ISSN 2349-4506
Impact Factor: 2.545

Global Journal of Engineering Science and Research Management

TEXT MINING: TEXT SIMILARITY MEASURE FOR NEWS ARTICLES BASED ON STRING BASED APPROACH

R. Kohila*, Dr. K. Arunesh
* M. Phil Scholar, Department of Computer Science, Sri.SRNM College, Sattur, Virudhunagar Dist
Department of Computer Science, Sri. SRNM College, Sattur, Virudhunagar Dist

DOI: 10.5281/zenodo.57373

KEYWORDS: Text Mining, Cosine similarity, News Article Dataset.

ABSTRACT

Now-a-days, the documents similarity measuring plays an important role in text related researches. There are many applications in document similarity measures such as plagiarism detection, document clustering, automatic essay scoring, information retrieval and machine translation. String Based Similarity, Knowledge Based Similarity and Corpus Based Similarity are the three major approaches proposed by the most of the researchers to solve the problems in document similarity. In this paper, the String Based Similarity measure Term Based algorithm Cosine Similarity is used to measuring the similarity between the documents. The nouns in the documents are extracted and context word synset are also extracted using WordNet. The bigram dataset is created based on Context words. In this proposed method the similarity measure between the documents is measured using cosine similarity algorithm. Preprocessing dataset, context word dataset and bigram dataset are used to measure the similarity. The context word document set measure gives a better similarity than bigram and pre-processing document set.

INTRODUCTION

Today, WWW has turned to be the largest information source available in the world. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information and there are an enormous amount of documents in different areas like literature, technology, science, medicine etc., many scholars abused these documents it enforced plagiarism. As the result of this issue, measuring the closeness of two documents is essential and it is very important to detect the plagiarism of a wide range of different documents. Measuring the similarity between the documents are that have wide range of purposes such as machine translation, checking plagiarism, information retrieval, document clustering, classifying text and automatic essay scoring.

Comparing the similarity between the documents is not a new field. There are many researchers have been proposed a lots of different algorithms. The techniques are string based, knowledge based and corpus based similarities.

String Based method defines the similarity by working on Term classification and Character Classification. Jarco Winkler, Needleman-Wunsch, N-gram, Longest Common substring and Smith Waterman are character based similarity algorithms. The term based similarity algorithms are Dice's coefficient, cosine similarity, jaccard similarity, Block distance, overlap coefficient and matching coefficient.

Corpus Based Method defines the similarity between words giving to information extended from large corpora. The algorithms are the cross-language explicit semantic analysis (CL-ESA), Latent Semantic Analysis (LSA), Explicit Semantic Analysis (ESA) and Generalized Latent Semantic Analysis (GLSA), [1].

Knowledge based method depends on classifying the degree of closeness between words utilizing information derivative from semantic systems.

LITERATURE REVIEW

Evaluating the semantic and string based similarities of texts and documents have been widely studied. Many scientific researchers have struggled for a long time to find a criterion for semantic similarities between two words

or short texts and also between two documents. Madylova, A [2] proposed a new technique to calculate the semantic similarities between the documents. This calculation is based on the cosine similarities. The documents are converted into content vectors which are based on calculating the cosine similarity. This documents including IS-A relations for classification. The time taken for to calculate the semantic similarity measure between a pair of documents is the drawback of this procedure. Mihalcea, R et al., [3] proposed a time consuming semantic similarity measure algorithm mihalcea to overcome the issues in [1]. Two short-texts are considered for calculating semantic similarity by wordnet. The short texts are referred as knowledge based and literature based measurement.

Strehl et al., [4] used the YACHOO datasets for similarity measures and clustering. This dataset was already categorized by manually. Euclidean, Pearson correlation, cosine, extended jaccard and the clustering algorithms like hyper group partitioning, generalized k-means, weighted graph, self-organizing feature map are considered by the authors to measure the similarity between the documents. Among those algorithms they found that extended jaccard and cosine similarity performance is closely to human manual work result.

