computer and human (natural) languages. NLP is related to the areaof human-computer interaction. NLP is the component of an Artificial Intelligence (AI). It is helpful to analyse the human languages so that computers can understand natural languages as humans do. The approaches to NLP is relies on machine learning, a type of artificial intelligence that examines and uses the patterns in data to improve a program's own understanding. The role of NLP in Text Mining is to provide the systems in the information extraction phase with linguistic data that they need to perform their task. NLP includes the tasks [3]

- **Part Of Speech tagging**: It is used to classify the words into categories such as noun, verb or adjective.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

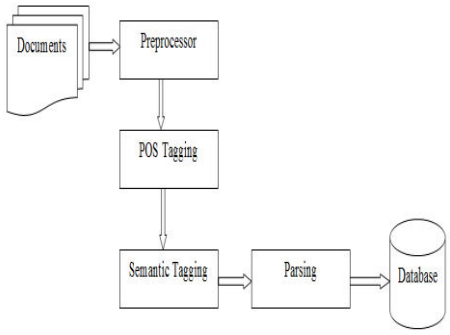
4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

- **Chunking**: It is also called as shallow parsing, used to identify only the main grammatical elements in a sentence such as noun phrases and verb phrases.
- **Semantic Role Labelling**: It is used to detection of semantic arguments associated with predicate or verb of a sentence.
- **Language Model**: It assigns a probability of sequence of words by means of probability distribution. ItProvides context to distinguish between words and phrases that sound similar.
- **Semantically Related Words**: This is the task of predicting whether two words are semantically related which is measured using the WordNet database.



3.2. Information Retrieval

It's the association and retrieval of information from a large number of text-based documents.
In database concurrency control, recovery, transaction management are not present. Also, some common information retrieval problems are usually not encountered in conventional database systems, such as unstructured documents, estimated search based on keywords, and the concept of relevance. Due to the huge quantity of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalogue systems, on-line document management systems, and the more recently developed Web search engines.[6]

3.3. Information Extraction

The information extraction technique identifies key words and relationships among the text. It does this by looking for predefined sequences in the text, a method known as patternmatching. This technique is very helpful when dealing with large volumes of text.
The major task of information extraction,

- Term Analysis**:** It identifies the term; the term may contain one or more words. It can be helpful forExtracting information from documents
- Named Entity Recognition: It identifies the textual information in a document relating the names ofPerson, place, organization or product
- Fact Extraction: It identifies and extracts the complex facts from the documents.

3.4. Summarization

Text summarization is to reduce the length and detail of a document while retaining most important points and general meaning. Text summarization is helpful for to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information hence summary can replace the set of documents. Humans first reads entire text section to summarize then try to develop a full understanding, and then finally write a summary, highlighting its main points.

3.5. Categorization

Categorizations engage identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often delight the document as a "bag of words." The categorization only count words that appear and, from the counts, identifies the main topics that the document covers.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

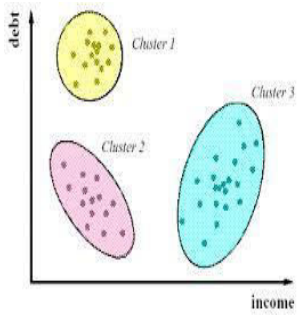
15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

3.6. Clustering

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Another advantage of clustering is that documents can emerge in multiple subtopics, thus ensuring that a useful document will not be absent from search results. Clustering technology can be useful in the organization of Management information systems, which may contain thousands of documents.[5]



3.7.Concept Linkage

It promotes browsing for information rather than searching for it. Concept linkage is a valuable idea in text mining, especially in the biomedical fields where so much study has been done that it is impossible for researchers to read all the material and make organizations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans cannot.

3.8. Information Visualization

It is helpful when a user needs to narrow down a broad range of documents and explore related topics. For example the government can use information visualization to find terrorist networks. It could provide them, with a map of attainable possible relationships between suspicious activities so that they can investigate connections that they might not have come up with on their own.

