International Journal for Modern Trends in Science and Technology, 9(01): 54-61, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0901010
Available online at: http://www.ijmtst.com/vol9issue01.html

A Comparative Analysis ofText Mining Techniques and Algorithms

M.Karthica[1] | Dr.K. Meenakshi Sundaram[2]

[1]Research Scholar, Department of Computer Science, Erode Arts and Science College(Autonomous), Erode,Tamilnadu, India. , karthica92@gmail.com
[2]Associate Professor, Department of Computer Science, Erode Arts and Science College(Autonomous), Erode, Tamilnadu,India. lecturerkms@yahoo.com

To Cite this Article
M.Karthica and Dr.K. Meenakshi Sundaram. A Comparative Analysis ofText Mining Techniques and Algorithms. International Journal for Modern Trends in Science and Technology 2022, 9(01), pp. 54-61. https://doi.org/10.46501/IJMTST0901010

Article Info
Received: 20 December 2022; Accepted: 15 January 2023; Published: 25 January 2023.

ABSTRACT

With the abundant technological progression and its colossal consumption develops the gigantic quantity of unstructured text data digitally. This type of data controlluxurious information as well as knowledge. Therefore, in order to extract such an amount of knowledge from unstructured text data, a data expert involve to perform mining techniques over textual data. Text mining is the procedure of extracting hidden, priory unidentified, as well asconsiderablyutilizeful information from unstructured textual data.Web browsers became an significantas well as implement to create the information available at our finger tips. World Wide Web became with information as well as it became tough to regaindata according to the required data. Text mining is a subdivision under web mining. This paper deals with a study of different techniques, pattern of content text mining and the areas which has been influenced by content mining. The web contains efficient, unstructured, partiallyprearranged and multimedia data. This paper focuses on text mining techniques and its algorithmswhich help to retrieve data information in huge data retrieval in content based method.

KEYWORDS: Text mining, Web Usage Mining, Summarization, Clustering, Information retrieval

1. INTRODUCTION

Data Mining is a set of techniques that aims to determine implicit utilizeful information as of big data Web mining helps to understand customer behaviour, estimate the presentation of a web site as well as the explore donein web content mining ultimately helps to enhanceproduction.Nowadaysmainly of the information in government, industry, business,as well as other institutions are stored electronically, in the formof text databases. Data stored in most text databases arepartially structured data in that they are neither completelyunstructured nor completely structured. Webcontent mining examines the search effect of search engine.Physicallyexploitbelongings consumes a assortment of instance. The datato be analyzed is in bulky quantities, and then it is unbreakable to discover outthe appropriateinformation. As now in every field of life manual workis replaced by technology, the overall process of discoveringpotentially utilizeful previously unknown information orknowledge from the web data. Web mining is utilized tocaptureapplicable information,

creating novelfamiliarity out ofthe relevant data, personalization of the information, learningabout Consumers or individual utilizes as well asnumerous others. Several data mining techniques consist of mining imperative patterns in text documents. However, how to successfully utilize and update exposed patterns is still an open research issue, especially in the domain of text mining. Text mining is a procedure to take outattractive as well asimportant patterns to investigate knowledge as of textual data sources s [3]. Text mining is a multiple departmentopiniondepends on information retrieval, as well as computational linguistics. Numerous text mining techniques like summarization, classification, clustering etc., can be functional to extort knowledge. Text mining can handle with natural language text which is stored in semi structured and unstructured format [4]. Text mining techniques are frequentlyuseful in industry, academia, web applications, and internet as well as technical fields.Application areas like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics utilize text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis [6]. Technically, text mining is the utilize of automated methods for exploiting the enormous amount of knowledge available in text documents. Text Mining represents text retrieval and it is a relatively novel and vibrant research area which is changing the emphasis in text-based information technologies from the level of retrieval to the level of analysis and exploration. Text mining, on occasionmoderately invoke to as text data mining, assign normally to the process of originate high quality in sequencebecause text. Analysers adore and others censoriousto facilitate text mining is in addition to known as Text Data Mining and knowledge Discovery in Textual Databases.

2. RELATED WORK

T. Chenet. al [5] described that assembly, extracting, pre-processing, text transformation, feature extraction, pattern selection, and evaluation steps are part of text mining process. In calculation, dissimilarexpansivelyutilized text mining techniques, i.e., clustering categorization, decision tree categorization, application in various fields are surveyed.

