IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Impact of Feature Selection Algorithms for Fake News Spreaders Detection

[1]K. V. Nageswari, [2]K. Raja, [3]K. Bhanuchand, [4]K. Rambabu

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

[1, 4] Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Narasapur, AP

[2, 3] Information Technology, Swarnandhra College of Engineering and Technology, Narasapur, AP

Abstract: In last decade of time, people are heavily relied on social media environments like Twitter, Facebook, Blogs, Whatsapp, Instagram etc., to know about the news about different concepts, famous people, products and services. Most of the people utilized these environments for sharing their opinions on different entities. Some people are spreading false or fake information in these environments to misguide the users. Identification of people who spreads the fake information becomes one important challenge for research community. PAN competition conducted a competition on fake news spreaders detection task in 2020. Several researchers proposed solutions for finding the authors who spreads the fake news in social media environments. In this work, we proposed an approach for fake news spreaders detection by using different feature selection algorithms. The content based features like words are most important features to differentiate the writing styles of fake news spreaders and real news spreaders. The identification of important words for experimentation is very important to improve the accuracy of fake news spreaders detection. In the proposed approach, feature selection algorithms are used for identifying most relevant words or features for experimentation. The identified features are used for representing the documents as vectors. These document vectors are trained with machine learning algorithms for generating the classification model. This model is used for predicting the accuracy of proposed approach as well as for predicting whether new author is fake news spreader or real news spreader. The PAN 2020 competition fake news spreader detection dataset is used in this experiment. Two machine leaning algorithms such as random forest and support vector machine are evaluating the accuracy of proposed approach. The proposed approach attained best accuracy for fake news spreader detection when compared with most of the approaches.

***Index Terms*- Fake News, Fake News Spreaders Detection, Feature Selection Algorithms, Machine Leaning Algorithms**

1. INTRODUCTION

Fake news spreading was become a concerning problem in online social media networks in recent years. Researchers was found that fake news is more likely to go viral than real news, spreading both faster and wider and is threatening public health, election outcomes, emergency management and response, and is responsible for a general decline in trust that citizens of democratic societies for online platforms. The recognition of fake news spreaders becomes a crucial challenge for social media platforms and research community to stop the spreading of fake news. For example, Twitter bots are capable of generating fake information and propagating it through their follower networks, which impact real-life entities such as stock markets and possibly even elections [1].

The fake news spreaders detection is a crucial step to prevent the dissemination of fake news through social media. PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection) organized International Workshops on different types of tasks like plagiarism detection, author profiling, authorship attribution, authorship verification, style change detection, bot detection, celebrity profiling etc., in every year. In 2020 competition, PAN organizers introduced the task of fake news spreaders detection by providing twitter dataset of 500 authors in two languages such as English and Spanish. The intention of PAN organizers was to detect fake news spreader which is a type of author profiling task. The fake news spreaders task is to detect whether the author of a text is spreading fake news or not.

Detecting fake news spreaders in Twitter was modelled as a typical binary Text Classification (TC) problem that labels a given news spreader as fake or real. TC is a Supervised Machine Learning (ML) technique that automatically assigns a label from the predefined set of labels to a given unlabelled input. It has wide applications in various domains, such as target marketing, medical diagnosis, news classification, and document organization.

In this work, we proposed an approach for fake news spreaders detection based on feature selection algorithms. In this approach, the experiment conducted with content based features. The content based features gives the difference among fake news spreaders and real news spreaders based on the words they used in their messages. The number of words is huge in the writings of fake and real news spreaders. The identification of important or specific words for fake and real news spreaders is one important task. The feature selection algorithms are used for recognizing the relevant and redundant words. In this proposed approach, different feature selection algorithms are used for fake news spreaders detection. The experimental results of proposed approach with different machine learning algorithms are presented for fake news spreaders detection.

