A Survey on Text Analytics and Text Mining

C.P. Thamil Selvi, Dr.PushpaLaksmi

***Abstract*: Text analytics is rapidly growing day by day in computing world. Text analytics and text mining is a necessary process integrated with the several recent research areas like information retrieval and computational linguistics. Generally natural language process is used for knowledge extraction from ontological data generated and written from various sources and human beings respectively. To proceed with text mining, text analytics is important to analyse the text data, because it is in the form of unstructured. Several recent research works were focused on implementing an efficient text analytics and mining approaches based on various dataset. But, still there is a need of improving the accuracy of classification and mining over text data. Before going to design and implement a text analytics and mining method, it is essential to understand the issues and challenges met by earlier research works. Also, social networks and other communication networks, people are sharing their own pattern of text which has spelling error, grammatical error and sentence error, needs to be corrected. Hence, text analyzation and mining are a complicated task from social network data. Some of the review documents discussed about information extraction and some emphasized about different applications of text mining. But, most of the earlier review and survey documents do not target especially on the social networking data sets. The goal of this paper is presenting a detailed survey of several text analytics, and text mining methods based on the social networks. Also, this survey examines and explores the recent emerging methods used in the text analytics domain. Finally, it has been provided a summarized report for easy and fast understanding the various issues and challenges in text analytics and mining. It helps to find out a solution for big data analytics.**

Index Terms: Text Analytics, Text Mining, Social Network Data, Twitter Dataset, Feature Extraction, Classification, Optimization Algorithms.

Revised Manuscript Received on December 22, 2018.

C. P Thamil Selvi, Associate Professor, Computer Science and Engineering Department,Sri Ranganathar Institute of Engineering and Technology, Athipalayam, Coimbatore.

Dr. PushpaLakshmi, Professor, Department of Information Technology PSNA College of Engineering and Technology, Dindigul.

I. INTRODUCTION

One of the types of data mining is text data mining or text mining. It is roughly equivalent to text analytics. Text Analytics (TA) is the process of extracting most essential information from text. It can be obtained by clustering and creating meaningful text patterns to learn the entire text. This follows the process of statistical pattern learning. The process of exploring the data mechanically that is accumulated in huge quantity is known as Data mining, it also helps to find out certain models also its process which could be carried out easily. The data are partitioned also the upcoming events are estimated in data mining by utilizing arithmetical algorithms. Data mining is an upcoming novel technology also it is referred as Knowledge Discovery in Data (KDD). In Data mining the huge quantity of data are accumulated in data warehouse with the help of several methods likely,

- Machine Learning
- Artificial Intelligence
- Statistical Learning Methods

Various data mining research works have been used the above said methods generally. To recognize the language normally, text mining also data, Text Analytics is the word that is currently used nowadays. Big Data is basically an unstructured text in the recent years in the commercial areas since 80%of these unstructured free texts are employed in production framework which consists of surveys, wikis and tweets (**Xerox Corporation, (2015)**). More research works are carried regarding the Text Analytics. The Text Analytics consists of novel methods which are not used earlier; it does this by fetching the details mechanically with the help of various used defined sources.

In the recent times, Text Analytics has achieved a huge response in various business as it extracts and obtain huge

amount of text which helps the entrepreneur to value the current trend in market areas also it helps them to easily locate the fraudulence. The Text Analytics has a unique feature which combines expertise in various fields such as, specialists in the fields of

- Law
- Medicine
- Data scientists
- Psychologists, also
- Computer engineers.

Because of these various experts, the various process in data mining research and development gets segmented.

II. LITERATURE SURVEY

Several research works have been focused on solving text analytics and mining problems. In order to understand the research problem regarding text analytics and text mining, it is necessary to do a detailed study about the various methods proposed in earlier research work and faced problems.

Searching in Internet

Internet contains a huge volume of knowledge also it has n number of journals that could be accessed easily by any common man and it becomes difficult while accessing the reliable data. **[Stringer et al. (2008)]** stated that "the development in research literature has leaded a huge impact upon researchers". The digital libraries should be enhanced so that the documents are recognized easily it is also employed in such a way that it assures different requirements is given in **[García-Crespo et al. (2011)]**. Author in, **[Lee et al. (2010) and Ma et al. (2012)]**, stated that the classification and analysis discussed in the earlier research works were highly focused on extending the existing research areas.

Mostly keywords are used by the search engines also by the bibliometric research; in order to find out which domain the research articles belong to, also the important documents need to be recognized, is discussed in **Hjørland, B. (2012)**. The natural language text is a key source of knowledge for text analysis and it is utilized mainly for grouping text together also to find out which kind of text they belong to. Concept mining is a similar form of text analytics, in which ordering of documents takes place by considering the similar model as a base. The documents could be easily compared with the help of concept mining, in **Tseng et al. (2010)**, depending upon the outcome of research it also helps in organizing the document is discussed in **Bichindaritz and Akkineni, (2006)**. Two different classification methods are used for text mining based on the similarity score and it is presented in Ma, L. (2012). The major reason involved neither for categorizing the research for different regions is that the research could be either a theoretically based research nor by applying the outcome of research. Once the research is published a greater number of viewers finds it easy when they employ it to attain the end outcome of published research as **Kellogg et al. (2007) and Susman et al. (1978)**.

Text Mining in Social Networks

A detailed discussion is presented in **Rizwana Irfan et al. (2015)** regarding text mining in social networks. Various methods used for text mining using text patterns are used for social network websites. A fast and effective communication medium to share the public opinion is social networking. In recent days social network becomes a backbone of the e-commerce. **Baumer et al. (2010)** told, social websites are making people to be engaged under various communities. People can share their valuable messages and various morals. **Sorensen (2009)**, presented that social websites are powerful communication medium where any individuals can share their valuable knowledge and learn mutually. Some of the famous websites such as Myspace, LinkedIn, Facebook and Twitter, where any user joins and link with one another under various communities. **Evans et al., (2010)** said, the coordination problems among people can be solved by social networks, due to the geographical distance, it is also said by **Li et al. (2011b)**. **Li & Khan, (2009a, 2009b)** reported that, making all social campaigns as effectiveness by broadcasting the essential information anywhere and anytime, it is also said by **Baumer et al. (2010)**. **Sorensen, (2009)** also said that, the data available in the social websites are unstructured in language. During conversation, people don't care about the spellings, grammar and accurate sentence formation. Hence, extracting the information from unstructured dataset is difficult.