Huang, A., [5] presents Similarity Measure for Text Document Clustering. For categorizing the documents similarity measurements are used and they are tested different datasets using the following similarity measurement algorithms are Cosine similarity, Euclidean distance, relative entropy and jaccard coefficient. Among the performance of these algorithms, authors concluded Euclidean distance did not suitable for their dataset when compared with other algorithms.

Masumeh Islami Nasab et al., [6] proposed method to calculate the semantic similarity between the articles. The similarities are calculated by the following points,
1. Article texts are separated into three parts as title, abstract and keywords.
2. Based on the contribution to the article weighting title, abstract, keywords.
3. Based on the title, abstract and keyword the weighted mean calculated.

To find the similarity between human and system scores using with the Pearson's correlation formula. In this proposed method have 87% accuracy.

Vikas Thada et al., [7] authors used the cosine similarity, dice coefficient, jaccard similarity algorithms. They conducted their experiment using first 10 pages GOOGLE search result. The number of keywords used to conducting several experiments. Finally concluded the cosine similarity was best fitness for this dataset compare with others.

Gang Qian et al., [8] proposed Euclidean distance measure and cosine angle distance measure to compare feature vectors of images and document retrieval. Euclidean distance measure was used for the purpose of comparing feature vectors of image like context of colour image database. Cosine angle distance measure was used for the purpose of document retrieval. From this experiments both the two metrics are given similar performance. They had done the same experiments using with Manhattan (L1) distance measure. The Manhattan (L1) result shows that differs from cosine and Euclidean distance.

Manan Mohan Goyal et al., [9] proposed method used to compare the cosine and fuzzy similarity using the k-means algorithm. Turned out time is calculated in fuzzy and cosine similarity measures. The fuzzy similarity measure has taken less time to calculate similarity measure.

Neha Agarwal et al., [10] proposed to find the similarities between two documents using the cosine similarity and jaccard coefficient. To generate the cluster, time is required where using the cosine similarity; it takes less time as compare to jaccard. In cosine similarity is the mathematical calculation used to measuring document similarity. Jaccard coefficient considered all the terms of one document to another to calculate the similarity which is taken large amount time to complete the process. So, cosine similarity gives more accurate result as compare to jaccard coefficient.

Ewees .A et al.,[11] "Comparison of Cosine Similarity and k-NN for "Automated Essays Scoring" the system analyzed the performance comparison of cosine similarity and k Nearest Neighbors algorithm with Latent Semantic Analysis. The following preprocessed steps are used such as entered text, unifying the form of letters, deleting the formatting, replacing synonyms, stemming and deleting "stop words". From this experiment cosine similarity with Latent Semantic Analysis give more accurate result as compare K Nearest Neighbors with Latent Semantic Analysis.

