International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 7, Issue 6 Ver. I (Jun 2018), PP. 23-27

Text Mining: Need, Tools and Application

Dr Archana Tushar Raje[1], Allan Fernandes[2]
[1](Assistant Professor, Information Technology, K. J. Somaiya Institute of Management Studies and Research, India)
[2](Student, Information Technology, K. J. Somaiya Institute of Management Studies and Research, India)
Corresponding Author: Dr Archana Tushar Raje

Abstract:- *Data these days is mainly categorized under two forms: structured and unstructured. While the structured data is stored in the databases, which makes it easier to work upon, unstructured data, on the other hand, include emails, text documents, and web page. Text mining thus comes in the picture when useful information needs to be retrieved from these form of unstructured text. This highly increases the need for text mining in the field of analysis and investigation.*
But how to go about with text mining. There are a number of tools and techniques present for text mining, so how does one go with the right selection of text mining technique in order to intensify the speed and at the same time reduce the time, effort to get the best and precise result. This paper focuses on providing the insights on tools that are available to work on the unstructured data along with techniques and fields where text mining can be applied.
Keywords: - *Text mining techniques, Text analysis, Classification, Summarization*.

Date of Submission: 17-08-2018 Date of acceptance: 31-08-2018

I. INTRODUCTION

A huge of amount of textual information exists that can be worked upon, however text data is unstructured and requires a means to analyze and extract needful information. Text mining, also knowing as Text analytic or Text data mining, is the means of retrieving high-quality data and patterns from the text. These high-quality data is retrieved through the crafting of trends and patterns. It can be considered as a three process step that involves: Processing the input text, deriving patterns and evaluation of the output

Typical text mining activities include text clustering, document summary, text categorization. Thus one can scrutinize cluster of words in a record, or do an analysis of the documents to figure out the similarities between them and find the relation between them. Text mining is thus a vast area that can be explored to find out high-quality information from all the unstructured data around and help in in various field such as social media, research and academic field, cybercrime prevention etc.

Thus it can be said that straightforward data mining tools can suffice the need to handle textual information and specialize means would be required with stronger algorithms that can analyze the text.

There is no field that text mining has not touched upon be it for social media analysis, security analysis or even business intelligence.

In this paper, there is an effort to understand the popularly accessible tools for text mining, what are the applications and fields that text mining play a major role in and also different types of tools that are available around.

II. TEXT MINING TOOL TYPES

Classification of text mining tools can be done into three categories:

1. **Online Tools:** These are website tools that can be run using any web browser. However, these tools are limited by their functionality. Some of the popular online tools available are:

A. Textalyser: It supports text as well as keyword analysis primarily used for text analysis.
Link: http://textalyser.net/

B. Ranks.nl: This supports keyword analysis technique with feature such as page, multi-page, article analysis.
Link: https://www.ranks.nl/

C. Voyant: Using keyword analysis technique for text analysis.
Link: https://voyant-tools.org/

2. **Proprietary tools:** Owned by companies, these text mining tools are commercial tools and hence these tools need to be purchased before use. However, demo versions are freely available with limited functionality and limited usage period. Some of the popular Proprietary tools available are:

A. Discovertext: It uses cloud-based analytics technique along with active machine learning.
   Link: https://discovertext.com/
B. Langsoft: Using natural language processing and artificial intelligence for analysis.
   Link: http://www.langsoft.ch/
C. Dtsearch: Supports advanced data classification with main feature for text analysis.
   Link: https://www.dtsearch.com/

3. **Open Source Tools:** These tools are at hand accessible for no cost. Also the source code these tools are readily available and thus any one can contribute to the same. Some of the popular open source tools are:

A. Datumbox: Uses keyword extraction and machine learning technologies with features including SEO, social media monitoring, text analysis as well as sentiment analysis.
   Link: http://www.datumbox.com/
B. R Programming: Graphical and statistical techniques with features such as text analysis and data transformation.
   Link: https://www.r-project.org/
C. Rapidminer: Making use of machine learning language for text processing and analytics as well as data mining.
   Link: https://rapidminer.com/

III. TEXT MINING TECHNIQUES

In text mining process, technologies such as extraction of information, summarization, categorization, information visualization and clustering are used.

