**Chapter- Methods and Techniques of Data Mining**

1. **Introduction**

Over the past several decades, the immense collection and use of information have been observed. Mismatch and overload are major problems with this information. The incompatibility is ignored by some useful and interesting data and duplications, but the collected data is not the one that user needs.

In this age of learning, most conceptual information, which includes worldwide web as the strongest repository of the world, is collected and stored in different databases worldwide.

It is not simple to discover information from this accessible repository. For the pre-processing of textual documents for data mining, the automated feature selection method is therefore required. The feature selection method focuses on defining relevant data, helps to understand and visualize data, often decreases learning and storage time for huge amounts of data and improves reliability for subsequent data mining tasks.

In this modern digitalized world, there is rapid growth in the digital data, which is in quasi-structured, or unorganized data format, unstructured data is material that does not adopt traditional data models in many different ways and is therefore generally not appropriate for a standard relational database. Today there is an increasing trend in using this data and attracted towards processing the unorganized information into organized information.

Text mining is an emergent area, which gained attention towards processing this kind of unorganized information into relevant information. It is defined as the method of processing and observing large volumes of software-assisted knowledge that can reorganize the data’s thoughts, models, patterns and keywords. It is also called text analytics it is in the sense of, although some people distinguish between two concepts, is a software to allow the use of text mining techniques to be processed over data sets.

In this contemporary digitized world, there is a rapid boom in digital records, i.e. in semi-established or unstructured record format, which has multiplied the trend of using these records and attracted the processing of unstructured data into based information. Today Text mining is an evolving area that has been giving more focus to turning this kind of unstructured data into relevant information.

Exploitation and text processing allows companies to seek valuable strategic insights into corporate records, client correspondence, customer care reports, paraphrased survey messages, web posts, medical records and other sources dependent on the script.

Content miner technologies are also increasingly being incorporated into AI Chat boxes and into digital agents which are being implemented by companies to provide consumers with automated responses to their advertising sales and service operations.

Content Mining is portrayed as the arrangement of running from unstructured content documents over a concealed, gainful and fascinating example. Content mining usually involves arranging the literary data content, inferring styles within the organized details, and lastly assessing and understanding the yield. Content mining produces the knowledge which is highly relevant and disposes of insights which are not essential. High nice throughout textual content mining, on the whole, alludes to a few mixes of meaning, curiosity and intriguing quality.

Recovery of data from unstructured documents can be a very difficult undertaking as it is made up of the vast details that need to be processed. There are several strategies for executing the mining operation of textual content which we described in this chapter.

* 1. **Functions of Data Mining**

Information farming is close to machine learning in nature but focuses more on data rather than organized knowledge. The main important focus in this process is to arrange and design the data so that qualitative and quantitative analysis can be carried out. In general, this includes the use of the natural language programming software related to the study and interpretation of data sets [1] using computational language principles.

Text Mining is described as “Scientific literature extraction of data”. It has three important components including Information recovery, Information processing and Information integration [2]. This means that text mining is compatible with extracting information and extracting knowledge from documents. It supports to extract the information from message-passing knowledge and consider a special yet previous unknown sequence.

The steps involved in the entire text mining processing can be defined as follows

**Text Documents Database**

**Text Pre-processing**

1. **Tokenization**
2. **Stop words Removal**
3. **Stemming**

**Document Conversion**

**(Generation of Characteristics)**

**Selection of Features**

**(Selection of Attributes)**

**Text Mining Techniques**

**Evaluation**

appl

**Figure 1 : Phase of Text Mining**

1. Text Document Database: It is the compilation of a set of large text documents in the form of a corpus that is mostly required for further processing to furnish with relevant facts to the user.
2. Text Preprocessing: These steps are necessary to convert human-readable text into machine-readable text. Essentially preprocessing involves standardizing text documents by eliminating noisy information, Upper case letters to lower case letter transition, white space reduction, abbreviation extension etc.
3. Document Conversion: This step involves the process of translating the entire document that can be used for most successful investigations into a word bag or vector space document notation.
4. Feature Selection: is the method where you pick the features that most relate to your prediction function or the output you are interested in, automatically or manually.
5. Text Mining Techniques: You may apply different text mining techniques to perform a specific task based on the problem statement. The task may include classification, clustering, summary formation, topic marking, retrieval of information etc.
6. Evaluation: This process includes assessing and analyzing results in terms of reliability, memory estimation, consistency etc.

