CATEGORIZING THE SIGNIFICANCE OF NEWS ARTICLES BY APPLYING THE DECISION STREAM

D.Nithya[1], Dr.S.Sivakumari[2]

[1]Asst. Prof., [2] Prof. & Head,

Dept. of Computer Science and Engineering, School of Engineering,

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

ABSTRACT

In today's era, online news is an emerging channel where the internet users can get news. In day to day life, news websites are flooded with plenty of news articles. Analyzing the huge volume of online news articles is a challenging one, because everyday the online news articles are generated and updated. Big data techniques are used to tackle this problem that process large volume of data within limited run times. An approach based on Evolving Fuzzy Systems (EFS) was used to classify the different news articles into various categories based on the text content of the article. EFS system is used to describe the changes in the content of corresponding articles. The selection of threshold value is not sufficient in EFS system. So, a Decision stream is introduced for merging of similar terms after splitting each iteration. The similarity is estimated using two-sample test statistics that is applied to compare the distribution of labels in each pair of terms. In the news significance validation process, the best Split is applied for relatively small information, where a precise selection of the split is crucial to organize the content similarity. The experimental results show that the proposed web news mining methods achieves better performance in terms of accuracy, precision and recall.

Keywords: Web News Mining, Evolving Fuzzy System, Decision Stream.

1. INTRODUCTION

In an emerging internet world, huge volumes of information are generated in web. The key objective of web mining is to mine useful information from web data. The amount of information in web is enormous and also simply accessible. It is multidisciplinary fields which include data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc. Web news mining is most frequently searched content by the internet usersusing both mobile phones and computers. In mobile internet, the user can read their information in their fragmented time and in computer, the user can read news through more channels [6].

The information from news website can be structure and unstructured text data. So, the feature of text mining is the combination of two words "text" and "mining". It is the process of structuring the input as text, derive the patterns with structured data and finally evaluate the output. There are many techniques to analyse and extract the information from on-line news sources.During last years, there have been many approaches related with

classification, clustering, summarization and categorization of news articles. The news articles can classified based on the categories obtained [3].

From fig.1 structural point of view, news article can be classified into two parts (i.e.) zones and fields. In Zones attribute the news articles can be represented as title, body and attached multimedia. In fields attribute such as the publication date and time, the journalist's name or code, and the source.

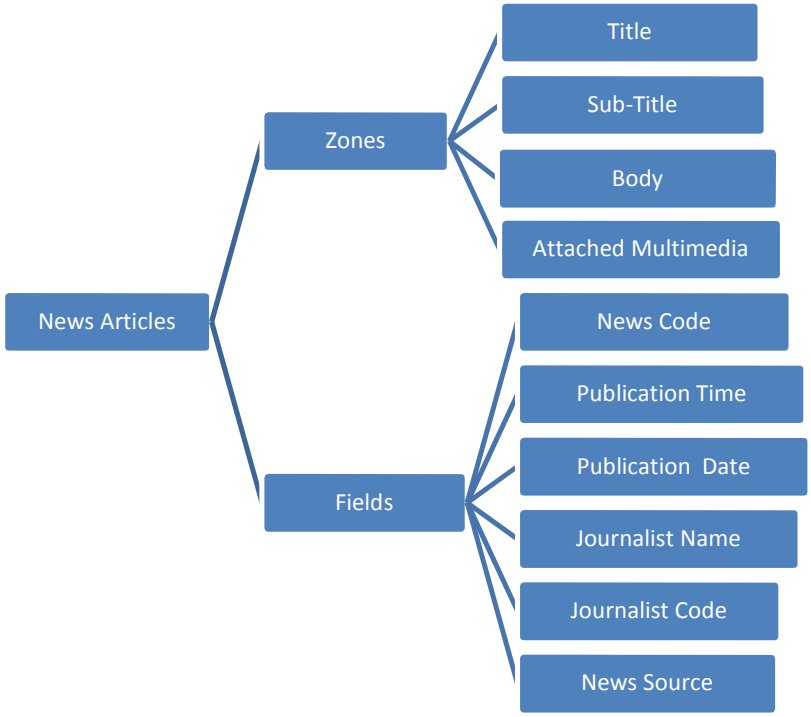

Fig.1. Semi-structured News content from structural point of view

The paper is organized as follows review of related work is discussed in section 2, section 3 discusses about the proposed work and section 4 presents the experimental results and analysis and section 5 deals with conclusion and future work.

