Introduction to NLP – Text Data Mining

Vaibhav Vilas Apraj

Abstract: -

The world is drowning in data and to handle this data driven applications artificial intelligence techniques are used. The data available is present is unstructured format and the humans cannot identity the important information from it. In artificial intelligence the processing of natural language present in the data has always been one of the biggest research concerns because of the two main reasons the key position language performs in human intelligence and due to the fact of the abundance of doable applications. NLP provides a solution for computers to brilliantly and correctly perceive, examine and obtain meaning from human language. The NLP uses text data analysis technique to represent the data in a structured format using the ideology of data mining. Text data mining is a technique of NLP which is widely used to organize the data and is really helpful in recent applications. This paper aims to grant an overview of the strategies of NLP and text data mining used in the systems and concludes with future possible areas of research interest.

1. Introduction: -

Natural Language Processing is a part of artificial intelligence techniques that deals with communicating with a high intelligence machine in the form of natural language. NLP is an associate degree knowledge domain field involved with the interactions between computers and human natural languages (e.g.: English) — speech or text [1]. NLP is divided into two major component i.e. natural language understanding and natural language generation [17]. Understanding refers to mapping of the given input into the natural language whereas generation is the process of producing meaningful sentences and phrases. The overall goal is here to essentially turn the text into meaningful data using data analysis via the natural language processing.

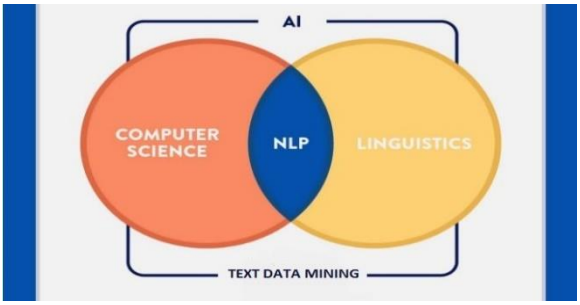

Fig. 1 - Types of fields in NLP [1].

NLP is divided into two fields: Linguistics and Computer Science as shown in Fig. 1 [1]. Linguistics is about language, it is about its formation, meaning, syntax, distinct phrases (noun or verb) and whatnot [1]. Computer Science deals with applying linguistic knowledge, by transforming it into computer programs with the help of technology such as Artificial Intelligence (Text data mining) [1]. Text data mining is the technique of extracting meaningful information from natural language text. This proposal gives an overview of Natural language processing as well as the

technique of text data mining used in NLP. The rest of the paper is as follows section 2 describes the challenging issues include in NLP and text data mining, section 3 explains the core methodology of text mining, section 4 explains the important techniques used in implementation of text data mining, section 5 explores the applications of text mining and NLP followed by the section 6 which gives the overall conclusion.

2. Challenging issues: -

The natural language processing has an important challenging issue which is the complexity and the complications which are arising due to use of text mining [13]. The question of uncertainty is still causing a problem of the natural language as it is not exempted from the ambiguity [13]. One phrase may also have different meanings and more than one words can have similar meaning [13]. If the understanding can be interpreted in two different meanings it is called ambiguity. When the information is extracted from the natural language the noise is present in it because of this ambiguity. It is not possible to completely remove the ambiguity because usability and flexibility out of ambiguity [13]. The words or sentences can be explained in various ways so that different meanings can be obtained [13]. Although various research has been done to this ambiguity issue but the issue is still not solved as the proposed works are not up to the mark and it is domain specific [13]. The semantic understandings of various identified words are not clear, so it is possessing as a challenge to answer them [13].

Merits of Text Mining: -

1. Information extraction is to be used to identify the names of various objects and the familiarity between from the raw data of textual document [13].
2. The most important and demanding problem controlling and organizing the unstructured data which is extracted from the raw data is resolved using text data mining [13].

Demerits of Text Mining: -

1. The data which is actually required is not at all written anywhere and is difficult to be found [13].
2. To extract a text for a particular data, knowledge or information the program which are required to understand and analyse the unformatted data is not easily available [13].