Throughout the earlier days, Natural language processing interfaces mostly are focused on mathematical models or guidelines that provide a reference for what to search for in datasets [3]. Furthermore, intense models of training that work in a less controlled environment appeared in the mid- 2010 as an alternative tactic to document review and other advanced applications lacking huge sets of data.As a result, the processing methods are supported, the basic similarities and correlation in text data are more clearly identified, even if data experts don’t understand what they will probably find at the outset of the project [4]. For example, without any guidance from an analyst, an unattended model can handle information in a community of subjects from text documents or e-mails.

**1.2. Knowledge Discovery of Text Mining**

Knowledge Discovery and Data Mining are better known by the oversimplified word Data Mining (DM), which develops methods for extracting information from a data set. KDD is capable of executing the process of data analysis, data cleaning, transitioning to Data mining Technique, checking and eventually transmitting the results of Data Mining Technique to the user.

Knowledge discovery database is the process which is involved with certain steps which can help the researchers to retrieve the useful patterns from the large corpus. The input to carry out this process is data and the desired output is useful information which is required for the users. Consulting with the domain experts and technical experts may be necessary to ensure that the result obtained using this process is reliable and consistent [5].

**Data**

**Knowledge**

**Selection**

**Pre-processing**

**Transformation**

**Data Mining**

**Interpretation**

**Target Data**

**Pre-processed Data**

**Transformed Data**

**Patterns**

**Figure 2: Process of Knowledge Discovery on Text**

The process of Knowledge discovery on text is discussed below

1. **Data Selection:**

This process is used to generate the desired set of data by selecting the data or subset of the large dataset which is essential for discovering the patterns. The input here is the huge data set and output is aim data that the customers require. For example among the numerous medical information sources available on the worldwide platform, only certain information may be obtained which is most client-friendly.

1. **Pre-Processing:**

This process involves scrubbing up data and get rid of noisy data from the huge corpus. It also includes data collected from the listed data fields, including appropriate methods for solving missing data and redundant accounting of data.

1. **Transformation:**

Based on the content shipping mission, the generated evidence must be translated into a predefined format. This step is to underway by selecting appropriate features to represent the data. However, the selection of features can be used for dimensionality reduction at this point. Finally, it provides relevant features that can be recognized as a specific data set.

1. **Data Mining:**

Analytics is a scientific process that is conducted to identify objects over transformed pieces of evidence. The trends found may be pairs of components from the same data set, a set of structured components that appear together or a set of maximum features based on the user's requirement.

1. **Interpretation:**

The found artefacts will be checked if they are real, innovative and potentially useful to consumers to satisfy their service needs. This service decides that weather the patterns discovered is fascinating enough to influence the conceptual intelligence.

## Techniques of Text Mining

#### Anomaly Detection

#### In this present global of massive records, it is very vital to shield the information from the intruders and attackers. As the statistics are always transferred and shared it is regular that there may be probabilities of it being exposed to attack. Techniques should be used to make the statistics much less prone. One such technique is Anomaly detection, which makes use of the ideas of records mining to locate the surprising conduct within the records. It identifies the surprising activities inside the information and compares it with regular activities. Mainly anomaly detection is the process of locating the styles in the facts, which aren't normal and usual. The records handled by way of these strategies are typically recorded records, which may be univariate and multivariate. These styles are referred to as anomalies or outliners. Detected anomalies might also or may not be harmful. They must be nicely studied and analyzed. The method used is first the normal profiles or styles have to be constructed which might be taken as preferred to locate the anomalies or outliners in the records. Immediately trigger is dispatched if any peculiar behaviour is observed. Relationship with the present everyday statistics has to be made. Analysis of records can include diverse strategies like class, clustering, and another system primarily based learning strategies to know the essential and substantial information [1].

