International Journal of Scientific and Technical Advancements
ISSN: 2454-1532

# Text Mining with Machine Learning Algorithms: A Review

Akankasha
Department of Computer Applications, Model Institute of Engineering and Technology (MIET)
Jammu, Jammu & Kashmir, India-181122
Email address: akankasha.mca@mietjammu.in

***Abstract-*** Text mining is a terminology which is used to analyze, process and extract appropriate information from unstructured data. With the encroachment in the digital era, a lot of organization gathers information and store massive amounts of data in data warehouse. To analyze such amount of data is multifarious job. Machine learning plays an essential role in text mining. It deals with natural language text which is either stored in semi-structured or unstructured manner. In this paper a brief introduction on text mining is given along with various phases and machine learning tools used for extracting data has been discussed.

***Keywords- Text Mining,Machine learning, K-Nearest Neighbour, Suport Vector Machine(SVM),Navie Bayes,K-mean classifier.***

.

## I. INTRODUCTION

Text Mining also referred as text analytics, is a practice of analysing large amount of data or information collected from larger organization and convert unstructured data into meaningful information [1]. The main aim of text mining is to identify facts and relationship and extraction of patterns and arranging text documents.

Text Mining or Text analytics uses Machine Learning algorithms. For Text Mining, information content can be taken from newspaper, articles, social media applications, emails, stories, reports etc. [2].

Machine learning has an immense role in context of text mining. It is set of statistical systems for recognizing grammatical forms, substances, opinions, and different parts of content or text. Internet has been gaining high priority now a day. It has changed the viewpoint of individuals on things [3].In researches, machine learning algorithms is achieving high momentum because of its flexibility and accuracy for automated text mining.

This paper is organized into six sections. After the prologue to text mining and machine learning in Section 1. Section 2 describes the review of literature of the recent trends in the text mining and machine learning algorithms and techniques. Text mining, its phases are described in Section 3. Section 4 defines the techniques for the text mining. Machine learning algorithms are elaborated in the Section 5. Finally, the paper conclusion and future work is presented in the Section 6.

## II. LITERATURE SURVEY

In [4], the author has discussed how text is mined using text mining techniques like Data Mining, Information Retrieval, Information extraction ,Natural Language Processing (NLP), clustering, information visualization and many more. The author has also discussed various phases of text Mining. In addition ,application of text mining like natural language queries or question answering (Q & A) which deals with how to find the best answer to a given questions has been discussed by the author.

In [5], the author has discussed a class of text classification issues that are described with numerous redundant features and build up a novel measure that captures feature redundancy, and use it to evaluate an expansive accumulation of datasets and when no feature selection is performed then performance of text categorization with SVM peaks. The author also developed a measure that capture feature redundancy and then use that measure to analyze the tremendous accumulation of information.

In [6], the author has discussed how online user surveys enables shoppers to manage data over-burdens and facilitates decision- making. Text mining procedures are utilized to expel semantic qualities from survey texts. The author has showed that the reviews with positive opinion are more favourable than neutral opinions. It also reveals insight into the circumspect of online clients' support casting a ballot exercises and the structure of an improved accommodation casting a voting mechanism for online client review frameworks.

In [7], the author has discussed how machine learning techniques are favourable to statistical models for text arrangement. Support Vector Machine (SVM) in combination with Kernel technique has been used for text-classification system. Only a linear programming problem desires to be resolved and it significantly reduces the computational costs.



Akankasha, "Text Mining with Machine Learning Algorithms: A Review," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 171-174, 2019.

In [8], the author has done comparison on K- Nearest Neighbors, Naïve Bayes and Term- Graph. The analysis showed that KNN shows the maximum accuracy as compared to the Naive Bayes and Term-Graph but gives high time complexity. Term-Graph with other methods rather than the traditional Term-Graph used with AFOPT. This hybrid shows a better result than the traditional combination. The author has also made an information retrieval application using Vector Space Model to give the result of the query entered by the client by showing the relevant document.

## III. TEXT MINING PHASES

The five principal steps associated with text mining are:

1. ***Data Collection***: The first and the foremost step is the collection of the data. The collected data by social media applications, emails, articles, blogs etc is very huge, of varied kind and is unstructured.

2. ***Pre-Processing***: The next and most important step is pre-processing after data collection. Pre-processing or cleansing aims to detect and remove impurities or anomalies from data. Cleansing process try to catch the genuine substance of content accessible and is performed to evacuate stop words stemming

3. ***Conversion***: In third step, convert all the pertinent data separated from unstructured information into structured information.

4. ***Pattern Analysis***: Pattern analysis is implemented by Management Information System (MIS).

5. ***Database:*** In the last step, all information is stored in secured database for further analysis[9]

## IV. TEXT MINING TECHNIQUES

The figure below shows techniques applied for mining data from text.

