ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

A Machine Learning Approach for Authorship Verification

Palacharla Ravikumar [1], Bobba Ramya Banu [2], Dr. Kamini Srinivasa Rao [3]

Assistant Professor, Bapatla Engineering College, Bapatla, A.P., India[1]

Assistant Professor, Pragati Engineering College, Surampalem, A.P. India[2]

Assistant Professor, Bapatla Engineering College, Bapatla,A.P. ,India [3]

ABSTRACT: The World Wide Web is exponentially increasing with textual data along with the problems of authorship also increases. Many authors claim the works of other authors. Authorship analysis is one technique attracted by the researchers to solve the problems of authorship. Authorship Verification is a type of Authorship analysis which is used to verify the authorship of a document by processing the documents of the suspected other. In existing approaches, the researchers extracted a set of stylistic features to discriminate the writing style of authors. We observed that the authors used same set of words, the same order of words and the same part of speech order in different documents. In this work, the experiment conducted with most frequent terms, most frequent word and POS N-grams. The documents are represented as vectors with the weights of the terms, word and POS N-grams. The vectors are forwarded to three machine learning algorithms to develop the classification model. The experiment carried on PAN competition 2014 and 2015 authorship verification datasets. It was found that our approach achieved good results for authorship verification and also observed that our approach efficiency is good when compared with several approaches of authorship verification.

KEYWORDS: Authorship Verification, Machine Learning Algorithms, Word N-grams, POS N-Grams, Accuracy

I. INTRODUCTION

The growth of data exponentially spreading across the web in different forms and it may result in manipulated, fraudulent, morphed, unidentified and stolen data. A wrong message can get viral and reach to the people very fast. The prime concentration of the researchers is finding the fraudulent data in forensic department, news authorities and government. In this process, finding the author of the text is very important. Authorship Analysis is one research area to find the authorship information of the authors of the text [1]. Authorship analysis is a technique of understanding the written document by analyzing the author whose roots are coming from stylometry. In the present research field the authorship analysis utilizes the knowledge of the text statistics and the researchers used the machine learning techniques in the process of identifying the author. Authorship analysis divided into three classes such as Author Profiling, Authorship Attribution and Authorship Verification.

Author Profiling determine the author's demographics such as age, gender, native language, education etc for a given text. This task focuses on detecting common characteristics among general classes of authors and it is associated with several applications in forensics and marketing. Authorship Attribution is used to detect the author of a given text by analyzing the documents of multiple authors. This is typically a text classification task and is directly related with the quantification of style of documents and more specifically the personal style of each author. The technique of comparing different parts of a written text of a single author and identifying the new text is written by the same author or not is called authorship verification. These verification techniques are important on various information technique applications like digital text forensics, digital humanities and social media analytics. In the cyber space, finding the suspect as a criminal or not is to be carried out by looking at the digital document which is used as an evidence and there by verifying authorship of a document. The process of authorship verification is shown in Fig. 1.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