3.9. Association Rule Mining

Association rule mining (ARM) is a technique used to discover relationships among a large set of variables in a data set. ARM determines variable-value combinations that frequently occur. It used to improve the accuracy and decrease the complexity.[9]

IV. TEXT MINING ALGORITHMS

4.1. K nearest neighboursisa classical and frequently used technique. In order to find a query text KNN classifier is outperforms. This method estimates the distance between two strings for comparison and classify the text on the basis of distance. Where x and y represents the data instances and d is distance between x and y. The main advantage of this algorithm is high accurate classification and demerit is resources consumption such as memory and time.[10]

4.2. Support vector machine: This approach is a one of most efficient and accurate classification algorithm. In this approach concept using hyper-plans and dimension estimation based technique are used to discover or classify the data. The main advantage of this algorithm is to achieve high accurate classification results. But that is quite complex to implement.[4]

4.3. Bayesian classifier: It is a probability based classification technique that uses the word probability to classify the text data. In this classification scheme based on previous text and patterns data is evaluated and the class possibility is measured. **That is some time slow learning classifier additionally that do not produces the more accurate results.**

4.4. K-mean clusteringthisis one of the text categorization. That uses the distance function as k nearest neighbour classifier to cluster data. That is an efficient method of text mining in order to preserve the resources, but accuracy of this cluster approach is susceptible due to initial cluster centre selection process.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

V. ADVANTAGE AND DISADVANTAGE

Text mining provides a valuable tool to deal with large amounts of unstructured text data. A major characteristic of the representation paradigm of text mining is high dimensionality of the feature space, which impose a big challenge to the performance of clustering algorithms. Text clustering is a technique that is used to group the text in similar groups. There are few advantage and disadvantage of the text mining they can be given as follows,

5.1 Advantage of text mining:

- It solved the problem of managing a great amount of unstructured information for extracting patterns easily
- Reduces the storage problems in the data base

5.2 Disadvantage of text mining:

1. 1.Programs cannot be in order to analyze the unstructured text directly to mine the text for information or knowledge.
2. Intermediate form
3. Multilingual Text Refining
4. Domain knowledge Integration

VI. TEXT MINING TECHNIQUES

Several text mining techniques are used in the text mining process some of them are given as follows,

- Supervised text categorization technique
- Pattern matching algorithm
- Support vector machine technique

6.1 Supervised text categorization technique

Supervised text categorization and clustering are closely related as both are concerned with "grouping" of objects. However, in the supervised setting, these groupings are given by common membership to a class that is assigned to sample documents before the training process starts. The training process then induces hypotheses of how the document space is shaped according to which new documents are assigned.

6.2Pattern matching algorithm

Text mining concerns looking for patterns in unstructured text. Pattern matching is to find a pattern, which is relatively small, in a text, which is supported to be very large. Documents contain vast amounts of data that cannot be easily examined one by a human. Mining patterns from a large data set is an important system management task.

6.3 Support Vector Machine Technique

A classification task usually involves separating data into training and testing sets. Support Vector Machine (SVM) is one of the most actively developed classification technique in data mining and machine learning. The goal of SVM is to produce a model based on the training data which predicts the target values of the test data given only the test data attributes. It has been successfully applied to a wide range of pattern recognition problems.