1. ***Information Extraction (IE):-*** Information Extraction is the phenomenon which extracts significant information from enormous chunks of textual data .IE system helps in automatically obtaining structured data from unstructured text. It depends on the data generated from Natural Language [4].The main focus of this method to analyze the mining of entities, aspects, and their relationships from semi-structured or unstructured content. Once the data is extracted is stored in database for further retrieval.
.

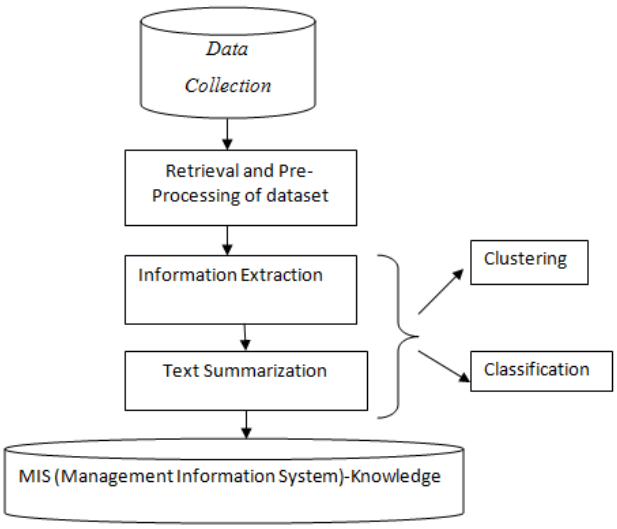


Fig. 1. Text Mining Phases

2. ***Information Retrieval (IR):-*** Information Retrieval is the process of extracting relevant information or in other words, it is a process of finding relevant data from the collection based on specific set of words or phrases. To track user's behavior or to search relevant data, IR system uses some algorithms. The most popular search engines which identify documents on the web are Google and Yahoo [4]. These search engines use inquiry based algorithms to follow the slants and accomplish progressively huge outcomes.

**3.** **Classification:-** Text Classification or categorization is a "supervised" learning process in which a pre-defined data is already given depending upon the document or text to be accessed. This method uses a bag-of-word approach. It also relies on dictionary, words, synonyms and other terms [10].This process gather content and handling and breaking down them to reveal the correct subjects from record.

4. **Clustering:-** Clustering is an "unsupervised" learning process. It tries to recognize inherent structures in literary data and arrange them into significant subgroups or 'groups' for further analysis. Ito frame significant groups from the unlabeled textual information without having any prior data offered [11].

5. **Summarization:-** Text summarization is also called pre-processing, is at process of mechanically generating a compacted version of a specific text that holds important information for the end user[11].In approach includes: removal of stop words like the, a , is etc., conversion of characters either in upper case or lower case ,stemming, punctuation (. , : ; " ' ? |), non-English text etc.



Akankasha, "Text Mining with Machine Learning Algorithms: A Review," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 171-174, 2019.

International Journal of Scientific and Technical Advancements
ISSN: 2454-1532

## V. Machine Learning Algorithms

Machine learning is the ability to automatically learn from the environment and experience without explicitly programmed [12].It is a process in which predefined dataset is given and calculation is prepared on it before applying it on real dataset.

Machine learning algorithms automatically mined text. And thus plays important role in text mining.

There are two types of machine learning:

1.*Supervised Learning*
2.*Unsupervised Learning*

Algorithms which are used in Text Mining are given in fig.2

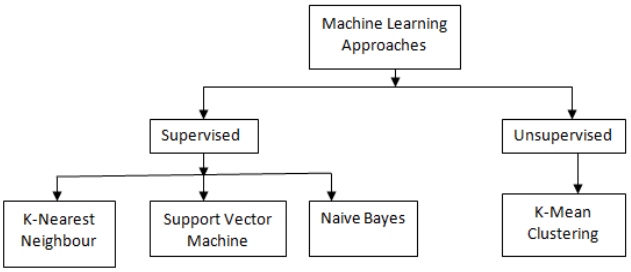


Fig.2 Machine Learning Algorithms

A. ***Supervised Learning***:- This technique considers the training data and produces a contingent function, which can be used for mapping new occurrences. An ideal scenario will allow the technique to correctly determine the class labels for hidden instances. For this learning algorithm is necessary to simplify the training data to unseen situations in a "reasonable" manner[13].

1. **K- Nearest Neighbor (KNN**):- K-NN is an instance-based learning which is also called lazy learning algorithm in which there is local approximation of function. This is a non-parametric method used to perform classification and regression. KNN is a supervised machine learning algorithm [3]. Method for finding distance is Euclidean distance.

$$d(x, y) = \sqrt[2]{\sum_{k=1}^{n}(x_k - y_k)^2}$$

*Equation 1. Euclidean distance*

where *x* and *y* represents the data instances and *d* is distance between x and y.