This paper is organized in 6 sections. Section 2 explains different research works proposed for fake news spreaders detection. Section 3 describes the characteristics of proposed approach. The section 4 explains the proposed approach and feature selection algorithms that are used in the proposed approach. Section 5 presents the experimental results of the proposed work. Section 6 concludes this work.

2. LITERATURE SURVEY

Feature selection techniques recognize best informative features and remove redundant features from a large set of features. Catherine Ikae et al., suggested [2] an approach based on a two-stage method by ignoring infrequent terms and ranking the others according to their occurrence differences between the two categories. After removing infrequent terms, they proposed a feature selection method that works in two stages. In the first stage, the term frequency (tf) information is taken into account. For each term, the discriminative power is computed by estimating the occurrence probability difference in both categories. After this step, one can stop the feature selection by considering the k terms (with k = 100 to 250) having the highest and smallest probD scores. The second step applies an additional feature selection procedure. In their work, the chi-square method was selected to reduce the feature space to a few hundred terms. The top 150 terms having the highest chi-square values was selected to define the feature set. They observed that doubling the number of features does not always improve the overall effectiveness of the approach.

Matteo Cardaioli et al., applied [3] KBest feature selection algorithm to reduce the dimensionality of data. Boško Koloski et al., performed [4] Dimensionality reduction via matrix factorization through sparse Singular Value Decomposition (SVD) [5]. TFIDF is Term Frequency and Inverse Document Frequency measure which assigns more weight to the features which are discussed in less number of documents. Nikhil Pinnaparaju et al., developed [6] a method by utilizing content analysis and more user modelling to capture who is more likely to share fake news. They used a very simple transformation like TF-IDF algorithm to transform the training data into numeric vector representations for Training. All the tweets of a given author were concatenated and consider it as a single big document corresponding to the author.

Ahmad Hashemi et al., developed [7] a model which contain three separate components. Each component is developed to process a separate set of features with same classifier. The features that are used by the three components are features extracted using TF-IDF, features extracted from word embeddings and combination of implicit and statistical features. This model utilized the Spacy package to extract [K19] the word embedding vectors.

3. DESCRIPTION ABOUT FAKE NEWS SPREADERS DETECTION DATASET

In PAN 2020 competition, the organizers of competition included a task of fake news spreaders detection. They provided the dataset of twitter tweets in two languages such as Spanish and English [8]. The English and Spanish language datasets contains training dataset of 300 author tweets and testing dataset of 200 author tweets. Each author file consists of 100 tweets. The training dataset was balanced in terms of equal number of documents in both classes such as 150 fake news profiles and 150 real news profiles. When compared with previous PAN competitions datasets this dataset hide the sensible information in the tweets to make profile anonymization like “user”, “rt” (re-tweet), “hashtag”, “URL” was obfuscated using some standardized keywords. Table 1 shows the dataset properties of fake news spreaders provided in PAN 2020 competition.

Table 1. The Dataset Characteristics

Language	Training		Test		Total
	Fake News Profiles	Real News Profiles	Fake News Profiles	Real News Profiles	
English	150	150	100	100	500

4. PROPOSED APPROACH

The steps in the proposed approach for fake news spreaders detection detection is displayed in Figure 1. In this approach, firstly, apply the pre-processing techniques such as removal of hashtags, @mentions, retweets, stop words and punctuation marks to remove unnecessary information from the dataset. After removing unnecessary information from the dataset, apply stemming technique on informative words. Stemming converts words into its root from and reduces the number of unique words in the dataset. Now, the dataset is ready to extract the useful features to differentiate the writing styles of fake and real news spreaders. Extract all terms from the cleaned dataset. Feature selection algorithms are used to determine the scores of all terms. Based on the scores, identify the top scored terms and considered these top scored terms as features to represent the documents as vectors. The TFIDF measure is used for finding the term value in the vector representation. The document vectors are trained with different machine learning techniques to produce a classification model. The classification model determines the accuracy of the proposed approach. The accuracy mainly depends on the features used in the document vector representation. The feature selection algorithms play a major role for identification of relevant features. In this work, various feature selection algorithms are used to identify the important features for document vector representation.