Text Analytics

For the execution of bibliometric analysis also for exploring literature such as, **Bragge et al. (2010)**, **Porter et al. (2002)** and **Raghuram et al. (2010)** the Business Analytics research, **Balakrishnan et al. (2010)**, make use of text analytics which is one of the most commonly used method. The text analytics provides a solution to these queries such as, what are the regions that has a scope for novel research space, also the name of the publisher who helps them as well as the region where the research work is being published in **Yang et al. (2008)**. The structure of the text analytics used in commercial sector is categorized into three different forms: they are,

If they could work with the document that is unstructured,

If they could work with the document that is structured with huge knowledge, and the final one is

Stick to a particular domain.

The academic article follows a structured framework but the kind of data that is obtain by bibliometric is not structured in such cases the common method of text analytics must be employed. During the knowledge discovery and pattern extraction several related information or certain features are always hidden. To enable the hidden features of the text patterns, the bibliometric text analytics is used in the large size datasets, **Bragge et al. (2010)**. Utilization of keywords for categorizing text has become a significant research method discussed in **Conway, M. (2010)**. **Glänzel, W. (2012)** specifies the most important step in recognizing the latest styles in literature by finding out the most generally used words inside the research area. In **Seol et al. (2011)**, it illustrates that how efficiently the keywords are employed in research area for categorization of article in text analytics, to categorize the present research work it also uses the keywords that are used in previous text analytics, **Ma, L. (2012)**. The research work carried out by this article specifies that "single-word" patterns are used by the theoretical research when compared to applied research work (e.g., Business Analytics).

Information Extraction

Information extraction, title/topic extraction, clustering and classification, context linkage, data visualization and ranking topics, deep learning and question answering are the various methodologies have been designed to do text mining. Information extraction is a process where it identifies the keywords and the associated text in the text-dataset. It searches for a keyword or a text pattern matching with dataset, using a regular expression. Information extraction is generally available in a specific form of Named Entity Recognition (NER), where it compares and identifies atomic elements in the text into predefined ontologies. It also extracts various features like person, location, organization, gene pattern, temporal expressions, etc. Various tools relevant for this IE and NER, Apache-Open-NLP **(http://opennlp.apache.org/)**, LingPipe **(http://alias-i.com/lingpipe/demos/tutorial/ne/read-me)** and Stanford Named Entity Recognizer **(http://www-nlp.stanford.edu/ software/CRF-NER.shtml, Finkel et al. (2005)**. In order to improve the accuracy of text mining, ranking, summarization and classification methods are used. A common architecture of information extraction is depicted in Figure.1.

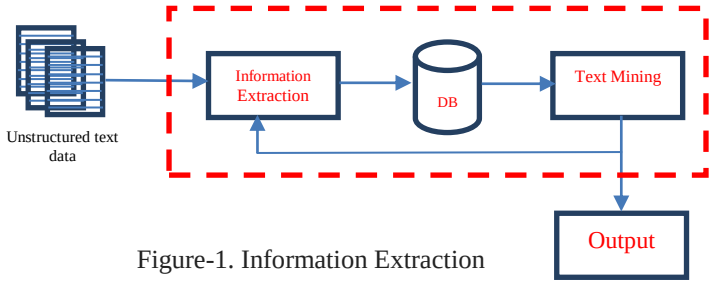

Figure-1. Information Extraction

Pre-processing

Dai et al. (2011) said that, during the text analytics, if the texts are loosely connected, then there will be more chances for data loss. Also, if the text-scanning process is not done carefully, then it will create garbage-in and garbage-out. Author in **Forman and Kirshenbaum (2008)** stated that, the accuracy of the text mining will be decreased when the text data is unstructured which creates a poor text analytics. The author also discussed that, pre-processing is an important for categorizing the text in to fixed number classes. Feature extraction and selection are the two essential process can improve the text analytics and mining process. From the above discussion and from other studied existing in internet sources, it is understanding that, pre-processing is an essential work for improving the data quality. But mining the data, Feature extraction and selection is most important.

Feature Extraction

There are three different ways of features extraction methods are used in general such as, syntactical analysis, morphological analysis and semantic analysis. **(Forman & Kirshenbaum, (2008))**, discussed about Morphological analysis, where it uses each word as a token in the input document. It starts with tokenization, removing the stop words, irrelevant characters, error elimination and stemming. Initially the document is converted into sentences, from the sentences the punctuations are eliminated **(Negi et al. (2010))**. The set of all stop words like, "is", "was", "the", "a", "that", "where", "how", "are" and etc., are removed. This punctuation and stop words elimination, the proficiency of the text analytics has been improved because analyzation process is only applied on the texts **(Shekar & Shoba (2009))**. **(Forman & Kirshenbaum, (2008) and (Shekar & Shoba (2009))** has been presented that stemming on the text document is a process of linguistic normalization method. Various stemming algorithms are:

- Brute-Force
- Suffix-Stripping
- Affix-Removal
- Successor Variety And
- N-Gram are discussed by (**Forman & Kirshenbaum, 2008; Shekar & Shoba, 2009**).