2. RELATED WORK

José Antonio Iglesias et al [4] proposed an approach based on Evolving Fuzzy System (EFS) used for web news mining.It will classify the huge amount of web news articles into different categories based on the text content of the articles. Initially a set of terms associated with each document are produced. Then the generated terms are pruned based on term frequency and inverse document frequency value (tf-idf). The generated terms are removed from the dataset based on the threshold value. Then, number of fuzzy rules is generated according to the text content of the news article. Based on the fuzzy rule the news articles are categorized.

Sukhpal Kaur and Er. Mamoon Rashid [5] discussed clustering based K-means and Back Propagation Neural Network for web news mining. For handling large amount of data and information is main problem. To overcome this problem they have used big data like K-means algorithm. In K- Means clustering, web news articles are classified and categorized based on the contents like sports, movies, politics etc. based on the Euclidean distance. Then Back Propagation Neural Network (BPNN) algorithm was used to classify the BBC news. K-means algorithm is used for clustering and BPNN algorithm is used in classification that checks the error rate and accuracy news with less running time.

Roya Hassanian-esfahani et al., discussed about news retrieval and mining. Author compared the four main categories of News Retrieval and Extraction, News Content Analysis, News Propagation Analysis, and News Visualization [7].

Raheja et al [10] discuss about web usage mining based on web log partition which provides less time and then clusters are intent to user searching results in the web logs. The searching time reduces by using this approach instead of the web log approaches. The web usage is the process of recording the user activity while browsing and navigating through web. The web log is the file which is crated based on the user while visit the internet in the web page or the web site.

Malhotra et al [15] explains about efficient technique for single document news article summarization. It helps to know the most important information in a short period of time. The proposed technique is query based approach, so that results on user query are filtered which are used for constructing the keyword table for the new article. The main aim of this technique is to cover all important information from the document.

Krishnalal et al [16] developed and experimented intelligent system for online news classification based on the Hidden markov model (HMM) and support vector machine (SVM). Intelligent system is used to extract the keyword from the news paper through online content and classify based on the pre defined categories. The different phases based on the news categories such as sports, finance and politics are classified the content of the online news paper such as the Pre-processing for the text, HMM based feature extraction and the classification using the SVM.

3. PROPOSED METHODOLOGY

3.1 WEB NEWS DATASET

The web news dataset from web hose are taken for the experimental analysis. The dataset contains are categorized into different topics like Study, Education, Economics Cinema, Business Trading, Health, Science, Sports, Technology and Travel, Film Media, War Military defence, Events, Political. Table 1.1 depicts the dataset layout.

Table 1.1 Dataset Description

Data Set Characteristics:	Text
Attribute Characteristics:	Categorical
Associated Tasks:	Classification
Number of Instances:	5000
Number of Attributes:	6

The columns included in this dataset are:

- ID : the numeric ID of the article
- TITLE : the headline of the article
- URL : the URL of the article
- PUBLISHER : the publisher of the article
- STORY : alphanumeric ID of the news story that the article discusses
- HOSTNAME : hostname where the article was posted

Fig. 2-Sample Web News Dataset without Pre-processing

3.2 PRE-PROCESSING

In pre-processing process the special characters like @, # and URL's tags that are removed. After removing the special characters then stop words are identified and removed. Once pre-processing is done filtering is implemented as feature extraction or called as tokenization. Tokenization is the act of breaking sequence of strings into pieces like keywords, symbols and phrases. In this method it will remove the punctuation marks. The figure 3 depicts the news dataset after pre-processing.

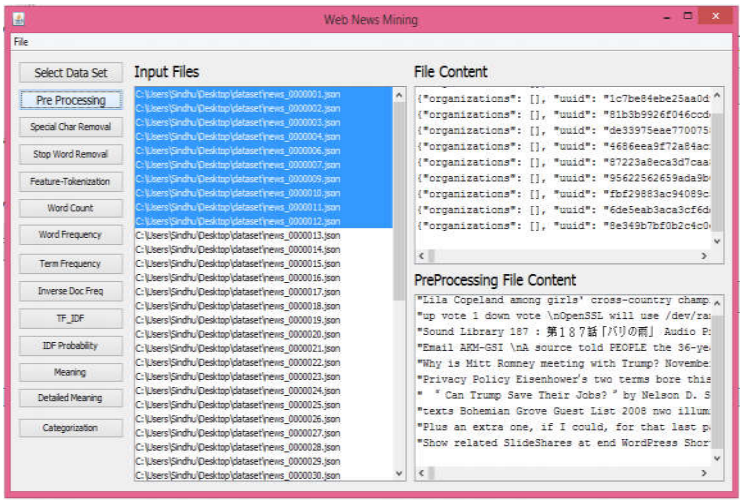

Fig.3After Pre-processing News Dataset

3.3 TERM GENERATION

In term generation module, the most relevant terms of each articles are obtained. This task has been done in any open-source tool used in web mining. This tool is used for executing the following steps:

