Distance Measures based Approach for Hate Speech Spreaders Detection

Archana Gelli[1], Karunakar Kavuri[2], T Raghunadha Reddy[3], Lakshmi Narayana M[4]

[1]Assistant Professor, [2,3,4]Associate Professor

[1]Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Narasapur, AP

[2]Computer Science and Engineering, Swarnandhra Institute of Engineering and Technology, Narasapur, AP

[3]Computer Science and Engineering, Matrusri Engineering College, Hyderabad, Telangana

[4]Information Technology, SRKR Engineering College, Bhimavaram, AP.

[1]ksj.archana@gmail.com

[2]karunakar.mtech@gmail.com

[3]raghu.sas@gmail.com

[4]lachi9866516918@gmail.com

***Abstract*—Due to the invention of different internet technologies in social media platforms like Review sites, LinkedIn, Facebook, Instagram, Twitter, people are shifting their communication method with other people. These platforms give permission to the people to freely exchange their knowledge, opinions to individual or group of people. Most of the people used these platforms for welfare of others, but some of them misused these environments by spreading hateful texts about a person, product or any other entity. Hate speech is a direct or indirect statement targeted towards a person or group of a people intended to degrade another on the basis of ethnicity, religion, disability, gender or sexual orientation. Identification of people or account that spreads hate speech becomes a challenging task to the research community. PAN organizers conducted competition on hate speech spreaders detection in 2021 by providing dataset. In this work, we proposed a distance measures based approach for hate speech spreaders detection. In this proposed approach, we used six different kinds of distance measures for finding the similarity among the documents. We used PAN 2021competition dataset of hate speech spreaders detection for experimentation. The training dataset is cleaned by using different pre-processing techniques. Identify most relevant terms by using the feature selection algorithm. The training document vectors are represented with identified terms. The term value is represented in vectors by using the value of a term weight measure. The test document vectors class is decided based on the similarity with the training documents. The class label is assigned based on which class contains more similar documents to test document. The proposed approach attained good accuracy for hate speech spreaders detection when compared with various approaches of hate speech spreaders detection.**

***Key Words*—Hate Speech, Hate Speech Spreaders Detection, Distance Measures, Feature Selection Algorithm**

I. INTRODUCTION

The online medium is an environment that allows people to easily communicate and freely express themselves. The rise of online social networks creates an environment to increase the user-generated content on the internet. Even though most of the generated content is respectful, social platforms also constitute a place where people can openly publish and share offensive, discriminatory messages in the form of hate speech [1]. Hate speech is not a new phenomenon but it has become more and more of a problem in recent years and has consequently attracted a lot of attention in the research community making hate speech detection a very active research field. In particular the growing impact of social media on the way people share and access information has demonstrated the need to tackle the problem systematically as issues such as cyber-bullying and other hurtful and anti-social behaviours have become a growing cancer that needs to be tackled broadly across many different platforms and applications.

Hate speech is defined as speech that attacks a person or a group of people based on attributes such as religion, race, ethnic origin, sex, national origin, disability, gender identity or sexual orientation [2]. From the mentioned categories, online discrimination is most prevalent for race, sexual orientation and ethnicity. However, other groups are targeted based on behaviour, physical aspects, class and disabilities [3]. The dynamics of online hate speech is influenced by real life events which can represent triggers for discrimination against a specific group [4]. Occasionally, hate speech on popular social platforms leads to cyber bullying, harassment and the creation of hate sites. Lately, there has been an increasing interest in regulating harmful user-generated content on social platforms and therefore, suitable hate speech detection tools are needed [1].

The hate speech messages contain short text. Short texts present several challenges associated to the lack of semantic and syntaxes as well as data sparseness and ambiguity [5]. In order to overcome these challenges, several solutions have been proposed to improve the text representation, enrich the semantics by using implicit or explicit information, use specific classification algorithms that are able to accurately classify short text, or combine these approaches to improve the overall results. Identification of people or accounts that are spreading hate speech becomes one challenging task to research community.

In this work, we developed a distance measures based approach for hate speech spreaders detection. In this approach, we experimented with six different distance measures to find the similarity among the documents. This paper is organized in 6 sections. Section 2 explains the existing approaches proposed for hate speech spreaders detection. The dataset characteristics are presented in section 3. Section 4 discuss about proposed approach and the distance measures used in the proposed approach. The experimental results are presented in section 5. The section 6 concludes this work with future directions.