Yuan, (2010) said that, a grammatical correction of a sentence can interpret the logical meaning of it. Syntactic analysis provides grammatical correction model of a predefined languages (English) which is called as syntax. Two processes are used in syntactic analysis such as parsing the text and tagging. Tagging is a technique used to add contextual information for each word in the document based on sentences. **Yoshida et al. (2007)** reported that, understanding the lexical class of the words make easy and fast of linguistic analysis operation. **Yuan, (2010)** said that several approaches given in the literature follows the dictionaries for implementing the POS-Tagging. One of the famous and accurate approach is HMM approach, where it follows rule-based morphological analysis after document is tokenized. HMM is called as stochastic tagging model since it obtains similar POS-TAGs from the input sequences **(Yuan, (2010))**. **(Ling et al., 2006)** said that, generally, the text documents use tree structure for examining and parsing the grammatical structure on the sentences. It is called as parse-tree used to arrange the sentences based on the accurate grammar order. The tree construction follows a top-down or bottom-up approaches.

Strapparava & Ozbal, (2010) presented that, any communication tools used for distributing the knowledge to the society it is necessary understand the meaning of the sentences. Keyword Spotting Method (KSM) extracts necessary text information from the dataset is discussed in **Ling et al., (2006)**. KSM follows a semantic lexicon for classifying the words using WordNet-Affect. It is also known that SentiWordNet used WordNet for meaning based text collection in emotion mining in text datasets **Esuli & Sibastiani, (2006)** over positive and negative emotions. **Ling et al. (2006)** stated that KSM is a state-of-the-art method used mainly for semantic analysis. Based on the syntax the emotions are analysed based on the context and structure of the pattern discussed by **Ling et al. (2006)**. KSM collects the set of all keywords based on the emotions, **Wollmer et al. (2009)**. In emotion mining, the set of all verb, noun, and adjectives are considered as keywords. Whereas, certain words do not represent the emotions.

To overcome this kind of issues, KSM is used for obtaining words by understanding them based on the semantics, **Ling et al. (2006)**, is called as semantic networks. It used to describe the relationship between the concepts of the words and events. **Ling et al. (2006)** concluded in his research article that emotion mining can be obtained effectively with improved performance using KSM. **Ling et al. (2006) and Li and Khan (2009a, 2009b)** said that emotion words extraction is obtained by semantic networks using contextual information of the texts from small size dataset. For large size datasets, WordNet-Affect and SentiWordNet combination for enhancing the accuracy. Only feature extraction cannot improve the efficiency of mining, since the dimensionality after pre-processing and feature extraction is same. Due to high dimensionality the time and cost complexity is always high. One of the methods to do dimensionality reduction to

decrease the overall computational complexity is selecting the appropriate features associated with maximum amount of dataset. Hence some of the authors in the earlier research works have been focused on feature selection.

Feature Selection

Basically, text analytics and mining process has more complexity regarding time and cost. Cost purely depending on the time taken for the entire process. In order to reduce the time complexity, the dimensionality of the dataset is reduced by feature extraction and feature selection. Some of the earlier research works extract the features and proceed for comparison. But selecting the appropriate feature set used for mining process can reduce the dimensionality. Feature selection method eliminate redundant and irrelevant information and features from the final dataset. Feature selection method assigns a score value to each word for easy and fast process, **Hua et al., (2009)**. He also stated that, text documents can also be used as vector model where it helps to reduce the dimensionality and parse the text. If the text data is represented as M X N matrix, then all the non-zero elements represent text present in the document (in case, M number of documents, with N number of texts). **Shekar & Shoba, (2009)** used feature vectors to represent the feature set extracted from the document.

For easy mining, text frequency and document frequency are two different parameters can be calculated. **Yoshida et al. (2007)** combined both text frequency and document frequency for calculating the relevant texts availability in the document.

Ma et al. (2005) and **Li et al. (2011b)** appli ed similarity-computation and semantic-relation for pre-processing. For calculating the similarity value between two entities, it is essential to have an adequate corpus. Also, the efficiency of the algorithms can be examined regarding data scalability requirements to be addressed **(Zhao et al. (2009)). Yoshida et al. (2007)** used various feature selection methods like random-mapping, semantic matching etc. In term of text mining, **Durga & Govardhan (2011)** used mining methods in analytics based on supervised and unsupervised clustering with classification. Once the dataset is cleaned, arranged, clustered, and the best features are selected then it is able to do mining process effectively and efficiently.

Sentiment Analysis

One of the data analysis method is sentiment analysis, where it is used in various real-time social network-based dataset. For example, the Sentiment analysis (SA) is carried out by extorting the sentiments of customers and the document is evaluated in the form of text **Pang and Lee (2008)**. In the present world people's vision upon specific matters is articulated with the help of media, also the internet which provides the social network, websites, blogs, moreover discussion forums. In order to explore this information that is provided by media as well the internet SA technology is required. For the consumers this technology is very useful particularly when he needs to sale the products, where novel consumer will try assessing the product quality based upon the remarks provided by the consumer who had used the product previously. The remarks provided by the consumers enable the other consumer to decide and to purchase the product. With the help of SA technology, the organization acquires its profit because of the remarks given by the consumers in the field of product quality, **Agarwal and N. Mittal (2013a, 2013b)**. This SA technology can also be used in various other fields like,

1. politics,
2. services,
3. organizations,
4. events and
5. issues.

This technology is a mixture of

1. text mining technology,
2. natural language programming (NLP) and
3. text classification

is given in **Seerat and Azam (2012)**. Text processing is very strongly associated with SA technology where the information obtained in the text documents are categorized likely the reviews on products, films, services and other disciplines, in order that the document has as positive or a

negative sentiment, **Chen and Lee (2011), Arafat et al. (2014) and Abbasi et al. (2011)**. The machine learning process is employed to categorize the sentiments, it is also pooled with linguistic process in which it recognizes the kinds of sentiment present in the document as a text form, **Abbasi et al. (2011)**. The SA technology consists of four major procedures:

Processing of data

Characteristics are selected

The relationship is recognized by its characteristic also by its sentiment words

Categorization of sentiments.

