International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

ELSEVIER SSRN

2020 IJIEMR.Personal use of this material is permitted. Permission from IJIEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper, all copy right is authenticated to Paper Authors

IJIEMR Transactions, online available on 27th Nov 2020. Link :http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-12

DOI: 10.48047/IJIEMR/V09/I12/83

Title: **A Novel Approach for Authorship Verification using Similarity Measure**
Volume 09, Issue 12, Pages: 437-445
Paper Authors

Dr. T. Raghunadha Reddy

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A Novel Approach for Authorship Verification using Similarity Measure

Dr. T. Raghunadha Reddy

Associate Professor, Matrusri Engineering College, Saidabad, Hyderabad
raghu.sas@gmail.com

Abstract: Authorship verification is a task of identifying whether two text documents are written by the same author or not by evaluating the veracity and authenticity of writings. Authorship Verification is used in various applications such as analysis of anonymous emails for forensic investigations, verification of historical literature, continuous authentication in cyber-security and detection of changes in writing styles. The Authorship Verification problem primarily depends on the similarity among the documents. In this work, a new approach is proposed based on the similarity between the known documents of the author and anonymous document. In this approach, extract the most frequent terms from the dataset for document vector representation. These most frequent terms are used to represent the train and test documents. The term weight measure is used to represent the term value in the vector representation. The Cosine similarity measure is used to determine the similarity among the training and test document. Based on the threshold value of similarity score, the author of a test document is verified whether the test document is written by the suspected author or not. The PAN competition 2014 Authorship Verification dataset is used in this experiment. The proposed approach achieved best results for Authorship verification when compared with various solutions proposed in this domain.

Key Words: Authorship Verification, Term Weight Measure, Similarity Measure, Accuracy, PAN Competition

1. INTRODUCTION

The textual data in the web is increasing tremendously through different sources like Blogs, Forums, Social Media, Reviews, Twitter Tweets etc. along with the authorship problems also increasing with the data. The Authorship Analysis is one research area attracted by the researchers to identify the author of an anonymous document. Authorship Verification (AV) and Authorship Attribution (AA) are related subfields within the larger field of Authorship Analysis [1]. The goal of AV is to recognize whether two separate text documents come from the same author or not [2]. AV seeks to determine if two different writings are from the same author, while authorship attribution's goal is to identify who wrote a given writing [3]. Both of these fields rely on the extraction of useful text features for discriminating between different authors. Traditionally, researchers have relied on linguistic style or

stylometric features, such as the counts and frequency of function words, average length of sentences, part-of-speech, characters, punctuation, whitespace usage and other low-level features.

Authorship verification is an active research area of computational linguistics that can be expressed as a fundamental question of stylometry, namely whether or not two texts are written by one and the same author. AV is traditionally performed by linguists who aim to uncover the authorship of anonymously written texts by inferring author-specific characteristics from the texts. Such characteristics are represented by so-called linguistic features. They are derived from an analysis of errors, textual idiosyncrasies and stylistic patterns. Nowadays, there are many unknown authorship letters such as email Fraud, suicide, and terrorism for which it is necessary to verify the authorship. Currently, to solve authorship verification problems there are different approaches such as distance based, machine learning based, and impostors which achieved great results in previous PAN tasks.

Automated (machine-learning-based) systems have traditionally relied on stylometric features. Stylometric features tend to rely largely on linguistically motivated/inspired metrics. The disadvantage of stylometric features is that their reliability is typically diminished when applied to texts with large topical variations. Most machine learning models that solve authorship attribution problems are author-specific that means the models are trained on a known set of authors. Authorship verification problems create a machine learning model that is generic, and thus applicable to any two given documents, even when a specific author is not known. Deep learning systems, on the other hand, can be developed to automatically learn neural features in an end-to-end manner. While these features can be learned in such a way that they are largely insensitive to the topic, on the negative side, they are generally not linguistically interpretable.