II. Related Work

The social media platforms like Facebook and Twitter provides a user-friendly environment and opportunity to raise their voice and opinion in the form of text, pictures and videos about different types of entities. Different age grouped people used these platforms to share every moment in their life with a community of people which causes these platforms flooded with data [23]. Due to the lack of controlling methods for restricting abusive comments, people are using these platforms for sending fake or false information and hate speech content in messages to damage the one’s image in our society. It is the responsibility of these platforms and governments to control such abusive language text before forwarding to a community of people. The Automatic hate speech detection faced a lot of problems because of non-standard differences in grammar and spelling of words in messages. This is more problematic when a country contains different language speaking people and the hate content created by these people contains code-mixed form of Multilanguage words. Sreelakshmi k et al., developed [6] a machine learning model for hate speech detection on the code-mixed dataset of Hindi-English in social media. The proposed method used fastText that is a pre-trained word embedding library from Facebook to represent data samples of 10000 that are gathered from various sources as hate or non-hate. The efficiency of proposed method is compared with doc2vec and word2vec feature techniques and observed that the fastText technique attained best feature representation when combined with (SVM)-Radial Basis Function (RBF). They also observed that the character level features gave best performance for code-mixed data.

In the work of Talha Anwar [7], feature embeddings of hundreds of tweets of a user are extracted using different transformers techniques such as BERT, BERTTweet, and RoBERTa for the English language and BETO for the Spanish language. They proposed a feature extraction technique using transformers embeddings and AutoML classifiers to classify hate speech spread users on Twitter. AutoML classifier is used to classify these embedding features. An accuracy of 75% and 85% is achieved using five-fold cross-validation and Accuracy of 72% and 82% is obtained for gold standard test data for English and Spanish, respectively. This author’s work is secured 4th position in PAN competition.

Claudio Moisés Valiense de Andrade et al., proposed [8] a solution which consists of exploiting several text representations and classifiers available in the literature. The authors exploited four representations such as word tfidf, char tfidf, vader and roberta’s word embeddings along with two classifiers such as SVM and RF. Finally, they depend on majority voting like most frequent decision for the final outcome. The experimental results show that the proposed solution attained an accuracy of 63% for English language and 79% for Spanish. This best combination of results in English and Spanish was obtained when experimented with the char tfidf representation along with Random Forests.

III. Description about Dataset

The PAN competition is conducting competitions on different research tasks every year [9]. In every competition, the organizers provide a dataset and request for submissions of research solutions. In 2021, they conducted a competition on Hate Speech Spreaders detection task [10]. In this task, they provided dataset in two languages such as English and Spanish. In this work, we concentrated on English dataset. in English dataset, class 1 indicates the profiles of authors who spread hate speech and class 0 indicates the profiles of authors who doesn’t spread hate speech. The training dataset consists of 40,000 tweets constituting a set of 200 tweets sampled per each of 200 anonymized users in XML format for two languages, English and Spanish. The test set contains tweets from 100 anonymized users per language. Note that the 200 tweets of hate speech spreaders may not all contain hate speech. The description about the dataset is represented in Table 1.

TABLE I
THE DATASET DESCRIPTION

Features / Classes	Class 0	Class 1
Training Profiles	100	100
Testing Profiles	50	50
Number of unique tweets in each profile	200	200

IV. PROPOSED APPROACH

In this work, we proposed an approach named as distance measures based approach for hate speech spreaders detection. The procedure of proposed approach is represented in Figure 1. In this approach, two pre-processing techniques like stop word removal and stemming applied on dataset to remove the irrelevant information. After cleaning data, extract all terms from the dataset. The feature selection technique of chi square measure is used to identify the best informative terms from all terms. Once relevant terms are identified, represent all documents in dataset are represented with identified terms. The term value in the vector representation is represented by using a term weight measure. After representing all documents of dataset as vectors with term weights of term, the test document vectors are prepared like the way the training document vectors are prepared. The class label of a test document is determined based on the distance score among the test document and all training document vectors. The class label is assigned to the test document based on the class which contains more number of similar documents. In the proposed approach, six distance measures such as Manhattan Distance, Minkowski Distance, Cosine, Jaccard and Dice are sued in the experiment to find the similarity among documents.