R. Bhayaniet al [8] highlighted the issues in text mining applications and techniques. Unstructured text is difficult as compared to structured or tabular data utilizing traditional mining tools as well as techniques. The applications of text mining process in bioinformatics, business intelligence as well as national security system. A natural language processing as well as individual recognition technique hascondensed the problems that occur during text mining process.

D Ramesh et.al [9] explored MEDLINE biomedical database by integrating a framework for named entity recognition, classification of text, hypothesis generation, testing, relationship, synonym extraction, extract abbreviations. This novel framework helps to remove unnecessary details as well asremove valuable information.

ShaileshPandeyet.al [10] analyzed the text using text mining patterns and displayed term based approaches cannot analyze synonyms and polysemy appropriately. Moreover, a samplerepresentation was calculated for measurement of patterns in terms of conveying weight according to their distribution. This approach helps to improve the competence of text mining process.

P. Monali et.al [11] obtainable a crime recognition system utilizing text mining tools and relation discovery algorithm was calculated to associate the term with contraction. Information mining is the expectation equipment for massive data sets it serves to huge associationcenter approximately the more considerable information. It's an appliance topredict the approaching patterns, qualifying association to dissolve on palm on informationambitious choices.

J. Bollen et.al[7], has explained in item surveys, it is seen that the circulation of limit appraisals overaudits collected by a variety of clients or assessed needy on diverse themes are regularly slanted inreality. Thusly, fusing client and item data would be utilizeful for the assignment of notioncharacterization of audits. In any case, existing methodologies overlooked the transient idea of surveysposted by a similar client or assessed on a similar item as well as to contend that the fleeting relations ofsurveys may be possibly valuable for learning client and item installing and subsequentlysuggestutilizing a grouping model to insert these sophisticatedrelationships into client and

item portrayals in command todevelop the exhibition of report level estimation examination.

3. TEXT MINING TECHNIQUES

3.1 Classification

Text classification is the progression of classifying documents into predefined division established on their contented. It is the programmedobligation of natural language texts to prearrangeddivision. Text classification is the crucialconstraint of text retrieval systems, recover texts in reaction to a utilize query, and text understanding systems, whatever transform text in a quantity of way such as developingtext summaries, answering questions otherwise extracting data. Available supervised learning algorithms to robotically classify text requireadequate documents to discoverprecisely. Categorization is to put things according to their characteristics. Given a set of class, classifier determines which classes a given object belongs to. Documents may be classified according to their subjects or the other attributes such as document type, author, printing year etc.

Classification [2] resourcesconveying a document otherwise object to one or more classes. This may be done manually or algorithmically. The intellectual classification of documents is mostly utilized in information science and computer science. Classification is preparedgenerallydepends on traits, performance or subjects. The Classification problem can be stated as a training data set agreeing of proceedings. Each record is identified by a unique record id, and consists of fields corresponding to the aspects. An element with a continuous domain is called a continuous attribute. An attribute with a finite domain of discrete values is called a categorical aspect. Classification is the process of discovering a model for the class in expressions of the continuing attributes. The objective is to utilize the training data set to build a model of the class label based on the other attributes such that the model can be utilized to classify novel data not from the training data set attributes.Other type of classification techniques are also utilized which comes under supervised classification and unsupervised classification.

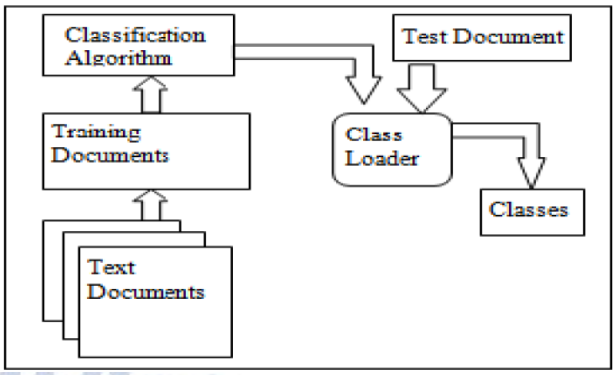

Fig 1.1 Work Flow of Classification

The above figure 1.1 explain the work flow of classification using text document for training data.