3. Methodology using Text data mining: -

Text data mining may be viewed as having two major phases**: text refining** that modifies text which is free formed in nature into a selected intermediate form, and **knowledge distillation** that gathers information, patterns and knowledges from the intermediate form [3]. Intermediate form (IF) are often semi-structured like the abstract graph illustration of the concepts or structured like the relative information illustration of data [3]. Intermediate form may be document-based whereby every entity in the form constitute a document or based on a concept whereby every entity constitutes an object or interests' concepts in an exceedingly specific area of a domain [3]. The techniques of clustering and categorization is used to mine data from this type of document based intermediate form. We get relationships and patterns within concepts or object by deriving a concept based IF using a data mining [3]. The techniques of associative discovery and predictive modelling of data mining falls into this category. By extracting the relevant information, we can transform a document based IF which is domain-independent into a concept based IF which is domain-dependant.

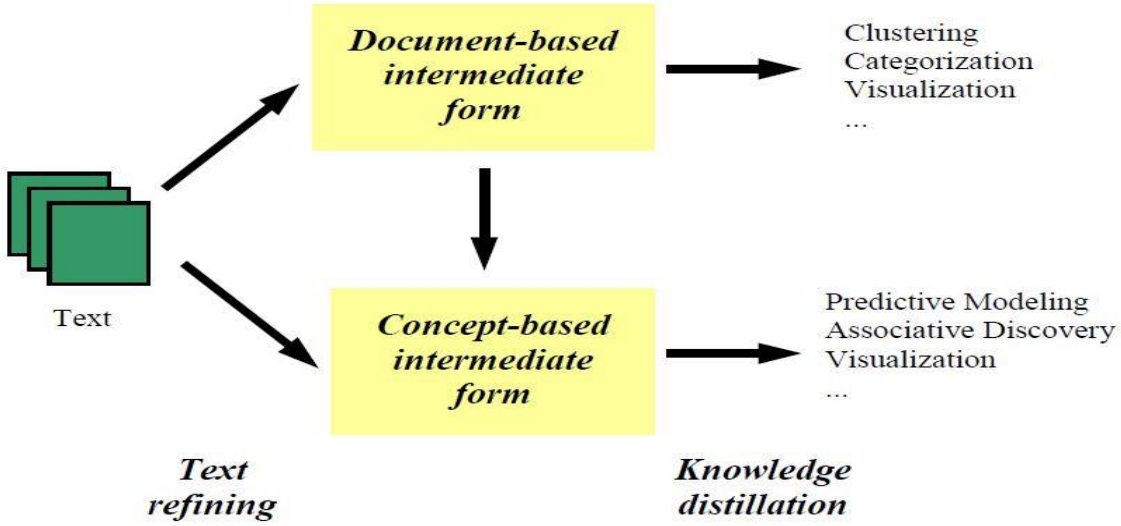

Fig. 2 - A text mining framework [3].

The text documents which are unstructured in nature are transformed into a structured intermediate form (IF) by using text refining. IF may be concept based or document based. The knowledge of data or patterns across the documents are deduced from a document based IF by Knowledge distillation. By extricating information related to specific objects which is useful to a particular domain a document based IF can be forecasted onto a concept based IF [3]. For example, consider if we have large number of news articles, the first task of text mining is to convert every type of document into a document based IF. After that the knowledge distillation is performed onto the document based IF to differentiate the articles based on the contents of the data and the for the purpose of easy understanding. If it is specifically required by a domain, the document based IF can b casted onto the concept based IF depending upon the requirements. For e.g. a student database can be created by extracting the data specific to "student". Then the student related knowledge can be extracted from the student database by using knowledge distillation [3].

4. Techniques in Text Data mining: -

To educate the computers the way to analyse, perceive and reproduce textual data technologies are implemented by natural language processing [13]. The techniques like information extraction, information retrieval, categorization, clustering, summarization is used in the text data mining process [13]. In the below sections we will talk about every one of those techniques and the key role they play in text data mining [13]. The kinds of situations wherever every technology could also be helpful so as to help the users are mentioned.

a. Information Extraction: -

The Fig. 3 shows the in-depth architecture of text extraction in text data mining framework. The important and relevant information or phrases and patters within the raw text is extracted by using Information Extraction software [2]. It uses pattern matching which is techniques based on regular expressions to look for a specific sequence of data in the raw text [2]. Entity recognition (NER) is the most common and popular form of IE in the market. NER tries to pinpoint and distinguish the elements which are atomic in nature into well-defined categories [2]. It selects key features such as location, names of entity, quantities, prices, etc. [2]. There are many tools used for this task such as Stanford Named Entity Recognizer [9] [10], Apache OpenNLP [11], LingPipe [12]. When we are dealing with large amount of textual data this technique can be implemented [13].