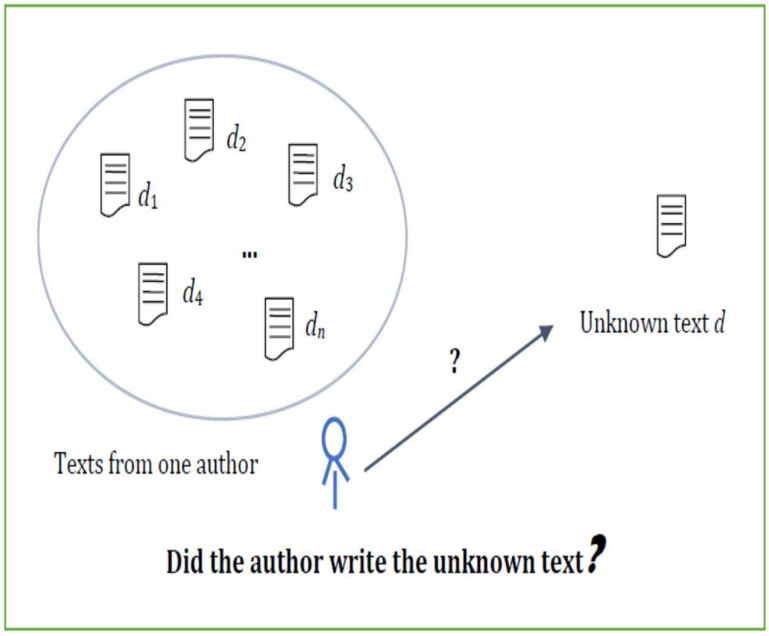

Fig.1. Authorship Verification

The fraudulent data is accessed as suspicious data or the text in dispute. Authorship verification is to check wether the disputed text was written by an author or not. As explain in fig.1 the unknown text 'd' is to be asigned to an author by looking at various scripts of the author. That means in the figure d1, d2, …dn are the known text of an author and we are verifying whether the unknown text 'd' belongs to the author or not.

In this work, the experimentation carried out with different machine learning algorithms to create the classification model. The classification model (classifier) is able to automatically assign authors to electronic texts by learning specific features/characteristics of each category of authors. We extracted various features like most frequent terms, most frequent word N-Grams and most frequent POS N-Grams for document representation. The weights of the features are computed by using a term weight measure.

This paper is structured in 6 sections. The previous works in Authorship Verification is discussed in section 2. The PAN competition 2014 and 2015 Authorship Verification corpus properties are presented in section 3. Section 4 explains the proposed approach with the explanation of term weight measure. The experimental results of the proposed approach is analyzed and presented in section 5. The section 6 concludes this work.

II. RELATED WORK

Authorship analysis is a technique of analyzing the anonymous chunks of text document and predicting the author based on the Stylometry which is a linguistic research area that involves different statistical approaches and machine learning techniques [1]. Authorship verification is a technique of processing the several written textual documents of one author and find out whether new document was written by this author or not, but it is not involving the identification of the author[1].

Nektaria Potha et al., proposed [2] two techniques such as instance and profile based and observed that their techniques performance is better than many popular approaches of extrinsic Authorship Verification . In instance based technique, they experiment with Ranking based Imposters by using the ranking information of known documents. In

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

profile based technique, they used Profile based imposter method by utilising the similar characteristics among the documents. Esteban Castillo et al., extracted [3] suitable linguistic features by using graph representation which is based on network analysis techniques. They experimented with PAN competition 2015 Author Verification dataset. They used four centrality measures such as betweenness, closeness, eigenvector and degree to detect the patterns in the writing styles of the authors. They observed that words with high closeness and high betweenness contributed more to verify the author of a document.

In the experiments of Darnes Vilarino et al., [4], the documents were represented with a vector graph based features, lexical and syntactic features. Graph based features were extracted by using data mining tool called Subdue. Classification model is built by using SVM. This experiment showed that the run time complexity is more when compared to other submissions for a competition. Victoria Bobicev proposed [5] a system, which depends on statistical n-gram model called prediction by partial matching (PPM) for automatic detection of author when a text document is given with corpus which contains known authors and a small training set. The training corpus contains 30 authors, for each author 100 posts are extracted and corpus is normalized to 150-200 words to maintain the uniformity. The experimental results are exhibiting that when the length of the document is increased, the accuracy measure (F-score) not increased.

Compression based dissimilarity measure were used by Cor J. Veenman et al., [6] to calculate the compression distance among the text documents. They used three approaches for author verification task. This experiment got the best accuracy for the authorship verification task over all the submissions in the competition of PAN 2013. A profile based approach called Common N-Gram (CNG) method was used by Michiel van Dam [7] which implements the normalized distance measure among imbalance text and short text. Every document is represented with character n-grams to implement CNG method. For Spanish and English language documents this method showed a good accuracy but failed for Greek language. A new machine learning technique was proposed by Alberto Bartoli et al.[8] which uses the combination of linguistic features. Many features were extracted like N-grams of character, word, POS, sentence lengths, word lengths, POS n-grams, word richness, text shape n-grams, sentence lengths n-grams and punctuation n-grams. In PAN 2015 conference this technique ranked with first in authorship verification.