As far as we know, there are few implementations of deep learning based methods to solve this problem. The main differences between traditional machine learning and a deep learning application from the development perspective are the hardware requirements, the amount of data to handle, and the ability to handle raw data as input. Conventional machine learning techniques need extensive feature engineering to transform the raw data in order to use it for training and testing a model. Existing AV algorithms can be taxonomically grouped based on their design and characteristics such as instance versus profile-based paradigms, intrinsic versus extrinsic methods or unary versus binary classification.

In this work, a new approach is proposed wherein similarity measure is used to determine the similarity among the test document and training document. The similarity score is used to check whether the new document is written by the suspected author or not. Most frequent terms are extracted to represent the document vectors for computing the similarity score. A term weight measure is used to calculate the vector value of a term in the vector

representation. Accuracy measure is used to know the number of test documents is verified correctly by the approach. The PAN 2014 competition AV dataset is used for this experiment.

This paper is planned in 6 sections. The literature survey of Authorship Verification is presented in section 2. Section 3 shows the dataset characteristics. Section 4 explains the proposed approach along with similarity measure and term weight measure. Section 5 expresses the experimental results of the proposed approach with different most frequent terms used to represent the document vector. Section 6 concludes this paper with conclusions and future scope.

2. Literature Survey

The AV task is used in various applications to solve the problems of authorship of documents. The researchers proposed different solutions for AV task based on machine learning and deep learning techniques. Emir Araujo-Pino et al., presented [4] an approach to the Authorship Verification task at PAN 2020. The Authorship Verification task is comparing two documents and automatically determines if they are written by the same author or not. They introduced a Siamese network architecture that is trained on character n-grams of the document pairs to be compared. They experimented with different hyper-parameters when training the model on a large and a small dataset. The hyper-parameters they tune are the minimum document frequency and the maximum document frequency. The model achieved best overall evaluation score of 0.804 which is the average of AUC, c@1, f_05_u, and F1 scores. Wei Feng Vanessa et al., developed [5] an unmasking approach to improve the quality of a feature when training the classifiers. They used combination of Stylometry and coherent features of 399, 538 and 568 features for Spanish, English and Greek languages respectively. They observed that the AV results are good for Spanish and English documents but the results are poor for Greek language documents.

Benedikt Boenninghoff et al., presented [6] a hierarchical fusion of two well-known approaches into a single end-to-end learning procedure. A deep metric learning framework at the bottom aims to learn a pseudo-metric that maps a document of variable length onto a fixed-sized feature vector. At the top, they incorporated a probabilistic layer to perform Bayes factor scoring in the learned metric space. They also provide text preprocessing strategies to deal with the cross-topic issue. The proposed method achieved excellent overall performance scores, outperforming all other systems that participated in the PAN 2020 Authorship Verification Task, in both the small dataset challenge as well as the large dataset challenge. Victoria Bobicev developed [7] a system named as statistical n-gram model based on prediction by partial matching (PPM). This system is used for detecting the author automatically when a corpus is given with text documents of known authors. The training corpus consists of documents of 30 authors. Each author document contains 100 posts and the corpus maintains uniformity by normalizing the size of document to 150-200 words. It was

observed from the results, the accuracy is not increased when the document lengths are increased.

Łukasz Gagała proposed [8] an approach to authorship verification (AV) task in PAN 2020 competition. They proposed a data compression method based on the widespread Prediction by Partial Matching (PPM) algorithm extended with Context-free Grammar character pre-processing. The fundamental principle of PPM is a context-dependent prediction of each subsequent character in a text string. The context is given by a window of preceding characters with a predefined length. The most frequent in-word bigrams in each text are replaced by special symbols and accepted by a modified version of PPM algorithm. For similarity measure between text samples, they selected Compression-Based Cosine (CBC) that performs slightly better than alternative measures. Darnes Vilarino et al., experimented [9] with lexical, vector graph based features and syntactic features for representing the documents. Subdue tool was used for extracting the graph based features. SVM is used to develop the model for classification. The results showed that the runtime complexity is high compared with other solutions in the competition.