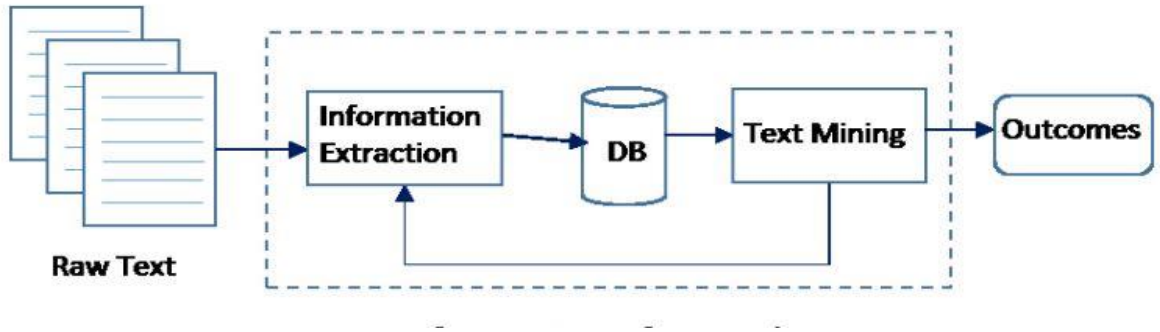

Fig. 3 – Overview of information extraction using text mining [2].

b. Information Retrieval: -

Information retrieval refers to the method of extracting the key features and patterns within the text which support a selected set of phrases and words [4]. Information retrieval system which is a part of text mining technique is implemented using various algorithms to supervise and detect the behaviours of users and to locate the important data [4]. The most famous information retrieval systems in the world are Google and Yahoo search engines [4]. The data which is retrieve in google should work in a real time environment so that the output is presented to the user in a substantial form. In information retrieval the data can summarized into particular subcategories. The data which is relevant to the system and can be used in the actual knowledge of the language to interpret the output is actually retrieved first. The sequence algorithms can be applied in the information retrieval process to examine the relevant data in a sequence which is required by the domain specific area

c. Categorization: -

Categorization mechanically allot one or many categories to the free textual document [13]. The textual document is placed into a predefined topic set by recognizing the important theme of the textual document [2]. Categorization solely counts words that show up and, from that, identifies the most important topics that is covered by the textual document [2]. Categorization regularly relies on the identities between narrow terms, related phrases, broad terms, synonyms, etc. [2]. It also has a method to rank the textual document based the most important contents it has on a specific topic [14]. Categorization is one of the supervised learning methods as it identifies the textual document based on the output and input of the examples [13]. The text documents are assigning predetermined classes based on their category [13]. It consists of classification, pre-processing, dimensionally reduction and indexing [15][16]. To categorize the textual data the techniques of statistical classification like Support Vector Machines, Support Vector Machines, Naïve Bayesian classifier and Decision Tree can be used [13].

d. Clustering: -

Clustering is a technique which is used to find blocks of textual data which has same content [13].It is based on the unsupervised learning technique. It is different from categorization because it does not have the predefined topics specified. By using clustering, we can view the textual document in a different and many subtopics thus making sure that an important document will never be obliterated from the search results of the search engine [2]. The very basic algorithm of clustering technique produces a topic of a vector of all documents and allocate the textual document to the topic cluster which is provided [2]. The K- means algorithm is the most commonly used algorithm in the data mining and hence it is also used in text mining because it provides god result [13].