Wei Feng Vanessa et al., proposed [9] a new technique called unmasking approach. This method is helpful in enhancing the feature quality when building the classifiers. In this work, they experimented with 399 Spanish language features, 538 English language features and 568 Greek language features. These features comprise of both coherent and Stylometry features. The experimental results shown that higher accuracy is achieved with English and Spanish documents and less accuracy was achieved for Greek documents. General impostor's technique was implemented by Shachar Seidman [10]. In this technique the similarity among the known document and number of external documents is calculated. This approach stood first in competition. Frequent significant features were extracted by Petmanson Timo [11] such as nouns, verbs, punctuations, and beginning words of lines or sentences and they calculated the Matthews Correlation Coefficient (MCC) for each pair of extracted features by using Principle Component Analysis (PCA).

Beyond PAN corpora, author verifiers can be evaluated using other benchmark datasets. The emails of Enron's employees were released by the Federal Energy Regulatory Commission and contain more than 200,000 messages covering various topics [12]. Noticeably, the Enron corpus is appealing to researchers because it is not only coming from a real organization but also it is a large scale email collection. Therefore, it is a challenging base for several authorship verification methods to capture the style of their authors. However, in a lot of cases, the amount of information provided is small enough, as several of Enron e-mails contain only a few words (one or two). Moreover, these data are not able to be used directly by the users. The total set of the email messages included in this corpus, demands a prepossessing step. The Enron collection has been used by several author verification studies [13, 14]. Recently, the Enron authorship verification corpus (EAV) [14] was released and includes emails of 80 authors. A specific methodology has been proposed to transform this collection to a set of verification instances compatible with the guidelines of the PAN Datasets where there is exactly one unknown document per verification problem. However, the EAV dataset is not balanced and contains much more negative verification instances [14].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

III. CORPUS CHARACTERISTICS

Since 2009 several authorship analysis tasks have been organized including author attribution, author profiling, author clustering and author obfuscation in PAN competition. In three consecutive competitions of PAN (2013, 2014, and 2015) a shared task in authorship verification was conducted and attracted the participation of multiple research teams. PAN organizers built new benchmark corpora covering several languages (English, Dutch, Spanish and Greek), genres (essays, novels, reviews, newspaper articles) and provided an online experimentation framework for software submissions and evaluation [15]. In this work, the experiment performed with two datasets taken from the PAN competition Authorship Verification track. Table 1 depicts the features of the PAN competition 2015 dataset. The PAN competition 2014 dataset characteristics are depicted in Table 2.

.

Table 1: PAN Competition 2015 Dataset

Features	Training data	Testing data
Number of problems	100	500
Documents count	200	1000
Average number of known Documents	1.0	1.0
Average words per document	366	536

Table 2: PAN Competition 2014 Dataset

Features	Training data	Testing data
Number of problems	200	200
Documents count	729	718
Average number of known documents per problem	2.6	2.6
Average words per document	848	833.2

IV. PROPOSED APPROACH

In Authorship Verification approaches, the researchers extracted a variety of stylistic features like character based, lexical, syntactic features etc to distinguish the authors style of writing. By analyzing the Authorship Verification corpuses, we observed that the authors used same style of writing while they are writing different documents. We also observed that the authors used same set of words in their writings. So the set of words are important to find the similarity among the two writings. The word order is also an important observation in the writing style of authors. The same author used a combination of words in same sequence in their writings. So the word N-grams plays a significant role in the detection of similarity among the texts. The grammatical order of words also plays a vital role in the detection of the similarity among the texts. The Part Of Speech (POS) N-grams are used to find the combinations of grammatical order of words. Based on these observations we experimented with most frequent words, word and POS N-Grams. The proposed approach is displayed in Fig. 2.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