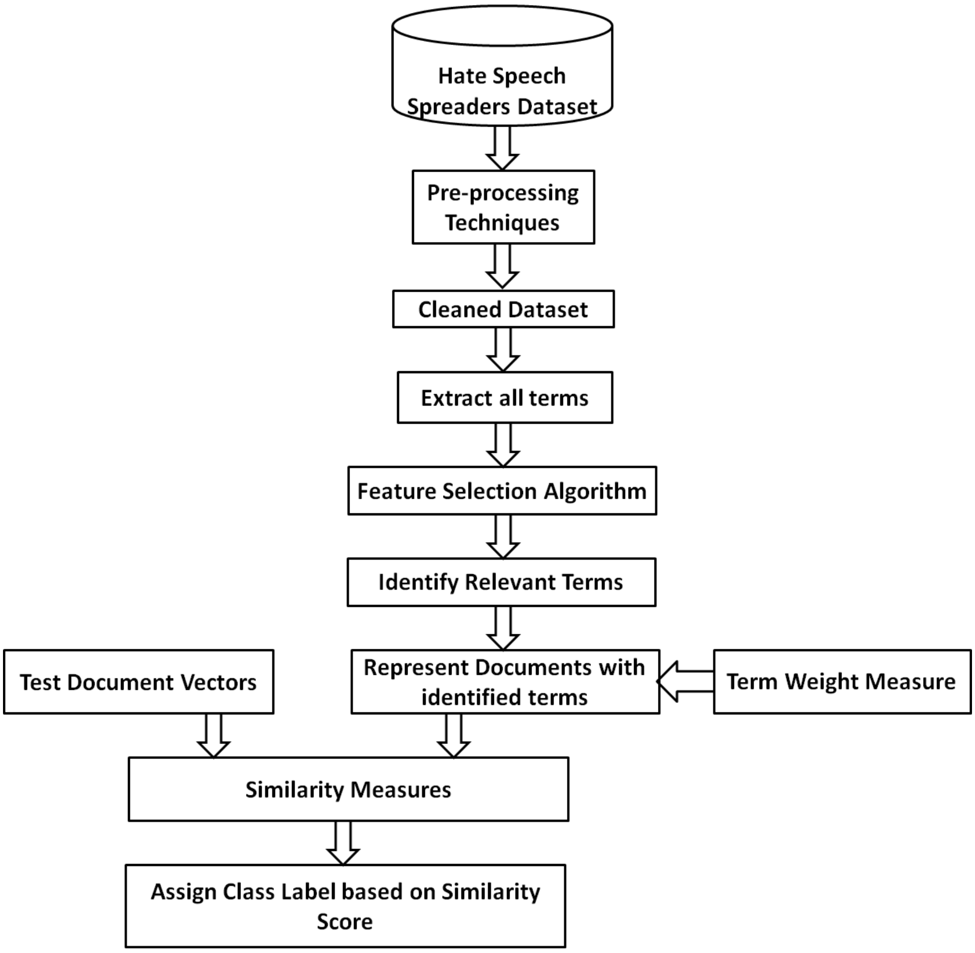
Hate Speech Spreaders Dataset
Pre-processing Techniques
Cleaned Dataset
Extract all terms
Feature Selection Algorithm
Identify Relevant Terms
Test Document Vectors
Represent Documents with identified terms
Term Weight Measure
Similarity Measures
Assign Class Label based on Similarity Score

Fig 1. The Steps in Proposed Approach

The performance of proposed approach mainly depends on the feature selection algorithm is used for finding important features, the term weight measure is used for computing the importance of term in vector representation, and the distance measures are used for determining the similarity among documents.

A. Feature Selection Algorithm - Chi Square (CHI2)

The chi square measure compares the actual value and expected value. This measure is proposed by Yang Y., et al., in 1997 [11]. If the actual and expected values are closer, then the chi square value is high which indicates the close association with term and class [24]. Equation (1) is used to compute the chi square measure.

$$CHI2\left(T_j, C_i\right) = \frac{N \times \left(a_{ij} d_{ij} - b_{ij} c_{ij}\right)^2}{\left(a_{ij} + b_{ij}\right) \times \left(b_{ij} + d_{ij}\right) \times \left(a_{ij} + c_{ij}\right) \times \left(c_{ij} + d_{ij}\right)} \tag{1}$$

Where, N is documents count in C_i class, a_{ii} and b_{ij} are the number of class C_i documents contain term T_j and doesn't contain term T_j respectively. c_{ij} and d_{ij} are the number of other than class C_i documents contain term T_j and doesn't contain term T_j respectively.