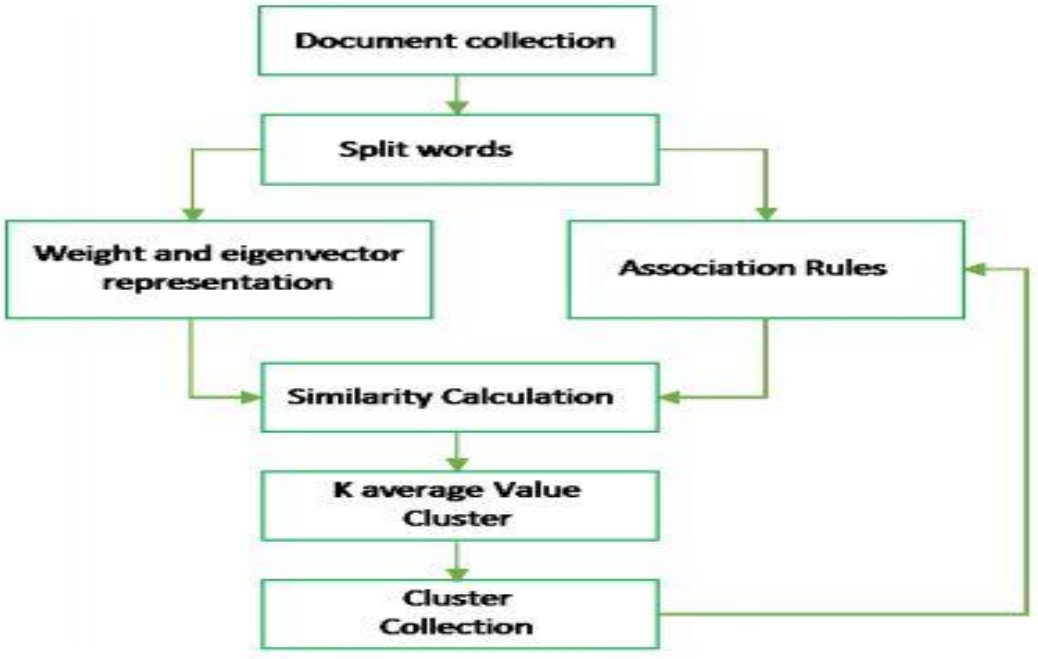

Fig. 4 – Document Clustering [2].

e. Summarization: -

Text summarization is to scale back the count and information of a textual document while keeping hold of the most critical information and without changing the meaning of the sentence [13]. Text summarization is really important to identify whether or not the user should read the lengthy document without actually reading just by referring to the summary of the document [13]. The sentence extraction is the most common implemented by text summarization [13]. It weighs and ranks the relevant sentences from the articles [13]. The tools implemented by summarization also tries to identify headlines and other forms of topics distinguishers to extract the relevant information from the textual document [13]. The summarization method has two groups:

- **Shallow analysis**, limited to the portrayal of syntactic level and always tries to get specific parts from the given text [13].
- **Deeper analysis,** which is semantic level of portrayal of the actual text.