1. Tokenization: This step breaks a text into phrases, words, symbols, or other meaningful elements called tokens. The result of tokenization will give meaningful keywords. At the same time, it will remove the meaningless strings such as symbols and spaces.
2. Stop word Elimination: Stop word is a list of meaningless words that is stored in the database. In general, stop words consists of a list of prepositions, articles, and pronouns. The result of this step is extracted meaningful words.

3. Stemming: Stemming is also known as base or root. It will identify the meaning of different words with same meaning by using stem form.

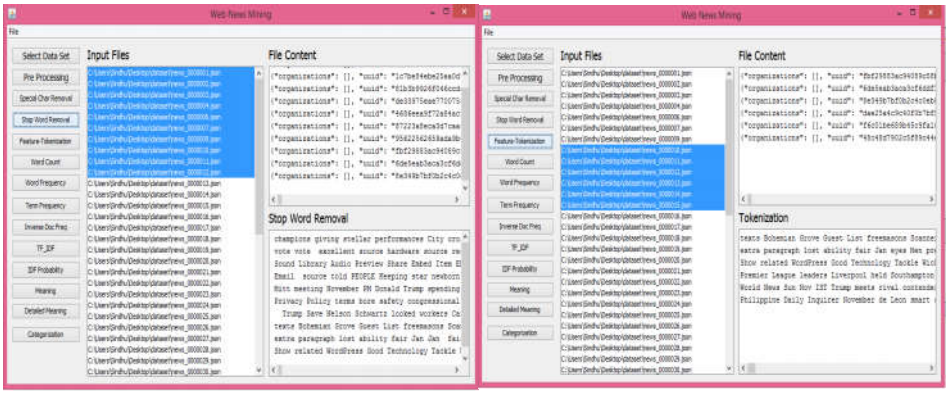

Fig.4Removal of Tokenization and Stop word Elimination

3.4 FEATURE EXTRACTION

In the feature extraction module, it will transform the input data into set of features. After identifying the features from collected news articles, calculate word count, word frequency, term frequency (tf) and inverse document frequency (idf). The word count is the number of occurrence words in a document or passage of text. Word frequency is the number of word occurred in the set. Then the sum tf and idf of a particular term is lower than a pruning threshold then it is removed from the dataset.

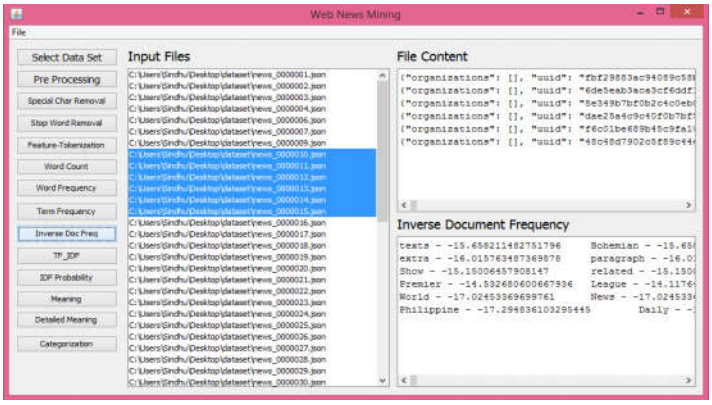

Fig.5 Probability of Term frequency and Inverse document frequency

3.5NEWS CLASSIFICATION

Classification is a general process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood.A classification is an ordered set of related categories used to group data according to its similarities.Web news classification techniques use diverse information to classify target web news: the content of the web news, web news URL and structure information on web news. Maximum news contents are organized as

(i) Home page that shows some headlines of all sections.
(ii) Numerous unit of pages that offer the headlines of diverse extents of interest like business, sports, entertainment, technology, politics etc. these different areas also contains some sub sections like national, international, market, cricket, football, science etc.
(iii) Pages that actually represent the news containing the title, author, related news link, date and body of the news.