Oren Halvani et al., proposed [10] simple effective distance-based authorship verification (AV) approach called TAVeer to address the AV shared task in the PAN 2020 competition. This approach considers only topic-agnostic feature categories based on punctuation marks, function words, contractions, transitional phrases as well as several subclasses of verbs and adverbs in its classification decision. The core of TAVeer is a distance function like Manhattan metric which is combined with a thresholding procedure and this combination act as the underlying classifier. On the official test set, their approach was ranked third out of all submitted approaches.

Catherine Ikae described [11] a simple model that performs Authorship Verification based on a Labbé similarity. They employed the most frequent tokens such as words and punctuation symbols as features from each author after including the most frequent ones of a given language. Such a representation strategy is based on words used frequently by a given author but not belonging to the most frequent in the English language. Evaluation based of authorship verification task with a rather small set of features shows an overall performance with the small dataset of F1= 0.705 and AUC = 0.840. Michiel van Dam used [12] Common N-Gram (CNG) method which is a profile based approach. In CNG method implementation, the character n-grams are used for document representation. They observed that this method achieved good results for English documents and Spanish documents, but it failed for Greek documents.

Janith Weerasinghe et al., described [13] an approach to create a machine learning model for the PAN 2020 AV Task. For each document pair, they extracted stylometric features from the documents and used the absolute difference between the feature vectors as input to our classifier. They created two models such as Logistic

Regression Model and Neural Network based model. Logistic Regression Model trained on a small dataset and a Neural Network based model trained on the large dataset. These models achieved AUCs of 0.939 and 0.953 on the small and large datasets respectively. Their approach got 2nd place in the competition for both datasets.

Juanita Ordoñez et al., described [14] a neural network that learns useful features for authorship verification from fanfiction texts and their corresponding fandoms. The proposed system used the Longformer which is a variant of popular transformer models that is pre-trained on large amounts of text. This model combines global self-attention and local self-attention to enable efficient processing of long text inputs. The pre-trained Longformer model is augmented with additional fully-connected layers and fine-tunes it to learn features that are useful for author verification. Finally, they incorporated fandom information through the use of a multi-task loss function that optimizes for both authorship verification and topic correspondence, allowing it to learn useful fandom features for author verification indirectly. On a held-out subset of the PAN-provided "large training" set, the Longformer-based system attained a 0.963 overall verification score. The proposed system attained a 0.685 overall score on the official PAN test set.

3. Dataset Characteristics

In this work, the experiment conducted with the dataset of Authorship Verification track of PAN 2014 competition. Table 1 displays the details of the PAN competition 2014 dataset.

Table 2: PAN Competition 2014 Dataset

Features	Training data	Testing data
Authors count	100	100
Documents count	500	100
size of Vocabulary	41583	12764
Documents count per author	5	1
Average words per sentence	25	21
Average words per document	1135	1121

The Accuracy measure is used for performance evaluation of the proposed approach. In this context, accuracy is the number of test documents are correctly verified from the set of test documents.

4. Proposed Approach

The AV task is to identify whether a pair of text documents are authored by the same person or not. In this work, this author verification problem is viewed as a similarity detection problem, where for an unknown document D_U and a known document D_K it has to be determined whether both were written by the same author A. The focus of the similarity determination in the context of AV is on the writing style and not on other factors such as the topic or genre. The similarity among the documents decides whether a suspected document written by the particular author or not. The similarity score among the training

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

documents of an author A and test document is less than threshold value then the document was written by that author A otherwise the document was not written by the author A. several factors are influencing the performance of the Authorship Verification such as feature sets, classification methods, text length, text topic and text language.

The proposed approach for authorship verification is displayed in Fig. 1. In this approach, first the dataset is cleaned by applying two pre-processing techniques such as stop words elimination and stemming. Then, frequent terms are extracted from the cleaned data based on their frequency. These features are used to represent the training documents and test documents as vectors. The term value is determined by using term weight measure. The similarity is determined between training document vectors of single author and test document vectors by using Cosine Similarity Measure (CSM). Based on the similarity scores, the author of a test document was verified. The proposed approach performance mainly depends on the term weight measure and similarity measure. The Normalization based Term Weight (NTW) Measure is used in this work to avoid the problem of different sized documents while assigning weight to the terms in the vector representation. The CSM is used to determine the similarity among the train and test documents.