VII. PERFORMANCE MEASURES

There are various methods to determine effectiveness or the performance of the algorithms. The metrics Precision, Recall, and F-measure are most often used. Precision is determined as the conditionalprobability that a random document d is classified under ci, or what would be deemed the correctcategory.

$$Precision = TP/TP+FP$$

Recall is defined as the probability that, if a random document (dx) should be classified under category (ci), this decision is taken.

$$Recall = TP/TP+FN$$

Where

True Positive (TP) - situation in text classification when the classifier correctly classifies a positive test case into the positive class;

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

True Negative (TN) – situation in text classification when the classifier correctly classifies a negative test case into the negative class;
False Positive (FP) – situation in text classification when the classifier incorrectly classifies a negative test case into the positive class;
False Negative (FN) – situation in text classification when the classifier incorrectly classifies a positive test case into the negative class;
Precision and recall are often combined in order to get a better picture of the performance of the classifier given as F-Measure
F − Measure = 2×Recall×Precision/Recall+Precision

VIII. TEXT MINING APPLICATION

The text mining technology is now mostly applied for a wide variety of government, research, and business needs. Application can be sorted into a number of categories by analysis type or by business function. Using this Approach to classifying solutions, application categories include:

- Enterprise Business Intelligence /Data Mining, Competitive Intelligence.
- Records management
- Sentiment analysis Tools, Listening Platforms.
- Natural Language/ Semantic Toolkit or service.
- Search/ Information Access.
- Social Media Monitoring

IX. RESEARCH AREAS OF TEXT MINING

1. **Information extraction**
   IE is the process of automatically obtaining structured data from unstructured data. This action includes natural language processing.

2. **Data Mining**
   Data Mining looks for pattern in data. It can be more described as the retrieval of hidden information from data. Text Mining in data mining tools can predict response and trends of the future. It enables businesses to make positive decisions based on knowledge and answer business questions.[8]

3. **Natural Language Processing**
   The purpose of NLP in text mining is to deliver the system in the knowledge retrieval phase as an input.

4. **Sentiment Analysis**
   One of the most important research area in text mining is sentiment analysis. This extracts text from articles, social media, survey, and reviews –essentially any source where suggestions, comments or feedback is given. Sentiment score is then applied, determining if the text is positive, negative, happy, sad or neutral. This gives the business or agency an idea of how users feel about a particular topic and
   Text Mining Tools
   1. Python
   2. R

X. CONCLUSION

Text mining is the process of seeking or extracting the useful information from the textual data. In our paper we have given an overview of text mining pre-processing, techniques for unstructured data. Future research directions. In future we can have NLP pre-processing and use transformation from machine learning to deep learning. We hope this paper will help the text mining researcher's community and they get good knowledge about various pre-processing techniques and algorithms.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569| || Volume 10, Special Issue 2, December 2021 ||

4th International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM '2021]

15th & 16th September 2021

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

REFERENCES

[1] "Preprocessing Techniques for Text Mining - An Overview" Dr.S.Vijayarani et al , International Journal of Computer Science & Communication Networks,Vol 5(1),7-16

[2] "Survey on Pre-Processing Techniques for Text Mining" Arjun Srinivas Nayak1 International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 5 Issues 6 June 2016, Page No. 16875-16879

[3] "Natural Language Processing", Aditya Jain International Journal of Computer Sciences and Engineering 2018

[4] "Text Classification Techniques",M.Sivakami, "Interdisciplinary Journal of information ,Knowledge,and Management 2018

[5] "A Survey on Text Mining and Sentiment Analysis for Unstructured Web Data", Dr. Prasad G R Journal of Emerging Technologies and Innovative Research (JETIR)2015

[6] "Text Mining Research: A Survey" R.Janani International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2016

[7] "Survey Paper On Natural Language Processing" Ms. Anuja Bharate ,International Journal of Computer Engineering and Applications, Volume VIII, Issue III, Part I, December 2014

[8] "A Survey of Data Mining Techniques and Tools" V.Sridevi , Dr. A. Kanagaraj, International Journal of Advanced Research in Computer and Communication Engineering,2017

[9] "A Literature Survey on Association Rule Mining Algorithms", P. Yazgan, A.O. Kusakci/ Southeast Europe Journal of Soft Computing Vol.5 No.1 March 2016

[10] "Performance Evaluation of Supervised Machine Learning Techniques for Efficient Detection of Emotions from Online Content", Computers, Materials & Continua 2020

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details