TEXT MINING MODEL: A REVIEW

[1]Gowri.J, [2]Suriya.M, [3]Ajay yeswanth.S

[1] Assistant Professor, [2,3] PG Scholar, Department of Software Systems,

Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India.

ABSTRACT: Text mining is one of the most critical ways of analysing and processing unstructured data which forms nearly 80% of world's data. Text mining also referred to as a data mining, large amounts of unstructured text data generated on the internet, text mining are believed to have high commercial value. Text mining often uses computational algorithms to read and analyse textual information. Text mining helps to identify patterns and relationships that exist within a large amount of text. It is an artificial intelligence technology that uses natural language processing to transform the free text in documents in databases into a normalized, structured data suitable for analysis or to drive machine learning algorithms. At last we classify text mining work as text categorization, text clustering, association rule extraction and trend analysis according to applications.

Keywords: **Text mining, data mining, technologies, applications.**

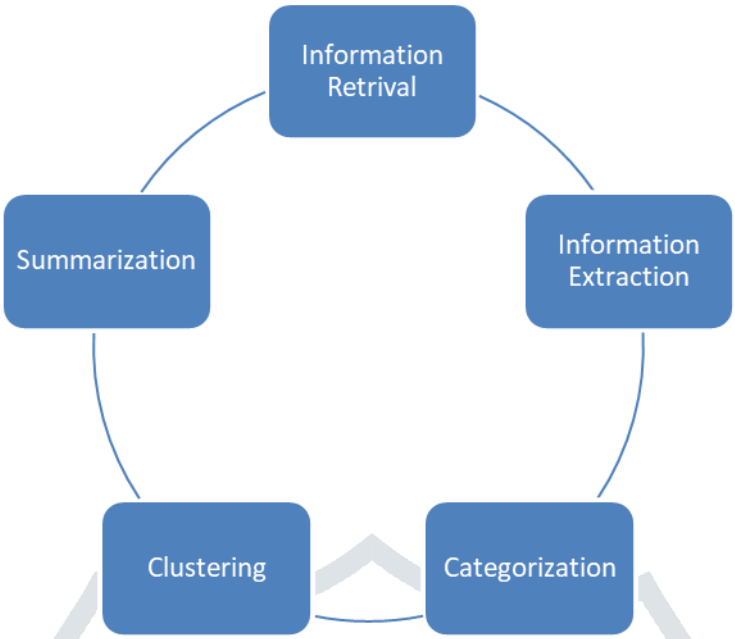
Contents:

I. Introduction:

Text mining also called as data mining. The size of data is increasing at exponential rates day by day. A huge amount of text is flowing over the internet in the form of digital libraries, repositories and other textual information such as blogs, social media networks and e-mails. The goal is data mining; data is analysis, via application natural language processing and analytical methods. The text mining main tools are QDA miner, WordStat, SimSat. A Business might want to extract specific information, like keywords, names or company information.

Hence you can analyse words, clusters of words used in documents. In data mining, text data was exchange for numbers. Particularly that has traditionally been too time-consuming to resolve. Data mining was also allowed to make for business positive knowledge-based decisions. It is a multi-disciplinary field based on information retrieval, data mining, machine learning, statics and computational linguistics.

A big amount of text is flowing over the internet in the form of digital libraries, social media networks and e-mails. Text mining techniques are continually applied in industry, academia, web applications, and internet. Text mining is stored by unstructured and semi-structured data (text). Applications deal with much more diverse and eclectic collections of systems and formats.



II. Text mining techniques:

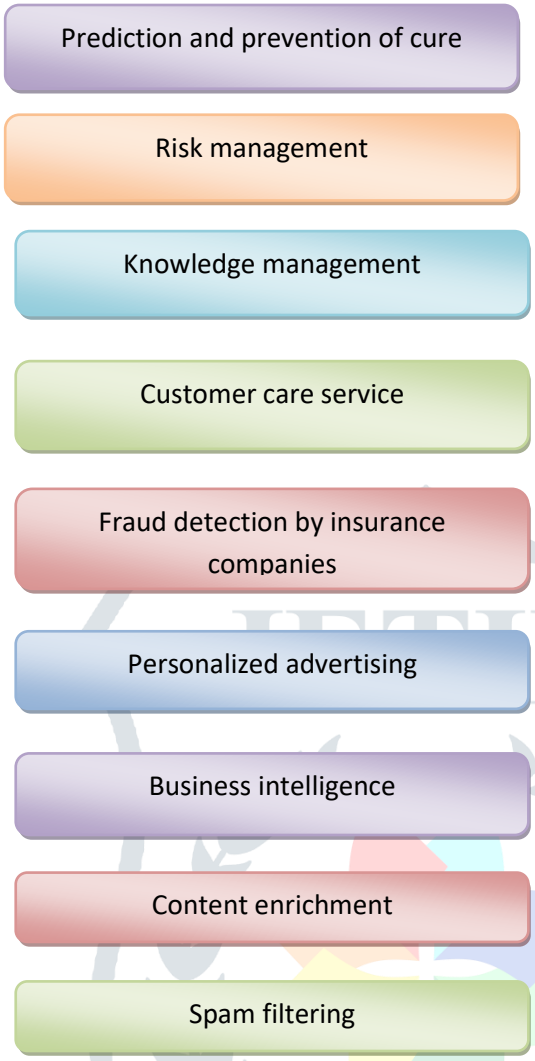
2.1. Information Retrieval: It is described in terms of predictive text mining**.** This system is a network of algorithms, search of relevant data or documents as per the user requirement. This is also tracking the utility of the displayed data as per user behaviour. The algorithms used by Yahoo and Google are much more comparable. Classical information retrieval of documents stored in databases to web or internet-based documents.