1. Extraction of information: This is the initial step for analyzing text that is unstructured by means of identifying key words and relationship among them. This includes use of pattern matching to observe any predetermined sequence in the text. It begins with parsing of sentences and phrases, followed by interpretation and then entering the pieces of information in the database. This can be illustrated as below:

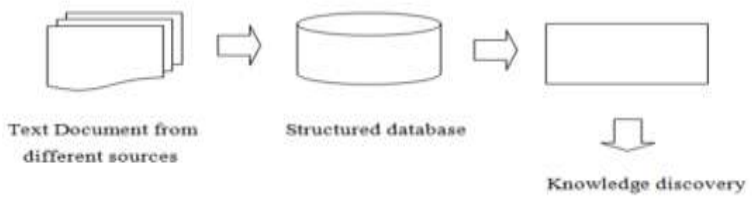

Fig. 1: Extraction of Information

2. Categorization: This step is responsible for assigning the category to the textual document. It can be considered as supervised learning methodology as it forms its basis from input output examples to classify the documents. It involves having predefined classes that are assigned to the documents. A basic categorization technique would involve pre-processing, followed by indexing, reduction and then classification. The goal is to train the classifier using known examples, so that the unknown examples can be automatically categorized.
3. Clustering: This method can be made in used for finding association of documents with similar content. The outcome of this includes a group called clusters and every cluster having number of documents. The document within each cluster have similar content and the differences lie between the content of documents of different clusters. Although, clustering sounds to be similar to categorization, but is differs as the documents are clustered on the go and not using any predetermined topic.
   In text mining, K-means is the most frequently used algorithm as it retrieves good results. To summarize, a basic algorithm would generate a direction of topics and calculate how well each document fits into a cluster.
4. Visualization: This method can used to disentangle the revelation of relevant information. Text flags are used to represent document category. Textual sources are jotted down in visual hierarchy allowing user interaction by scaling and zooming. Following are the steps involved in visualization process.

Information visualization can be branched into three steps:

A. First step includes obtaining and deciding visualization data.
B. Second step includes analyzing and extracting the data that is needed from the original data and forming a visualization space.
C. This involves use of certain algorithm to outline the visualization data space to target space.

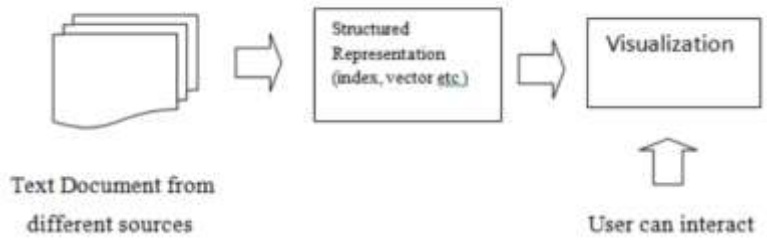

Fig. 2: Visualization of Information

5. Summarization: As the name suggest, it involves reducing the length and details of the document while ensuring retention of important points and overall meaning. Helpful in cases where in it is necessary to figure out if a lengthy document is useful enough and worth reading for any information, and thus the summary set can replace the documents. It is however a difficult task to train a software to perform analysis on the semantics and to interpret the text meaning.

This technique includes the following steps:

A. Pre-filtering to get the original text in a structured format.
B. Then to transform the text structure to a summary structure using an algorithm.
C. Finally, obtaining the final summary from the summary structure.

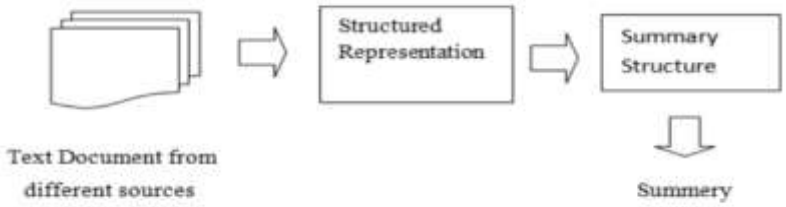

Fig. 3: Summarization of Information

IV. MODELS AND METHODS

Text mining is based around four methods that are:

1. Concept Based Method (CBM): This method involves analysis on the basis of sentences. This revolves around the idea of statistical analysis of word frequency so as to capture the importance of the word without the document. It may be possible that two terms may have the same frequency within the document, but as one contributes more relevantly than the other and thus should be given higher priority. This then leads to a new concept-based mining. This model is helpful in differentiating between important and non-important terms.
2. Pattern Taxonomy Method (PTM): This method involves analysis on the basis of patterns. Data mining techniques such as association rule, sequential pattern, closed pattern mining can be used for pattern discovery. Using patterns in the text mining field is tough and not always effective, as some long patterns which prove to be useful may lack in support i.e., less frequency problem. Also, not all recurring patterns are useful which can also be considered as pattern misinterpretation, leading to ineffective performance. But research and experimental studies have shown that pattern-based model tends to perform better than the other text mining models that are present.
3. Phrase Based Method (PBM): In this model, the document is analyzed more on phrases basis as phrases are more discriminative and less ambiguous than individual terms.
   The reason for their intimidating performance include: Occurrence frequency is low, Presence of huge number of noisy and redundant phrases. Also when compared to terms, they have lesser statistical properties.
4. Term Based Method (TBM): As the name suggest, this method involves analysis on the basis of terms. Also, this method has an advantage of computational performance for term weighing. This method however faces issue from synonym and polysemy. Polysemy meaning same word having multiple meaning and synonym meaning multiple words have the same meaning.