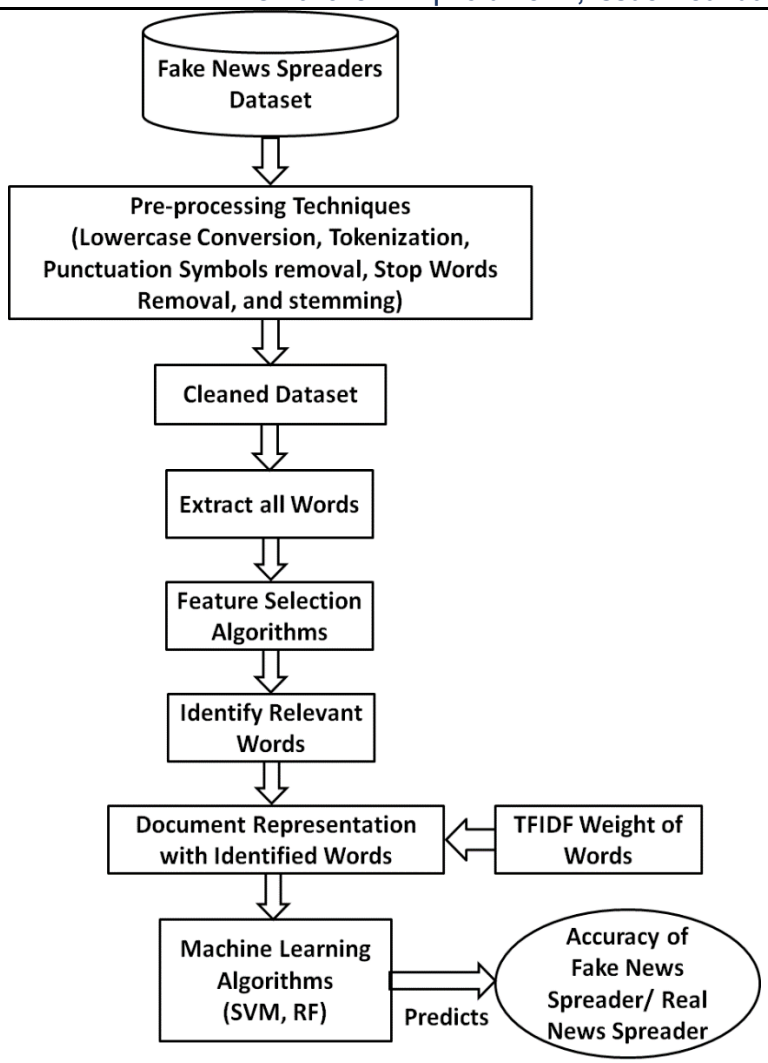

Figure 1: Proposed Approach for Fake News Spreaders Detection

4.1 Feature Selection Algorithms

In these days, High dimensionality is a bigger problem in machine learning approaches. Several researchers do experiments on this problem to lessen the features count in the experiment. Researchers used statistical methods to avoid redundant data and to reduce noise data. Nevertheless, all features are not used to train the model. The feature selection techniques play major role to enhance the efficiency of a model by identifying the most relevant and non-redundant features.

The feature selection algorithms play an important role in the recognition of relevant features for predicting the fake news spreaders detection. The feature selection algorithms are majorly divided into three classes such filter based, wrapper based and embedded based feature selection algorithms [9]. The filter based feature selection algorithms identify the important features by computing the scores of features. The score of a feature is high means the feature is more relevant for the class. The wrapper based feature selection algorithms divide the features into different subsets of features. The machine learning algorithms evaluate the performance of each feature subset. The best feature subset is returned based on the highest performance score of a feature subset. The embedded based methods used machine learning algorithms directly to identify the important features from a set of features.

In this work, different filter based feature selection algorithms are used in the experiment. All the FSAs determine the importance of a feature based on the way the terms are distributed in positive and negative classes. The notation used for distribution of term in different classes is represented in Table 2.