PROPOSED METHODOLOGY

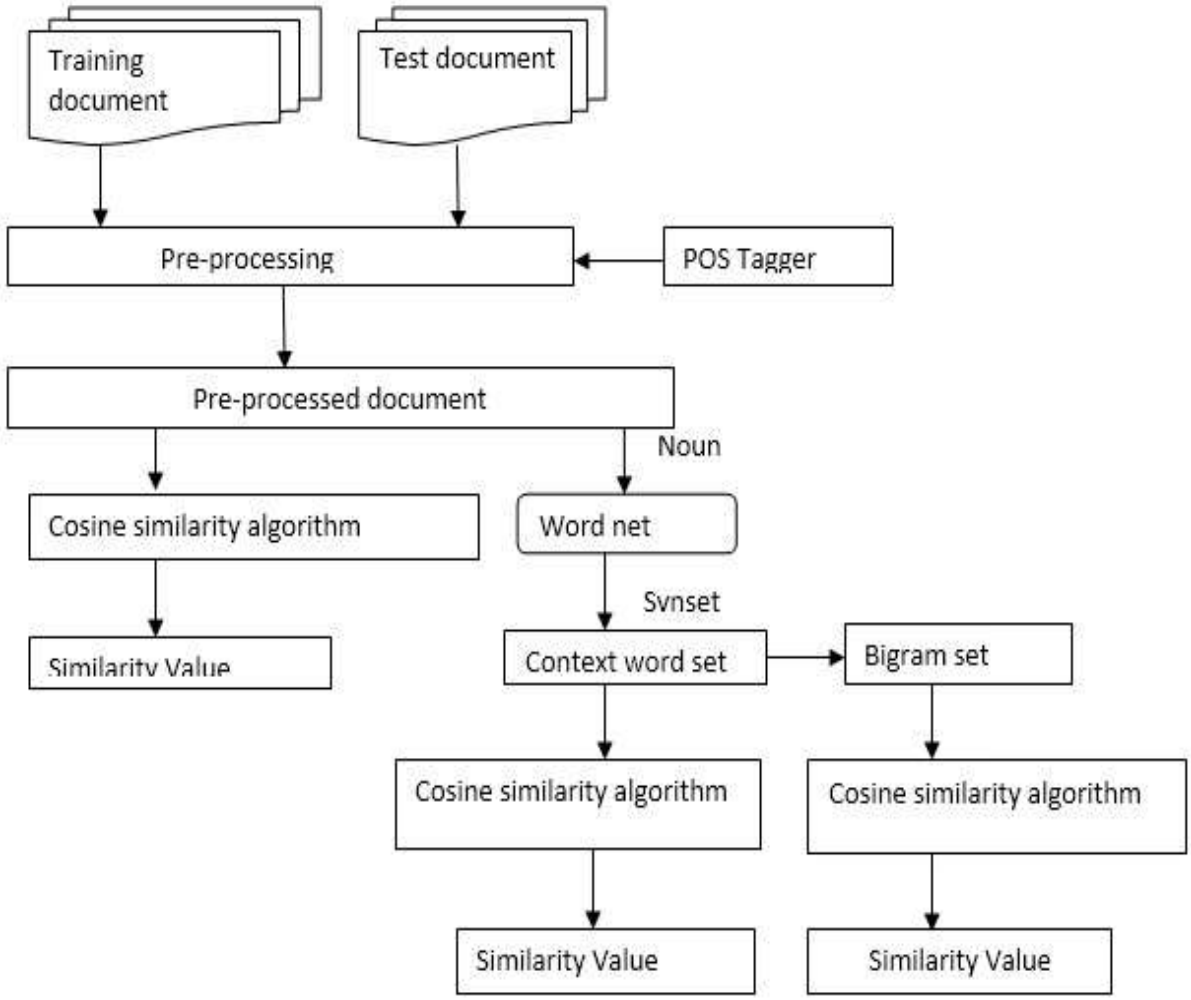

Figure1: System Architecture for document similarity

The similarity between the documents will be measured using cosine similarity algorithm. The system architecture for to find the significance between the documents are shown in Figure 1. The dataset is in .txt extension. Training and test documents are initially considered for pre-processing. The documents are pre-processed using POS tagger. The pre-processed document will be considered for to find similarity measure using Cosine Similarity algorithm. The highest similarity measure is considered to compare the results of other modules.

In the Second module, the nouns are extracted from the document to create a context word. These synonyms (synset) are extracted using WordNet. The document similarity is calculated and the similarity between the documents will be predicted. Third, a bigram document set is created based on the context word set to calculate the similarity value. The highest similarity value is considered for decision making.

ISSN 2349-4506
Impact Factor: 2.545

Global Journal of Engineering Science and Research Management

EXPERIMENTAL RESULT

The primary data news article was collected from the web. The websites are YOUGOTTHENEWS.COM and NATUERALSNEWS .COM. 80 articles have been collected from the website. The news article consists of cancer, crime, health and terror related information. In YOUGOTTHENEWS.COM Website, in the New York Times option, from the India news cancer, crime, health and terror related articles are collected. Cancer and health related articles have been collected from NATUERALSNEWS .COM. All the documents are in .txt format.