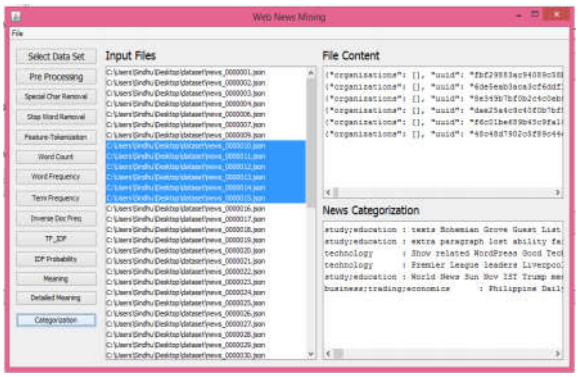

Fig 6 Categorization of News

3.6 SIGNIFICANCE OF THE NEWS USING DECISION STREAM

After classifying the news category, the tree interpretation is applied based on the terms identified from the relevant feature selection. The decision stream mainly process with the merging of similar terms after splitting each iteration. The similarity is estimated using two-sample test statistics that is applied to compare the distribution of labels in each pair of terms. In the news significance validation process, the best Split is applied for relatively small information, where a precise selection of the split is crucial to organize the content similarity.The news classification and organization system is improved by using the decision stream algorithm which identifies the potential of the news has to be published. The model is collaborated with the EFS system which determines the distinct term with the unique meaning to identify possible combination of the news category. After identifying the news category the information retrieval system is assigns the relevance value to the news which describes the importance of the news.

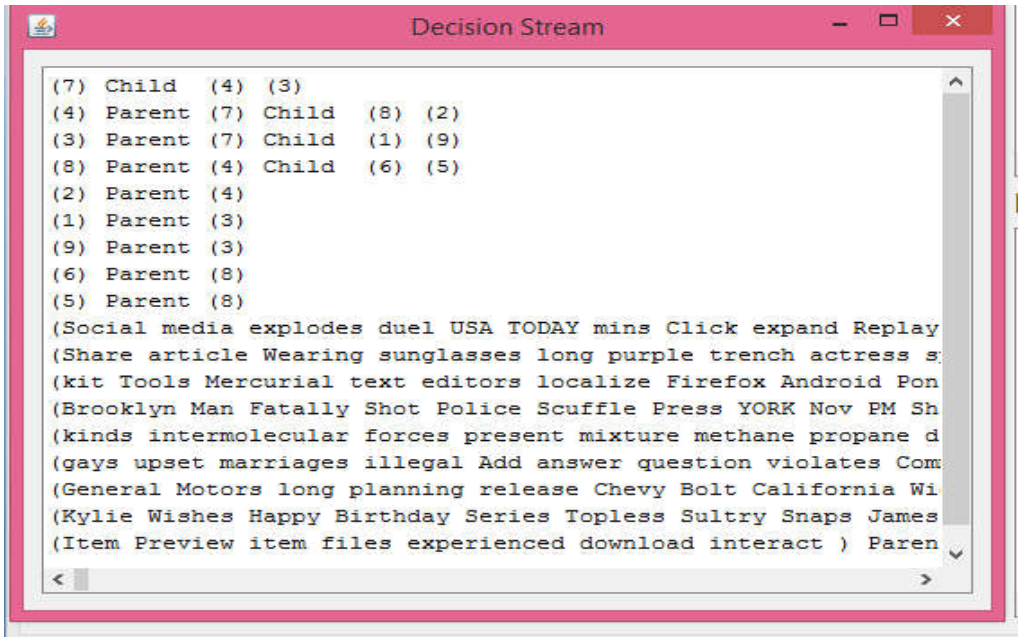

Fig 7Representation of decision stream

4. EXPERIMENTAL RESULTS

In this section, the results of the existing and proposed web news mining methods are analyzed in terms of accuracy, precision and recall. For the experimental purpose, five

different sets of data are created which combines two or more categories. The five datasets are Health vs. Science (H-Sc), Science vs. Technology (Sc-Te), Health vs. Science vs. Sports (H-Sc-Sp), Business vs. Health vs. Science vs. Sports (B-He-Sc-Sp), and Arts vs. Business vs. Health vs. Science vs. Sports vs. Travels (A-B-H-Sc-Sp-Tr).