2.2. Information Extraction: It is the role of automatically extracting structured information from semi-structured machine-readable documents, content extraction out of image, audio, video and documents could be seen as information extraction. Information Extraction is the part of greater puzzle deals with problem of devising automatic methods for text management, beyond its transmission, storage and display.

2.3. Categorization: This is the task of assigning predefined categories of free-text documents. Geographical codes: academic papers are often classified by technical domains and sub-domains; news stories are typically organised by subject categories. Email messages are classified into two categories of spam and non-spam. Patient reports in health-care organizations are often indexed from multiple aspects.

2.4. Clustering: It is similar to classification in that text is grouped. They are not predefined. Clustering has been utilized in numerous application areas, including science, drug, humanities, advertising and financial aspects. Clustering applications include plant and animal classification, disease classification, image processing, pattern recognition and document retrieval. Recent uses, including examining web log data to detect usage patterns

2.5. Summarization: Summarization is also called characterisation or generalisation. Text summarization refers to the techniques of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document. Automatic text summarization is a common problem in machine learning and natural language.



III. Applications:

3.1. Prediction and prevention of crime: It is a difficult task to pinpoint messages that might be considered a threat. This is easily done using advanced text analysis software that cans communication sources in real time and sound different levels of threat alert on finding different types of text. These technologies are used to prevent terrorist attacks catch sleeper cells and stop people from carrying out other unlawful activities.

3.2. Risk management: The text mining technologies used by such high-end software absorb pet bytes of data and present information in a consumable format. This helps in risk mitigation. Software is helping financial institutions all over the world, to decrease their percentage of non-performing assets.

3.3. Knowledge management: In many industries like the healthcare industry, managing a huge amount of textual information has become a problem. Such a huge exercise would be impossible without the help of proper text analytics systems in place that would manage the data and information and keep them in structured tree like format. They would lead to people being able to access the data in any way they need-region-based, disease based, gender based.

3.4. Customer care services: Text mining and natural language processing are frequently being used in customer care services, be it over chat or voice call. Most banks and e-commerce companies are using natural language processing-based chat bots that try to mimic a human customer care officer when talking to customers. Automating customer care services, companies are providing customers a better experience while at the same time, saving money.

3.5. Fraud detection by insurance companies: Text analytics has proved effective in going over huge collections of case files to understand the chances of an insurance claim being fraud. It greatly reduces the workload of the company officials since the fraud recognition software would automatically flag cases where a high probability of fraud is determined.

3.6. Personalized advertising: Digital advertising has been revolutionized by text and web mining. Text data related to all that you type, view, or sold to other companies to show you advertisements that you have a higher probability of clicking on. This is one of the latest and most widely used applications of text analytics and mining.

3.7. Business intelligence: Decision making is difficult. Text mining really helps gather evidence and draw up charts and graphs to put information to back your gut feeling. Only relevant information and data is extracted so that the people who lead can take the best decisions by going through only a few pages of information.

3.8. Content enrichment: This makes a significant difference when writing on topics that have huge volumes of pre-existing data on the internet. This helps make your content information and connect to previous articles and studies in the same field.

3.9. Spam filtering: E-mails are considered as the most official way of communication in mot organizations. Spam's not only fill up space but also serve as an entry point for viruses, scams. Companies are pushing hard to filter more and more spam by using intelligent text analytics as compared to the keyword matching used earlier.

IV. Issues:

Human interaction: since text mining problems are often not precisely stated, interfaces may be needed with both domain and technical experts [1]. **Overfitting**: Over fitting occurs when the model does not fit future states [2]. **Outliers:** There are often many text entries that do not fit nicely into the derived model [3]. **Interpretation of results:** Currently text mining output may require experts to correctly interpret the results [4]. **Visualization of results:** To easily view and understand the output of text mining algorithms, visualization of the results is helpful [5]. **Large datasets:** Create problems when applying algorithms designed for small datasets [6]. **High dimensionality:** The problem here is that not fact, all attributes may be needed to solve a given text mining problem [7]. **Multimedia text:** The use of multimedia data such as is found in GIS databases complicates or invalidates many proposed algorithms [8].

V. Advantage:

- It is helpful to predict future trends...
- It signifies customer habits...
- Helps in decision making...
- Increase company revenue...
- It depends upon market-based analysis...
- Quick fraud detection...

Conclusion:

A big collection of documents may provide useful information to people. But it is also a challenge to find out the useful information from a large collection of documents. Successfully implemented text mining techniques help to identify the category, of each text document, where they fit best. At last, most refer to that the field of text mining is still in the research phase and still its applications limited operation at the present time but the possibilities that can provide, which helps to understand the huge amounts of text and extract the core of which information is important and useful prospects in many areas.

References

[1] Jiawei Han and Michelin Kampar, Data Mining: Concepts and techniques, Second Edition, Morgan Kaufamann publications, United States of America, 2006.

[2] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature Scope", arXiv preprint arXiv: 1211.5723, 2012.

[3] A.M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," Briefings in bioinformatics, vol. 6, no. 1, 57-71, 2005.

[4] S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher Education system," European Journal of scientific Research, vol. 43, no. 1, pp. 24-29, 2010.

[5] N. Samsudin, M. Puteh, A. R. Handan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in proceedings of the world congress on Engineering, vol. 3, 2013, pp. 3-5.

[6] Julia Itskevitch, "Automatic hierarchical e-mail classification using association rules, "Belorussian State Polytechnic Academy, 1997.

[7] Yuen-Hsien Tseng, "FJU Test Collection for Evaluation of Chinese Text Categorization" 2004.