In SA every procedure is significant which helps in obtaining high quality sentiments precisely. The challenges that we face while categorizing the sentiment data are large in volume, has overlapped features and irrelevant, said by **Arafat et al. (2014), Abbasi et al. (2011), and Vinodhini and Chandrasekaran (2013)**. A process for categorizing the text is required. Text document signifies the model known as Bag-of-Words (BOW) which is employed while categorizing the sentiment by utilizing a technique called machine learning, **Agarwal and N. Mittal (2013a) and Arafat et al. (2014)**. Future vector is generated by the words that are found in documents which have the text in it. While categorizing the sentiment this future vector affects the vectors dimension by making it larger which in turn it influences the precision rate, **Agarwal and N. Mittal (2013a) and Arafat et al. (2014)**. By utilizing an apt FS this hitch could be controlled by a technique called machine learning where it removes the overlapping characteristics also the area which has a larger dimension, **Agarwal and N. Mittal (2013a), Arafat et al. (2014) and Abbasi et al. (2011)**. A feature subset is generated from a novel feature set by removing the overlapping characteristics and it is said to be one of the major goals of FS. Based on, **Yusta et al. (2009) and Uguz (2011)**, the process of finding the feature is split into two such as, selection of feature and the other one is removal of feature. From a large sized feature group, a single feature is selected which is the goal of selection feature.

Many researchers have been researching the field of "Sentiment analysis". The research has carried out century ago in which the primary phase involves in binary categorization, where a positive or negative comment is specified in bipolar group. **Turney (2002)** reported that parsing the positive and negative phrases can be done accurately by unsupervised learning methods. In **Ch.L.Liu et al. (2012)**, depending upon the filtering mechanism latent semantic analysis (LSA) is utilized by feature product in order to recognize how the words are placed efficiently so that an elegant review is attained. In sentences the positivist and negativist is compared and utilized in paper, **Luo and Huang (2011)**. The knowledge is revealed with the help of internet also the words are positioned manually. The author **Liu et al. (2010)**, utilizes a technique called rule-based that depends upon Base Line it also employs SVM for analysing the sentiments of Chinese document whose précised words are extracted from a sentiment word dictionary of the entire document. In **khan and Baharudin (2011)**, based upon the context of sentences, either the sentence is a positive or negative, the polarity words is evaluated including every word that is present in the sentence. The data is pre-processed, so that the quality of unstructured sentence framework needs to be enhanced and it was stated by **Lakshmi and Edward (2011)**. The sentiments are tested by a method called LSA. The sentiments are categorized by utilizing a phrase model and it was suggested by **Basant Agarwal et. al. (2013)**. In **Zhu et al. (2011)**, author specified a method in which the queries need not be answered by the customers. **M. Karamibekr and A.A. Ghorbani (2012)** they both devised a technique especially for document that are in social domain where verbs are utilized for categorization of sentiment. In Paper **A.Neviarouskaya et al. (2011)**, a sentiment wordlist is produced and it is known as SentiFul and it gets extended via:

Synonyms

Antonyms

Hyponyms Relations

Derivation and

Compounding.

Depending upon the character in which they are portrayed in characteristic sentiments, a novel technique is projected in order to differentiate four forms of affixes, such as, issues in propagation, weakening, reversing, and intensifying. In

order to enhance the sentiment analysis, the above stated techniques assist in increasing the list of words. The researchers have worked on major areas also employed certain techniques such as fuzzy logic and visual works intended for sentiment analysis. In **Liu et al. (2012) and Srivastava and Bhatia (2013)** it contains illustrations depending upon a strategy called fuzzy logic. Author in **Liu et al. (2012)** said that the fuzzy domain sentiment ontology (FDSO) plays a major role in text analytics and mining. Initially a fuzzy rule set is created as a tree based on the sentiment ontology where it includes the set of all sentiment words, product features, and the relation between features used to predict the polarity words. In **Srivastava and Bhatia (2013)** depending upon the role of membership the authors have proposed a method known as fuzzy inference. With the existence of adverbial modifier, the reviewer's view has gained power with the help of membership functions. For adverbial modifiers a technique known as tri-gram model is pertained.

Classification for Text Mining

Classification process can be obtained by supervised or unsupervised learning methods. These methods learn the data, apply rule-set and extract the information in training process. Each text in a document is classified based on a category is applied for text mining **(Yin et al., 2007)**. There are two different classification methods are used for text mining, are:

Machine learning-based text classification (MLTC) and
Ontology-based text classification (**Xu et al., 2008**) and is illustrated in Figure 2.

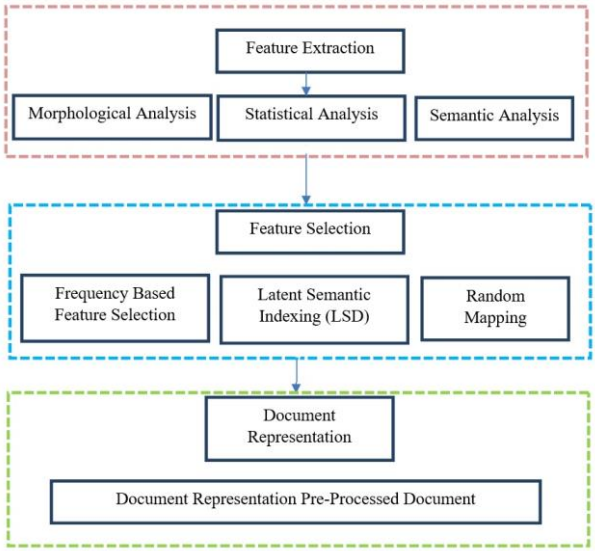

Figure-2. Data Pre-Processing

Opinion Mining

In the recent years opinion mining is a current method that put together the process of retrieving the knowledge also Computational Linguistics (CL) which specify the opinion of document and does not only it describe the topic present in the document. Opinions could be easily fetched from various sources likely, Online forums, newsgroups, blogs, from which huge knowledge could be ascertained.

