University of Mumbai
Anjuman-I-Islam's Kalsekar Technical campus

Text Mining

Retrieve valuable information from document

Khan Aftab
Information Retrieval
Prof. Salim G. Shaikh

1. Introduction

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. By applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms, companies are able to explore and discover hidden relationships within their unstructured data.

Text is one of the most common data types within databases. Depending on the database, this data can be organized as:

- Structured data: This data is standardized into a tabular format with numerous rows and columns, making it easier to store and process for analysis and machine learning algorithms. Structured data can include inputs such as names, addresses, and phone numbers.
- Unstructured data: This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like video and audio files.
- Semi-structured data: As the name suggests, this data is a blend between structured and unstructured data formats. While it has some organization, it doesn't have enough structure to meet the requirements of a relational database. Examples of semi-structured data include XML, JSON and HTML files.

Since 80% of data in the world resides in an unstructured format, text mining is an extremely valuable practice within organizations. Text mining tools and natural language processing (NLP) techniques, like information extraction (PDF, 127.9 KB), allow us to transform unstructured documents into a structured format to enable analysis and the generation of high-quality insights. This, in turn, improves the decision-making of organizations, leading to better business outcomes.

2. Literature Review

The amount of text that is generated every day is increasing dramatically. This tremendous volume of mostly unstructured text cannot be simply processed and perceived by computers. Therefore, efficient and effective techniques and algorithms are required to discover useful patterns. Text mining is the task of extracting meaningful information from text, which has gained significant attention in recent years. In this paper, we describe several of the most fundamental text mining tasks and techniques including text pre-processing, classification and clustering. Additionally, we briefly explain text mining in biomedical and health care domains.

One of the first tasks of an innovative project is delineating the scope of the project itself or of the product/service to be developed. A wrong scope definition can determine (in the worst case) project failure. A good scope definition becomes even more relevant in technological intensive innovation projects, nowadays characterized by a highly dynamic multidisciplinary, turbulent and uncertain environment. In these cases, the boundaries of the project are not easily detectable and it’s difficult to decide what it is in-scope and out-of-scope.

The present work proposes a tool for the scope delineation process, that automatically defines an innovative technological field or a new technology. The tool is based on a Text Mining algorithm that exploits Elsevier's Scopus abstracts in order to extract relevant data to define a technological scope. The automatic definition tool is then applied on four case studies:

Artificial Intelligence and Data Science.The results show how the tool can provide many crucial information in the definition process of a technological field. In particular for the target technological field (or technology), it provides the definition and other elements related to the target.

3. Research Methodology

Theoretical research-Gather more information about a particular topic without considering its practical working

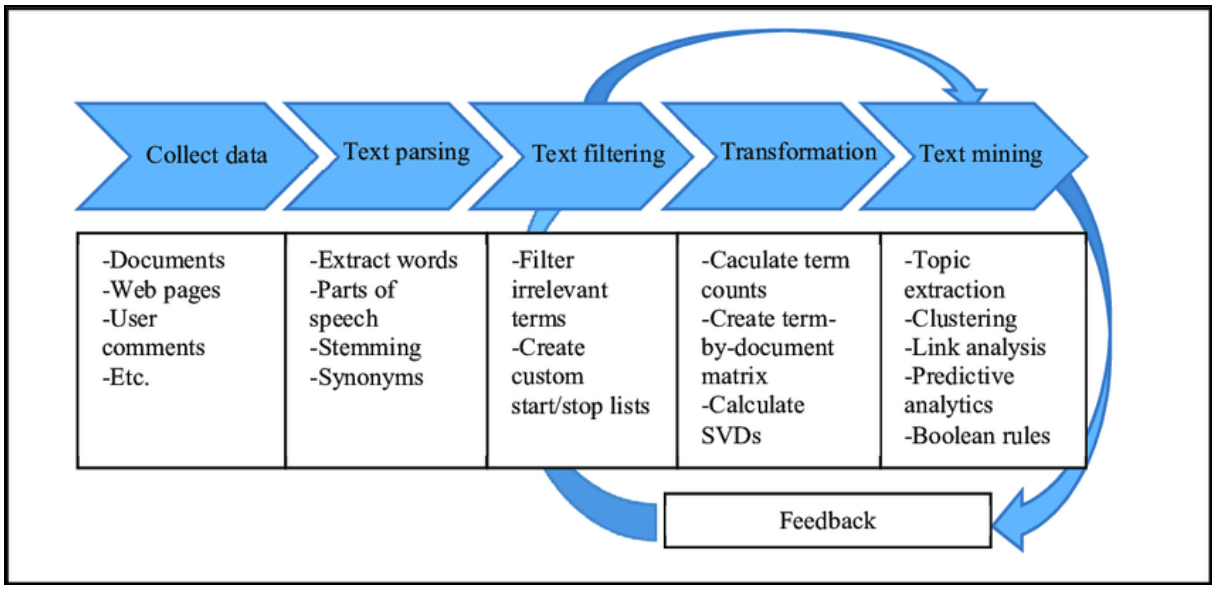
Non-experimental research-Observational study

Inductive research-Observational study that focuses on achieving generalized results

Secondary research-Available material like research papers, interviews, documentaries

as a source of data and information

4. Data Presentation and Analysis



Technique	Characteristics	Tools
Retrieval	Retrievals valuable information from unstructured text	Intelligent Miner, Text Analyst
Extraction	Extract information from structured database	Text Finder, Clear Forest Text
Summarization	Reduce length by keeping its main points and overall meaning as it is	Tropic Tracking Tool, Sentence Ext Tool
Categorization	Document based categorization	Intelligent Miner
Cluster	Cluster collection of documents, Clustering, classification and analysis of text document	Carrot, Rapid Miner

5. Conclusion

Data mining is the process of identifying patterns and extracting useful insights from big data sets. This practice evaluates both structured and unstructured data to identify new information, and it is commonly utilized to analyze consumer behaviors within marketing and sales. Text mining is essentially a sub-field of data mining as it focuses on bringing structure to unstructured data and analyzing it to generate novel insights. The techniques mentioned above are forms of data mining but fall under the scope of textual data analysis.

The process of text mining comprises several activities that enable you to deduce information from unstructured text data. Before you can apply different text mining techniques, you must start with text preprocessing, which is the practice of cleaning and transforming text data into a usable format. This practice is a core aspect of natural language processing (NLP) and it usually involves the use of techniques such as language identification, tokenization, part-of-speech tagging, chunking, and syntax parsing to format data appropriately for analysis. When text preprocessing is complete, you can apply text mining algorithms to derive insights from the data.

Text analytics software has impacted the way that many industries work, allowing them to improve product user experiences as well as make faster and better business decisions.

6. References

a. Charu C Aggarwal and ChengXiang Zhai. 2012. Mining text data. Springer.
b. Mehdi Allahyari and Krys Kochut. 2015. Automatic topic labeling using ontology-based topic models. In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, 259–264.
c. Mehdi Allahyari and Krys Kochut. 2016. Discovering Coherent Topics with Entity Topic Models. In Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on. IEEE, 26–33.
d. Mehdi Allahyari and Krys Kochut. 2016. Semantic Context-Aware Recommendation via Topic Models Leveraging Linked Open Data. In International Conference on Web Information Systems Engineering. Springer, 263–277.
e. Mehdi Allahyari and Krys Kochut. 2016. Semantic Tagging