B. TERM Weight Measure - TF-IDF-ICSDF (Term Frequency – Inverse Document Frequency – Inverse Class Space Density Frequency)

The IDF measure says that the terms which are discussed in less documents attained good weight. Like IDF measure, ICF (Inverse Class Frequency) measure says that the terms which are discussed less number of classes attained good weight. The TF-IDF-ICSDF measure was developed in the works of [13] by combining TF, IDF and ICSDF factors. The ICSDF is a variant of ICF, which gives the average number of classes contain the given term. The ICSDF is determined by aggregating the probabilities of documents count in individual classes. The equation (2) is used to determine the weight of term T_i using TF-IDF-ICSDF measure.

$$TF - IDF - ICSDF\left(T_i, D_k\right) = TF\left(T_i, D_k\right) \times \left(\log\left(\frac{N}{DF\left(T_i\right)}\right)\right) \times \left(\log\left(\frac{m}{\sum_{j=1}^{m}\left(\frac{n_{cj}\left(T_i\right)}{N_{cj}}\right)}\right)\right) \tag{2}$$

Where, m is classes count, $n_{cj}(T_i)$ is count of documents in class C_j contain term T_i, N_{cj} is total documents count in class C_j.

C. Distance Measures

The distance measures calculate the similarity among two text documents. In this work, six distance measures such as Manhattan Distance, Euclidian distance, Minkowski Distance, Cosine, Jaccard and Dice are sued in the experiment to find the similarity among documents.

1) Manhattan Distance (MHD) Measure

The Manhattan distance also known as the City Block distance or Taxicab distance which is used to compute the similarity among two documents which was proposed by Cha, S.H. 2007 [14]. MHD also defined as the sum of absolute differences among the two document vectors. The Equation (3) is used to compute the MHD.

$$MHD\left(D_1, D_2\right) = \sum_{i=1}^{n}\left|\left(W\left(T_i, D_1\right) - W\left(T_i, D_2\right)\right)\right| \tag{3}$$

Where, MHD(D1, D2) is Manhattan distance among D1 and D2 documents,

2) Euclidian Distance (ED) Measure

Euclidian distance measure is used to compute the similarity among two documents. This measure mainly used to group similar documents into clusters based on the similarity among the documents (Komal Maher et al., 2016 [15]). Equation (4) is used to determine the ED measure among two documents.

$$ED\left(D_1, D_2\right) = \sqrt{\sum_{i=1}^{n}\left(W(T_i, D_1) - W(T_i, D_2)\right)^2} \qquad (4)$$

Where, ED(D$_1$, D$_2$) is Euclidian distance among D$_1$ and D$_2$ documents, n is the number of terms considered in the experiment, W(T$_i$, D$_1$) and W(T$_i$, D$_2$) are the weights of term T$_i$ in document D$_1$ and D$_2$ respectively.

3) Minkowski Distance (MND)

Minkowski distance determines the similarity among two real-valued document vectors (Cha, S.H. 2007 [14]). This measure is a generalized distance measure for Manhattan and Euclidian or Chebychev distance measures, which adds a ‘p’ parameter in the equation. Based on the p values the equation is used to compute either Manhattan or Euclidian or Chebychev distance. Equation (5) is used to calculate the Manhattan distance among two document vectors.

$$MND\left(D_1, D_2\right) = \left(\sum_{i=1}^{n}\left|\left(W\left(T_i, D_1\right) - W\left(T_i, D_2\right)\right)\right|^P\right)^{1/p} \qquad (5)$$

Where, p is order parameter. If p value 1, the equation act as Manhattan distance measure. If p value 2, the equation act as Euclidian distance measure. If p value infinitive, the equation act as Chebychev distance measure.