Table 2: Term Distribution in different classes

	C	$\overline{C}$
T1	A	C
$\overline{T1}$	B	D

In Table 2, A and C are the documents count in positive class and negative class respectively which contain the term T, B and D are the documents count in positive and negative class respectively which never contains the term T.

4.1.1 Document Frequency Difference (DFD)

The DFD feature selection technique was developed by Nicholls and Song [10]. The DFD computes the frequency difference of a feature in positive and negative class. The difference is normalized with number of documents in the total dataset. The DFD value is high means the feature appearred in more positive class documents. The Equation (1) is used to determine the DFD measure of a term Ti in class C_j.

$$DFD(T_i, C_j) = \frac{DF_{t,\,pos} - DF_{t,\,neg}}{N} \quad (1)$$

Where, $DF_{t,\,pos}$, $DF_{t,\,neg}$ are the number of documents in positive and negative classes respectively which contain term Ti, N is number of documents in total dataset.
The DFD of a term Ti in total dataset is determined by using Equation (2).

$$DFD(T_i) = \max_{j=1}^{m} \left(DFD(T_i, C_j) \right) \quad (2)$$

4.1.2 Gain Ratio (GR)

GR is introduced in decision tree algorithm of C4.5 that works based on the concept of information gain [11]. The GR value of an attribute is computed by normalizing the attribute value of information gain with entropy [12]. The entropy is high means there is a uniform distribution of feature values. The entropy is low means the feature values are distributed around a point [13]. The GR value is high for a feature indicates the feature is a good representative for classification. The GR score of term is same in all classes. The GR is computed for a term Ti in any class by using Equation (3).

$$GR(T_i) = \frac{IG(T_i)}{-\frac{(A+B)}{N} \times \log \frac{(A+B)}{N} - \frac{(C+D)}{N} \times \log \frac{(C+D)}{N}} \quad (3)$$

4.1.3 Information Gain (IG)

IG measures the contribution of a term presence or absence in a document to predict the class of a document [11]. IG assigns best score to a term when it is a good representative for a class. The IG of a term specific to any class is computed by using Equation (4).

$$\begin{aligned} IG(T_i) = {} & \frac{A}{N} \times \log \left(\frac{A * N}{(A+B) \times (A+C)} \right) \\ & + \frac{B}{N} \times \log \left(\frac{B * N}{(A+B) \times (B+D)} \right) \\ & + \frac{C}{N} \times \log \left(\frac{C * N}{(A+C) \times (C+D)} \right) \\ & + \frac{D}{N} \times \log \left(\frac{D * N}{(B+D) \times (C+D)} \right) \end{aligned} \quad (4)$$

4.1.4 Chi-Square (CHI2)

Chi-Square measure is a feature selection algorithm that determines the level of correlation among a term and a class [14]. The zero value of a CHI2 measure indicates the term is not having any relationship with a class. The higher value of CHI measure indicates the term is a good representative for a class. The CHI2 value of a term in all classes is equal. The CHI2 of a term in any class is calculated by using Equation (5).

$$CHI(T_i) = \frac{N \times (A * D - B * C)^2}{(A+B) \times (A+C) \times (B+D) \times (C+D)} \quad (5)$$

4.1.5 Mutual Information (MI)

MI is a supervised feature selection method that determines the amount of correlation among feature and a class [15]. The MI of a term Ti in a specific class Cj is determined by using Equation (6).

$$MI(T_i, C_j) = \log \left(\frac{A * N}{(A+B) \times (A+C)} \right) \quad (6)$$

Where N is documents count in total dataset. The MI of a term in whole dataset is determined by using Equation (7).

