Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7923016

A STUDY ON TECHNIQUES AND TOOLS ASSOCIATE WITH WEB CONTENT

M. Karthica [1], Dr. K. Meenakshi Sundaram[2]
[1] (Ph.D. Research Scholar, Department of Computer Science, Erode Arts and Science College(Autonomous), Erode, Tamilnadu, India, karthica92@gmail.com,)
[2] (Associate Professor and Head, Department of Computer Science, Erode Arts and Science College(Autonomous), Erode, Tamilnadu, India, lecturerkms@yahoo.com)

ABSTRACT
Machine learning, as well as data mining are research areas of computer technology whose quick development is due to the advances in data analysis research, growth in the database industry as well as the resulting market needs for methods that are capable of extracting valuable knowledge from large data stores. It is a powerful platform that stores as well as retrieves mass information and it becomes a time-consuming uncomfortable task to search the information due to the unstructured as well as heterogeneous nature of data on the web development. In the recent past, the advancement in computer and multimedia technologies has led to the production of digital images and cheap large image depositories. The size of image collections has increased fleetly due to this, comprising digital libraries, medical images, etc. To attack this rapid-fire growth, it's needed to develop an image reclamation system that operates on a large scale. The primary end is to make a robust system that creates, manages, and query image databases in an accurate manner. The World Wide Web delivers a great platform that stores and retrieves mass information. It becomes a time-consuming and uncomfortable assignment to search the information due to its unstructured and heterogeneous nature of data on the World Wide Web. Web mining is one of the widespread techniques of data mining that is used to determine and extract useful information from web documents and their facilities. Web usage mining, web structure, and web content are three different types of web data mining. Each of these categories has numerous methods, tools, and approaches to excerpt data from the volume of information over the web. This review paper states various issues while encountering information from the web and also states several problems that occurred while finding appropriate information from the web.
Keywords: Text mining, Web Usage Mining, Summarization, Clustering, Information retrieval.

I. 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, and estimate the presence of a website as well as the exploration is done in web content mining ultimately helps to enhance production. Nowadays mainly information in government, industry, business, as well as other institutions is stored electronically, in the form of text databases. Data stored in most text databases are partially structured data in that they are neither completely unstructured nor completely structured. Web content mining examines the search effect of search engines. Physically exploit belongings consumes an assortment of instances. The data to be analysed is

in bulky quantities, and then it is unbreakable to discover the appropriate information. Now in every field of life manual work is replaced by technology, the overall process of discovering and potentially utilizing previously unknown information or knowledge from web data. Web mining is utilized to capture applicable information, create novel familiarity out of the relevant data, personalization the information, learning about Consumers or individual utilizes as well as numerous 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 out attractive as well as important patterns to investigate knowledge of textual data sources [3]. Text mining is a multiple department opinion that depends 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 natural language text which is stored in semi structured and unstructured formats [4]. Text mining techniques are frequently useful in industry, academia, web applications, and the internet as well as technical fields. Application areas like search engines, customer relationship management systems, 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 utilization 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 occasion moderately invoke as text data mining, assign normally to the process of originating high quality in sequence because of text. Analysers adore and others censorious to facilitate text mining are in addition known as Text Data Mining and Knowledge Discovery in Textual Databases.

II. LITERATURE REVIEW

T. Chen et. al [9] described that assembly, extracting, pre-processing, text transformation, feature extraction, pattern selection, and evaluation steps are part of text mining process. In calculation, dissimilar expansively utilized text mining techniques, i.e., clustering categorization, decision tree categorization, application in various fields are surveyed.

Pang et.al [2] 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 has condensed the problems that occur during text mining process.

D Ramesh et.al [4] 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 as remove valuable information. Shailesh Pandey et.al [10] analyzed the text using text mining patterns and displayed term based approaches cannot analyze synonyms and

polysemy appropriately. Moreover, a sample representation 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 association centre approximately the more considerable information. It's an appliance to predict the approaching patterns, qualifying association to dissolve on palm on information ambitious choices.

