A SURVEY ON THE TECHNIQUES FOR TEXT MINING

SHWETIMA, GURPREET KAUR, MD. ATAULLAH

Abstract: In today's highly sophisticated technological era, we are highly dependent on data; be it structured, semi-structured or unstructured. Text mining serves as an indispensable research field with focus being on extraction of hidden useful information. Text mining is the analysis of data contained in natural language text. With the advancements in technology, several different techniques and methods have been presented for Text Mining, each having its own advantages and disadvantages. This paper presents an analysis of text mining procedures and a comparison of various techniques used therein.

Keywords: Text classification, Text Mining, Text Mining Framework.

Introduction: The twenty first century is an era of technology and competency where economy and business highly depend upon accurate and precise information. Due to the advent of internet and World Wide Web, we come across huge volumes of easily available data every day; be it structured, semi-structured, or unstructured, which has no significance unless we have techniques to extract the required information from them. As communication plays a critical role for the exchange of information in all the spheres and text acts as an indispensable vehicle for the same, the need of the hour is to have an intelligent tool which extracts useful information from abundant text at optimal costs, in acceptable time limits and with minimal storage requirements. Here is when text mining comes into picture. Text mining, which has emerged as one of the most vibrant research fields, solves the problem by using automated methods for exploiting massive amounts of knowledge available in the form of text. The available text is in human understandable form i.e. analogue which a machine (computer) would not understand as it works in a constrained environment with limited capabilities. The term 'text' we referred here is in context of natural language which often refers to the concepts, senses and meanings which a computer simply cannot straightway deal with. But the same text also contains interesting patterns, trends, rules, models, etc. which facilitate the text mining techniques. Jusoh et al. [7] in his work discussed that text mining tools have the capability to analyse massive amounts of natural language text by determining lexical, syntactical and semantic components. Lexical analysis is the process where each word is tagged with its part of speech (discussed later in section 2). Syntactic analysis is the process of assigning syntactic structure or parse tree to a given natural language text. Semantic analysis is the process to relate syntactic structures to represent relevant meanings.

Text mining is an interdisciplinary field which draws techniques from various fields as discussed in figure below:

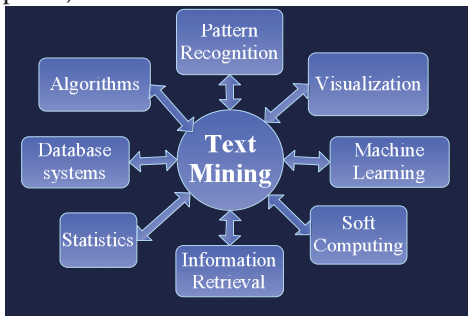


Figure 1: Domains from which Text Mining adopts techniques

The rest of this paper is organized as follows. In Section 2 the review of the literature related to the topic is being discussed. Section 3 gives an overview of Text mining Framework. Section 4 outlines the basic text mining approaches. In Section 5 we provide an impression of text classification techniques that can be used in text mining along with a comparison of the advantages and disadvantages of all these techniques. In Section 6 we conclude.

Literature Review: From the time the concept of text mining was introduced, much advancement in the field has taken place. Several techniques came into existence since then. They are described below in details with their strengths and limitations as follows:

Zhong et al. [5] in the work "Effective Pattern Discovery for Text Mining" talks about effective pattern discovery techniques using processes of pattern deploying and pattern evolving in order to improve the effectiveness of discovered patterns for finding relevant and interesting information.

The techniques proposed in last few decades comprise of association rule mining, closed pattern mining, sequential pattern mining, maximum pattern mining, and frequent item set mining. To implement these in text mining is ineffective and difficult. The author argues that not all frequent patterns are useful. Thus, misapprehensions and misconceptions of patterns derived from these techniques lead to the ineffective performance.

Jusoh et al. [7] in their work "Techniques, applications and challenging issues in Text Mining", describes that text mining system analyses large quantities of natural language text to extract meaningful and useful information.

They discussed the fundamental methods for text mining being natural language processing (NLP) and information extraction (IE) techniques.

A brief review on application domains of text mining has been presented. The paper also addressed the most challenging issue in developing text mining systems today.

Sukanya et al. [4] in the paper titled "Techniques on Text Mining" describes text mining as an extension of Data mining which discovers previously unknown information from different sources of data. Further the framework of text mining with techniques is discussed as well as the limitations and benefits have been taken into consideration.

Kumar et al. [3] in the journal "Text mining: concepts, process and applications ", talks about text analysis, and say it involves information retrieval, information extraction, data mining techniques including association and link analysis, visualization and predictive analytics.

