ICIC Express Letters
Part B: Applications
Volume **12**, Number **11**, November **2021**

 ISSN 2185-2766
pp. **1011–1017**

A MALAYSIAN SCHOLAR IDENTIFICATION MODEL BASED ON WORD2VEC-BASED-STYLOMETRY COMPUTATIONAL APPROACH

Muhammad Syafiq Mohd Pozi[1], Muhamad Azwan Abd Rahman[2] and Abdul Rafiez Abdul Raziff[3]

[1]School of Computing
Universiti Utara Malaysia
Sintok 06010, Kedah, Malaysia
syafiq.pozi@uum.edu.my

[2]Institute of Malaysian and International Studies
Universiti Kebangsaan Malaysia
UKM Bangi 43600, Selangor, Malaysia
azwanrahman@ukm.edu.my

[3]School of Digital Transformation
Veritas University College
Petaling Jaya 46200, Selangor, Malaysia
abdul.r@veritas.edu.my

Received March 2021; accepted June 2021

Abstract. *Team matching is considered one of the essential processes among many people to find the best connectivity between unique personnel. Finding the best of both parties is crucial, especially in making the best value worthy for time, money, and other resources. For example, in academic collaboration, in either intradisciplinary or interdisciplinary context, finding the right collaboration partner is an essential process in publishing a high-quality collaborative-based output. Hence, in this paper, we proposed a new stylometry model in a data mining framework to identify the best researcher based on their research publications. Based on our experiments, after comparing with several machine learning classifiers, we found that our framework combined with neural network classifier managed to give a good classification accuracy in determining the best authors based on past publications with 78.38% classification accuracy.*
Keywords: Stylometric, Word2Vec, Research interest, Collaboration

1. **Introduction.** National education, especially the higher learning institution, plays an important role in shaping the society through knowledge empowerment and community services [3]. Almost all developed countries spent high percentages of the annual budget in their respective education sector. Specifically, based on the World Bank, Malaysia spent about 4.535% of GDP on education in 2018. Despite that, many critics said that local universities failed to provide sufficient education to their graduates resulting in increasing youth unemployment rate in the country[1].

As one of the main university products is to supply capable high skilled workers to the industry, the trend that has been shown nowadays is really vulnerable. It seems that university graduates have either no desire to get a job or simply the jobs are not there. One reason is due to job mismatch between the graduates' skills and the job supply. With the continuous industry revolution that keeps revolutionizing in accelerating rate, job mismatch seems to be the main issue to university vision and mission while at the same time, being constrained by industry needs [5].

DOI: 10.24507/icicelb.12.11.1011

[1]https://www.statista.com/statistics/812222/youth-unemployment-rate-in-malaysia/

Hence, in this paper, we are trying to reduce the degree of this vulnerability by proposing and implementing a team matching framework based on proof-of-works. Here, proof-of-works could be many things, such as past projects, patent, publications, programming codes, and legal proceedings. For example, a group of lawyers could be formed based on how many cases they won, which can be obtained from legal proceedings. Our paper, however, will be focusing on how to match researchers based on similar interest.

Finding the best research collaborators in any specific or multidisciplinary domains is very crucial in the academic world. Alongside with the current COVID-19 pandemic, the needs of the research collaboration are very high with many organizations offering research grants in exploring the insights, cure, prevention and mitigation across many fields as this pandemic affects the people in every sector all over the world. Some have said that high-quality research output is a product of collaborative work, consisting of multiple experts from various research domains [9]. Thus, the importance of having a research team that can synergize with each other is a must in contributing to a good project.

Author identification also known as authorship attribution or authorship verification is a process of identifying the true author from a group of candidate authors based on their writing sample [18]. Each author had their own writing style, just like their fingerprint, which is unique [8]. Hence, based on the writing style, one can determine whether the document is written by the one particular author (or ghostwriter) [22], multiple authors [18] or no authors at all (machine-generated text) [12].

Specifically, the method of measuring the similarity between author writing styles with publication is called as stylometric. In 1851, stylometric was first suggested and developed by Augustus De Morgan (in his letter to a friend) based on word length to resolve the author disputes (as described in his memoir [4]). Fast forward to second pandemic era [2], stylometric has become more crucial in detecting fake news [1], hate speech [16] and any other disinformation services.

There are two main methods in performing author identification task through stylometric measurement. The first one is through syntactic features. Syntactic features are work based on sentences structure [19]. Examples of the syntactic features are frequency of function words, part of speech (POS), morphology of words and retaining stop-words. For example, in [21], the author is represented through the most frequent character n-grams of the training corpus, representing the text sample of particular author. Not limited to English, especially in under-resourced language [7, 20, 24], specific syntactic features of those languages are used to represent the author documents.