####

#### Figure 3: Anomaly detection

* + 1. **Regression**

This training data is generally depicted as a bunch of matrices for inputs and outputs. the input matrix is similar to our trait matrix A, and the output matrix B comprises of point-outs indicating the category subscription of the corresponding document in matrix A. Matrix B has rows equal to column A (namely m) and c where c seems to be the estimated number of specified groupings. Regression methods are language-dependent and they are used with single and multiple category issues [2].

#### Classification

It is the method a model is built that can classify the objects to guess the future class. It can be done in two steps. In the first, the model is constructed based on the training set that defines the characteristics of classes and concepts the data in the training set. In the next step, the built model is used to predict potential groups. In other words, the grouping of content can be termed as the division of a series of input documents into two or more classes in which each document belongs to one class or several classes [3].

**Figure 4: Classification**

* + 1. **Clustering**

Clustering of the document is identified in different areas of text mining. This has become popular in recent days as its wide application is identified in the areas of web mining, search engine, information retrieval and many more. Clustering is a mechanism used to bundle uniform documents into groups. The process of organizing the documents into clusters or groups, documents within the clusters are referred to as Document clustering which has the highest similarity compared to documents within other clusters. Representation of documents in the form of clusters necessarily loses some finer details but achieves the simplification. It results from clustering articles on the run rather than using predefined topics by categorization. The primary purpose of this clustering is to maximize intra-cluster resemblance and diminish inter-cluster resemblance. The benefit of clustering is that records will appear in several subtopics so that a valuation report is not excluded from the search results. The biggest challenge of clustering is to recognize the appropriate groups effectively. Clustering technologies can help organize data management frameworks which may include millions of documents [4].



#### Figure 5: Clustering

* + 1. **Association Rule Mining**

Association Mining is an essential element of the data analysis. The connections which data mining has revealed are conveyed as association principles. These Principles of the association are an essential subset of pattern-seeking approaches in records. The primary objective is to find an exciting connection or variance among a wide variety of data items. Since there is a massive collection of data which increases continuously, most of the companies are involved in association mining standards. In addition to the rule, the association mining also calculates some statistics about the rule. Association mining has many application domains. Traditionally, the technique has been used to perform market basket analysis. Association rule mining uses include Wal-Mart analytics, psychological disorder, science, domain navigation assessment, security systems, schooling, economy, corporation etc [5-6].

#### Feature Selection and Extraction

The focal point of Feature Selection approach is distinguishing important information, assists with comprehension and imagine the information, it likewise diminishes the preparation and handling time of gigantic measures of information just as increment the exactness for the resulting information mining errands. Highlight determination is a long-existing, novel technique that means to evacuate that immaterial and loud data by centring and form just the significant and educational information for use in content mining. Highlight strategy of choice makes a new entrance for content mining analysis. There are two enquiries in highlight determination approach first enquiry is' what are the highlights for AI that can effectively talk to the content? Also, the following one is: 'What is the most suitable way to prune a large collection of highlights down to a fair arrangement of the highlights most segregating? 'We would conclude that for the primary inquiry it depends on the preparedness, language and organization that deals with and above all on the specific problem that you are concerned with. Response for the subsequent inquiry: we can seek various methodologies to prune capabilities like classifiers that order and produce the relevant data, point skilled Mutual Data to measure the relationship between a class ascribed, figure out the highlights that are firmly important and superfluous and chi-square strategy can be used to determine the contrast between class ascribe and class. More thoroughly, there are two methodologies for picking the 'best' highlights, and they are close draws for channels and wrappers. Determination of the subset is handled in channel strategy without considering any calculation that is usually performed before planning. In the wrapper method, evaluation of a list of capabilities is finished by taking into account the algorithm[7].