4) Cosine Distance Measure (CDM)

Cosine distance measure popularly used in different applications to compute the proximity among the document vectors (Moheb Ramzy Girgis et al., 2014 [16]). CDM determines the angle among the vectors. If the angle is small, the two vectors are more similar. If the angle is larger, the two vectors are more dissimilar. The angle values vary from 0 to 180 degrees and the CDM values ranging from +1 to -1. The cosine angle 0^0 gives CDM value of 1 which indicates the documents are similar. The angle 90^0 gives CDM value of 0 which means that the two documents are dissimilar [17]. The CDM is computed by using equation (6).

$$CDM\left(D_1, D_2\right) = \frac{\sum_{i=1}^{n}\left(W\left(T_i, D_1\right) \times W\left(T_i, D_2\right)\right)}{\sqrt{\sum_{i=1}^{n} W\left(T_i, D_1\right)^2} \times \sqrt{\sum_{i=1}^{n} W\left(T_i, D_2\right)^2}} \qquad (6)$$

Where, CDM(D$_1$, D$_2$) is the Cosine distance among D$_1$ and D$_2$ documents..

5) Dice Distance Measure (DDM)

Dice Distance measure compare the similarity among two document vectors (E man Al Mashagba et al., 2011 [18]). The DDM is computed by using equation (7).

$$DDM\left(D_1, D_2\right) = \frac{\sum_{i=1}^{n}\left(W\left(T_i, D_1\right) \times W\left(T_i, D_2\right)\right)}{\alpha\sum_{i=1}^{n} W\left(T_i, D_1\right)^2 + (1-\alpha)\sum_{i=1}^{n} W\left(T_i, D_2\right)^2} \qquad (7)$$

Where, α is a parameter which control the magnitude of false positive versus false negative errors. The range of α values is from 0 to 1. If $\alpha < 0.5$, more significance is given to recall by DDM measure. If $\alpha > 0.5$, more significance is given to precision by the DDM measure.

6) *Jaccard Distance Measure (JDM)*

Jaccard Distance measure is generally used to compute the similarity among query and document, document and document (Abhishek Jain et al., 2017 [19]). The JDM value range is from 0 to 1. The JDM value of 0 indicates the documents are completely dissimilar, the value of 1 indicates the documents are completely similar [20]. JDM measure is computed by using Equation (8).

$$JDM\left(D_1, D_2\right) = \frac{\sum_{i=1}^{n}\left(W\left(T_i, D_1\right) \times W\left(T_i, D_2\right)\right)}{\sum_{i=1}^{n} W\left(T_i, D_1\right)^2 + \sum_{i=1}^{n} W\left(T_i, D_2\right)^2 - \sum_{i=1}^{n}\left(W\left(T_i, D_1\right) \times W\left(T_i, D_2\right)\right)} \tag{10}$$

V. Experimental Results

The experiment conducted on the dataset of PAN 2021 hate speech spreaders task [21]. The dataset is divided into 70% of data for training and 30% data for testing. In this work, the training documents are represented as vectors and test document vectors class labels are identified based on the similarity among the test document vector and training document vectors. The distance measures are used to determine the similarity among the documents [22]. In this work, six distance measures are used to identify accuracy of hate speech spreaders detection. The accuracy is number of test documents are correctly predicted their class label divided by total number of test documents considered. Table 4 displays the accuracies of Hate Speech Spreaders Detection when experiment conducted with different distance measures.

Table II
The accuracies of Hate Speech Spreaders Detection when distance measures are used

Distance Measure	**Accuracy**
Euclidian distance	0.6637
Manhattan Distance	0.6309
Minkowski Distance	0.6821
Cosine	0.5754
Jaccard	0.7356
Dice	0.6087

From Table 2, the Jaccard similarity measure attained best accuracy of 0.7356 for hate speech spreaders detection when compared with other distance measures.

VI. Conclusions and Future Scope

Now a days people are using social media platforms to know about many issues like news, about famous people etc. some people are using these platforms to send hate messages to defame the reputation of products, services, people etc. The identification people or accounts that spread hate speech are one important research in recent times. PAN competition conducted competition on hate speech spreaders detection task in 2021 competition. In this work, we proposed a distance measures based approach for hate speech spreaders detection. In the proposed approach, we used six different types of distance measures to find the distance among two documents. The proposed approach attained an accuracy of 0.7356 for hate speech spreaders detection.

In future work, we are planning to propose a new distance measure to find the similarity among documents. We also planned to implement deep learning techniques to improve the accuracy of hate speech spreaders detection.