Heonho Kim et.al [8], has explained in item surveys, it is seen that the circulation of limit appraisals over audits collected by a variety of clients or assessed needy on diverse themes are regularly slanted in reality. Thusly, fusing client and item data would be utilizeful for the assignment of notion characterization of audits. In any case, existing methodologies overlooked the transient idea of surveys posted by a similar client or assessed on a similar item as well as to contend that the fleeting relations of surveys may be possibly valuable for learning client and item installing and subsequently suggest utilizing a grouping model to insert these sophisticated relationships into client and item portrayals in command to develop the exhibition of report level estimation examination.

III. TEXT MINING TECHNIQUES

Classification

Text classification is the progression of classifying documents into predefined division established on their contented. It is the programmed obligation of natural language texts to prearranged division. Text classification is the crucial constraint 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 developing text summaries, answering questions otherwise extracting data. Available supervised learning algorithms to robotically classify text require adequate documents to discover precisely. Categorization is to put things according to their characteristics. Assumed a set of class, classifier defines which classes a given object belongs to. Documents may be classified allowing to their subjects or the other attributes for instance document type, author, printing year etc.
Classification resources conveying a document otherwise object to one or more classes [2]. This may be done manually or algorithmically. The intellectual classification of documents is mostly utilized in information science and computer science. Classification is prepared generally depends 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. The following Fig. 1.1 explains the work flow of classification using text document for training data.

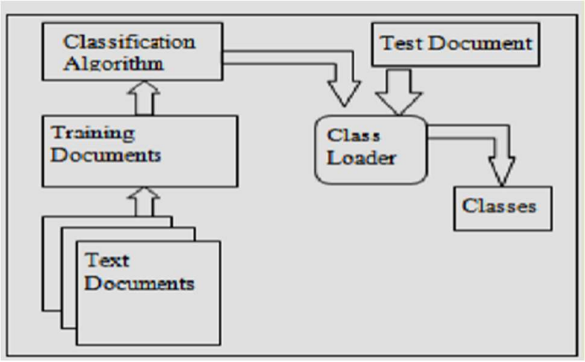

Fig. 1.1 Work Flow of Text Classification

Clustering

Clustering is individual of a large amount ordinary investigative data analysis technique utilized to acquire a perception concerning the construction of the data. The situation is able to be definite as the charge of classifying

subgroups in the data such that data points in the matching subcategory cluster are enormously similar although data points in altered clusters are exclusively miscellaneous. The decision of which similarity measure to utilize is application-specific. The clustering process fragment is deliberate to cluster the documents with reference to its connection. 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 and semantic cluster groups the document using semantic weights. Clustering documents can also in addition be done by looking at every document in vector format. But documents infrequently contain context. The furthermost procedure to script is to offer every word in the dictionary its hold vector measurement and then just count the occurrences for each word from the entire article.

Information Retrieval

Information retrieval is countryside so as to have been budding in parallel with database systems for many years. Unlike the field of database systems, which has focus on query and operation processing of structured data, information retrieval is disturbed with the organization and retrieval of information from a large number of text- based documents. Since information retrieval as well as database systems each handle different kinds of data, some database system struggle are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, estimate search based on keywords, and the notion of relevance. Outstanding to the abundance of text information, information retrieval has established various applications. Around happen numerous information retrieval structures, such as on-line library sequence systems, on-line manuscript organization schemes, as well as the more freshly established web search engines. A typical information retrieval problem is to locate relevant documents in a

document collection based on a utilizer's query, which is often some keywords describing an information need, although it could be relevant document. In such a search problem, a utilizer takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a utilizer has some ad hoc information need, such as finding information to buy a utilized car. When a utilizer has a long-term information need, a retrieval system may also take the initiative to "push" any novelty arrived information item to a utilizer if the item is judged as being relevant to the utilizer's information need. Since a practical lookout, search as well as clarifying segment several collective techniques.