The second one is through word embedding. Word embedding such as Word2Vec [14], FastText [25], including contextual word embedding such as BERT [27] and XLNet [26], has been used in author identification task with good performance results over syntactic approaches. This is because, word embedding method provides a semantic meaning of word (or context) with other words in corpus. For example, this semantic value can be used to accurately distinguish two different authors (or more), as their writing can be represented in form of semantic changes between two authors.

Two main questions that guide our research are as follows:

1) How to model the characteristics between publications and individual author, in terms of data representation, and;
2) What is the best machine learning classifier that can approximate those characteristics with high accuracy metric.

The purpose of this paper is to investigate the viability of the data mining approach in matching the best researcher based on past publications. This is done by approximating a model that can accurately correlate each author with every publication that the author

has ever published, in which those correlations, validities are measured and compared against several machine learning classifiers.

The remainder of this paper is structured as follows. In Section 2, we introduce our proposed Malaysian scholar identification model within the research interest matching framework. Next, in Section 3, we analyze the results of this study. Finally, we conclude the paper and outline future directions in Section 4.

2. **Research Interest Matching Framework.** The framework consists of three components. The first component is data collection and preparation. The second component is text embedding and the third component is the classification task. This section discusses these components in detail. For reference, the framework flowchart is also illustrated in Figure 1.

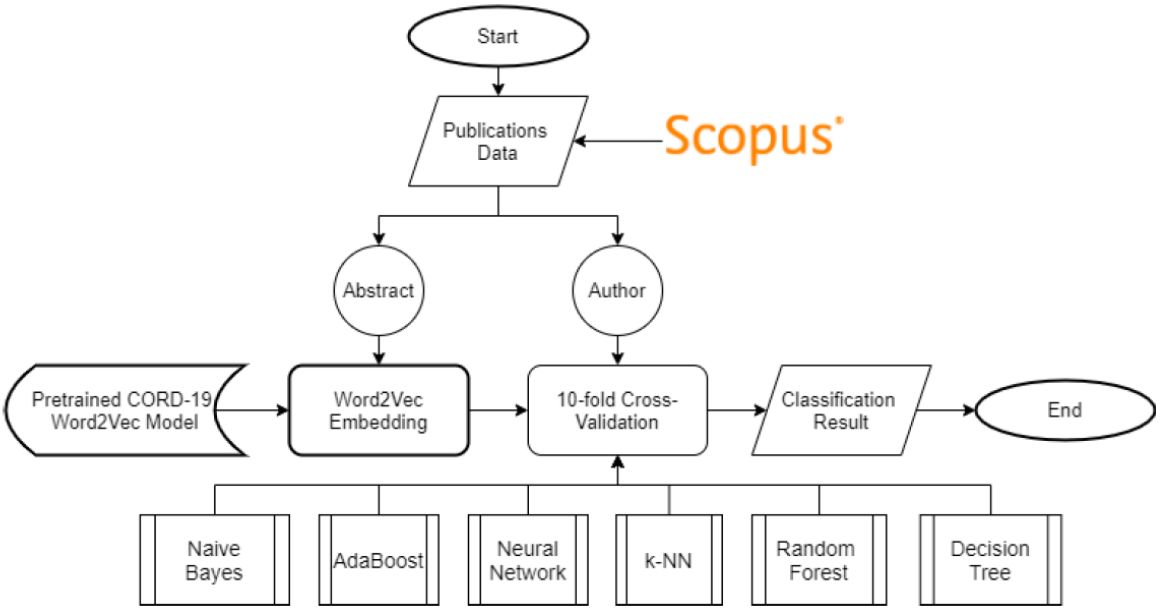

FIGURE 1. The experiment flowchart, starting from data collection, text embedding and finally classification task, validated over 10-fold cross-validation

2.1. **Data collection and preparation.** We used Scopus[2] repository as the main source of data for our proposed framework. The collected dataset consists of every publication indexed in Scopus from 2008 to 2018, with a total of 245,618 publications. Every publication is characterized by its publication title, abstract, keywords, authors and authors' affiliation. Figure 2 shows the non-cumulative distribution of the publications, published in each year, spanning from 2008 to 2018, that are affiliated with Malaysian research institutions.