In this paper, three types of document sets are generated such as pre-processing documents, context word documents and bigram documents.

To compare the similarity value of pre-processing documents, context word documents and bigram document sets, Cosine Similarity algorithm is used. In Cosine Similarity algorithm first, calculating a vector space model by tf-idf then document similarity is measured.

The each training documents are converted based on the above steps and the tf-idf is calculated. The same techniques are also used to calculate the test documents. The new test document is used calculate the similarity value between the training documents (already created). The similarity value is calculated with all the training documents. From that similarity measures the highest similarity measure document is considered as resultant document.

i) After pre-processing, the similarity values and the context words for testing document are taken out.
ii) The similarity value is calculated from the test context word document and training set. The similarity value is calculates.
iii) Create the bigram test document and find the similarity value.

RESULT AND DISCUSSION

In this research work, the pre-processed cancer related test articles are classified as 60% accuracy, the crime related articles are classified as 80% accuracy, the health related articles are classified as 40% accuracy and terror related articles are exactly classified as 80%.

The context word cancer related test articles are classified as 100%, the crime, health and terror related articles are classified in 80%, 40% and 80% respectively.

The bigram set cancer related test articles are classified as 100%, the crime, health and terror related articles are classified in 80%, 40% and 80% respectively.

From this research work we observe that cancer related articles of context word documents are correctly match with the training set documents. But crime, health and terror related articles are not exactly matching with the related training documents.

For example, a new test article of crime that can be a possibility of matching with terror and vice versa and also health related new test article can be matched with cancer related documents.

The similarity values in the context word documents are higher than the pre-processed documents similarity values. In bigram documents the similarity values are almost near is context word documents. The difference is at most 0.001 to 0.003.

The Figure 2 to 5 shows the result of pre-processing documents. The Figure 6 to 9 shows result of context word documents. In this result the x axis represents as training document and y axis represents as the calculate similarity value.

ISSN 2349-4506
Impact Factor: 2.545
Global Journal of Engineering Science and Research Management

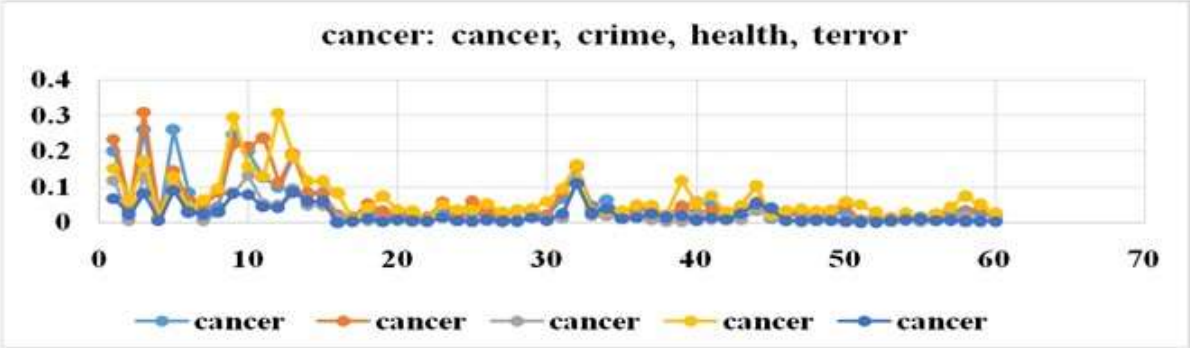

Figure 2 Similarity Value for Cancer Article

The results for similarity value between the pre-processing document of training sets and cancer test sets as shown in Figure 2, based on the cosine similarity method. It is observed that cancer article incorrectly classified into health related articles.