$$MI(T_i) = \max_{j=1}^{m} \left(MI(T_i, C_j) \right) \quad (7)$$

4.1.6 Odds Ratio (OR)

The OR algorithm finds the odds of a feature that are present in the positive class [16]. OR assigns more score to the features that are occurred more in positive class than negative class. OR gives less weight to the terms occurred more in negative class than positive class. OR allocate zero score to the feature when the feature occurred equally in both positive and negative class. OR measure gives equal score to the term in any class. The OR of a term Ti in any class is determined by using Equation (8)

$$OR(T_i) = \log\left(\frac{A * D}{B * C}\right) \quad (8)$$

4.1.7 Correlation Coefficient (CC)

The CC measure is a variant of a CHI measure which is obtained by taking the square root of a CHI measure [17]. The CC value of term is higher indicates that the term is good representative for a class. The CC value of a term in a specific class Cj is computed by using Equation (9)

$$CC(T_i, C_j) = \frac{\sqrt{N} \times (A \times D - B \times C)}{\sqrt{(A+B) \times (A+C) \times (B+D) \times (C+D)}} \quad (9)$$

The CC of a term in whole dataset is determined by using Equation (10).

$$CC(T_i) = \max_{j=1}^{m} (CC(T_i, C_j)) \quad (10)$$

4.2 Machine Learning Algorithms

The machine learning algorithms generates the classification model by training the machine with training data. The classification model is used to predict the class label of test documents as well as to determine the efficiency of the proposed methods. In this work, two machine learning algorithms such as SVM [K18] and RF [19] are used to build the classification model for fake news spreaders detection. In this work, the experiment conducted on the dataset of 300 profiles of both fake news profiles and real news profiles. The machine learning algorithms splits this 300 profiles dataset into training data and testing data. 70% of dataset is used for training the algorithm and 30% of dataset is used for testing purpose.

The researchers used various measures precision, recall, f1-score and accuracy to evaluate the performance of the proposed approach. In this work, accuracy measure is used to display the results of fake news spreaders detection. The accuracy is defined in Equation (11).

$$Accuracy = \frac{Number\ of\ \text{test profiles correctly predicted their fake news spreader (TP+TN)}}{Total\ number\ of test\ profiles\ (TP + FP + FN + TN)} \quad (11)$$

5. Experimental Results

The experiment is carried out with the features selected by the feature selection algorithms. Seven feature selection algorithms are used in this work to identify the important features for experimentation. Two machine learning algorithms are used to generate the classification model. This model is used to predict the accuracy of the fake news spreaders detection.

The Table 2 shows the accuracy values of fake news spreaders detection when experimented with the features identified by the different feature selection algorithms and support vector machine classification algorithm.

Table 3: Accuracies of Support Vector Machine classifier for FNS Detection

FSA's / High Scored Terms	DFD	GR	IG	CHI2	MI	OR	CC
2000	72.56	72.92	73.93	74.56	77.91	78.56	79.56
4000	73.56	73.56	74.56	74.92	78.56	79.91	79.91
6000	73.93	74.56	75.57	75.93	78.93	80.56	80.56
8000	74.56	74.93	75.91	76.56	79.56	80.92	81.92
10000	75.23	75.91	76.56	77.94	80.92	81.23	82.93

In Table 2, the CC feature selection algorithm obtained highest accuracy of 82.93% for fake news spreaders detection when experiment conducted with most relevant terms of 10000 and support vector machine classifier.

The Table 3 shows the accuracy values of fake news spreaders detection when experimented with the features identified by the different feature selection algorithms and support vector machine classification algorithm.

Table 4: Accuracies of Random Forest classifier for FNS Detection

FSA's / High Scored Terms	DFD	GR	IG	CHI2	MI	OR	CC
2000	76.62	77.63	77.61	78.94	80.61	81.61	82.61
4000	76.94	77.94	77.94	79.94	80.94	81.94	83.94
6000	77.94	78.62	78.63	80.62	81.62	82.62	84.62
8000	78.61	78.94	79.94	80.94	81.94	82.94	84.94
10000	78.94	79.94	80.62	81.27	83.63	83.94	85.63

In Table 3, the CC feature selection algorithm obtained highest accuracy of 85.63% for fake news spreaders detection when experiment conducted with most relevant terms of 10000 and random forest classifier.