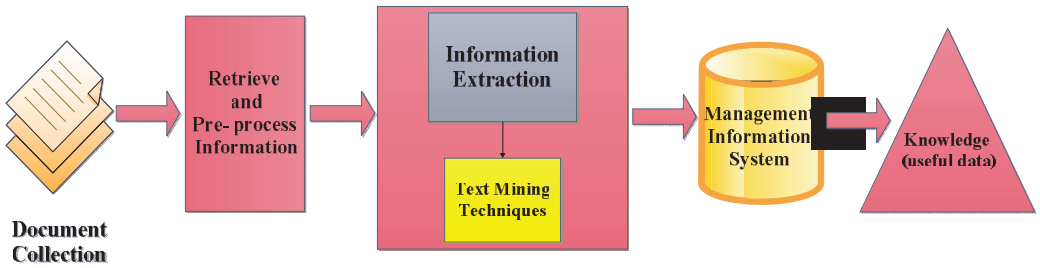
The paper discussed the concept, process and applications of text mining, which can be applied in multitude areas such as web mining, medical, resume filtration, etc. It also enlightens the hidden potential that lies in the field of text mining and motivated to explore it further.

Agrawal et al. [6] in the journal titled "A Detailed Study on Text Mining Techniques", puts focus on the study of the concept of Text Mining and various techniques used in it. It also describes the major ways in which text is mined when the input is plain natural language, rather than partially-structured Web documents. In structured text they have discussed how internal documents structure and external structure is mined which gives explicit hypertext links between documents.

Text Mining Framework: Text Mining techniques allow us for mining the actual information present within the text. The information extraction software identifies the relationships with in the text as natural language text contains information that cannot be directly used for mining.

Waegel [1] defined text mining as "the non-trivial extraction of hidden, previously unknown, and potentially useful information from (large amount of) textual data". Text mining is the process of collecting data from variable resources, retrieving and pre-processing the information, and applying text mining techniques to extract Knowledge. Jusoh et al. [7] described it as a relatively new and vibrant research area which is changing the emphasis in text-based information technologies from the level of retrieval to the level of analysis and exploration.

In recent years, exponential data increase that comes in the form of emails, text or word documents and web postings on shared media streams like Facebook, Twitter, etc. made text mining a necessity. Text mining uses such random, unstructured textual data and transforms it following a series of crucial steps to obtain potential information from text which are discussed in Figure2:



a) **Document collection:** In these recent exceedingly complicated technological years, data has increased exponentially which usually comes in the form of emails, text or word documents and web postings on shared media streams like Facebook, Twitter and LinkedIn, etc. Text mining uses such random, unstructured textual information and transforms it into useful information and knowledge by performing the steps discussed as below.

b) **Retrieve and Pre-process Information:** This step involves retrieval of the pure text from the available sources of data, for example, extracting text from HTML pages by removing the HTML tags, etc. This retrieval process is followed by a series of pre-processing steps that take place as follows:

- *Tokenization*, breaking down the continuous text into discrete tokens.
- *Part-of-speech tagging*, assigning to each token its respective grammatical category.
- *Lemmatization*, converting the tokens into their base forms (e.g. was → be)
- *Chunking*, dividing the words into syntactical categories (e.g., [The Queen]NP [was beautiful]VP, respectively the noun phrase and the verb phrase of the example)
- *Parsing*, to recognize the sentence structure.

c) **Information Extraction using Text Mining Techniques:** IE can be described as the creation of a structured representation of selected information drawn from texts. Jusoh et al. [7] explained that natural language texts are mapped into a structured representation, or templates which represent an extract of key information from the original text. Text Mining techniques allow us to mine the actual information present within the text.

d) **Management Information System:** The retrieved information is fed into the management information systems. These management information systems are the computer systems that use this information to improve efficiency and effectiveness of decision making.

e) **Knowledge:** All we have now is the useful information and when this information meets our requirements, brings to us the unknown facts and statistics, and helps us in taking better decisions, it is then called Knowledge.

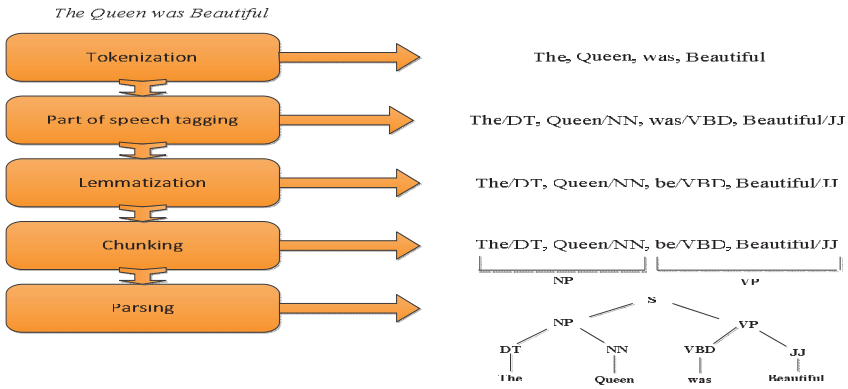


Figure 3: Pre-Processing steps

Text Mining Approaches: The section underneath highlights the concepts, advantages and disadvantages of various text mining approaches been discussed in the following categories:

A) Term Based:

The oldest known Text mining technique is the term based search. On the basis of the terms extracted in the pre-processing step, we now need to structure the text by finding the potential terms which may be used as a tool to dig into the text. Out of the big number of terms found, we use following techniques to rate these terms:

1. tf.idf: This is the classical information retrieval metric defined as the number of occurrences of a term in a given document multiplied with the (often log smoothed) inverse of the number of documents where the term appears. This is a measure of phrase importance, which promotes candidates that fulfill the basic requirements of term selection.