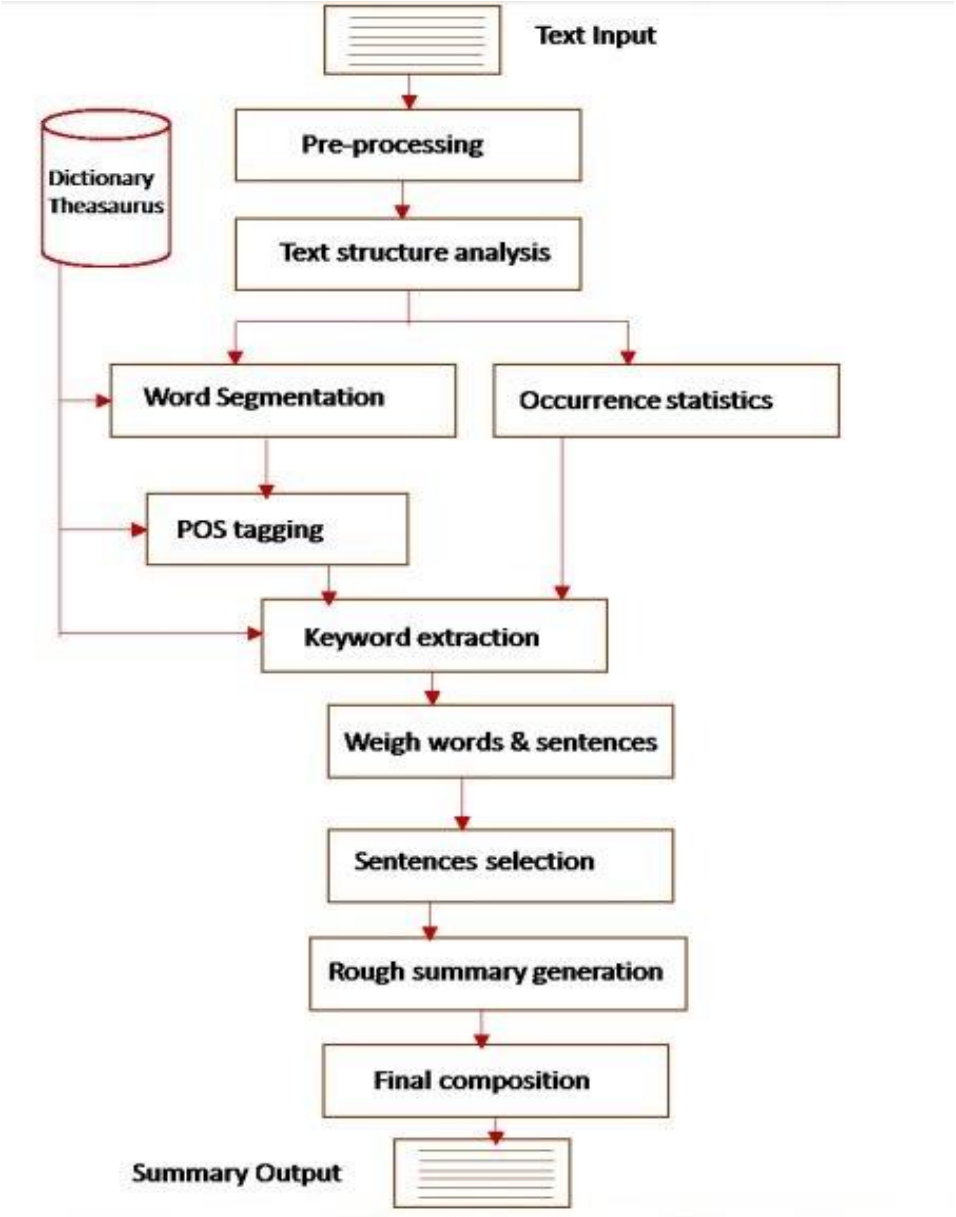

Fig. 5 – Text Summarization [2].

5. Applications of Text data mining: -

a. Sentimental Analysis: -

The sentimental analysis method is used to classify the data set of the twitter which consists of English language into negative and positive sentiments [5]. The implementation of this was usually carried on in the Mahout and the Naïve Bayes algorithm was the mainly used for classification [5]. Before using the Naïve Bayes algorithm many types of linguistics techniques and Natural Language Processing methods was implemented on this tweet before classifying them [5]. The utilization of linguistics pre-processing and NLP techniques targets to increase the probability of Naïve Bayes in identifying the sentiments [5]. The Fig. 6 shows the steps of sentimental analysis. The tweets were pre-utilized before the implementation of training and testing was carried out [5]. After the pre-utilization the tokenization, the POS tagging and the Lemmatization were carried out on this tweets to get the required sentiments out of it [5].