3.2 Clustering

Clustering is individual of a large amountordinaryinvestigative data analysis technique utilized to acquire an perceptionconcerning the construction of the data. The situationbe able to be definite as the charge of classifying subgroups in the data such that data points in the matchingsubcategorycluster are enormously similar although data points in altered clusters are exclusivelymiscellaneous. The decision of which similarity measure to utilize is application-specific.The clustering process section is designed to cluster the documents with reference to its relationship. The clustering process groups the documents. The clustering process is alienated into two primary modules. They are term cluster and semantic cluster. The term cluster module is considered to cluster the manuscript with the term weights. The semantic cluster groups the document with semantic weights. Clustering **documents** can also in addition be done by looking at every document in vector format. But documents infrequentlycontaincontext. The furthermostprocedure to script is to offerevery word in the dictionary its hold vector measurement and then just count the occurrences for each word from the entirearticle.

3.3Information Retrieval

Information retrieval is a countrysideso as tohave been budding inparallel with database systems for many years. Unlike thefield of database systems, which has focus on query andoperation processing of structured data, informationretrieval is disturbed with the organization and retrieval ofinformation from a

large number of text-based documents.Since information retrieval as well as database systems eachhandle different kinds of data, some database systemstruggle are usually not present in information retrievalsystems, such as concurrency control, recovery, transactionmanagement, and update. Also, some common informationretrieval problems are usually not encountered in traditionaldatabase systems, such as unstructured documents,estimate search based on keywords, and the notion ofrelevance. Outstanding to the abundance of text information,information retrieval has establishvarious applications. Aroundhappennumerous information retrieval structures, such as on-linelibrary sequence systems, on-line manuscriptorganizationschemes, as well as the more freshlyestablishedweb searchengines. A typical information retrieval problem is to locaterelevant documents in a document collection based on autilizer's query, which is often some keywords describing aninformation need, although it could be relevant document. In such a search problem, a utilizer takesthe initiative to "pull" the relevant information out from thecollection; this is most appropriate when a utilizer has some adhoc information need, such as finding information to buy autilized car. When a utilizer has a long-term information need, aretrieval system may also take the initiative to "push" anynovelly arrived information item to a utilizer if the item isjudged as being relevant to the utilizer's information need. Since a practicallookout, search as well asclarifyingsegmentseveralcollective techniques.

3.4 Information Extraction

Information extraction (IE) is the assignment of automatically withdrawprearranged information beginningshapeless or semi-structuredtext. In supplementarydisputesevidenceabstraction can be measured asa imperfectarrangement of occupied natural language sympathetic, where theinformation remainobservingaimed at are recognized beforehand. IE is one of the disapprovingresponsibilities in text mining as well asextensivelydeliberatein diverse research societies such as information retrieval,natural language processing and Web mining. Information extraction includes two fundamental tasks, namely,name entity recognition and relation extraction. The states of the art in both tasks are statistical learning methods. The commonpersistence of Knowledge Detection is to "extractcontained, previously unknown, and potentially utilizefulinformation from data". Information Extraction IE mostlyagreements with classifying words or mouthlanguagesas ofinside adocumentary file. Feature positions can be well-defined as those which aresprightlyconnected to the province.

Categorization,Summarization, Prioritization, Survey Searching
Feature Term Identification
Token Creation & Assembly
Document Input

Figure 1.2 A layered model of the Text Mining Application.

3.4.1 Stemming

Stemming mentions to detecting the derivation of a definite word.To hand are essentiallydual types of stemming techniques, initial oneis inflectional and second one is derivational. Derivationalstemming can create a novel word from an existing word,sometimes by simply changing grammatical category. The categoryof stemming continuedable to scheme is calleddeclensionreducing. A frequentlyutilized algorithms is the'Porter's Algorithm' for stemming. The normalizationis restricted to normalizinglinguisticvariations such assingular/plural or past/present, it is referred to curvaturestemming.To minimalize the belongings of variation as well as structural disparities of words, attitude has reprocessedrespectivelydiscussion utilizing a deliveredvariability of the Porter stemming algorithm with a fewfluctuationsconcerning the end in which have omitted somecases.