####

#### Figure 6. Feature Selection

* + 1. **Information Extraction**

Analysis of the unstructured information is very much essential. This requires information extraction. Much information is available in natural language rather than structural databases. this retrieval of knowledge recognizes the words or phrases, the connections within the text through the evaluation process. It transforms the corpus of textual documents into a structural form of databases. Once the information is extracted, it can be processed throughout the mining repositories, Concise overview consulted informal language etc. The first step in processing is extracting data which is linguistic pre-processing. There are other linguistic strategies like tokenization, component labelling of expressions, lemmatization, and so on. To feed the machine to extract information. IE is very important when dealing with a large volume of data [8].



#### Figure 7: Single Pipeline architecture of Information Extraction

* + 1. **Categorization**

Generally, the data in natural language form needs processing. text Taxonomy assigns these computational transcripts based on the content to the predefined categories. Taxonomy assists in defining a document's central themes by positioning the document in a predefined set of concepts. These sets of predefined categories are called a controlled vocabulary. Categorization essentially relies on the terminology under which the subjects are user-defined and relationships are identified via the analysis for concrete terms, narrow definitions, synonyms and acronyms. Categorization computes the phrases that occur and interprets from the count of major topics enclosed by the report. Categorization helps in grading the documents such that which documents have the most data on specific topics [9,10].

* + 1. **Summarization**

Summarization is a mode which is helpful in the very large documents, providing the summary of the data makes the user to easily understand. It benefits the viewer by analyzing tremendously and figuring out whether the text satisfies the user's expectations and it is worth to read further. The primary concept of the description is to shorten the length, but retaining its main points and sense. But the task is to examine the syntax and the context [11].



#### Figure 8: Summarization Process on Text Data

#### Visualization

Optical Text mining or facts Conception makes the visual hierarchy or map of the large textual sources. Visual mining is the mechanism of revealing tuples of tacit yet useful information using visual techniques. Popularly used tool in this technique is DocMiner which allows the user to analyze the content visually. User interaction with the document by zooming, scaling and creating sub-maps can also be done. visualization is often more effective as consumers need a wider range of documentation to discuss the relevant topics [12].

#### Topic Tracking

The Topic following framework helps the client in finding the data by the catchphrases and illuminates them regarding the related data. In light of the client profiles, the client sees it distinguishes the archives that are important to the client. There are several provinces wherein the following theme can be introduced to the business sector. it might very well be used to alarm organizations whenever there is a competitor in the news. This permits them to remain fully cognizant of critical components and market crashes. Organizations should also check up on news and articles about their organization. It can also be used by specialists and others in the pharmaceutical field searching for new disease drugs and having to keep up to date developments. individuals in the practice field could also use the following patterns to be confident that they have the current sources for investigation in their general vicinity of intrigue [13].

#### Concept Linkage

Concept linkage is another technique of text mining where it attempts to relate the documents by identifying the common shared idea or concept between the documents. This idea of the technique is very valuable in text mining, where it helps the user to browse the information rather than searching [14].

#### Question Answering

Question noting is a region of handling normal language inquiries or question replying. It attempts to discover how to respond to the inquiry in the most ideal manner. It utilizes the idea of data extraction to separate the substances and attempts to arrange the inquiries for relegating them to the suitable structure (Who, What, When, Where, How). The Question Answering system receives a customer's query in a distinctive vocabulary (NL). This investigation is transmitted to a POS tagger that transcribes the query and identifies POS from each term correlated with the query. The query generators then use the named inquiry to generate various kinds of queries that can be passed on to a web critter. Such investigations are then sought out in equal measure by an internet searcher. The web critter offers the documents that we are searching for possibly will have the right answers. These records are checked for this by the appropriate response extractor. Piece Extractor extracts snippets that include phrases/terms of inquiry from the records. These bits are passed to the ranker which sorts them as indicated by the positioning calculation. Question Answering has different applications, causes the client to discover answers to basic inquiries and some Frequently Posed to Questions (FAQ) .