References

[1] Banks, J. (2010). Regulating hate speech online. International Review of Law, Computers & Technology, 24(3), 233–239.

[2] Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1–10.

[3] Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the Targets of Hate in Online Social Media. Retrieved from http://arxiv.org/abs/1603.07709

[4] Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-muslim movements in the Czech Republic and Germany. International Journal of Cyber Criminology, 11(1), 143–160.

[5] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, “A Survey on Author Profiling Techniques”, International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.

[6] Sreelakshmi k,Premjith B, Soman K.P, “Detection of Hate Speech Text in Hindi-English Code-mixed Data”, Procedia Computer Science 171 (2020) 737–744

[7] Talha Anwar, “Identify Hate Speech Spreaders on Twitter using Transformer Embeddings Features and AutoML Classifiers”, CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

[8] Claudio Moisés Valiense de Andrade, Marcos André Gonçalves, “Profiling Hate Speech Spreaders on Twitter: Exploiting Textual Analysis of Tweets an Combinations of Multiple Textual Representations”, CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

[9] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, “Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling ”, International Journal of Intelligent Engineering and Systems, 9 (4), pp. 136-146, Nov 2016.

[10] https://pan.webis.de/clef21/pan21-web/author-profiling.html

[11] Yang Y., J.O. Pedersen, A comparative study on feature selection in text categorization, in: Fourteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, 1997, pp. 412–420.

[12] Raghunadha Reddy T, Vishnu Vardhan B, GopiChand M, Karunakar K, “Gender prediction in Author Profiling using ReliefF Feature Selection Algorithm”, Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 169-176, 2018.

[13] Ren, F., & Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. Information Sciences, 236 , 109–125. http://doi.org/10.1016/j. ins.2013.02.029.

[14] Cha, S-H. (2007). Comprehensive survey on distance similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences, 1 (4), 300–307.

[15] Komal Maher, Madhuri S. Joshi, “Effectiveness of Different Similarity Measures for Text Classification and Clustering”, International Journal of Computer Science and Information Technologies, Vol. 7, No.4, pp.1715-1720, 2016.

[16] Moheb Ramzy Girgis, Abdelmgeid Amin Aly & Fatima Mohy Eldin Azzam, “The Effect Of Similarity Measures On Genetic Algorithm-Based Information Retrieval”, International Journal of Computer Science Engineering and Information Technology Research, Vol. 4, Issue 5, pp. 91-100, Oct 2014.

[17] K. Pradeep Reddy, T. Raghunadha Reddy, G. Apparao Naidu, B. Vishnu Vardhan, “Impact of Similarity Measures in Information Retrieval”, International Journal of Computational Engineering Research, Vol 8, Issue 6, pp. 54-59, 2018.

[18] E man Al Mashagba , Feras Al Mashagba and Mohammad Othman Nassar, “Query optimization using genetic algorithm in the vector space model”, International Journal of Computer Science, vol. 8, no. 3, pp.450-457, Sept. 2011.

[19] Abhishek Jain, Aman Jain, Nihal Chauhan, Vikrant Singh, Narina Thakur, “Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model”, International Journal of Computer Applications, Volume 164, No 6, PP.28-30, 2017.

[20] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, “Author profile prediction using pivoted unique term normalization”, Indian Journal of Science and Technology, Vol 9, Issue 46, Dec 2016.

[21] Swathi Ch, Karunakar K, Archana G, T. Raghunadha Reddy, “A New Term Weight Measure for Gender Prediction in Author Profiling”, Proceedings in Advances in Intelligent Systems and Computing, Volume 695, PP. 11-18, 2018.

[22] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, “A Document Weighted Approach for Gender and Age Prediction”, International Journal of Engineering -Transactions B: Applications, Volume 30, Number 5, pp. 647-653, May 2017.

[23] Karunakar Kavuri, Kavitha, M. (2020). “A Stylistic Features Based Approach for Author Profiling”. In: Sharma, H., Pundir, A., Yadav, N., Sharma, A., Das, S. (eds) Recent Trends in Communication and Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-0426-6_20

[24] Karunakar. Kavuri and M. Kavitha, "A Term Weight Measure based Approach for Author Profiling," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), 2022, pp. 275-280, doi: 10.1109/ICESIC53714.2022.9783526.