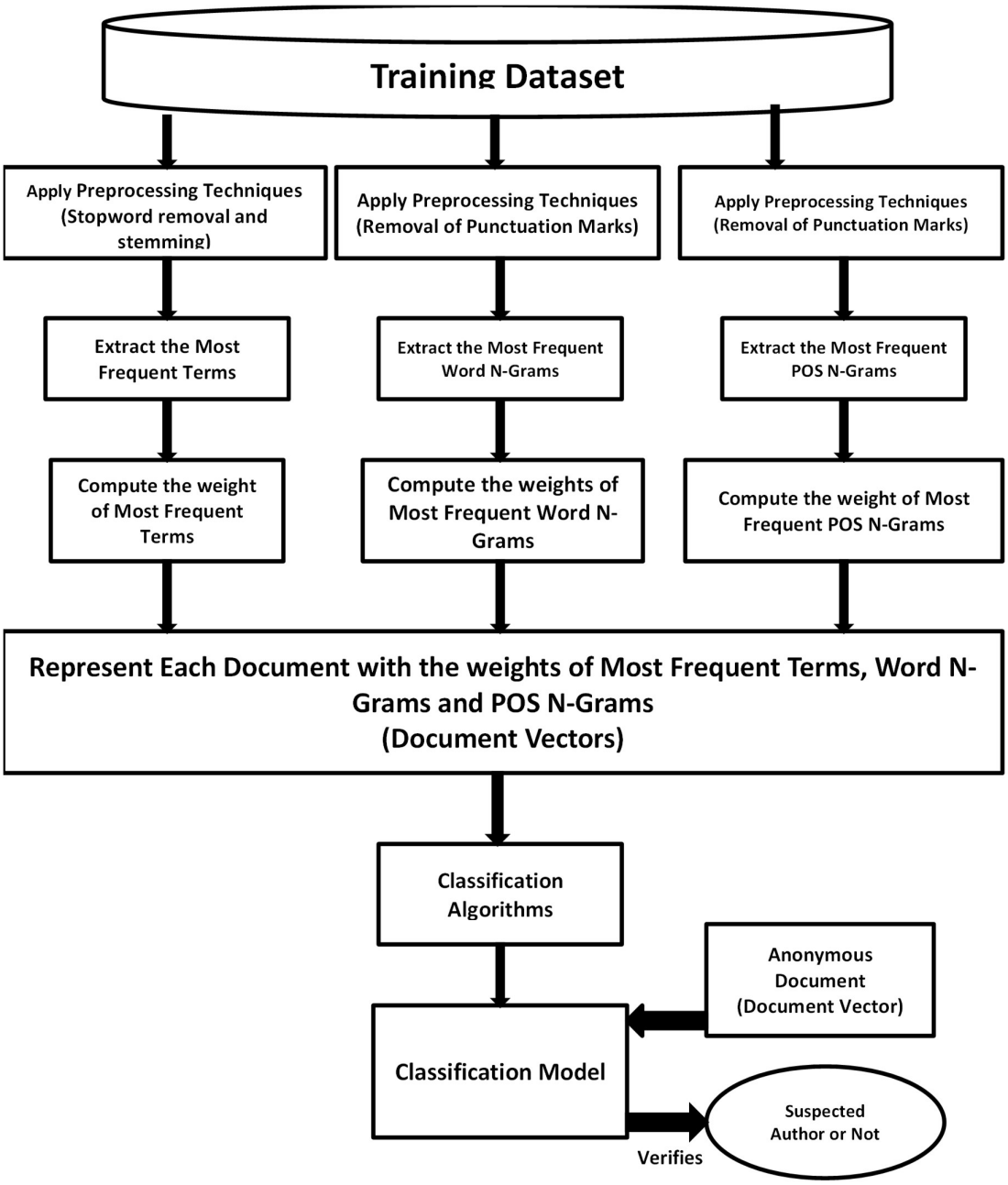

Fig. 2 The Proposed Approach

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

Initially, it is very common that many text preparation methods are used to remove information from documents and facilitate further processing. In the proposed approach, first we extract most frequent terms from the training corpus. Before extraction of terms, we apply preporcessing methods such as stopwords removal and stemming. Stop words are words that are removed before processing texts. In our case we remove words that are frequently occurring in texts and which also not carries any information on their own. Some common stop words in the English language are determiners, prepositions, articles, conjunctions etc. The stemming process task is to remove the suffixes. A prefix can also be added to a word, but this will often change the meaning of the word, so those will not be removed in the stemming process. In addition, word stemming aims at transforming word occurrences to their stems (e.g., all "play", "plays", "playing" have the same stem). The most commonly used stemmer which strips English words of its suffixes is the Porter stemmer algorithm. Next extract the most frequent word and POS N-Grams from the training corpus and considered as features. Before extraction of word and POS N-Grams, we apply the preprocessing technique of removal of punctuation marks.