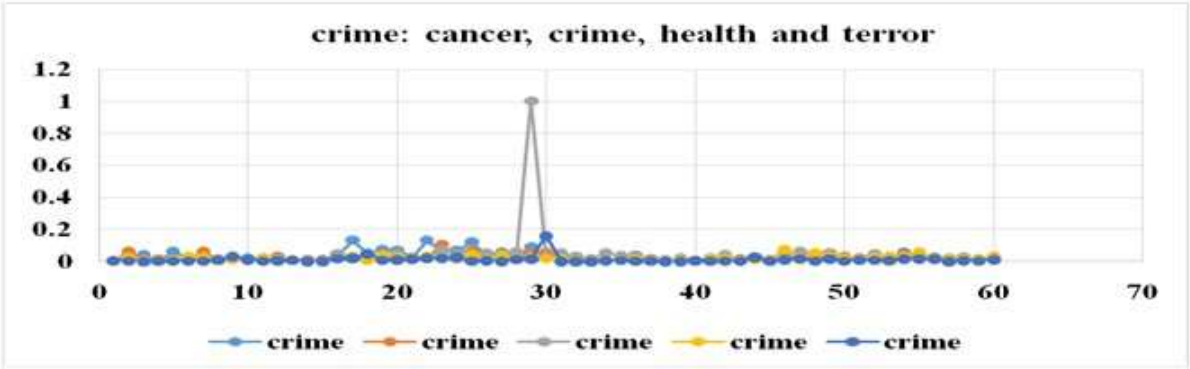

Figure 3 Similarity Value for Crime Article

The results for similarities between the pre-processing documents of training sets and crime test sets as shows in Figure 3, based on the cosine similarity method. The crime article is incorrectly classified into terror related article.

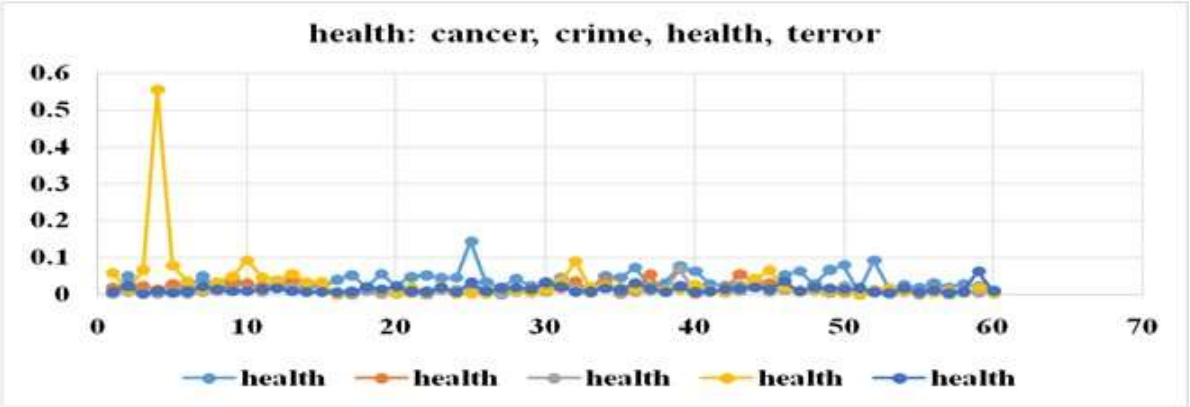

Figure 4 Similarity Value for Health Article

The results for similarities between the pre-processing documents of training sets and crime test sets as shows in Figure 4, based on the cosine similarity method. We observed that he three articles are incorrectly classified.