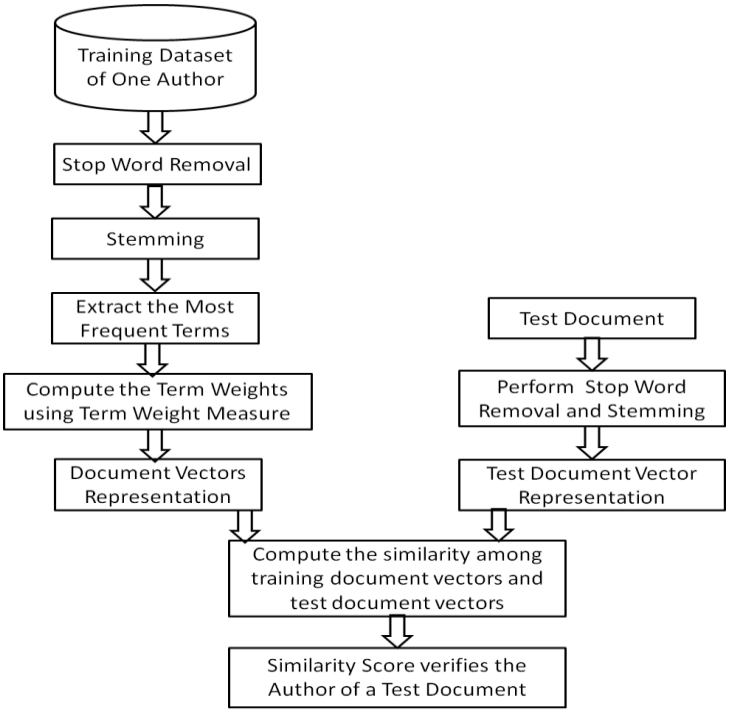

Fig.1 The Architecture of our Proposed Approach

The next subsections explain the similarity measure and term weight measure used in this approach.

4.1 Cosine Similarity Measure (CSM)

The similarity measures are used to compute the degree of similarity among a pair of textual objects. Similarity measures are very crucial in several information processing applications like text mining and document clustering. In general, the similarity measures were used popularly in the area of information retrieval domain to calculate the similarity among the query and documents, similarity among queries, similarity among documents. Similarity measures were also used to assign the ranks to the documents based on the amount of similarity among the query and document. In this work, the experiment conducted with CSM.

CSM is used for computing the similarity value among the train and test document

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

vectors [15]. The CSM measure is represented in equation (1).

$$CSM(D_U, D_K) = \frac{\sum_{i=1}^{m} W(T_i, D_U) \times W(T_i, D_K)}{\sqrt{\sum_{i=1}^{m} W(T_i, D_U)^2} \times \sqrt{\sum_{i=1}^{m} W(T_i, D_K)^2}} \quad (1)$$

Where, $W(T_i, D_U)$ and $W(T_i, D_K)$ are the term T_i weights in unknown document D_U and known document D_K respectively.

CSM determines the angle among the two document vectors. The angle is small when two vectors are close and the angle is large when two vectors are distant. The Cosine angles vary from 0 to 180 degrees for values +1 to -1 respectively. The documents are similar when CSM value is 1 and the documents are dissimilar when CSM value is 0.

4.2 Normalization based Term Weight (NTW) Measure

Term weighting is a process used to rank each term in a piece of text and determine each term's level of importance to the document. As a result, documents can be easily classified. For example, term weighting is used in information retrieval to retrieve the most relevant documents. In this work, we used a term weighting scheme namely Normalization based Term Weight (NTW) to compute the weight of terms.