6. Conclusions

In this work, we proposed an approach fake news spreaders detection by using content based features. In this approach, different feature selection algorithms are used for identifying important words from the dataset. Two machine learning algorithms are used to determine the efficiency of the proposed approach. The terms that are selected by the feature selection algorithms are used as features for representing the documents as vectors. The proposed approach with random forest classifier attained highest accuracy

of 85.63% for fake news spreaders detection. It was observed that the features selected by the feature selection algorithms are helped more to improve the accuracy of fake news spreaders detection.

REFERENCES

[1] Brigida, M., Pratt, W.R.: Fake news. The North American Journal of Economics and Finance 42, 564–573 (2017)

[2] Catherine Ikae, Jacques Savoy, "UniNE at PAN-CLEF 2020 Profiling Fake News Spreaders on Twitter", Notebook for PAN at CLEF 2020, 2020, 22-25 September 2020, Thessaloniki, Greece.

[3] Matteo Cardaioli, Stefano Cecconello, Mauro Conti, Luca Pajola, and Federico Turrin, "Fake News Spreaders Profiling Through Behavioural Analysis", Notebook for PAN at CLEF 2020, 2020, 22-25 September 2020, Thessaloniki, Greece.

[4] Boško Koloski, Senja Pollak, and Blaž Škrlj, "Multilingual Detection of Fake News Spreaders via Sparse Matrix Factorization", Notebook for PAN at CLEF 2020, 2020, 22-25 September 2020, Thessaloniki, Greece.

[5] Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions (2009)

[6] Nikhil Pinnaparaju, Vijaysaradhi Indurthi, and Vasudeva Varma, "Identifying Fake News Spreaders in Social Media", Notebook for PAN at CLEF 2020, 2020, 22-25 September 2020, Thessaloniki, Greece.

[7] Ahmad Hashemi, Mohammad Reza Zarei, Mohammad Reza Moosavi, and Mohammad Taheri, "Fake News Spreader Identification in Twitter using Ensemble Modeling", Notebook for PAN at CLEF 2020, 2020, 22-25 September 2020, Thessaloniki, Greece.

[8] Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (Sep 2020), CEUR-WS.org

[9] K. R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos.1-2, pp. 273-324, 1997.

[10] C. Nicholls, F. Song, "Comparison of Feature Selection Methods For Sentiment Analysis," in AI'10 Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence, Vol. 10, No. 3, pp. 286–289.

[11] J. R. Quinlan, Induction of decision trees, Machine Learning 1(1) (1986) 81−106.

[12] Duch, W. (2006). Filter methods. In Feature extraction (pp. 89–117). Springer.

[13] 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

[14] 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.

[15] H. Dağ, K. E. Sayin, I. Yenidoğan, S. Albayrak and C. Acar, "Comparison of feature selection algorithms for medical data," 2012 International Symposium on Innovations in Intelligent Systems and Applications, Trabzon, 2012,pp1-5.doi: 10.1109/INISTA.2012.6247011

[16] Yang Y. and Pedersen J., "A Comparative Study on Feature Selection in Text Categorization," in Proceedings of the 14th International Conference on Machine Learning, Nashville, pp. 412-420, 1997.

[17] Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning (Ph.D. thesis). University of Waikato, Hamilton, New Zealand.

[18] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi. "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods". Applied Soft Computing, vol. 86, p. 105836, 2020.

[19] 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.

[20] 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

[21] 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.

[22] Z. Zheng and R. Srihari. "Optimally combining positive and negative features for text categorization", in ICML 2003 Workshop, 2003.

[23] Cortes, C. & Vapnik, V. Machine Learning (1995) 20: 273. https://doi.org/10.1023/A:1022627411411

[24] Breiman, L. (2001). Random forests. Machine Learning, 45(1),5–32, https://doi.org/10.1023/A:1010933404324