ISSN 2349-4506
Impact Factor: 2.545

Global Journal of Engineering Science and Research Management

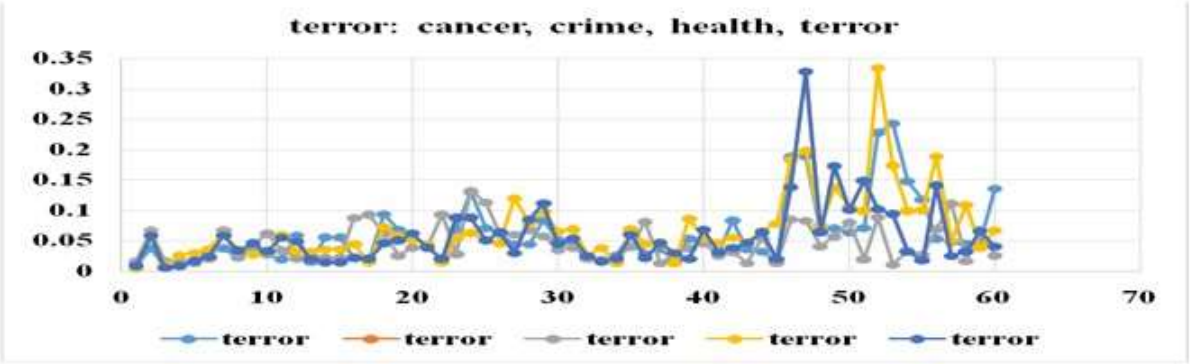

Figure 5 Similarity Value for Terror Article

The results for similarities between the pre-processing documents of training sets and terror test sets as shows in Figure 5, based on the cosine similarity method. Here only one article is incorrectly classified into crime related article. Actually new test terror article is compare with training document.

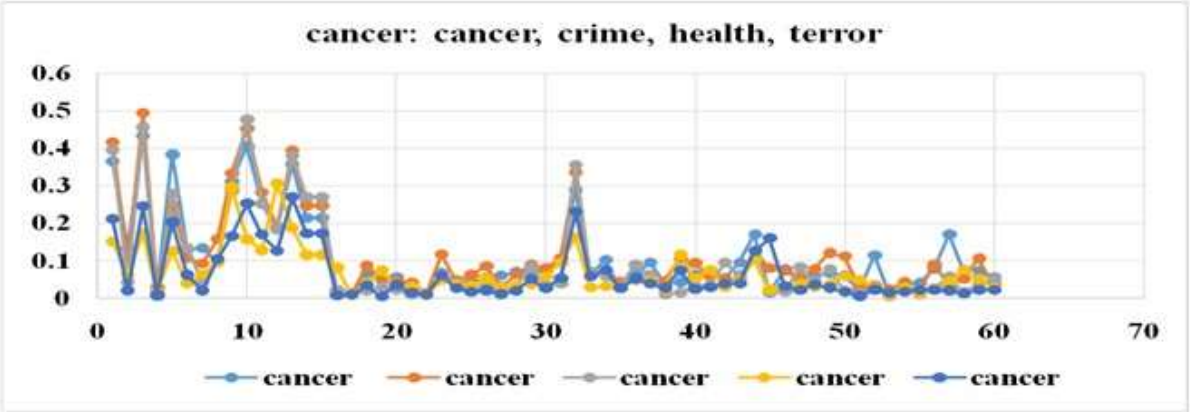

Figure 6 Similarity Value for Context Word Cancer Article

The results for similarities between context word of training sets and cancer test sets applying proposed method based on the cosine similarity method are shows in Figure 6. The all test articles are correctly classified in context word documents and it gives a high similarity value when compare with pre-processing documents.