From Figure 2, we can see that publications in 2018 are approaching eight times increment in terms of publications that have been published in 2018 compared to 2008. There are many reasons of this behaviour, though the most logical reason could be the increasing number of academicians in Malaysia due to various research incentives available in Malaysia[3]. Nevertheless, based on the collected data, we structured the data into a set of features, such that

$$x = \textbf{abstract of each publication}$$
$$y = \textbf{author of that publication}$$

[2]https://www.scopus.com/
[3]https://tinyurl.com/y7gx7kug

Here x consists of a set of keywords that made of that abstract while y is the unique identifier of the author of that publication. In addition, we only selected the top 10 most published authors to avoid any computing and modelling complexity that might occur due to the large size of publication records. With that, our new samples now consist of 6,688 publications.

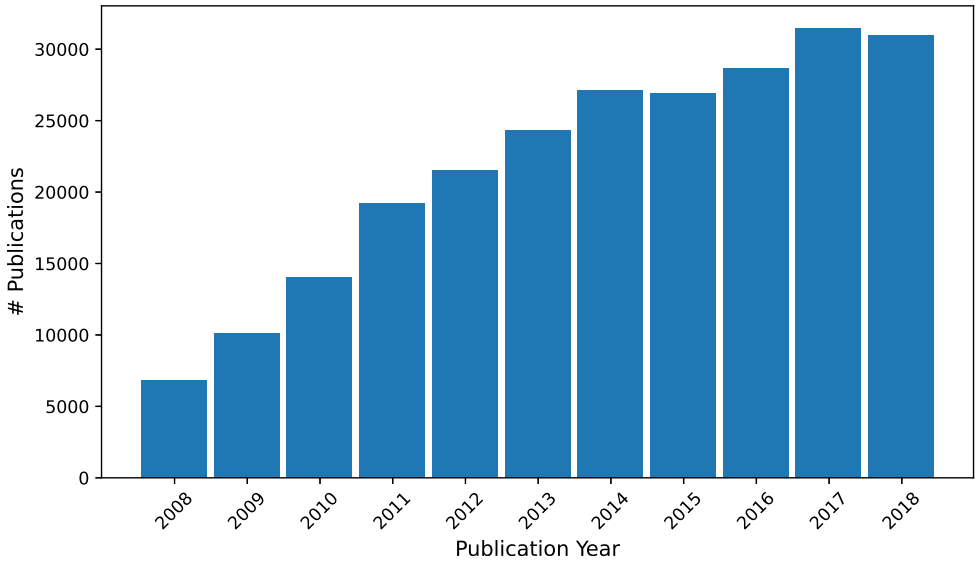

FIGURE 2. The distribution of yearly based publications, affiliated with Malaysian institutions that are indexed in Scopus database, from 2008 to 2018

2.2. **Text embedding.** Referring to Figure 1, the next step is to vectorize the abstract of each publication. The vectorization process is done by counting its co-occurrence counts with other words to represent the meaning of a term [17]. This method, however, suffers from the sparse matrix due to vocabulary expansion. The other approach is to learn the dense representation of each word in the form of neural embeddings that form a vector space [10]. This is more appealing as the only vector of a word with respect to other words is stored in memory. Here, we used Word2Vec for creating word embedding on each abstract.

For this task, we used a CORD-19 pre-trained Word2Vec model [11] to derive the Word2Vec vector of size 300, every word in the abstract. This is because CORD-19 dataset consists of thousands of scholarly publications spanning from 1920 to 2020. With a large size of the corpus, we assume that it is improbable for a word not to have a vector representation in the nearest future.

Hence, for every Word2Vec representation of every word in the abstract, we compute the mean of all vectors, such as follows:

$$x_a = \frac{1}{n} \sum_{i=1}^{n} f(w_i) \tag{1}$$

where x_a in Equation (1) is the vector that represents each abstract, n is the total words in the abstract and $f(w_i)$ is the Word2Vec vector representation of word w_i that formed the abstract.

2.3. **Classification task.** The data now consists of abstract (in textual format) of each publication, a single Word2Vec 300-dimension vector size that represents the publication's abstract, and the author that associates with the publication. Hence, for the classification

task, we removed the textual abstract from the data, and only make use of the Word2Vec vector as the features for classification task and author as the respective class, such as follows:

$$\{x_1, x_2, \ldots, x_{300}\} = \textbf{abstract representation in Word2Vec}$$
$$y = \textbf{author of that publication}$$

Then, the selected data is modelled through 6 different machine learning classifiers, that is, Neural Network, k-Nearest Neighbour (k-NN), Random Forest, AdaBoost, Naive Bayes and Decision Tree. The experimentation flowchart followed in this paper is illustrated in Figure 1. The classification performance of each classifier is compared through 10-fold cross-validation, in which the classification results are tabulated in Table 1.