3.4.2 Domain dictionary

Trendydirectiveis to progress tools of this category, it is important toafford them with a data base. A cooperativeusual of allthe feature terms is the Domain dictionary. The assembly of the Domaindictionary implemented contained of three levelsin the hierarchy. Namely, Parent Grouping, Sub-categoryas well as word. Starting groupingsdescribe the centralgroupingfurther downwhich some sub-category otherwiseexpression falls. A categorywill be exceptional on its level in the hierarchy. Additional categoriesgo to

a convincedinitial category as well as every subcategorywill involve of all the words related with it. A lot of words in a text file can be preserved as undesirableclatter. To eradicate the invented a distinct file adding all related words. These containdisputes such as the, a,an, if, off, on etc.

3.4.3 Text Indexing Techniques

Around are frequentcommon text retrieval indexing techniques,as well asoverturneddirectories as well as signature files. An overturnedcatalogue is an index structure that preservesbinary hash indexedor B+-tree indexed tables: document table and term table,where document boardcontains of a set of manuscriptproceedings,everycomprising two fields: doc id and posting list, whereposting list is a list of terms otherwise pointers to terms that occurin the document, sorted according to some relevancemeasure. This involves of a set of duration records, respectivelycomprising twice in a fields: term id and posting list, whereposting list requires a list of manuscript identifiers the term performs. Throughan organization, it is informal toresponsequestions like "Novelty all of the formsconnectedwith anassumed set of terms," or "Discovery all of the languagesconnected with a given set of forms." Toinvention all of the formsrelated with a set of terms. First identify the slope of manuscript identifiers in term table on behalf ofrespectively. Then overlap them to attain the established ofrelevant documents. Inverted indices are widely utilized inindustry. They are informal to device as well as the posting listscould be rather long, manufacture the loading requirement quitelarge. They are easy to implement, but are not satisfactory at handling synonymy like where two very dissimilar words canrequire the equivalentdenotation as well as polysemy where an separateword might have many meanings. Each signature has a secure size of *b* bitsrepresenting relations. A humbleencrypting scheme goes asfollows. Each bit of a manuscriptname is modified to 0.A bit is established to 1 if the term it signifiesseems in thedocument. Such multipletosingle mappings make the search expensive because adocument that matches the signature of a query does notautomaticallycomprise the set of keywords of the query. Thedocument consumes to be recovered, analysed, stemmed, as well as checked. Enhancementsbe capable to be complete by first executionoccurrence analysis, stemming, as well as by straining stop words,as well as utilizing a hashing technique as well as coveredcoding technique to encrypt the list of terms into bitrepresentation. However, the problem of multipletoonemappings still consists the major disadvantage ofthis approach. Researchers can declaration to for supplementaryconversation of indexing techniques, containingexactly how tocompress an index.

3.5 Natural Language Processing

NLPutilizes some level of underlying linguistic representation of text, to formulate sure that the generated text is grammatically correct and fluent. Most NLP systems include a syntactic releaser to ensure that grammatical rules such as subject-verb agreement are observed, as well as text planner to resolve how to assemblestretches, paragraph, as well as other parts coherently. In tokenization, a verdict is segmented into a list of tokens. The token symbolizes a word or a special symbol such an exclamation mark. Morphological otherwise lexical examination is a procedureuniversallycorrespondinglyappearance is identifiedfinished its quantity of dialogue. The difficulty arises in this regulates, charity for instance, how a sentence is broken down into phrases, and in what way the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words utilized.

Table 1.1 Comparative Analysis for Text Mining Techniques

TECHNIQUES	ADVANTAGES	DISADVANTAGES
Classification	• Training is very fast. • Easy to understand and implement	• Perform very poorly when features are highly correlated
Clustering	• clustered solution is automatic recovery from failure • recovery without utilizer intervention • No training Data needed	• Clustering are complexity and inability to recover from database corruption. • Not to explicit as supervised classification.
Information Retrieval	• The most practical for indexing and retrieving large	• Low level features are not able to describe and interpret semantically.

	amount of images. • Textual induction	
Information Extraction	• Non-toxic • Statistically clear	• Unguided analysis • Statistically dependent
Natural Language Processing	• Relieves burdenof learning • No Training	• Require more clarification • Unpredictable • May not show context

4. TEXT MINING ALGORITHMS

4.1 Naive Bayes Classifier

Probabilistic classifiers consumeincreasedan allocation of popularity freshly as well asto completeunusually well. Theseprobabilistic approaches varietyexpectations about how the data(words in documents) are generated and propose a probabilisticmodel based on these assumptions. Then utilize a set of training examplesto estimate the parameters of the model. Bayes rule is utilizedto classify novel examples and select the class that is most likely has generated.The Naive Bayes classifier is perhaps the simplest and the mostwidely utilized classifier.