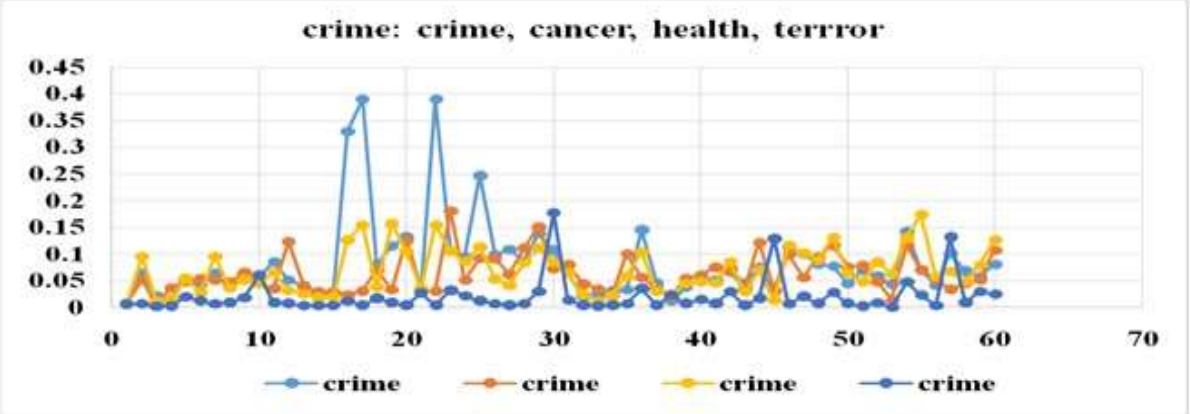

Figure 7 Similarity Value for Context Word Crime Article

ISSN 2349-4506
Impact Factor: 2.545
Global Journal of Engineering Science and Research Management

The results for similarities between the context word of training sets and crime test sets applying proposed method based on the cosine similarity method are shows in Figure 7. The similarity value is high when compared with Figure 3.

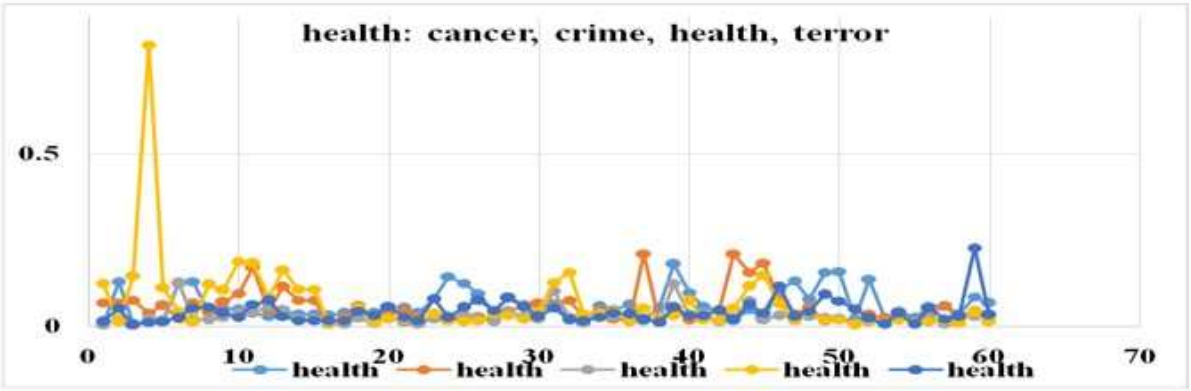

Figure 8 Similarity Value for Context Word Health Article

The results for similarities between context word of training sets and health test sets as shows in Figure 8, when applying proposed method based on the cosine similarity method. Here we observed that the 2 articles are incorrectly classified into cancer and terror.