V. TEXT MINING FEATURES

The main uses of text mining are as follows:

- Analysis of Sentiment: From the text, helps to discover subjective information. Such tools that work on sentiment analysis are also known as Opining Mining.
- Processing of Text: This includes manipulating and transforming unstructured textual information so as to perform analysis method on the same.
- Text Analytic: Extracting needful information and trends from the text. This is the primary feature of all the tools.
- Knowledge Discovery: This works with recognizing needful information from large amount of text. This feature is also included by most of the mining tools.
- Categorization/Classification: Number of tools support this feature of categorizing and classifying documents.

VI. APPLICATIONS

A. Academic Field

Analysis of educational trends in specific region, employment ratio, student's interest are the primary driving force for using text mining in the field of academics. Also how students perform in different subjects and what are the attributes for selection of subjects can also be analyzed.

B. Social Media

Analysis of posts, comments, reviews can be easily done with the help of text mining tools. There are packages available that extensively provide the functionality for analyzing and monitoring plain text from email, blogs, and internet news. This analysis focuses primarily on the sentiments of the user and the people, analyzing the reaction of the people on different news, post. Further analysis can be done by grouping the people in specific age groups or by communities having similar and variation in their views

C. Life science

Health care and life science industries generate huge amount of numerical and textual data regarding the records of the patients, symptoms, diseases, medicines and many more. This thus poses a challenge to percolate relevant and appropriate text from the huge biological repository.

Again, the records are highly varying in nature, complexity, length and technical vocabulary. Text mining tools can thus provide a means or opportunity to get valuable information, associations among them and inferring relationships among various species, genes and diseases. These tools can also help in evaluating effectiveness of the treatments by inspecting different diseases, their symptoms and the course of treatment. Text mining can also contribute highly in the pharmaceutical industry, mapping of genes diseases, clinical trade analysis and many more.

D. Human Resource Management

These mining techniques can be used to manage the human resources critically, mainly with the aim to analyze employee satisfaction, staff opinions and also reading the CVs for selection purpose. Referencing the human resource management, these are often used to guide the health of the company.

VII. ISSUES FACED IN TEXT MINING

In text mining before applying a pattern analysis on the document, there arises a need to convert the unstructured data into intermediary forms. Also there are times when real data importance is broken down due to the alteration in the sequence of texts.

Another issue arises from the multilingual dependency as there are only a few tools available out there that support multiple languages. Different algorithm and techniques are in use to support multilingual textual data. Again, the use of synonyms, polysemy and antonym create problems for the tools as they take both in the same context.

Hence it becomes difficult to perform categorization on the documents, when the collection of document is huge and generated from different fields having the same domain. Abbreviations also act as an issue when it comes to text mining tools are they give changed meaning in different situations.

Also, the varying granularity concept change the text context according to domain knowledge and condition. Developing plug-ins would also requires proper and depth knowledge about specific domains. A lot of complications are itself created by natural languages in the text refinement methods and while identifying relationships among the entities. There are certain words that have the same spelling but give out different and contrasting meaning such as fly. A text mining tool would then consider both as the same while on can be a noun and the other a verb.

Apart from this grammatical rules based on the nature and context ads up to the list of issues faced by text mining tools.

VIII. CONCLUSION

The presence of large amount of data which is text based need to be analyzed to derive needful information. Text mining can thus be used to get relevant and interesting information efficiently and effectively from any unstructured data. The presented paper contains a brief explanation of why text mining is needed. It then focuses on the different types of tools that are available around, thus facilitating proper selection of tools based on our requirements. Some important techniques, models and methods are also the highlight in this paper. Also coming to the issues faced by text mining, there are granularity concepts, multilingual text, domain knowledge integration, presence of synonyms, polysemy are some of the major issues faced by the field of text mining. Thus, this paper concludes by explaining the importance of text mining in various other fields.

REFERENCES

[1]. R. J. Mooney and U. Y. Nahm. Text Mining with Information Extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002.
[2]. Feldman, R., Sanger, J. (2006). The text mining handbook. Cambridge University Press.
[3]. www.google.com
[4]. https://datascience.stackexchange.com
[5]. Talib, Ramzan & Kashif, Muhammad & Ayesha, Shaeela & Fatima, Fakeeha. (2016). Text Mining: Techniques, Applications and Issues. International Journal of Advanced Computer Science and Applications. 7.
[6]. C. C. Aggarwal and C.-X. Zhai, Mining Text Data, Springer, 2012
[7]. Seminar Report on Text Mining Submitted By Jisha A.K.

Dr Archana Tushar Raje "Text Mining: Need, Tools and Application "International Refereed Journal of Engineering and Science (IRJES), vol. 07, no. 06, 2018, pp. 23-27