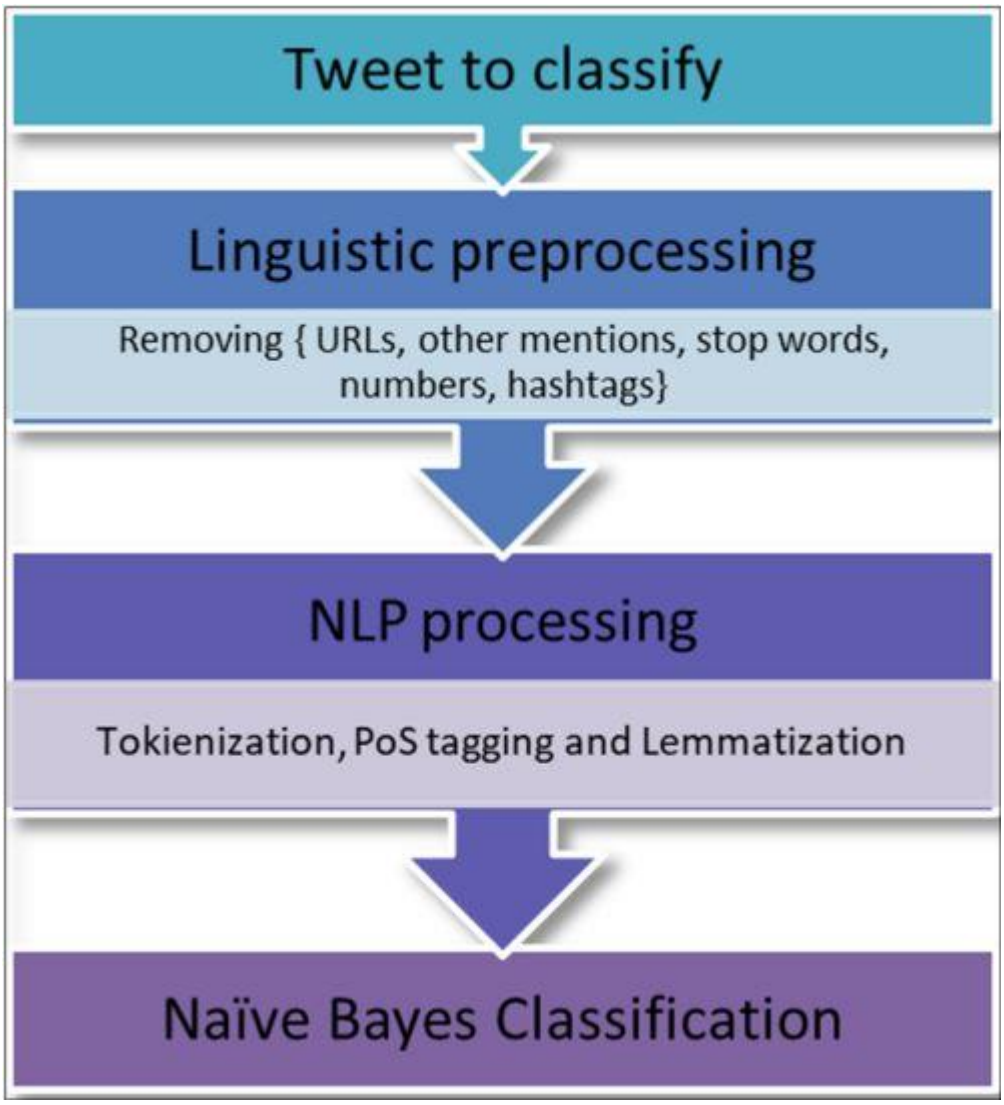

Fig. 6 – Sentimental analysis method [5].

b. Job Market Analysis: -

Text mining is used to extract important and relevant information to identify the skills which are required and the criteria by analysing the current needs of the job market [7].
The keywords are the criteria on which the analysis was done in the database. The applicants name and the relevant job skills required was extracted from the database and the information was made available to the different companies based on their requirements.