Nowadays customers shop through online in which they get everything easily by just selecting the product and paying it via online it also helps them to save time. If a customer wants to buy a product, they clearly read the review which consist of the entire description of the product including its colour, size, price and so on, which is given by the online retailer. It also gives the customer a clear decision whether to purchase it or not. The retailers request the customers to express their reviews as their views are very truthful also it helps them in delivering a better product than before.

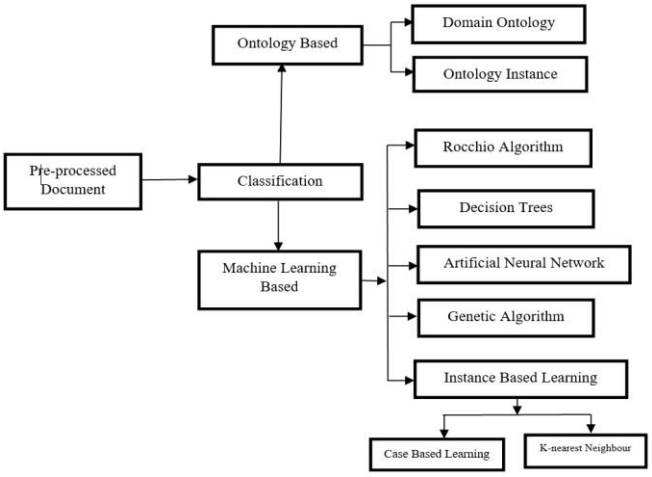

Figure-3. Various Classification Algorithms for Text Mining

According to **Score and Kelsey group (2017)**, since many reviews are expressed by online customers it has created an impact in future buyers. The reviews suggested by the customers increase gradually which becomes difficult for an individual customer to go through all the reviews and then to take a proper decision. When a customer reads very few reviews then in such cases opinion are iniquitous. Hence, automatically mining the reviews also summarization has become the present research issue. The easy access of customer reviews has paved a way for opinion mining, which gives the summary of customer opinion about the

products **[Popescu and Etzioni, (2005), Hu and Liu, (2004a, 2004b), He et al. (2008), Kim et al. (2006), Liu et al. (2005)]**. Mostly the current work emphasizes upon the reviews attained by the product.

The OPINE method was stated by **Popescu and Etzioni, (2005)** in which this method helps in leisurely locating words. The Pulse structure was projected by **Gamon et al. (2005)**, in which a method called bootstrapping instructs a sentiment classifier. Depending upon the key words the sentences are categorized and so removal of feature take place. A technique was developed by **Hu and Liu, (2004)** in which they make use of quality words, also the number of times the words occur and the part of speech. Once the features are removed the sentences are rebuild and then an overview is generated. The author in, **Bing Liu et al. (2005)**, presented a list of opinion words, which comprises of nouns describing the review comments of the products in a specific format. Considering the previous works of sentiment analysis, its main aim was to generate a review which obtains a positive or negative feedback. The author, **Pang et al. (2005)**, suggest that the opinion of the author's should be decided by considering all the views. For categorizing the sentiment **Whitelaw et al. (2005)**, devised a new technique.

One of the groups in data mining referred as appraisal group, which categorize the set of all attributes in the semantic taxonomy. With the help of semi-automated method, a word list considering adjectives is created. These features are employed depending upon the general feature known as "bag-of-words". The reviews of movie are graded and the outcome attained is 90.2% the exact precision. Hence the projected technique has improvised compared to the traditional outcome. Another two methods were devised by **Kennedy et al. (2006)** were the sentiment was determined while articulating the movie review. According to the orientation of semantic, its reviews could be either a positive one or negative one.

There are three different classified results are produced such as, intensifiers, diminishers and negations during the classification process. In order to quash a specific word in semantic polarity, diminishers also intensifiers have been applied moreover the degree of a word might be neither positive nor negative. By applying General Inquirer, we could easily identify whether the term is negative or positive including the intensifiers, negation and diminishers. The similarity score is calculated for finding the semantic based texts. By using term-counting method also by contextual valence shifters, were in these two improvises the accuracy categorization. The next technique makes use of Machine Learning algorithm such as SVM.

In general, the unigram features are firstly utilized then comes in by adding bigrams which consists of an extra word were it also includes the valence shifter. High accuracy rate is attained with the help of bigrams present in valence shifter and the rate is high because of the characteristics of words that are presented in the lists which may be either a positive word or a negative word. A novel technique is devised for sentiment analysis by **Mullen et al. (2004)** were it combines various sources of knowledge by deeming suitable assess for adjectives and phrases also the information in the form of text that is available in support vector machines (SVMs). In the past many features and prototypes are been applied more to the prototype of unigram which has proven to be competent. Depending upon the SVMs the unigram features are combined with the hybrid SVM in such case it provides a better performance value also produces an efficient outcome which is ascertained from a movie review known as Epinions.com.

Table-1. Summary of the Survey for Text Analytics and Mining

Author and Year	Method Proposed	Merits	Limitations
Hjørland, B. (2012)	Bibliometric research	Recognizing the Documents	More Time complexity
Tseng et al. (2010)	Concept Mining	Retrieve the similar document	More irrelevancy
Rizwana Irfan et al. (2015)	Detailed survey on text mining	Concluded that Social networking is a sharing media to the public	Discussed only about opinion mining.
Sorensen (2009)	Discussed about social networks	Said, social networks are powerful communication medium	Information based on people communication.