2. χ^2 independence test: This test helps us determine whether two events occur together more often than by chance.

Count (phrase in document)	Count (all other phrases in document)
Count(phrases in other document)	Count (all other phrases in all other documents)

3. Key phraseness: This method exploits the vast information contained in the already annotated documents. We estimate the probability of a term W to be selected as a keyword in a new document by counting the number of documents where the term was already selected as a keyword (count(D_{KEY})) divided by the total number of documents where the term appeared (count(D_W)).

$$P(\text{Keyword} \mid W) \approx \frac{\text{count } (D_{KEY})}{\text{count } (D_W)}$$

Table1: Advantages and disadvantages of text mining approaches

	ADVANTAGES	DISADVANTAGES
TERM BASED	• Efficient computational performance. • Mature theories for term weighting.	• Polysemy (words with multiple meaning). • Synonymy (multiple words have same meaning). • Lack of precision and recall.
PHRASE BASED	• Less ambiguous. • More discriminative.	• Have inferior statistical properties to terms. • Low frequency of occurrence. • Contain more redundant and noisy phrases.
PATTEN BASED	• Good statistical properties. • Free from polysemy and homonymy.	• Low frequency (noisy data). • Misinterpretation (not understanding what users want).

B) Phrase Based:

The next known approach is phrase based approach in which, instead of using terms, we use phrases to classify the text. Phrases are preferred to terms as these are more precise in context. Again we can use the same approaches as used in term based search to rate the key phrases.

C) Pattern Based:

The last one is the pattern based search. In this, we search for the most frequently occurring patterns in the text may that be sequentially ordered, or randomly associated term set or itemset.

Checking on the above mentioned techniques we try to figure out their pros and cons:

Text Classification Techniques: We use the extracted terms, phrases or patterns to classify the text in hand. So, in the Information extraction process we need to identify certain classes for this text to be a part of, thus structuring the text. For this we need to be familiar with the classification techniques.

We may choose any of the following techniques:

a) Bayesian belief network

A Bayesian Belief Network has a directed acyclic graph and a set of conditional probability tables. The nodes of the graph represent either attributes of given data or hidden variables that form relationships. Each variable is conditionally independent of its non- descendants in graph

b) Back propagation algorithm:

Back propagation is a neural network learning algorithm. It learns by adjusting the weights to be able to predict the correct class label of input tuples. It is also called connectionist learning.

c) Support Vector Machines:

This is a kind of supervised learning model. SVM is used for pattern recognition and classification. It can be used for linearly as well as non-linearly separable pattern classification. It searches a hyper plane to distinguish between positive and negative patterns.

d) Frequent Patterns:

Frequent patterns highlight interesting relationships between the attribute-value pairs in a given set of data, for which the frequency of occurrence is high. For an example, age= youth, credit= OK occur in 20% of data tuples that describe customer data if he may or not purchase a computer.

e) Lazy Learners:

In contradiction to eager learners, the lazy learners are not ready and eager to classify the tuples which are not seen before. It rather waits until the last minute before constructing a model to classify the given test tuple.

Examples:

i. k-Nearest Neighbor Classifiers: It is based on learning by analogy. It compares the test data with training data and searches for 'k' data items from training data which are closest to the test data in terms of distance metric.

ii. Case Based reasoning: It relies on the previously stored problem solutions to solve the new problems. It store the information is the form of cases or complex symbolic descriptions. Given a test case, the CBR check the data store to relate if an identical training exists.

f) Genetic Algorithm:

It is based on the concept of natural evolution. A starting population is created which contains some

randomly generated rules. According to the notation of survival of fittest a new population is formed that has the fittest rule in the current population as well as offspring of their rules.

g) Rough set approach:

The rough set approach can use noisy or imprecise data to develop unstructured relationships among them. It forms the equivalence classes of the given training data. It does not only classify the distinguishable classes but can also approximately or roughly define the undistinguishable classes.

h) Fuzzy Set approach

Fuzzy approach is a possibility approach and is used against the traditional two-value logic or probability approach. Using Fuzzy theory we can deal with vague or inexact facts.