Information Extraction

Information extraction (IE) is the assignment of automatically withdraw prearranged information beginning shapeless or semi-structured text. In other words information extraction can be deliberated as a limited form of full natural language understanding, where the information is considering for is known beforehand. IE is one of the censorious tasks in text mining and widely studied in different research communities such as information retrieval, natural language processing and Web mining. Information extraction comprises two essential tasks, namely, name entity recognition and relation extraction. The states of the art in both tasks are statistical learning methods. The general purpose of Knowledge Discovery is to "extract implicit, previously unknown, and potentially utilizeful information from data". Information extraction is mostly agreements with classifying words or mouth languages as of inside a documentary file. Feature terms can be demarcated as those which are directly associated to the domain. These are the positions which can be recognized by the tool. In order to accomplish this function optimally, we had to look into few more features which are as follows:

Stemming

Stemming mentions to detecting the derivation of a definite word. To hand are essentially dual types of stemming techniques, initial one is inflectional and second one is derivational. Derivational stemming can create a novel word from an existing word, sometimes by simply changing grammatical category. The category of stemming continued able to scheme is called declension reducing. A frequently utilized algorithms is the 'Porter's Algorithm' for stemming. The normalization is restricted to normalizing linguistic variations such as singular/plural or past/present, it is referred to curvature stemming. To minimalize the belongings of variation as well as structural disparities of words, attitude has reprocessed respectively discussion utilizing a delivered variability of the Porter stemming algorithm with a few fluctuations concerning the end in which have omitted some cases.

e.g. apply – applied – applies print – printing – prints – printed

In both the cases, all words of the first instance will be treated as 'apply' and all words of the second example will be preserved as 'print'.

Domain dictionary

Trendy directive is to progress tools of this category, it is important to afford them with a data base. A cooperative usual of all the feature terms is the Domain dictionary. The assembly of the Domain dictionary implemented contained of three levels in the hierarchy. Namely, Parent

Grouping, Sub-category as well as word. Starting groupings describe the central grouping further down which some sub-category otherwise expression falls. A category will be exceptional on its level in the hierarchy. Additional categories go to a convinced initial category as well as every subcategory will involve of all the words related with it. A lot of words in a text file can be preserved as undesirable clatter. To eradicate the invented a distinct file adding all related words. These contain disputes such as the, a, an, if, off, on etc.

Text Indexing Techniques

Around are frequent common text retrieval indexing techniques, as well as overturned directories as well as signature files. An inverted index is an index structure that maintains two hash indexed or B+-tree indexed tables: document table and term table, where document table consists of a set of document records, each containing two fields: doc id and posting list, where posting list is a list of terms otherwise pointers to terms that occur in the document, sorted according to some relevance measure. This involves of a set of duration records, respectively comprising twice in a fields: term id and posting list, where posting list requires a list of manuscript identifiers the term performs. Through an organization, it is informal to response questions like "Novelty all of the forms connected with an assumed set of terms," or "Discovery all of the languages connected with a given set of forms." To invention all of the forms related with a set of terms. First identify the slope of manuscript identifiers in term table on behalf of respectively. Then overlap them to attain the established of relevant documents. Inverted indices are widely utilized in industry. They are informal to device as well as the posting lists could be rather long, manufacture the loading requirement quite large. They are easy to implement, but are not satisfactory at handling synonymy like where two very different words can have the same meaning and polysemy where an individual word may have many meanings. A signature stores a signature record for every document in the database. Each signature has a secure size of b bits demonstrating relations. A humble encrypting scheme drives as follows. Each bit of a document signature is set to 0. A bit is set to 1 if the term it denotes appears in the document. Such multiple to single mappings make the search expensive because a document that matches the signature of a query does not necessarily contain the set of keywords of the query. The document consumes to be recovered, analysed, stemmed, as well as checked. Enhancements be capable to be complete by first execution occurrence analysis, stemming, as well as by straining stop words, as well as utilizing a hashing technique as well as covered coding technique to encrypt the list of terms into bit representation. However, the problem of multiple to one mapping still consists the major disadvantage of this approach. Researchers can declaration to for supplementary conversation of indexing techniques, containing exactly how to compress an index.

Natural Language Processing

NLP utilizes 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 obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently. The most distinguished NLG application is machine translation system. The system examines texts from

a source language into grammatical or conceptual representations and then produces corresponding texts in the objective language. NLU must follow any one of the field which given underneath. Now tokenization, a sentence is segmented into a list of tokens. The token signifies a word or a special symbol such an exclamation mark. Morphological otherwise lexical examination is a procedure universally correspondingly appearance is identified finished its quantity of dialogue. The difficulty arises in this process when it is possible to tag a word with more than one part of speech. Syntactic analysis is a process of conveying a syntactic structure or a parse tree, to a given natural language sentence. It 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 analyses 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 amount of images • Textual induction	• Low level features are not able to describe and interpret semantically
Information Extraction	• Non-toxic • Statistically clear	• Unguided analysis • Statistically dependent
Natural Language Processing	• Relieves burden of learning • No Training	• Require more clarification • Unpredictable • May not show context