4.1 ACCURACY

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Accuracy is calculated as follows:

ACC = (TP+TN)/(P+N)

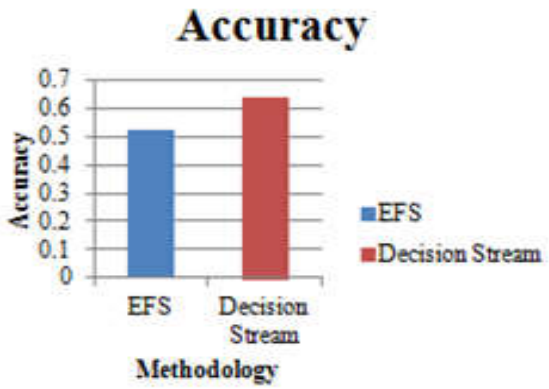

Fig 8. Comparison of Accuracy between EFS and Decision Stream

Fig8, shows the comparison of accuracy between Evolving Fuzzy System (EFS) and proposed Decision Stream. X axis represents the different datasets and Y axis denotes the accuracy in terms of %. From the Fig8 it is proved that the proposed EFS-Decision Stream has high accuracy than the existing EFS.

4.2 PRECISION

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

PPV = TP / (TP+FP)

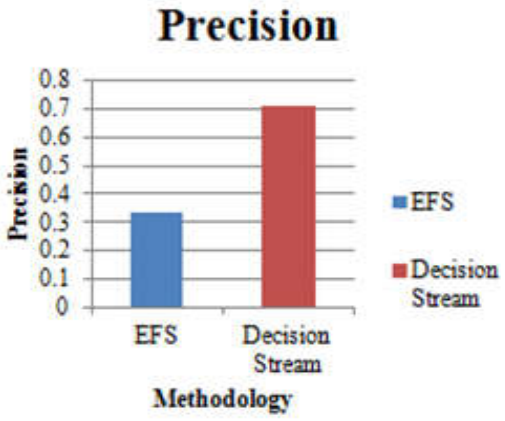

Fig 9. Comparison of Precision between EFS and Decision Stream

Fig9, shows the comparison of precision between Evolving Fuzzy System (EFS) and proposed EFS with Decision Stream (EFS-Decision Stream). X axis represents the different

datasets and Y axis denotes the precision. From the Fig9 it is proved that the proposed EFS-Decision Streamhas high precision than the existing EFS.

4.3 RECALL

The recall value is evaluated according to the classification of web news articles at true positive prediction and false negative prediction.

RECALL = TP / TP+TN)

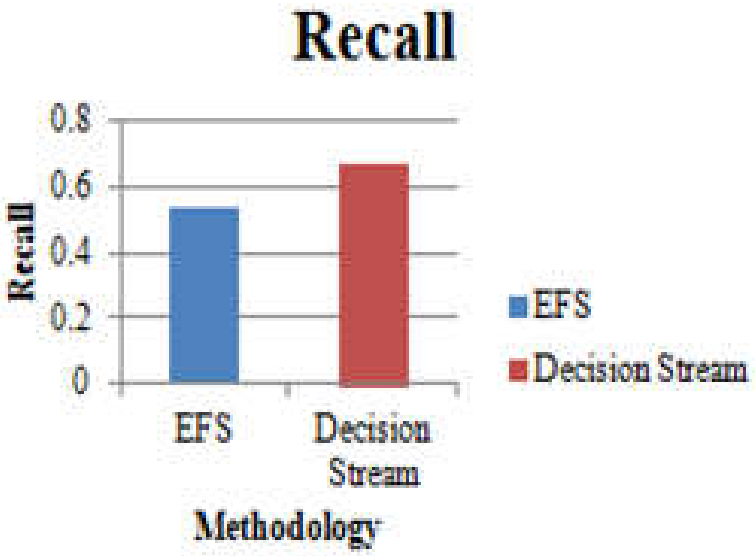

Fig 10. Comparison of Recall between EFS and Decision Stream

Fig10, shows the comparison of recall between Evolving Fuzzy System (EFS) and proposed EFS with Decision Stream (EFS-Decision Stream). X axis represents the different datasets and Y axis denotes the recall. From the Fig10 it is proved that the proposed EFS-has high recall than the existing EFS.