After extracting the most frequent features from the training corpus, the next significant step is to represent documents as numerical vectors. One very common technique is simply represents a document by the weights of features. It is a simple and efficient approach since it does not require any deep linguistic analysis of documents. This method ignores both the context of words and semantic relationships between words. The context of words can be captured when phrases are considered. Usually, this is done by considering sequences of n words known as word n-grams. This process produces a high dimensional and sparse representation of documents. We identified 1000 most frequent terms, 500 most frequent word N-grams (N range from 1 to 3) and 500 most frequent POS N-Grams.

To compute the weight of the features, we used a Nonuniform Distributed Term Weight (NDTW) Measure. This measure is used [16] to compute the term weight that was extracted in the previous step. The equation (1) represents the NDTW measure.

$$W\left(f_i, D_k\right) = \log\left(TOFF_{fi}\right) - \sum_{k=1}^{m}\left(\frac{FF\left(f_i, D_k\right)}{TOFF_{fi}} \log\left(\frac{1+FF\left(f_i, D_k\right)}{1+TOFF_{fi}}\right)\right) \quad (1)$$

Where $\{f1, f_2, ...f_n\}$ is a set of features, $\{D_1, D_2, ..., D_m\}$ is a set of training documents in the corpus. $FF(f_i, D_k)$ is the feature f_i frequency in D_k document. $TOFF_{fi}$ represents the total occurrence of the feature f_i in all the documents of the corpus.

After representing the documents as vectors, these document vectors are passed to the machine learning algorithms. In this work, we experimented with three machine learning algorithms to create the classification model. This model is used to verify whether the test document is written by the original author not. The efficiency of the model is evaluated with the measure of accuracy. Accuracy is the ratio among the number of test documents correctly verified and the total test documents.

V. EXPERIMENTAL RESULTS

In this work, we experimented with three machine learning algorithms such as Naive Bayes Multinomial (NBM), Support Vector Machine (SVM) and Random Forest (RF). A most popular probability based model which uses Bayes theorem internally is called Naive Base Classifier (NBC). Three variants of NBC is available. Multinomial NBC is applied if the data is discrete, Gaussian NBC is used if the data is continuous and Bernoulli NBC. To classify the text Multinomial NBC is extensively used. Support Vector Machines (SVM) is a widely used algorithm capable to analyze high dimensional and sparse data. SVM can handle only binary classification problems by detecting the optimal separator (hyperplane) among two classes. To this end, it maximizes the margin between the hyperplane and the closest instances belonging to the two classes (the support vectors) in the training set. SVM can be extended to handle multi-class problems by building a series of binary classifiers and then combine them. SVM is very robust to outliers and avoids the local minima problem with a few parameters to tune. The main drawback of SVM is the long training time needed when large training data are available.

A decision tree is a simple algorithm that learns a set of IF-THEN rules forming a tree structure that can assign documents to classes. Generally, decision trees are fast and easy to interpret. However, they are negatively affected by

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

the sparseness of data and overfitting. On the other hand, random forests attempt to avoid such problems by building a rich set of trees, selecting a random subset of features and training instances in each step. The combination of these trees has been demonstrated to be a very competitive classification method with few parameters to tune.

Supervised machine learning algorithms do not provide optimal results without being properly tuned. In the majority of times, a classification model involves a set of parameters that have to be properly optimized. In addition, some parameters that relate to document representation should be also tuned (e.g., the type of features, the removal of certain terms and the size of the vocabulary). To this end, the optimization of hyperparameters of the model can be performed using a validation set. We used 2000 features combination of most frequent terms, word and POS N-Grams for representing the document vector. The accuracies of three machine learning algorithms on PAN 2014 and 2015 corpuses are depicted in table 3.