In Authorship Verification technique, the number of documents in dataset influences the accuracy of the authorship verification. If the document contains more text, it was easy for the approaches of Authorship Verification to distinguish the authors writing styles. For small sized documents, it was difficult to identify the importance of terms in a document. The Normalization based Term Weight (NTW) measure is one such unsupervised term weight measure to analyse small sized documents and assign suitable weight to the terms [16]. Equation (2) shows the NTW measure. This measure considers the information of the terms with in a same class of documents.

$$W(T_i, D_k) = \sum_{k=1}^{m} \frac{(1+\log(TF_i)) / (1+\log(AVGTF_i))}{(1-slope) * AVGUT_k + slope * UT_k} \quad (2)$$

Where, TF_i is the term T_i frequency in document D_k, UT_k is the count of unique terms in document D_k, $AVGTF_i$ is the ratio among TF_i and term count in document D_k, $AVGUT_k$ is a ratio of unique terms count to total number of terms in document D_k. The researchers obtained good results when slope value is set to 0.2.

5. Experimental Results

In this work, we proposed a new approach for AV problem based on term weight measure and similarity measure. We experimented with 4000 most frequent terms for document vectors representation. In every iteration, the number of terms is increased by 1000 for document vectors representation.

Table 2: The accuracies of proposed approach for Authorship verification

Number of Terms used to represent the document vector	Accuracy
1000	83.79

2000	87.58
3000	89.12
4000	92.37

The proposed approach achieved an accuracy of 92.37% for Authorship Verification when 4000 most frequent terms are used for document vectors representation. It was found that the accuracy was increased when the number of terms is increased for representing the document vectors.

6 Conclusion

In this work, we proposed a new approach based on the idea of similarity measures and term weight measures. The AV is a technique of verifying the author of an unknown document by analysing the documents of a single author. The training documents and test documents are represented as vector with the features of most frequent terms. The vector values are computed by using a suitable term weight measure. In this work, some important step was taken toward developing a robust authorship verification approach that works on different lengths of text by using normalization based term weight measure. A similarity measure was used in this experiment to compute the similarity among the training documents of author and test document. The similarity measure attained best accuracy value of 92.37% for authorship verification.

In our future work, we are planned to implement the deep learning techniques in Authorship Verification. It is also planned to propose a novel similarity measure to enhance the accuracy of authorship Verification.

References

1. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Survey on Author Profiling Techniques", International Journal of Applied Engineering Research, March 2016, 11 (5), pp. 3092-3102.
2. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology 65(1), 178–187 (2014).
3. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. J. Assoc. Inf. Sci. Technol. (2009)
4. Emir Araujo-Pino, Helena Gómez-Adorno, and Gibran Fuentes-Pineda, "Siamese Network applied to Authorship Verification", CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.
5. Vanessa Wei Feng and Graeme Hirst, "Authorship Verification with Entity Coherence and Other Rich Linguistic Features", Proceedings of CLEF 2013 Evaluation Labs, 2013
6. Benedikt Boenninghoff, Julian Rupp, Robert M. Nickel, and Dorothea Kolossa1, "Deep Bayes Factor Scoring for Authorship Verification", CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.
7. Victoria Bobicev, "Authorship Detection with PPM", Proceedings of CLEF 2013 Evaluation Labs, 2013

8. Łukasz Gagała, "Authorship Verification with Prediction by Partial Matching and Context-free Grammar", CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.
9. 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
10. Oren Halvani, Lukas Graner, and Roey Regev, "Cross-Domain Authorship Verification Based on Topic Agnostic Features", CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.
11. Catherine Ikae, "UniNE at PAN-CLEF 2020: Author Verification", CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.
12. Michiel van Dam, "A Basic Character N-gram Approach to Authorship Verification", Proceedings of CLEF 2013 Evaluation Labs, 2013
13. Janith Weerasinghe and Rachel Greenstadt, "Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification", CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.
14. Juanita Ordoñez, Rafael Rivera Soto, and Barry Y. Chen, "Will Longformers PAN Out for Authorship Verification?", CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.
15. 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.
16. M. Sreenivas, Raghunadha Reddy T, Vishnu Vardhan B, "A Novel Document Representation Approach for Authorship Attribution", International Journal of Intelligent Engineering and Systems, 11 (3), pp. 261-270, DEC 2018.