* + 1. **Stemming**

The stemming process is applied before the data can be recovered to decrease the size of the archive which can build the adequacy of data recovery framework. Words showing up in the reports or questions frequently contain the morphological variations.

These morphological variations have the comparative semantic understandings and can be considered as equal. Stemming is where these various types of words are diminished to a typical structure. It lessens bent words to their stem, base or root structure. There are various stemming Algorithms or stemmers, which endeavour to decrease a word to its stem or root structure. Consequently, the key terms of an inquiry or report are spoken to by stems as opposed to by the first words. This not just implies that various variations of a term can be conflated to a solitary agent structure – it additionally diminishes the lexicon size. It brings about the littler extra room and handling time. [15][16] There are stemming calculations which vary in execution and precision. The various traditional stemming algorithms are:

* Husk\Pair
* Sorter
* Krovetz
* Dawson

## Methods of Text Mining

#### Term Based Method

Term-based model is a methodology wherein the term is used to validate the text. This word has the terminological sense in the text. It enhances the successful statistical efficiency and advanced measurement hypotheses for the term. the challenge facing this technique is it continues to suffer from the polysemy and synonymy problems. Polysemy is a term with many significations and synonymy is numerous terms which has the same context [17].

#### Phrase-Based Method

Expression approach is a system in which the sentence-based text is analyzed. There is more terminology in the expression, less vague and less restrictive than in individual words. Challenges faced in this approach are when sentences have lower statistic properties of terms, the frequency range of incidence, even when there is a huge number of redundant and distracting phrases [18].

#### Concept-Based Method

In this case, the words and definitions are extracted from the paper. The majority of text mining strategies are the focused onset of words data review. This data review takes into consideration the meaning of the term or definition without the text. In the document, two concepts may have the same intensity, but only one term correctly contributes to the sense of the other [19]. The concept-based model consists of three factors, first interpreting the linguistic structure of the sentences, A second factor explains semantic systems, and the last component derives the principles focused on the first two modules. Such concepts can be used for the development of feature vectors using the standard vector space model. The idea-oriented model is typically based on natural language processing. This can be used effectively to differentiate between meaningful words that define the meaning of the sentence and non-significant words.

#### Pattern Taxonomy Method

The documents are evaluated based on the similarities in the pattern taxonomy system. Patterns can only be organized using a connection. Many techniques such as association rule mining, regular itemset mining, sequential pattern mining and closed pattern mining will uncover patterns [19].

There are two crucial steps in this approach i.e. extracting meaningful patterns from text documents and how to make use of such patterns discovered to improve efficacy. The approach to produce the pattern taxonomies are implemented by breaking the text documents into paragraphs and treating each paragraph as an individual transaction. The pattern-based approach uses both methods of distributing patterns and changing patterns.

We have to decipher the found examples. Highlight space which comprises of a set of individual terms created by the utilization of customary record ordering systems made example scientific categorizations and highlight space can be utilized to speak to the idea of reports by applying an information mining-based strategy like SPM. By conveying designs into the component space, PDM utilization of successive examples to keep the helpful semantic data, yet also improve the framework productivity by forestalling the tedious example revelation draws near.

The PTM has been essentially improved after the reception of an example conveying technique, which utilizes the system of mapping found examples into a speculation space for taking care of the low-recurrence issue to the particular long examples. There is at some point negative archives contain some valuable data to distinguish vague examples in the idea. A negative record and is a report that the framework at some point recognized as a positive archive. The guilty party of nd is a conveyed design which acquires at any rate one segment that shows up in nd. At the point when a negative archive is recognized, DPE begins to discover guilty parties and actualizes design advancing at "Hypothesis Space" state. Interestingly, IPE executes a similar activity at the "Pattern" state. Besides, the "Space Hypothesis" and "Pattern" structures are special, an elective interpretation and IPE calculation are needed [20].