Table 3. The Accuracies of Authorship Verification

	Naive Bayes Multinomial	Support Vector Machine	Random Forest
PAN 2015 Corpus	81.21	87.78	89.45
PAN 2014 Corpus	84.38	88.56	92.82

The Random Forest classifier obtained good accuracies of 89.45% and 92.82% for Authorship Verification on PAN 2015 and PAN 2014 corpuses respectively. The RF classifier shows good performance than other two classifiers. The accuracy is good for PAN 2014 corpus because the training data and test data almost equal. The accuracy for PAN 2015 corpus is less because the test dataset contains more number of documents when compared with training data. It was observed that our proposed approach achieved best accuracies for Authorship Verification.

VI. CONCLUSION

In this work, a new approach was proposed for Authorship Verification. In this approach, we used 2000 feature combinations of most frequent terms, word and POS N-Grams for document vector representation. The NDTW measure is used to calculate the feature weight in the document vector representation. The proposed approach performance is evaluated by using three machine learning algorithms. The random forest classifier achieved best accuracies when compared with other classifiers. We experimented on the corpuses of PAN competition 2014 and 2015.

In future, we are planning to propose a new feature set and a new feature weight measure for computing weights of the features and to improve the accuracy of Authorship Verification. It was also planning to apply our approach to multiple languages corpuses.

REFERENCES

1. Moshe Koppel, Jonathan Schler, and Shlomo Argamon, “Computational methods in authorship attribution”, Journal of the American Society for Information Science and Technology, 60(1):9–26, 2009
2. Nektaria Potha, Efstathios Stamatatos, “Improved algorithms for extrinsic author verification”, Knowledge and Information Systems, OCT 2019.
3. Esteban Castillo, Ofelia Cervantes and Darnes Vilarino, “Authorship Verification using a Graph Knowledge Discovery Approach”, Journal of Intelligent & Fuzzy Systems 36 (2019) 6075–6087
4. Darnes Vilariño, David Pinto, Helena Gómez, Saúl León and Esteban Castillo, “Lexical-Syntactic and Graph-Based Features for Authorship Verification”, Proceedings of CLEF 2013 Evaluation Labs, 2013
5. Victoria Bobicev, “Authorship Detection with PPM”, Proceedings of CLEF 2013 Evaluation Labs, 2013
6. Cor J. Veenman and Zhenshi Li, “Authorship Verification with Compression Features”, Proceedings of CLEF 2013 Evaluation Labs, 2013
7. Michiel van Dam, “A Basic Character N-gram Approach to Authorship Verification”, Proceedings of CLEF 2013 Evaluation Labs, 2013
8. Alberto Bartoli, Alex Dagri, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao, “An Author Verification Approach Based on Differential Features”, Proceedings of CLEF 2013 Evaluation Labs, 2015
9. Vanessa Wei Feng and Graeme Hirst, “Authorship Verification with Entity Coherence and Other Rich Linguistic Features”, Proceedings of CLEF 2013 Evaluation Labs, 2013
10. Shachar Seidman, “Authorship Verification Using the Impostors Method”, Proceedings of CLEF 2013 Evaluation Labs, 2013
11. Timo Petmanson,” Authorship identification using correlations of frequent features”, Proceedings of CLEF 2013 Evaluation Labs, 2013

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 2, February 2020

12. Bryan Klimt and Yiming Yang. "Introducing the Enron Corpus." In: CEAS. 2004.
13. Marcelo Luiz Brocardo, Issa Traore, and IsaacWoungang. "Authorship verification of e-mail and tweet messages applied for continuous authentication". In: Journal of Computer and System Sciences (2015), pp. 1429–1440.
14. Oren Halvani, Lukas Graner, and Inna Vogel. "Authorship Verification in the Absence of Explicit Features and Thresholds". In: European Conference on Information Retrieval. Springer. 2018, pp. 454–465.
15. https://pan.webis.de/clef15/pan15-web/author-identification.html
16. Dennis, S.F., "The Design and Testing of a Fully Automated Indexing-Searching System for Documents Consisting of Expository Text",in Informational Retrieval: A Critical review, g. Schecter, editor,Thompson Book Company, Washington D.C., 1967, pages 67-94