Table2: Advantages and disadvantages of text mining technique

Technique	Advantages	Disadvatages
Bayesian Belief Network	• Can take human inputs as prior knowledge to improve learning rate. This is possible through the explicit representation of casual structures of belief networks.	• Computationally intensive
Back Propagation Algorithm	• It is highly tolerant to noisy data • It is suitable for continuous valued input and output • Can be successful for real world data e.g. handwritten character recognition, pathology, pronouncing languages, etc. • It is inherently parallel which makes computation process really fast	• Poor interpretability • Can take human inputs as prior knowledge to improve learning rate. This is possible through the explicit representation of casual structures of belief networks.
Support Vector machines	• SVMs produce highly accurate results. • Can model complex non-linear decision surfaces. • There are less chances of over fitting problem in SVMs than the other methods.	• The speed of computation of SVMs is a big concern as even the fastest of the SVMs can be extremely slow.
Frequent Patterns Approach	• It has greater accuracy and scalability than some of the traditional techniques like C4.5. • It can also explore the highly confident associations among more than one attributes.	• Representing all the frequent patterns takes a lot of memory • Support counting takes very long for large transactions.
Lazy Learners Approach	• Naturally support incremental learning • Can be implemented on parallel hardware • They can model complex decision spaces having hyper polygonal shapes.	• Computationally expensive • Require efficient storage techniques • Don't offer details of data structures
Genetic Algorithm	• It can easily be parallelized • Can be used for optimization as well	• No assurance of a global optimum • Some poorly known fitness functions hinder some optimization problems
Rough Set Approach	• Can induce rules by analysing incomplete datasets • Uses three regions: Acceptance, Rejection and Deferment for decision making	• High space and time complexity
Fuzzy Set Approach	• Allows imprecise or contradictory inputs • Rule base can easily be modified • Increased robustness • Simplify knowledge acquisition and representation	• It's difficult to develop a model from a fuzzy system • Requires more simulation for being operational

Conclusion: A detailed study on the text mining reveals that the disadvantage of term based approach

has a severe effect on the performance of the system which cannot be ignored, although it could have provided the most noise-free results if the problems of polysemy and homonymy could have been removed. Also the discussion about the advantages and disadvantages of the text classification techniques used in text mining process are discussed in Table1. We infer from this table that the fuzzy set approach and rough set approach are more flexible to the noisy data. Where fuzzy system is capable of dealing with inexact and imprecise data, the rough set approach can be used to roughly or approximately categorize the indistinguishable data.

Text mining still remains a potential field of research when it comes to bringing about a meaningful interpretation of natural language text.

Acknowledgement

This is a humble effort to express our sincere gratitude towards our advisor **Md. Ataullah** (Assistant Professor, Department of Computer Science, Lovely Professional University) who suggested many related points, is always very helpful and constructive and under whose firm guidance, motivation and vigilant supervision we could present this paper.

Also we are thankful to **Mr Kamal Deep Garg** (Assistant Professor, Department of Computer Science, Lovely Professional University) who has guided and helped us to explore the field of Text Mining. We are grateful to him for imparting so much valuable knowledge and for all his encouragement and words of kindness.

References

1. Daniel Waegel, "The Development of Text-Mining Tools and Algorithms", Ursinus College, 2006.
2. Jiawei Han, Micheline Kamber, Jain Pei, "Data Mining: concepts and techniques", Morgan Kaufmann Publishers, 3rd edition, 2012.
3. Kumar Lokesh, and Bhatia Parul Kalra (2013), Department of IT, Amity University, Noida, U.P., India, "Text mining: concepts, process and applications ", JGRCS Journal of Global Research in Computer Science, Volume 4, No. 3, March 2013.
4. Sukanya M., and Biruntha S. (2012),"Techniques on Text Mining" , Department of Computer Science and Engineering (PG) SNS College of Technology, Sathy Main Road, Coimbatore-641035, India , 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).
5. Zhong Ning , Li Yuefeng, Wu Sheng Tang, and, "Effective Pattern Discovery for Text Mining" by IEEE transactions on knowledge and data engineering, VOL. 24, NO.1 , JANUARY 2012
6. Agrawal Rashmi and Batra Mridula (2013), Department of Computer Applications, Manav Rachna International University, Faridabad, India, "A Detailed Study on Text Mining Techniques", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-6, January 2013.
7. Jusoh Shaidah and Alfawareh Hejab M., "Techniques, applications and challenging issues in Text Mining", College of Computer Science & Information Systems, Najran University, Saudi Arabia, IJCSE International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, December 2012.

***

Final Year studentS M.Tech (CSE), Department of Computer Science Engineering,
Lovely Professional University, Punjab, India,
shwetima0290@gmail.com
gurpreet11307@gmail.com

Asst. Professor, Department of Computer Science Engineering,
Lovely Professional University, Punjab, India, mdataullah.khan@gmail.com