Bragge et al. (2010).	Bibliometric text analytics	It enables the hidden features of the text patterns	Verified with Large size dataset.
Conway, M. (2010)	Text Classification using keywords	80% of classification accuracy is obtained.	Time complexity is high
Dai et al. (2011) and Forman and Kirshenbaum (2008)	Pre-processing on text data	Unstructured data is converted into structured data	Not common, Suits only for certain number of classes.
Forman & Kirshenbaum, (2008)	Morphological analysis-based feature extraction	Convert the input document into tokens speedily.	It needs pre-processing.
Negi et al. (2010)	Carried out a sequence of data analytical process	Tokenization is effective.	Need a template for stop words and punctuation.
Yuan, (2010)	Syntactic analysis using HMM model	Interpret the sentences into logical meaningful.	Need to include Grammatical Correction.
Ling et al. (2006)	Parse-tree based text analysis	Top-down/Bottom-up approach is used for improving the efficiency.	Can not applicable for all kind of documents.
Ling et al. (2006), Esuli & Sibastiani, (2006) and Wollmer et al. (2009)	Keyword Spotting Method	Emotion mining is efficient	Needs to apply for common text mining.
Li and Khan (2009a, 2009b)	WordNet-Affect and SentiWordNet	Accuracy is high for large size dataset	Suits only for emotional mining.
Hua et al., (2009).	Feature Selection using text score.	Accuracy is more	Based on the text score, mining can be done.
Shekar & Shoba, (2009)	Feature vector for feature representation	Dimensionality is reduced.	More computation.
Yoshida et al. (2007)	Combination of TF and DF is used for mining	Dimensionality is reduced.	More computation.
Ma et al. (2005) & Li et al. (2011b)	A similarity calculation and semantic relation for data pre-processing.	Make speedy in text mining with improved accuracy.	Takes more time.
Durga & Govardhan (2011)	Supervised and Un-supervised classifiers for mining	Efficiency of the mining process is improved.	Time and cost complexity is high.

III. CONCLUSION

The main objective of this research work is to study and understand various stages of text analytics and text mining to find out a research problem in text mining. To do that, it has been reviewed text analytics, information retrieval, pre-processing, feature extraction, feature selection and different mining methods. It is done to notice the merits and demerits of the existing approaches. Some of the research works specifically mentioned that pre-processing is important. Few of the research works discussed about combining two or more processes for improving the mining efficacy. Some of the algorithms discussed feature extraction and selection is more important for fast and accurate mining. But still the efficiency of the mining need to be improved and still need a common mining method for various kinds of text data.

It can be done by creating a novel framework which can behave as a text analytics tool for efficient mining using optimized feature selection method. Optimal features can be selected by heuristic and meta heuristic approaches, and hence this paper also conclude that metaheuristic or hybrid-metaheuristic algorithms can provide improved efficiency in text analytics and mining.

In future work, the author designs a framework comprising of two stages such as metaheuristic approach-based text analytics and hybrid metaheuristic-based text mining for improving the overall efficiency of the text data, whichever the data format is.

APPENDIX

It is optional. Appendixes, if needed, appear before the acknowledgment.

REFERENCES

1. Apache OpenNLP (2015), “http://opennlp.apache.org/”.
2. Abbasi, S. France, Z. Zhang, and H. Chen (2011), “Selecting Attributes for Sentiment Classification Using Feature Relation Networks,” IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 447–462.
3. B.Agarwal, V.K.Sharma, and N.Mittal, ”Sentiment Classification of Review Documents using Phrase Patterns,” International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1577-1580, . 2013.
4. B. Agarwal and N. Mittal, (2013a), “Sentiment Classification using Rough Set based Hybrid Feature

Selection," in Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, June, pp. 115–119.

5. B. Agarwal and N. Mittal, "Optimal Feature Selection for Sentiment Analysis," in Computational Linguistics and Intelligent Text Processing, 2013b, vol. 7817, pp. 13–24.
6. H. Arafat, R. M.Elawady, S. Barakat, and N. M.Elrashidy, "Different Feature Selection for Sentiment Classification," Int. J. Inf. Sci. Intell. Syst., vol. 1, no. 3, pp. 137–150, 2014.
7. Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In Proceedings of EMNLP 2005, pp.339-346.
8. Baumer et al. (2010), "America is like metamucil: fostering critical and creative thinking about metaphor in political blogs", International Conference on Human Factor in Computing Systems, ACM, 34–45.
9. Balakrishnan, R., Qui, X.Y. and Srinivasan, P. (2010) On the Predictive Ability of Narrative Disclosures in Annual Reports. European Journal of Operational Research, 202, 789-801.
10. Bragge, J., Thavikulwat, P. and Töyli, J. (2010) Profiling 40 Years of Research in Simulation & Gaming. Simulation & Gaming, 41, 869-897.
11. Bing Liu, Minqing Hu and Junsheng Cheng. Opinion Observer: Analyzing and comparing opinions on the web. In Proceedings of WWW 2005, pp.342-351.
12. Bichindaritz, I. and Akkineni, S. (2006) Concept Mining for Indexing Medical Literature. Engineering Applications of Artificial Intelligence, 19, 411-417.
13. Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of ACL 2005, pp.115-124.
14. Y. Y. Chen and K. V. Lee, "User-Centered Sentiment Analysis on Customer Product Review," World Appl. Sci. J. 12 (Special Issue Comput. Appl. Knowledgement Manag., pp. 32–38, 2011.
15. Comscore and Kelsey, (2017), http://www.shop.org/c/journal_articles/view_article_content? groupId=1 & articleId702 & version=1.0.
16. Conway, M. (2010) Mining a Corpus of Biographical Texts Using Keywords. Literary and Linguistic Computing, 25, 23-35.
17. Y.Dai et al. (2011), "MinEDec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis method", International Journal of Computer Information System and Industrial Management Applications, Vol. 3, pp.165–173.
18. Durga & Govardhan (2011), "Ontology based text categorization-telugu document", International Journal of Scientific and Engineering Research, Vol. 2, No. 9, pp. 1–4.
19. Esuli & Sibastiani (2006), "SentiWordNet: a publicly available lexical resource for opinion mining", International Conference on Language Resources and Evaluation, pp. 417–422.
20. Evans, B. M., Kairam, S. & Pirolli, P. 2010. Do your friends make you smarter: an analysis of social strategies in online information seeking. Information Processing and Management 46(6), 679–692.
21. Forman, G. & Kirshenbaum, E. 2008. Extremely fast text feature extraction for classification and indexing. In Proceedings of 17th ACM Conference on Information and Knowledge Management, 26 –30.
22. García-Crespo, Á., Gómez-Berbís, J.M., Colomo-Palacios, R. and García-Sánchez, F. (2011) Digital Libraries and Web 3.0. The Callimachus DL Approach. Computers in Human Behavior, 27, 1424-1430.
23. Glänzel, W. (2012) Bibliometric Methods for Detecting and Analysing Emerging Research Topics. El professional de la information, 21, 194-201.
24. B. He, C. Macdonald, J. He, and I. Ounis, An Effective Statistical Approach to Blog Post Opinion Retrieval, CIKM., 10 (2008), pp. 1063-1069.
25. Hjørland, B. (2012), "Is Classification Necessary after Google? ", Journal of Documentation, 68, 299-317.
26. Hua, J., Tembe, W. D., Dougherty, E. R. & Edward, R. D. 2009. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition 42(3), 409–424.
27. M. Hu and B. Liu, Mining and Summarizing Customer Reviews, Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD-2004), 8 (2004a), pp. 168–174.
28. M. Hu and B. Liu, Mining Opinion Features in Customer Reviews, Proceedings of the 19th National Conference on Artificial Intelligence., 7 (2004b), pp. 755-760.
29. J. R. Finkel, T. Grenager, and C. Manning (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005).
30. M. Karamibekr, A.A.Ghorbani, "Verb Oriented Sentiment Classification," Processed of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp. 327-331, 2012.
31. Kellogg, D.L. and Walczak, S. (2007) Nurse Scheduling: From Academia to Implementation or Not? Interfaces, 37, 355-369. http://dx.doi.org/10.1287/inte.1070.0291.
32. Kennedy, Alistair and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2):110–125.
33. A.khan,B.Baharudin, "Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs," Processed on National Postgraduate Conference (NPC), pp. 1 – 7, 2011.
34. S. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, Automatically Assessing Review Helpfulness, EMNLP., 7 (2006), pp. 423-430.
35. Lee, J.Y., Kim, H. and Kim, P.J. (2010) Domain Analysis with Text Mining: Analysis of Digital Library Research Trends Using Profiling Methods. Journal of Information Science, 36, 144-161.
36. LingPipe (2011): http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.