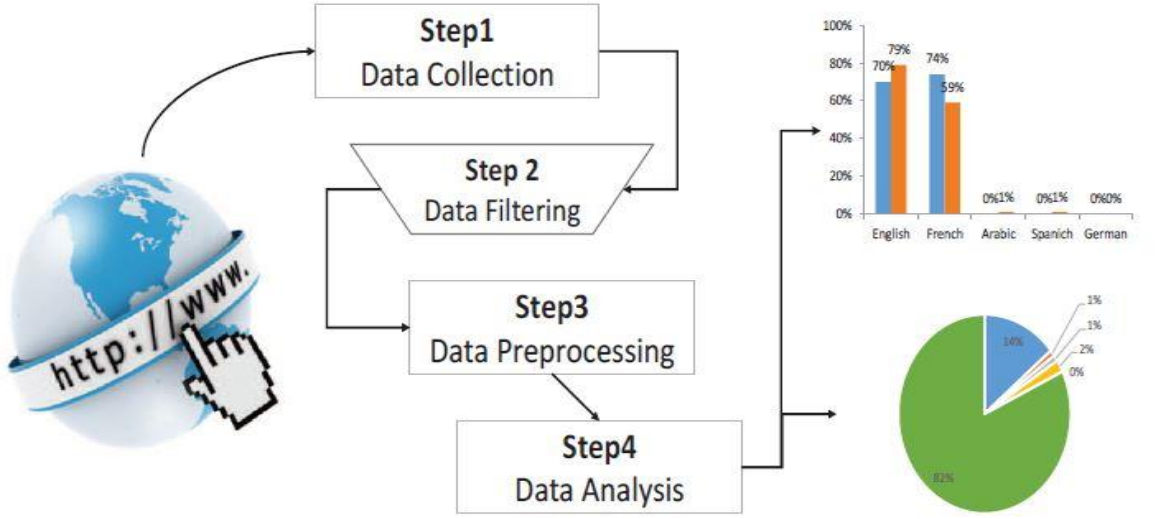

Fig. 7 – Job Market analysis approach [7].

c. Database Analysis: -

The text mining techniques were implemented on the IEEE international journal database in order to extract the research papers based on the relevant information and to organise the papers based on the contents they are currently holding.

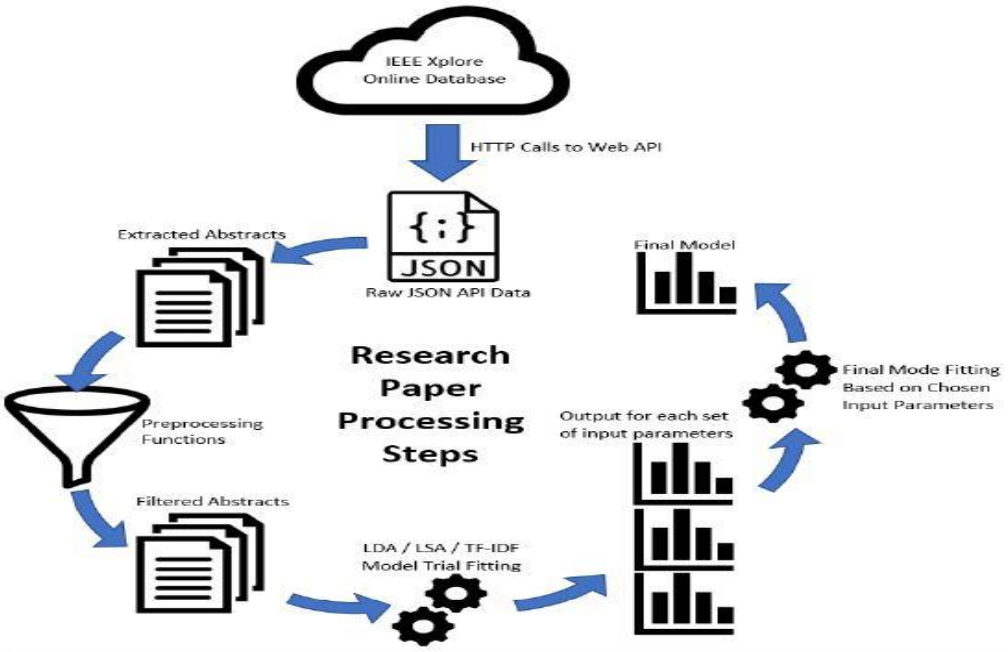

Fig. 8 - Document processing flowchart of text mining in IEEE Xplore database [8].

Company/ Organization	Product/ Application	Text Refining Functions	Intermediate Form	Knowledge Distillation Functions
Cartia	ThemeScape		Document-based	Clustering, visualization
Canis	cMap		Document-based word histograms	Clustering, visualization
IBM/ Synthema	Technology Watch		Document-based	Clustering, visualization
Inxight	VisControls		Document-based Hyperbolic tree	Visualization
Semio Corp	SemioMap		Concept-based	Visualization
Knowledge Discovery System	Concept Explorer	Info retrieval	Concept-based	
Inxight	Linguist	Info retrieval, text analysis, summarization	Document-based	
IBM	iMiner	Info retrieval, summarization	Document-based	Clustering, categorization
TextWise	DR_LINK CINDOR CHESS	Info retrieval Info extraction	Concept-based	
Cambio	Data Junction	Info extraction	Concept-based	
Megaputer	TextAnalyst	Info retrieval, summarization	Document-based semantic net	Classification

Table 1 - Text mining products used in current applications [3].