4.2 Decision Tree classifiers

Decision tree is basically a hierarchical tree of the training instances,in which a complaint on the quality value is utilized to divide thedata hierarchically. Respectively node of the treeis a test of aroundcharacteristic of the traning occurrence, as well as respectively branchdescendantsince the node resembles to one the value of thisaspect. An occurrence is confidential by establishment at the root node,testing the characteristic by this node as well as moving down the tree branchconforming to the value of the characteristic in the specifiedoccasion. Foroccurrence a node may be segmented to its nonexistence of a particular term in the document. Decision trees have been utilized in combination with boostingtechniques.

As soon as decision tree is utilized for text classification it contain tree internal node are label by term, divisions departing from labeled by test on the weight, as well as leaf node are characterize corresponding class labels. Tree be able tocategorize the document by consecutivelycomplete the query structure from root to awaiting it scopes a convinced leaf, which characterizes the area for the classification of the document.

4.3 Support Vector Machines

Support Vector Machines (SVM) are supervised learning classificationalgorithms where have been extensively utilized in text classificationproblems. SVM are a form of Linear Classifiers. The context of text documents are models that manufacture a classificationdecision is constructedarranged the assessment of the linear arrangementsof the documents features. Thus, the output of a linear predictoris defined to be y = Ra · Rx + b, where Rx = (x1, x2, . . . , xn) is thenormalized document word frequency vector, Ra = (a1, a2, . . . , an)is vector of coefficients and b is a scalar. We can interpret the predictory = Ra · Rx + b in the categorical class labels as a separatinghyperplane between different classes.A single SVM can only separate two classes, a positive class and a negative class. SVM algorithm attempts to find ahyperplane with the maximum distance fromthe positive and negative examples. The documents with distance from the hyperplane are called support vectors and specify theactual location of the hyperplane. Uniqueimprovement of the SVM method is that, it is moderatelystrong tohuge dimensionality, for learning is almost autonomous of the dimensionalityof the feature space.

4.4k-means Clustering

K-means clustering is one the partitioning algorithms which iswidely utilized in the data mining. The k-means clustering partitionsn number of documents in the environment of manuscript data into k number clusters. Representativearound which the clusters are built. The basic form of k-meansalgorithm is:Finding an optimal solution for k-means clustering is computationallydifficult (NP-hard), however, there are efficient heuristics suchas that are employed in order to converge rapidly to a localoptimum. The main difficulty of k-means gathering is that it iscertainlyprecisesearching to the preliminaryoptimal. Thus,there are some techniques utilized to determine the initial k, usinganother lightweight clustering algorithm such as agglomerative clustering algorithm.

4.5 Hierarchical algorithms

Hierarchical clustering denotes to an unsupervised learning procedure that regulatesconsecutive clusters depends on formerlydemarcated clusters. The final point discuss to a various set of clusters, where each and every cluster is various from the other type of cluster,

and the objects within each cluster are the same as one another.

There are two different types of hierarchical clustering

- ❖ Agglomerative Hierarchical Clustering
- ❖ Divisive Clustering

4.5.1 Agglomerative hierarchical clustering

Agglomerative clustering is one of the most common types of hierarchical clustering utilized to group similar objects in clusters. Correspondingly data point performance as an individual cluster as well as at each step, data objects are assembled in a bottom-up technique in Agglomerative clustering. Primarily, each data object is in its cluster. At each iteration, the clusters are collective with different clusters until one cluster is formed.

4.5.2 Divisive Hierarchical Clustering

Disruptive hierarchical clustering is accurately the contrasting of Agglomerative Hierarchical clustering. In disruptive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster.