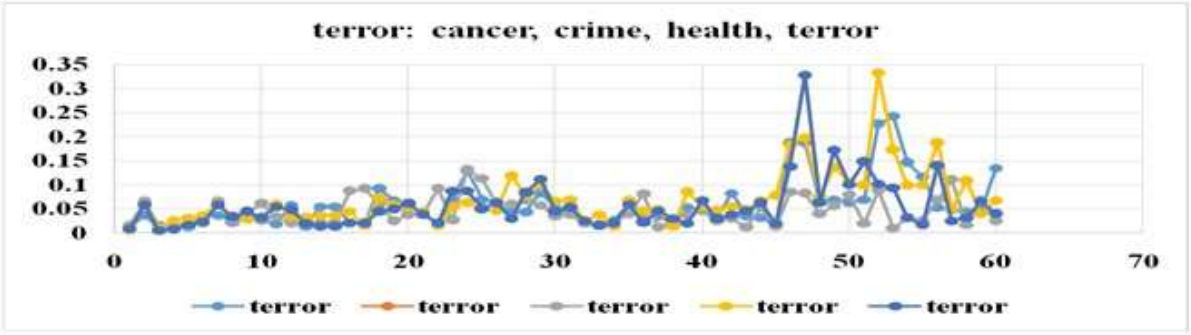

Figure 9 Similarity Value for Context Word Terror Article

The results for similarities between context word of training sets and terror test sets as shows in Figure 4.9, when applying proposed method based on the cosine similarity method. We observed the result only one article is incorrectly classified into crime.

CONCLUSION

In this paper, we introduced a method for context word document using Cosine Similarity algorithm to find the similarity value between the news articles. The present work uses three important pre-processing techniques namely, Stop Word Removal, Extracting the Noun and TF-IDF on news article dataset.

From the experimental evaluation and analysis some important observation has been made. We analyze the use of WordNet feature, synonyms, the context words are created based on WordNet synonyms. The nouns in the text are considered to create the context words. From the context words the bigram document sets are created.The similarity values in the Context Word documents are higher than the pre-processed document similarity values. In Bigram documents the similarity values are almost nearer in context word documents. The difference is at most 0.001 to 0.003.From this research work we observed that cancer related articles of context words document are matched with the training set documents. But crime, health and terror articles are not exactly matching with the related training documents.

There are two main directions for the future work in this research, The first direction is to implement the other WordNet's features and functionalities such as hypernyms, hyponyms, meronyms holonyms and troponym.

The second direction for the feature work is to establish greedy based disambiguation mechanism to calculate the score of synsets by measuring the similarity between documents. Moreover, more datasets can be used to show success of the proposed methodologies.

REFERENCES

1. Wael H. Gomaa, Aly A. Fahmy, "A Survey of Text Similarity Approaches," International Journal of Computer Applications, pp. 13-18, 2013.
2. Madylova, A., "A Taxonomy based Semantic Similarity Documents Using Cosine Measure", Computer an Information Sciences, IEEE, Iscis 2009. 24th, International Symposium. (2009).
3. Mihalcea, R., Corley, C, Strapparava, C., "Corpus-based and Knowledge-based Measures of Text Semantic Similarity", Proceeding of th National Conference on Artificial Intelligence ,pages:775-780. (2006).
4. Strehl etal., "Impact of similarity measures on web-page clustering". In AAAI-2000: Workshop on Artificial Intelligence for Web Search, July 2000.
5. Huang, A., "Similarity Measure for Text Document Clustering", Department of Computer Science The University of Waikato, Hamilton, New Zealand, pp:49-56, (2008).
6. Masumeh Islami Nasab etal., "A New Approach For Finding Semantic Similar Scientific Articles". Journal of Advanced Computer Science and Technology (JACST) , 4 (1) (2015) page no: 563 - 59.
7. Vikas Thada etal., "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm". International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013, SSN: 2319-1058 pg: 202-205.
8. Gang Qian etal., " A Comparative Analysis of Two Distance Measures In Color Image Database".
9. Manan Mohan Goyal etal., "Comparison Clustering using Cosine and Fuzzy set based Similarity Measures of Text Documents" publication: 275836208, may 2015.
10. Neha Agarwal etal., "Comparative Analysis of Jaccard Coefficient And Cosine Similarity For Web Document Similarity Measure". International Journal For Advanced Research In Engineering And Technology, Volume 2, Issuse X, October 2014, ISSN 2320-6802.
11. Ewees .A etal., "Comparison of Cosine Similarity and k-NN For Automated Essays Scoring". International Journal of Advanced Research in Computer and Communication Engineering ISSN (Online) : 2278-1021, ISSN (Print) : 2319-5940, Vol. 3, Issue 12, December 2014.