2. **Support Vector Machine (SVM)**:- It is a supervised machine learning system. An intense instrument for 2-class order and relapse challenges. It is a non-probabilistic straight classifier. It plots the training information in multidimensional space and separate class with Hyperplane. The initial step of Support vector machine is to investigate the information and after that it recognizes the choice boundary and then finally utilizes portions for evaluating execution on input zone. The information is 2 sets of vectors about m measure each. By then every data which deciphered as a vector gets grouped as a class [13].

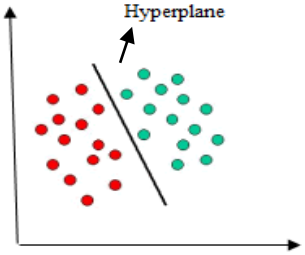


Fig. 3 Linear Classifier [14]

3. ***Naive Bayes:-****This algorithm uses probability based classifier technique. It predicts enrollment probabilities for each class, for example, the likelihood that given record or information point has a place with a specific class. The class with the most astounding likelihood is considered as most likely class [15].*

$$P(c|x) = P(x|c)P(c)/P(x)$$

*Equation 2. Naïve Bayes[15]*

*where P(c|x) is Posterior Probability,*
*P(x|c)is Likelihood,*
*P(c) is class prior probability*
*P(x) is Predicator Prior probability [15].*

B. ***Unsupervised Learning:*** -This technique is used to find hidden structure in unlabeled data. It is a method in which there is no predefined dataset. There is no right target and hence depends on assumptions [3].

*1.* ***K-Mean:-*** It is one of the finest method for clustering. From the given arrangement of n information, k distinctive clusters; each group portrayed with a one of a kind centroid (mean) is apportioned using the K-mean calculation. The components having a place with one cluster are near to the centroid of that specific group and dissimilar ones belongs to other cluster [16].

## VI. Conclusion

In this paper, a brief introduction to Text Mining is given. Phases of the text mining and techniques like Information Extraction, Information Retrieval, classification, clustering and summarization has been discussed. Also Machine learning algorithms like supervised and unsupervised that are used for automatically mining of the text has been discussed. This paper gives only brief review of the above mentioned but in



Akankasha, "Text Mining with Machine Learning Algorithms: A Review," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 171-174, 2019.

future implementation of the algorithms will be done and comparison will be presented.

## REFERENCES

[1] https://www.linguamatics.com/what -is-text-mining-nlp-machine-learning.
[2] U.Nahm and R. Moone , Text minig with information extraction. In *Proceedings of the AAAI 2002 Spring Symposium on mining Answers from Texts and Knowledge Bases*,2002.
[3] Akankasha and Bhavna Arora " Sentimental Analysis on Twitter: Approaches and Techniques", An International Journal of Engineering Science, Special Issue March 2018,Vol. 27,UGC Approved Journal (S.No.63019) ISSN: 2229-6913(Print), ISSN: 2320-0332
[4] S.sathya and N. Rajendran, "A Review of Text Mining Techniques", International Journal of Computer Science Trends and Technology (IJCST)- Vol. 3 Issue 5, Sep-Oct 2015.
[5] Shivani Sharma and Saurabh kr. Srivastava,"Review on Text Mining Algorithms",International Journal of Computer Applications-Vol 134-No. 8, Januray 2016.
[6] Qing Cao, Wenjing Duan, Qiwei Gan, "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach", 0167-9236/$-see front matter © 2010 Elsevier B.V. All rights reserved. doi:10.1016/j.dss.2010.11.009
[7] Liwei Wei, Bo Wei, Bin Wang,"Text Classification Using Support Vector Machine with Mixture ofKernel", A Journal of Software Engineering and Applications, 2012, 5, 55-58, doi:10.4236/jsea.2012.512b012 Published Online December 2012
[8] Bijalwan, V., Kumari, P., Pascual, J., & Semwal, V. B. (2014). KNN based Machine Learning Approach for Text and Document Mining, (June). https://doi.org/10.14257/ijdta.2014.7.1.06
[9] https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/.
[10] Rahul Patel and Gaurav Sharma, "A survey on text mining techniques," International Journal of Engineering and Computer Science(IJECS)- Vol-3 issue 5 May, 2014.
[11] https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/
[12] https://www.expertsystem.com/machine-learning-definition/
[13] http://www.statsoft.com/Textbook/Support-Vector-Machines
[14] Ryan M. Eshleman and Hui Yang," A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints", 2014, IEEE, 978-1-4799-6719-3
[15] https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf.
[16] Prabin Lama "Clustering System Based On Text Mining Using The K-Means Algorithm", Bachelor's thesis (UAS) Information Technology Text Mining and Clustering 2013



Akankasha, "Text Mining with Machine Learning Algorithms: A Review," *International Journal of Scientific and Technical Advancements*, Volume 5, Issue 1, pp. 171-174, 2019.