6. Conclusion: -

Natural language processing and text mining, with its everlasting and esteemed history is the most influential area which is in constant evolution [2]. It is one of the core methods in Artificial Intelligence and sits in the center of the AI techniques. With the constant increasing data and new technology emerging in this 21st century Natural language processing and text mining will be always in depend. The latest advanced developments in the text analytics and NLP leaving behind the simplest search technique are the secret to finding and supporting data and information in all areas. In this Paper the technique, application, methodology and challenging issues are addressed. The challenging issues which we are facing in natural language processing are discussed and to cope with that merit and demerit of text mining are explained. Further this paper explains in depth the methodology and the five important techniques of text data mining. Based on that the current application which are implementing text mining are explained. So, we come to conclusion that the NLP and text mining goes hand in hand and are really important techniques for the upcoming data age in which most operations and workflows will depend on the understanding of data. In the future we should focus more on the extracting and visualization techniques of text data mining so that it can be implemented in NLP which on the other hand will be beneficial in development of artificial intelligence-based applications.

References

1.Elden, I., "Introduction to Natural Language Processing (NLP)" Towards Data Science, Towards Data Science, 17 Sep. (2019), https://towardsdatascience.com/introduction-to-natural-language-processing-nlp-323cc007df3d [Accessed 30-11-2019 13:04].

2.Moreno, A., & Redondo, T., "Text Analytics: the convergence of Big Data and Artificial Intelligence", IJIMAI, 3, pp 57-64 (2016).

3.Tan, A.-H., "Text mining: The state of the art and the challenges", In Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, pp 65-70 (1999).

4.Rai, A., "What is Text Mining: Techniques and Applications", upGrad Blog, 1 Jun. (2019), https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/ [Accessed 30-11-2019 13:04].

5.Khader, M., Awajan, A. and Al-Naymat, G. "The Effects of Natural Language Processing on Big Data Analysis: Sentiment Analysis Case Study", 2018 International Arab Conference on Information Technology (ACIT), Werdanye, Lebanon, 2018, pp. 1-7

6. Ogudo, K. A. and Nestor, D. M. J. "Sentiment Analysis Application and Natural Language Processing for Mobile Network Operators' Support on Social Media", 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Winterton, South Africa, 2019, pp. 1-10

7. Rahhal, I., Makdoun, I., Mezzour, G., Khaouja, I., Carley, K. and Kassou, I. "Analyzing Cybersecurity Job Market Needs in Morocco by Mining Job Ads," 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, United Arab Emirates, 2019, pp. 535-543.

8.Miller, S., "Text Mining the IEEE Xplore Database".

9.Stanford Named Entity Recognizer (2015): http://www-nlp.stanford.edu/software/CRF-NER.shtml [Accessed 30-11-2019 13:07].

10. Finkel, J.R., Grenager, T. and Manning, C., "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).

11.Apache OpenNLP (2015): http://opennlp.apache.org/ [Accessed 30-11-2019 13:20].

12.LingPipe (2011): http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html [Accessed 30-11-2019 13:21].

13.Gaikwad, S.V., Chaugule, A. & Patil, P., "Text mining methods and techniques.", International Journal of Computer Applications 85, no. 17 (2014).

14.Wen, G., Chen, G., & Jiang, L., "Performing Text Categorization on Manifold", IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan, IEEE(2006)

15.Lam, W., Ruiz, M.E. & Srinivasan, P., "Automatic Text Categorization and Its Application to Text Retrieval", IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999

16. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. & Watkins, C., "Text Classification Using StringKernels", J. Machine Learning Research, vol. 2, pp. 419- 444, 2002.

17.Khurana, D., Koli, A., Khatter, K., & Singh, S.,"Natural language processing: State of the art, current trends and challenges." arXiv preprint arXiv:1708.05148 (2017).