37. Li,J.,Li,Q.,Khan,S.U.&Ghani,N.2011b.Community-based cloud for emergency management. In Proceedings of the 6th IEEE International Conference on System of Systems Engineering (SoSE), 55 –60.
38. Li, J. & Khan, S. U. 2009a. MobiSN: semantics-based mobile ad hoc social network framework. In Proceedings of IEEE Global Communications Conference (Globecom), Zomaya, A. Y. & Sarbazi-Azad, H. (eds). John Wiley & Sons, Hoboken, NJ, USA, 2013, ISBN: 978-0-470-93688-7.
39. Li, J. & Khan, S.U. 2009b. On How to Construct a Social Network from a Mobile Ad Hoc Network. Technical report, NDSU-CS-TR-09-009, North Dakota State University.
40. Ch.L.Liu, W.H. Hsaio, C.H. Lee,and G.C.Lu, and E. Jou," Movie Rating and Review Summarization in Mobile Environment," IEEE Transactions on Systems, Man, and Cybernetics, Part C 42(3):pp.397-407, 2012.
41. Ling, H. S., Bali, R. & Salam, R. 2006. Emotion detection using keywords spotting and semantic network. In Proceedings of International Conference on Computing and Informatics IEEE (ICOCI),1 –5.
42. L.Liu, X.Nie,and H.Wang," Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis," Processed of the 5th Image International Congress on Signal Processing (CISP), pp. 1620 – 1624, 2012.
43. R.Liu, R.Xiong, and L.Song, "A Sentiment Classification Method for Chinese Document," Processed of the 5th International Conference on Computer Science and Education (ICCSE), pp. 918 – 922, 2010.
44. Liu, M. Hu, and J. Cheng, Opinion Observer: Analyzing and Comparing Opinions,WWW., 5 (2005), pp. 342-351.
45. Y.Luo, W.Huang, " Product Review Information Extraction Based on Adjective Opinion Words," Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp.1309 – 1313, 2011.
46. Ma, L. (2012), "Principles of Classification", ALCTS Webinar.
47. Ma, J., Xu, W., Sun, Y., Turban, E., Wang, S. and Liu, O. (2012) An Ontology-Based Text Analytics Method to Clus- ter Proposals for Research Project Selection. Transactions on Systems, Man, and Cybernetics—Part A, 42, 784-790.
48. Ma, C., Helmut, P. & Mitsuru, l. 2005. Emotion Estimation and Reasoning Based on Affective Textual Interaction, 3rd edition. Springer.
49. Michael Gamon, Anthony Aue, Simon Corston-Oliver and Eric Ringger. 2005. Pulse: Mining customer opinions from free text. In Proceedings of IDA 2005, pp.121-132.
50. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of ACM-KDD 2004, p.168-177.
51. Negi,P.S.,Rauthan,M.M.S.&Dhami, H.S.2010.Languagemodelforinformationretrieval. International Journal of Computer Applications 12(7), 13–17.
52. Neviarouskaya, H.Prendinger, and M.Ishizuka," SentiFul: A Lexicon for Sentiment Analysis," T. Affective Computing 2(1), pp.22-36, 2011.
53. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Found. Trends® Inf. Retr., vol. 2, no. 2, pp. 1–135, 2008.
54. Popescu and O. Etzioni, Extracting product features and opinions from reviews, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing., (2005), pp. 339-346.
55. Porter, A.L., Kongthon, A. and Lu, J.C. (2002) Research Profiling: Improving the Literature Review. Scientometrics, 53, 351-370.
56. Raghuram, S., Tuertscher, P. and Garud, R. (2010) Mapping the Field of Virtual Work: A Cocitation Analysis. Information Systems Research, 21, 983-999.
57. L.Ramachandran,E.F.Gehringer, "Automated Assessment of Review Quality Using Latent Semantic Analysis," ICALT, IEEE Computer Society, pp. 136-138, 2011.
58. Rizwana Irfan et al. (2015), "A survey on text mining in social networks", The Knowledge Engineering Review, Vol. 30:2, 157–170.
59. B. Seerat and F. Azam, "Opinion Mining: Issues and Challenges (A Survey)," Int. J. Comput. Appl., vol. 49, no. 9, 2012.
60. Seol, H., Lee, S. and Kim, C. (2011) Identifying New Business Areas Using Patent Information: A DEA and Text Mining Approach. Expert Systems with Applications, 38, 2933-2941.
61. Shekar, C. B. H. & Shoba, G. 2009. Classification of documents using Kohonens self-organizing map. International Journal of Computer Theory and Engineering (IACSIT) 1(5), 610–613.
62. Sorensen, L. 2009. User managed trust in social networking comparing Facebook, MySpace and LinkedIn. In Proceedings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic System Technology, (Wireless ITAE 09), 427–431.
63. R. Srivastava, M. P. S. Bhatia," Quantifying Modified Opinion Strength: A Fuzzy Inference System for Sentiment Analysis," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1512-1519, 2013.
64. Stanford Named Entity Recognizer (2015): http://www-nlp.stanford.edu/ software/CRF-NER.shtml.
65. Strapparava, C. & Ozbal, G. 2010. The color of emotion in text. In Proceedings of 2nd Workshop on Cognitive Aspects of the Lexicon, 28 –32.
66. Stringer, M.J., Sales-Pardo, M. and Amaral, L.A.N. (2008) Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. PLoS ONE, 3, 1-8.
67. Susman, G.I, and Evered, R.D. (1978) An Assessment of the Scientific Merits of Action Research. Administrative Science Quarterly, 23, 582-603.
68. Tseng, Y., Chang, C., Rundgren. S.C. and Rundgren, C. (2010) Mining Concept Maps from News Stories for Measuring Civic Scientific Literacy in Media. Computers & Education, 55, 165-177.
69. Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In