5. CONCLUSION AND FUTURE WORK

In order to get the relevance of the different terms of the web news by using the term frequency and inverse document frequency. This classifier is based on Evolving Fuzzy Systems (EFS) and the model that describes a specific topic area changes according to the change in the text content of their articles. Since news websites are daily overwhelmed with plenty of news articles, one of the main advantages of the proposed approach is that it can cope with huge amounts of news in real-time. The web news processing system has to be designed to classify and categorize the news article based on dimensional model as collects news articles, identify the keywords concerning influence factors and term frequency with inverse document frequency. The significance of the news article has to be identified by applying the relevancy validation using the decision stream model. Also, the proposed classifier is one pass, non-iterative, recursive and it can be used in an interactive mode. Thus, this method can cope with huge amounts of news and process them quickly. Although the amount of terms from the articles is huge, the most important terms can be extracted and there is no need to store all the news in memory.

In future, this work can be applied for trend and sentiment analysis with analysis of the web link in social media and other related aspects. The context modelling can be used for the web news recommendation based on the behaviour analysis of the web user and the requirements.

REFERENCES

[1]. Xindong Wu, Gong-Qing Wu, Fei Xie, Zhu Zhu, and Xue-Gang Hu, 2010. News filtering and summarization on the web. IEEE Intelligent Systems, Vol. 25, No. 5, pp. 68-76.

[2]. Lian'en Huang and Xiaoming Li, 2010. HisTrace: A system for mining on news-related articles instead of web pages. Web Society (SWS), 2010 IEEE 2nd Symposium, pp. 30-37.

[3]. D.Nithya, Dr.S.Sivakumari. (2017). State of the Art of Web News Mining. International Journal of Computer Engineering and Applications. 9(8), 122-129.

[4]. Jos´e Antonio Iglesiasa, Alexandra Tiembloa, Agapito Ledezmaa and Araceli Sanchis, 2016. Web news mining in an evolving framework. Information Fusion, Vol. 28, pp. 90-98.

[5]. Sukhpal Kaur and Er. Mamoon Rashid, 2016. Web news mining using back propagation neural network and clustering using K-Means algorithm in Big data. Indian Journal of Science and Technology, Vol. 9, No. 41, pp.1-8.

[6]. D.Nithya, Dr.S.Sivakumari. (2017). A Study on Web Mining Tools. International Journal of Research in Electronics and Computer Engineering. 5(2). 135-137.

[7]. Roya Hassanian-esfahani and Mohammad-javad Kargar. (2016). A Survey on Web News Retrieval and Mining. IEEE Second International Conference on Web Research (ICWR).90-110.

[8]. Huang, W., & Li, Y. (2012). Bell-Shaped Probabilistic Fuzzy Set For Uncertainties Modeling. *Journal of Theoretical & Applied Information Technology*, *46*(2), 875-882.

[9]. Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., & Kompatsiaris, I. (2014, November). News articles classification using Random Forests and weighted multimodal features. In *Information Retrieval Facility Conference* (pp. 63-75). Springer, Cham.

[10]. Raheja, N., & Katiyar, V. K. (2014). Efficient web data Extraction using Clustering approach in Web usage mining. IJCSI, ISSN, 1694-0814.

[11]. Longe, H. O. D. (2014). A Text Classifier Model for Categorizing Feed Contents Consumed by a Web Aggregator. *International Journal of Advanced Computer Science and Applications (IJACSA),* 5(9), 95-100.

[12]. Sharma, N., & Kaur, P. (2015). Categorize Online news Using Various Classification Techniques. *International Journal of Advanced Research in Computer Science & Technology (IJARCET),* 4 (2), 337-340.

[13]. Roya Hassanian-esfahani and Mohammad-javad Kargar. (2016). A Survey on Web News Retrieval and Mining. IEEE Second International Conference on Web Research (ICWR).90-110.

[14]. Weixin Li, Jungseock Joo, Hang Qi, and Song-Chun Zhu, 2017. Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph. IEEE Transactions on Multimedia, Vol. 19, No. 2, pp. 367-381.

[15]. Malhotra, S., & Dixit, A. (2013). An effective approach for news article summarization. International Journal of Computer Applications, 76(16).

[16]. Krishnalal, G., Rengarajan, S. B., & Srinivasagan, K. G. (2010). A new text mining approach based on HMM-SVM for web news classification. International Journal of Computer Applications, 1(19), 98-104.

[17]. Gheraibia, Y., & Moussaoui, A. (2013). Penguins search optimization algorithm (PeSOA). In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Berlin, Heidelberg. 222-231.

[18]. Maria Carla Calzarossa and Daniele Tessera. Modeling and predicting temporal patterns of web content changes. *Journal of Network and Computer Applications* 56 (2015): 115-123.