IV. TEXT MINING ALGORITHMS

Naive Bayes Classifier

Probabilistic classifiers consume increased an allocation of popularity freshly as well as to complete unusually well. These probabilistic methods variety expectations about how the data (words in documents) are produced and propose a probabilistic model based on these expectations. Then utilize a set of training instances to estimate the parameters of the model. Bayes rule is developed to classify novel samples and select the class that is most

likely has created. The Naive Bayes classifier is feasibly the simplest and the most extensively utilized classifier. It models the distribution of documents in each class using a probabilistic model supposing that the distribution of different terms are independent from each other. Whereas this so called “naive Bayes” assumption is obviously false in many real world applications, naive Bayes performs unexpectedly well.

Decision Tree classifiers

Decision tree is basically a hierarchical tree of the training instances, in which a condition on the attribute value is utilized to divide the data hierarchically. In other words decision tree recursively partitions the training data set into smaller subdivisions based on a set of tests defined at each node or branch. Respectively node of the tree is a test of around characteristic of the training occurrence, as well as respectively branch descendant since the node resembles to one the value of this aspect. An occurrence is confidential by establishment at the root node, testing the characteristic by this node as well as moving down the tree branch conforming to the value of the characteristic in the specified occasion. For occurrence a node may be segmented to its nonexistence of a particular term in the document. Decision trees have been utilized in combination with boosting techniques.

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 to categorize the document by consecutively complete the query structure from root to awaiting it scopes a convinced leaf, which characterizes the area for the classification of the document.

Support Vector Machines

Support Vector Machines (SVM) are supervised learning classification algorithms where have been extensively utilized in text classification problems. SVM are a form of Linear Classifiers. The context of text documents are models that manufacture a classification decision is constructed arranged the assessment of the linear arrangements of the documents features. Thus, the output of a linear predictor is defined to be $y = Ra \cdot Rx + b$, where $Rx = (x1, x2, \ldots, xn)$ is the normalized document word frequency vector, $Ra = (a1, a2, \ldots, an)$ is vector of coefficients and b is a scalar. We can interpret the predictor $y = Ra \cdot Rx + b$ in the categorical class labels as a separating hyperplane between different classes. A single support vector machines can simply separate two classes, a positive class and a negative class. SVM algorithm attempts to find a hyperplane with the maximum distance from the positive and negative examples. The documents with distance from the hyperplane are called support vectors and specify the actual location of the hyperplane. If the document vectors of the two classes are not linearly distinguishable, a hyperplane is determined such that the minimum number of document vectors is located in the erroneous side. Unique improvement of the SVM method is that, it is moderately strong to huge dimensionality, for learning is almost autonomous of the dimensionality of the feature space.

K-means Clustering

K-means clustering is one the partitioning algorithms which is widely utilized in the data mining. The k-means clustering partitions n number of documents in the environment of manuscript data into k number clusters. Representative around which the clusters are built. The basic form of k-means algorithm is: Finding an optimal solution for k-means clustering is computationally difficult (NP-hard), however, there are efficient heuristics such as that are employed in order to converge rapidly to a local optimum. The main difficulty of k-means gathering is that it is certainly precise searching to the preliminary optimal. Thus, there are

some techniques utilized to determine the initial k, using another lightweight clustering algorithm such as agglomerative clustering algorithm.

Hierarchical algorithms

Hierarchical clustering denotes to an unsupervised learning procedure that regulates consecutive clusters depends on formerly demarcated 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 categories of hierarchical clustering

- Agglomerative Hierarchical Clustering
- Divisive Clustering

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.

Table 1.2 Comparison Table of Text Mining Algorithms

ALGORITHMS	PROS	CONS
Naive Bayes Classifier	• Work well on numeric textual data • Easy to implement andcomputation • Easily modified Compare withdifferent algorithm	• Perform very poorly when features are highlycorrelated
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 overfitting.
Support Vector Machines	• Work well on numeric or textualdata • Easy to implement andcomputation • Work for linear and nonlineardata • More capable to solve multi-labelclassification	• Perform very poorly when features are highly corrected.