Proceedings of EMNLP 2004, pp.412-418.

70. P.D. Turney," Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424, July 2002.
71. H. Uguz, "A Two-Stage Feature Selection Method for Text categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm," Knowledge-Based Systems, vol. 24, no. 7. pp. 1024–1032, 2011.
72. Vinodhini G. and R. Chandrasekaran, "Effect of Feature Reduction in Sentiment Analysis of Online Reviews," Int. J. Adv. Comput. Eng. Technol., vol. 2, no. 6, pp. 2165–2172, 2013.
73. Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM-2005), pages 625–631.
74. Wollmer, M., Eyben, F., Keshet, J., Graves, A., Schuller, B. & Rigool, G. 2009. Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional networks. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 3949–3952.
75. Xerox Corporation (2015): http://www.xrce.xerox.com/ResearchDevelopment/Industry-Expertise/Finance.
76. Yang, Y., Akers, L., Klose, T. and Yang, C.B. (2008) Text Mining and Visualization Tools—Impressions of Emerging Capabilities. World Patent Information, 30, 280-293.
77. Yoshida, K., Tsuruoka, Y., Miyao, Y. & Tsujii, J. 2007. Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers. In Proceedings of 20th International Conference on Artificial Intelligence, 1783–1788.
78. Yuan, L. 2010. Improvement for the automatic part-of-speech tagging based on Hidden Markov Model. In Proceedings of 2nd International Conference on Signal Processing System IEEE (ICSPS), 744–747.
79. S. C. Yusta, "Different Metaheuristic Strategies to Solve the Feature Selection Problem," Pattern Recognition Letters, vol. 30, no. 5. pp. 525– 534, 2009.
80. Zhao, P., Han, J. & Sun, Y. 2009. P-Rank: a comprehensive structural similarity measure over information networks. In Proceedings of 18th ACM Conference on Information and Knowledge Management, 233–238.
81. J.Zhu, H.Wang, M.Zhu, B.K.Tsou, and M.Ma,," Aspect-Based Opinion Polling from Customer Reviews," T. Affective Computing2(1):pp. 3749, 2011.

AUTHORS PROFILE

C.P.ThamilSelvi is an Associate Professor of Computer Science and Engineering College of Engineering & Technology at (Anna University), TamilNadu, India. She received her B.E degree in Computer science and Engineering from R.V.S College of Engineering Technology-Dindigul in the year 1993 and M.E degree in Computer Science and Engineering during the year 2011 from Anna University Tiruchirappalli, TamilNadu, India. Additionally she has finished MBA (Production & System),M.S(Information Technology),PGDMM (Materials Management).She has 15 years of teaching experience and 11 years of industry experience. She has published nearly 15 papers in International & National Conferences. Her main areas of research interest are Artificial Intelligence, Net Working, Big Data, Network and Information Security, Machine Learning, Deep learning. She is a Life member of Indian Society for Technical Education (ISTE) and Computer Society of India (CSI).

. Technology at PSNA College of Engineering & Technology (Anna University), TamilNadu, India. She received her PhD in Information and Communication Engineering from Anna University, Chennai, India, in 2014. She received the B.E. degree in Computer Science and Engineering from Madurai Kamaraj University, TamilNadu, India, in 2001, and the M.E. degree in Computer Science and Engineering from Anna University, Chennai, India, in 2004. Her main areas of research interest are Wireless Networks, Network Security, Soft Computing and Data Mining. She is a Life Member of the Indian Society for Technical Education (ISTE).