3. **Results and Analysis.** In this section, we show the experimentation results based on 6 different machine learning classifiers and perform comparison analysis between the performance of each classifier based on AUC-ROC curve [6], classification accuracy, F1 score, Precision and Recall metrics. The classification performance of each classifier based on those metrics is tabulated in Table 1.

TABLE 1. A 10-fold cross-validation results obtained from 6 different classifiers, ranked from classifier with highest AUC value to classifier with lowest AUC value

Model	AUC	CA (%)	F1	Precision	Recall
Neural Network	0.9764	0.7838 (78.38%)	0.7836	0.7837	0.7838
k-NN	0.9587	0.7347 (73.47%)	0.7309	0.7305	0.7347
Random Forest	0.9374	0.6876 (68.76%)	0.6858	0.6867	0.6876
AdaBoost	0.9364	0.7171 (71.71%)	0.7152	0.7196	0.7171
Naive Bayes	0.9162	0.6036 (60.36%)	0.5986	0.6045	0.6036
Decision Tree	0.8198	0.5970 (59.70%)	0.5976	0.5997	0.5970

Through 10-fold cross-validation, Table 1 shows that all tested machine learning classifiers managed to give the higher area under ROC curve, with the highest value (0.9764) obtained from Neural Network, and the lowest value (0.8198) obtained from Decision Tree. It means that each classification model has a good capability in distinguishing between distinct author based on the Word2Vec vectors that represent each author.

Regardless, looking closely at Table 1, Neural Network is the one that provides good and consistent classification performance compared to other machine learning classifiers, in every tested metric. In addition to that, for every class in the dataset, Neural Network together with k-NN, and AdaBoost gives a balanced classification performance on each class while Random Forest, Naive Bayes and Decision Tree give an imbalanced classification performance between classes.

From the experiment, it is noticeable that our framework together with Neural Network produced the highest score and accuracy among other classifiers. This is because the Neural Network is capable of learning from very high dimensional data. This situation is totally opposite for Decision Tree as it is more suitable for dataset that has very few features in hand, due to the nature of decision tree that is likely to produce an overfitted solution when dealing with high-dimensional dataset [23].

Regardless, there is much improvement that can be done on this framework. For example, we only test on top 10 authors from our dataset out of more than 1000 authors in the dataset, as modelling more than ten authors requires more data and more computing resources. In addition to that, when dealing with multiple classes, various fundamental

issues need to be addressed such as pattern overlapping, outliers, imbalanced class, and dataset shift, in which those issues will degrade the classification performance [13, 15].

4. **Conclusion and Future Work.** In this paper, a framework to match research interest between researchers has been implemented using enhanced data mining framework, based on the abstract of authors' past publications. The dataset used in this experiment is obtained from Scopus repository ranging from 2008 to 2018 with the total of 245,618 publications and arranged in a specific order of features. From that, we selected only top-10 authors that have published the most as part of our experimentation settings. For the feature vectorization process, we used CORD-19 pre-trained Word2Vec model to obtain a single text embedding Word2Vec vector for every word that formed the abstract. Finally, for the classification process, Word2Vec vector is used as the features. Six different machine learning classifiers were applied and tested in order to find the classifier that provides the best classification results. Generally, all classifiers provide good classification performance with Neural Network that scores the highest accuracy.

As for future work, the framework should be enhanced to support more classes (authors) that are inside the dataset, together with fair ranking system. For instance, there must be two separated models that can be integrated into the framework to distinguish between senior researchers and junior researchers. In addition, a slightly deviation from this paper is to include temporal aspect modelling in the framework, so that the interest matching is based on recent proof-of-work, instead of based on past research interest that has been ditched long time ago.

Acknowledgement. This research is funded by Ministry of Higher Education of Malaysia, under grant FRGS-RACER/1/2019/SS09/UUM//2.

REFERENCES

[1] A. Y. A. Amer and T. Siddiqui, Detection of COVID-19 fake news text data using random forest and decision tree classifiers, *International Journal of Computer Science and Information Security (IJCSIS)*, vol.18, no.12, 2020.

[2] K. R. Choi, M. V. Heilemann, A. Fauer and M. Mead, A second pandemic: Mental health spillover from the novel coronavirus (COVID-19), *Journal of the American Psychiatric Nurses Association*, vol.26, no.4, pp.340-343, 2020.