K-Means Clustering	• Easy to implement and identifyunknown groups of data from complex datasets. • The results are presented in anEasy and simple manner.	• No-optimal set of clusters • Lacks of consistency • Breaks large clusters. • It is sensitive to noise andoutliers.
Hierarchical Algorithms	• It is robust and impervious tonoise • Better speed and accuracy	• Handles only numerical data

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. The separated data points are treated as an distinct cluster.

V. EXPERIMENTAL ANALYSIS

Precision is a good measure to determine, when the costs of False Positive is high. For instance, email spam detection. In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall actually calculates how many of the Actual Positives our model captures through labeling it as Positive (True Positive). Recall is the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. The true positive rate the numbers of instances are relevant the model correctly identified as related data.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Table 1.3 Experimental Results of Metric values for various Text Mining Algorithms

Text Mining Algorithms	Precision	Recall	Accuracy
Naive Bayes Classifier	**81.56**	**80.67**	**80.89**
Decision Tree classifiers	**82.78**	**80.45**	**81.23**
SupportVector Machines	**87.61**	**86.80**	**87.90**

K-Means clustering	**87.75**	**86.89**	**85.90**
Hierarchical algorithms	**87.81**	**87.73**	**88.79**

Accuracy can be a misleading metric for imbalanced data sets.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. The above table 1.3 explains and compared various text documents. Using the above equation find the experimental results of metric values like Precision, Recall, Accuracy values are compared using various text mining Algorithms.

V. CONCLUSION

Text Mining can be defined as a technique which is utilized to extract interesting information or knowledge from text documents which are usually in the unstructured form. Text Mining is discussed with its various techniques which can be utilized such as Classification, Clustering, Summarization, and various techniques and methods discussed for efficient and accurate text mining. In this short survey, compare the notion of text mining techniques has 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. Compare the various metric values of the text mining algorithms for web content mining for texts in different type of document is a major problem ever since text mining tools should be able to work with various document and multilingual documents.

REFERENCES

[1] 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.

[2] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs Up?: Sentiment Classification using Machine Learning Techniques”, in Proc. ACL Conf. Empirical Methods Natural Language Process, pp. 79–86, 2019.

[3] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis”, Found. Trends Inf. Retrieval, Vol. 2, Issue 1, pp. 1–135, 2018.

[4] 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.

[5] 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.

[6] 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

[7] 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.

[8] Heonho Kim, Unil Yun, Yoonji Baek, Jongseong Kim, Bay Vo, Eunchul Yoon, and Hamido Fujita, "Efficient List Based Mining of High Average Utility Patterns with Maximum Average Pruning Strategies", Information Sciences, Vol. 543, 3, pp: 85–105, 2021. DOI: https://doi.org/10.1016/j.ins.2020.07.043,

[9] 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.

[10] Shailesh Pandey, Sandeep Kumar, "Enhanced Artificial Bee Colony Algorithm and its 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] Lior Shabtay, Philippe Fournier-Viger, Rami Yaari, and Itai Dattner, "A Guided FP-Growth Algorithm for Mining Multitude-Targeted Item-Sets and Class Association Rules in Imbalanced Data", Information Sciences, Vol. 553, pp. 353–375, DOI: https://doi.org/10.1016/j.ins.2020.10.020, 2020.

[13] Shashi Raj, Dharavath Ramesh, M. Sreenu, and Krishan Kumar Sethi, "EAFIM: Efficient Apriori-based Frequent Itemset Mining Algorithm on Spark for Big Transactional Data", Knowledge and Information Systems, Vol. 62, Issue 4, DOI: https://doi.org/10.1007/s10115-020-01464-1, 2020.

[14] Unil Yun, Hyoju Nam, Gangin Lee, and Eunchul Yoon, "Efficient Approach for Incremental High Utility Pattern Mining with Indexed List Structure", Future Generation Computer Systems, Vol. 95, pp:221–239, DOI: https://doi.org/10.1016/j.future.2018.12.029, 2019.

[15] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, "Opinion Flow: Visual Analysis of Opinion Diffusion on Social Media", IEEE Trans. Vis. Comput. Graph., Vol. 20, Issue. 12, pp. 1763–1772, Dec. 2019.