####

#### Figure 9: Pattern Evolving and Deployment Process

## Use of found examples is troublesome and inadequate, as long examples with low recurrence may need support. However, all regular examples are not helpful and may prompt misinterpretations promotion inadequate execution. For the successful execution, powerful example disclosure strategy must be picked to defeat the low recurrence and confusion issues.

## Applications of Text Mining

##  Content mining is an interdisciplinary field with wide applications. Content mining is the client-driven procedure with influence examination systems and figuring capacity to get to the important data from the unstructured information. It disposes of the need to physically peruse the unstructured information. Today take a shot at content mining is done on different spaces [21].

##  A portion of the uses of content mining:

## Bioinformatics: Bioinformatics writing is the objective for content mining in this field. The main objective is to permit analysts with successful and valuable information disclosure.

## Business Intelligence: Text mining is utilized by numerous associations in basic leadership. It makes experts to straightforwardly land at the arrangements by removing and giving them just significant information .

## National Security: Text mining is utilized as a reconnaissance instrument. It gives the apparatuses, strategies, systems, ideas to Figure it against the fear-based oppression, wrongdoing and aides in recognizing future specialized and operational difficulties.

## Knowledge Management: As the data develops rapidly, it is continually testing to deal with this gigantic measure of information. Information the board programming is given dependent on content mining to give a solid arrangement.

## Customer Care Service: It is one of the conventional use of content mining. Numerous wellsprings of data of client are utilized, for example, audits, studies and give an opportune acceptable reaction.

## Fraud Detection through cases examination: Combines the impacts of content investigation and organized information to counteract the fakes.

## Content improvement Helps in successfully dealing with the huge volumes of data. Content mining procedures advance the substance, arrange and outline the accessible substance that makes it reasonable for an assortment of purposes.

## Spam sifting: Spam is a big problem for network access providers today and it is a section point where infections can be modified for Text mining systems to boost the adequacy of factually based separation strategies.

## Examination of social media knowledge: Today online existence is one of the most active sources of unstructured information Text investigation will tackle both through dissecting massive amounts of unstructured information, through extracting perceptions, opinions and assumptions and their relationships with brands and products.

## Challenges of Data Mining

## Significant difficulties happen from basic language itself in content mining. The mining in unstructured material is some of the time checked because it is conflicting and huge.

## In the content records single word may have numerous implications just as various words have similar importance in explanations settling such sort of vagueness in the archive is one of the difficult issues of content mining (Ambiguity Problem).

## • There is no uniform access to overall sources. for example, Web, email, databases, and so on, Preprocessing of information is required i.e., important to reformat the content which is costly and tedious.

## • Complex and Subtle relationship exists between the ideas in content mining.

## • Learning methodologies for getting ready substance need clarified planning. It is fundamental to amass the substance refining counts that methodology with multilingual substance records and produce language self-ruling transitional structure.

## • Another enormous test is how to make semantic investigation progressively effective and adaptable for each huge corpus content report.

## • Integration of area information in the content mining instrument which assumes a significant job during the time spent finding the information on archives is additionally one more challenge of content mining.

## • Another issue with content mining is cleaning the extricated information of online writings since the reference address of the picture interface is hard to expel.

## Conclusion

Text Mining is developing, developing innovation with its very own innate properties which can be utilized to take care of the assortment of issues today. It doesn't have an especially certain system. Content Mining incorporates a wide assortment of procedures and strategies dependent on the issue considered and some of them talked about in this paper. These systems and strategies have experienced with different research studies. There are loads of difficulties and openings found here. Content mining procedures have different applications and it is a powerful method for information revelation without the specific space information.