[3] E. Cicognani, C. Pirini, C. Keyes, M. Joshanloo, R. Rostami and M. Nosratabadi, Social participation, sense of community and social well being: A study on American, Italian and Iranian university students, *Social Indicators Research*, vol.89, no.1, pp.97-112, 2008.

[4] S. E. De Morgan and A. De Morgan, *Memoir of Augustus De Morgan*, Longmans, Green, and Company, 1882.

[5] P. Esposito and S. Scicchitano, *Educational Mismatches, Routine Biased Technological Change and Unemployment: Evidence from Italy*, SSRN 3620909, 2020.

[6] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters*, vol.27, no.8, pp.861-874, 2006.

[7] S. Kale and R. Prasad, Author identification on imbalanced class dataset of Indian literature in Marathi, *International Journal of Computer Sciences and Engineering*, vol.6, pp.542-547, 2018.

[8] S. Khedkar, S. Agnihotri, A. Agarwal, M. Pancholi and P. Hande, Author identification using stylometry, *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pp.1-4, 2018.

[9] S. Lee and B. Bozeman, The impact of research collaboration on scientific productivity, *Social Studies of Science*, vol.35, no.5, pp.673-702, 2005.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, pp.3111-3119, 2013.

[11] M. S. Mohd Pozi, A. Jatowt and Y. Kawai, Temporal summarization of scholarly paper collections by semantic change estimation: Case study of CORD-19 dataset, *Proc. of the ACM/IEEE Joint Conference on Digital Libraries*, pp.459-460, 2020.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019.

[13] M. S. M. Pozi, M. N. Sulaiman, N. Mustapha and T. Perumal, Improving anomalous rare attack detection rate for intrusion detection system using support vector machine and genetic programming, *Neural Processing Letters*, vol.44, no.2, pp.279-290, 2016.

[14] M. Rahgouy, H. B. Giglou, T. Rahgooy, M. K. Sheykhlan and E. Mohammadzadeh, Cross-domain authorship attribution: Author identification using a multi-aspect ensemble approach, *CLEF (Working Notes)*, 2019.

[15] A. A. Raziff, M. Sulaiman, N. Mustapha and T. Perumal, Smote and OVO multiclass method for multiple handheld placement gait identification on smartphone's accelerometer, *J. Eng. Appl. Sci.*, vol.12, no.2, pp.374-382, 2017.

[16] P. Rosso, Profiling bots, fake news spreaders and haters, *Proc. of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, European Language Resources Association (ELRA), Marseille, France, p.1, https://www.aclweb.org/anthology/2020.restup-1.1, 2020.

[17] M. Sahlgren, *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*, Ph.D. Thesis, 2006.

[18] R. Sarwar, C. Yu, S. Nutanong, N. Urailertprasert, N. Vannaboot and T. Rakthanmanon, A scalable framework for stylometric analysis of multi-author documents, *International Conference on Database Systems for Advanced Applications*, pp.813-829, 2018.

[19] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh and L. Chanona-Hernández, Syntactic dependency-based n-grams as classification features, *Mexican International Conference on Artificial Intelligence*, pp.1-11, 2012.

[20] D. D. Sinaga and S. Hansun, Indonesian text document similarity detection system using Rabin-Karp and confix-stripping algorithms, *International Journal of Innovative Computing, Information and Control*, vol.14, no.5, pp.1893-1903, 2018.

[21] E. Stamatatos, Author identification: Using text sampling to handle the class imbalance problem, *Information Processing & Management*, vol.44, no.2, pp.790-799, 2008.

[22] M. S. Tamboli and R. Prasad, Author identification with feature transformation method, *Digital Scholarship in the Humanities*, vol.35, no.3, pp.642-651, 2020.

[23] P. N. Tan, M. Steinbach and V. Kumar, Classification: Basic concepts, decision trees, and model evaluation, *Introduction to Data Mining*, vol.1, pp.145-205, 2006.

[24] N. Tarmizi, S. Saee and D. H. A. Ibrahim, Author identification for under-resourced language KadazanDusun, *Indonesian Journal of Electrical Engineering and Computer Science*, vol.17, no.1, pp.248-255, 2020.

[25] T. van Tussenbroek, *Who Said That? Comparing Performance of TF-IDF and FastText to Identify Authorship of Short Sentences*, Bachelor Thesis, 2020.

[26] A. Uchendu, T. Le, K. Shu and D. Lee, Authorship attribution for neural text generation, *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.8384-8395, 2020.

[27] C. Zhang and M. Abdul-Mageed, Bert-based Arabic social media author profiling, *arXiv Preprint*, arXiv: 1909.04181, 2019.