Table 1.2Comparison Table of Text Mining Algorithms

ALGORITHMS	PROS	CONS
Naive Bayes Classifier	• Work well on numeric textual data • Easy to implement and computation • Easily modified Compare with different algorithm	• Perform very poorly when features are highly correlated
Decision Tree classifiers	• Easy to understand • Easy to generate rule • Reduce problem complexity	• Training time is expensive • A document only connected with one branch • May Suffer from over fitting.
Support Vector Machines	• Work well on numeric or textual data • Easy to implement and computation	• Perform very poorly when features are highly corrected.
	• Work for linear and non linear data • More capable to solve multi-label classification	
K-Means Clustering	• Easy to implement and identify unknown groups of data from complex datasets. • The results are presented in an Easy and simple manner.	• No-optimal set of clusters • Lacks of consistency • Breaks large clusters. • It is sensitive to noise and outliers.
Hierarchical Algorithms	• It is robust and impervious to noise • Better speed and accuracy	• Handles only numerical data

5. CONCLUSION

Text Mining can be defined as a technique which is utilized to extract interesting information or knowledge from the text documents which are usually in the unstructured form. Text Mining is discussed with its various techniques which can be utilized such Classification, Clustering, Summarization and various techniques and methods are discussed for efficient and accurate text mining. In this short survey, compare the notion of text mining techniques have been analysed and algorithms available have been presented. Due to its novelty, there are many potential research areas in the field of Text Mining, which includes finding better intermediate forms for representing the outputs of information extraction, an XML document may be a good choice. Mining texts in different languages is amajor problem, since text mining tools should be able towork with many languages and multilingual documents.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] B. Pang and L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retrieval, Vol. 2, Issue.1, pp. 1–135, 2018.

[2] J. Bollen, H. Mao, and A. Pepe, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, in Proc. Int. AAAI Conf. Weblogs Social Media, pp. 17–21, 2019.

[3] D. Devikanniga, A. Ramu, and A. Haldorai, Efficient Diagnosis of Liver Disease using Support Vector Machine Optimized with Crows Search Algorithm, EAI Endorsed Transactions on Energy Web, p. 164177, Jul. 2018. doi:10.4108/eai.13-7-2018.164177

[4] H. Anandakumar and K. Umamaheswari, Supervised machine learning techniques in cognitive radio networks during cooperative spectrum handovers, Cluster Computing, vol. 20, no. 2, pp.1505–1515, Mar. 2017.

[5] T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang, Learning utilizer and product distributed representations using a sequence model for sentiment analysis, IEEE Comput. Intell. Mag., 11(3), pp. 34–44, Aug. 2018.

[6] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, OpinionFlow: Visual analysis of opinion diffusion on social media, IEEE Trans. Vis. Comput. Graph., Vol. 20, Issue. 12, pp. 1763–1772, Dec. 2019.

[7] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment classification using machinelearning techniques, in Proc. ACL Conf. Empirical Methods Natural Language Process., pp. 79–86, 2019.

[8] A.Go, R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision,Stanford Univ., Stanford, CA, USA, Project Rep. CS224N, pp: 1–12, 2018.

[9] D Ramesh , B Vishnu Vardhan, " Analysis of Crop Yield Prediction using Data Mining Techniques", IJRET: International Journal of Research in Engineering and Technology, eISSN: 2319-1163, pISSN: 2321-7308, Volume: 04 Issue: 01, Jan-2015.

[10] ShaileshPandey, Sandeep Kumar, "Enhanced Artificial Bee Colony Algorithm and its a. Application to Travelling Salesman Problem", International Journal of Technology Innovations and Research, HCTL Open IJTIR, Volume 2, e-ISSN: 2321-1814, March 2013.

[11] P. Monali , K. Sandip, "A Concise Survey on Text Data Mining" in proceeding of the International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 9, pp: 8040- 8043, September 2014.

[12] Unil Yun, Hyoju Nam, Gangin Lee, and Eunchul Yoon. 2019. Efficient approach for incremental high utility patternmining with indexed list structure. Future Generation Computer Systems95, pp:221–239. DOI: https://doi.org/10.1016/j.future.2018.12.029, 2019.

[13] Lior Shabtay, Philippe Fournier-Viger, Rami Yaari, and Itai Dattner. 2020. A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data