## References:

*[1]. S. Agrawal and J. Agrawal, “Survey on Anomaly Detection using Data Mining Techniques”, 19th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Procedia Computer Science, vol. 60, (2015), pp.708-713.*

*[2]. G. R. Banu and V. K. Chitra, “A Survey of Text Mining Concepts”, International Journal of Innovations in engineering and technology, ISSN-2319-1058, vol. 5, no. 2,(2015).*

*[3]. Q. Zhao and S. S. Bhowmick, “Association Rule Mining: A Survey, Technical Report”, CAIS, Nanyang Technological University, Singapore, No. 2003116,(2003).*

*[4]. R. Patel and G. Sharma, “A survey on text mining techniques”, International Journal of Engineering and Computer Science ISSN:2319-7242, vol. 3, no. 5, (2014), pp.5621-5625.*

*[5]. M. Tiwari, R. Singh and S. K. Singh, “Association–Rule Mining Techniques: A general survey and empirical comparative evaluation”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, no. 10,(2012).*

*[6} I. Tudor, “Association Rule Mining as a Data Mining Technique”, Universitatea Petrol-Gaze din Ploieşti, vol. LX, no. 1,(2008).*

*[7]. K. Nirmala and M. Pushpa, “Feature-based text Classification using Application Term Set”, International Journal of Computer Applications (0975 – 8887), vol. 52, no. 10,(2012).*

*[8]. H. Karanikas, C. Tjortjis and B. Theodoulidis, “An Approach to Text Mining using Information Extraction”.*

*[9]. N. S. Giridhar, K. V. Prema and N. V. S. Reddy, “A Prospective Study of Stemming Algorithms for Web Text Mining”, Ganpat University Journal of Engineering & Technology, vol. 1, no. 1,(2011).*

*[10]. S. V. Gaikwad, A. Chaugule and P. Patil, “Text Mining Methods and Techniques”, International Journal of Computer Applications (0975 – 8887), vol. 85, no. 17,(2014).*

*[11]. R. Bhaisare and V. Nayyar, “Analysis of Effective Pattern Discovery with Text Mining in Business Based Application”, International Journal of Research (IJR) ISSN 2348-6848, vol. 1, no. 11,(2014).*

*[12]. S. Jusoh and H. M. Alfawareh, “Techniques, Applications and Challenging Issue in Text Mining”, IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814* [*www.IJCSI.org,*](http://www.IJCSI.org/) *vol. 9, no. 2,(2012).*

*[13]. V. Gupta and G. S. Lehal, “A Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1,(2009).*

 *[14]. Akhgar, Babak, et al. Application of big data for national security: a practitioner’s guide to emerging technologies. Butterworth-Heinemann, 2015.*

*[15]. Shehu, Harisu Abdullahi, et al. "Deep sentiment analysis: a case study on stemmed Turkish twitter data." IEEE Access 9 (2021): 56836-56854.*

*[16]. Berko, Andrii, et al. "The text classification based on Big Data analysis for keyword definition using stemming." 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). Vol. 1. IEEE, 2021.*

*[17]. Guo, Weilin, et al. "Machine-Learning based methods in short-term load forecasting." The Electricity Journal 34.1 (2021): 106884.*

*[18]. Minaee, Shervin, et al. "Deep learning--based text classification: a comprehensive review." ACM computing surveys (CSUR) 54.3 (2021): 1-40.*

*[19]. Hussain, Musarrat, et al. "Text classification in clinical practice guidelines using machine-learning assisted pattern-based approach." Applied Sciences 11.8 (2021): 3296.*

*[20]. Flores, Christopher A., Rosa L. Figueroa, and Jorge E. Pezoa. "Active learning for biomedical text classification based on automatically generated regular expressions." IEEE Access 9 (2021): 38767-38777.*

*[21]. Kumar, Sunil, Arpan Kumar Kar, and P. Vigneswara Ilavarasan. "Applications of text mining in services management: A systematic literature review." International Journal of Information Management Data Insights 1.1 (2021): 100008.*