PAPER • **OPEN ACCESS**

Author Identification: Performance Comparison using English and Under-Resourced Languages

Author Identification: Performance Comparison using English and Under-Resourced Languages

Nursyahirah Tarmizi, Suhaila Saee and Dayang Hanani Abang Ibrahim

Department of Information System, Universiti Malaysia Sarawak (UNIMAS), Kota Samarahan, 94300 Sarawak, Malaysia

syahirahmizi93@gmail.com

Abstract. This paper presents Author Identification (AI) task using different language which are English and Under-Resourced Languages (U-RL) (i.e. KadazanDusun and Iban). In this paper, the performance of AI task is analysed using English and the U-RL datasets in terms of accuracy. Different stylometric features and emerging machine learning algorithms (i.e. SVM and Random Forest) are examined to obtain optimal results in AI task. The approach used in AI task is based on supervised machine learning. Cross-validation is used to evaluate the performance of AI task. The findings include the performance comparison of different stylometric feature and classifiers between the three datasets based on their accuracy values. The combination of word n-grams with character 3-grams achieved the highest accuracy with almost 75% using English dataset. For classifier, SVM gained better result for all three datasets compared to Random Forest.

1. Introduction

In recent years, cyberbully activities have grown significantly against social media users. The anonymity circumstance provides an illusion to the cyber-bullies that they will never get caught. The situation has toughened the law enforcement in gathering the evidence to identify the bullies. In order to assist the forensic investigator to gather the digital evidence, Author Identification (AI) task is carried out to identify the most likely suspect that convict cyberbullying.

AI task has been used in a small but diverse number of application areas such as identifying authors in literature [5], in program code [6] and forensic analysis for criminal cases [1]. Due to the increasing number of available documents in digital form in social networks, AI task is crucial in analysing digital document to solve cybercrime issue such as cyberbullying. Yet, the problem of AI has always been harder for short text. Short text possesses the difficulty to be analysed because of the limited text-length and insufficient amount of content. Short text issues make the identification process much harder and complicated. For instance, social media i.e. Twitter allows users to post tweets with the restriction of 280 characters or less in length. The characteristics of tweets as stated above make Twitter data appealing to be used as testbed to overcome short text issues in AI.

There are many studies done in AI for social media using other languages such as Japanese [2], Portuguese [3] and Arabic [4]. However, less research is carried out for Malaysia indigenous languages. To the best of our knowledge, there is no study in AI is done using Under-Resourced Languages(U-RL) in Malaysia. Inadequate corpora and tool for UR-L are the key attribute to lack of research progress in U-RL AI. This paper focuses on AI task using U-RL languages such as KadazanDusun of Sabah and Iban of Sarawak besides English. KadazanDusun and Iban are both

indigenous languages from eastern Malaysia which makes them a part of U-RL. Besides, this study could help in promoting, preserving and revitalizing the indigenous languages of Malaysia as promoted by United Nation recently [13]. Thus, this paper examines stylometric features and emerging machine learning algorithms that could obtain optimal results in AI task using under-resourced language datasets.

2. Related Works

As the current trend in information technology encourages an abundance of short, informal writing that can be found in micro-blogging site such as Twitter. Micro-blogging site becomes increasingly important due to its impact on marketing, security, and forensic linguistics especially in barricading the cyber-criminal activity in social media. This section will cover previous studies that used various approaches to close up the gap of short text issues in AI by using micro-blogging text as a testbed.

A study done by [10] analysed the temporal changes of words used by the Twitter user over a period of time. The approach used was inspired by the time-based language model which employed a decay factor technique. The model utilized character n-grams as features and extended SCAP method with feature sampling that resulted in a large difference between the slopes. Another study that utilized character n-grams as features is, [11] which presented CNN model. The paper evaluated the model using 1,000 tweets per user of 50 users with an accuracy of 76.1% and the model outperformed state-of-art model, SCAP. While, a study done by [7] used word n-grams, character n-grams, flexible patterns and word embeddings to collect the stylometric information on tweets. By using a Multi-Layer Perceptron (MLP) as the classifier, they achieved the highest accuracy of more than 82% using TFIDF word embeddings feature sets on 1000 tweets for every 50 authors.

In a paper by [2], a combination of selection technique is proposed to train the dataset in handling wide range topic. POS-tag-combined-n-grams feature sets are used to shorten the execution time by decreasing the number of features. They used a big number of users as a challenge with 10,000 users each carry 120 tweets and their system managed to obtain 53.2% of accuracy. A study done by [12] had used a various combination of stylometric features such as word-, character-level n-grams, POS n-grams, and word frequency distribution feature sets. Among six methods, Linear Regression has shown good accuracy with 71.25% that combine all sets of features by using only a small dataset of 200 tweets per user (10 users). While, [1] had investigated the combination of features including character- and word-level n-grams feature sets with POS-tag n-grams which turned out to be effective.

As discussed above, proposed solutions for short text issues in AI have been sought actively to increase the performance of AI system. Therefore, it is important to study the issues and apply the state-of-the-art techniques to U-RLs so that a link between U-RL and AI field is established.

3. Research Methodology

In this section, the methodology of AI task will be discussed. The task is divided into three stages as shown in Figure 1.

3.1. Pre-processing

Twitter is a social media platform that focused on micro-blogging. The messages posted by the users are so-called the tweets which are made up of short messages (up to 280 characters) that combines with other elements such as photographs, videos, and/or web links as well. All tweets were crawled based on their respective vulgar words list using a crawler equip by Twitter API. After the tweets were crawled, extra tweets from each user were crawled too based on their *user_id* (unique attribute) and stored for pre-processing later. The tweets that were crawled undergone pre-processing to eliminate the meta-data and sparse characters while, the tweets and author column were remained kept. Tweets that shorter than four words were discarded too. Subsequently, the tweets undergone normalisation process. Throughout normalisation process, the original content of the tweets is substituted with standard tags which represent the replaced content. For instance:

Before normalisation: **@MissQilah** okay ba ok kak ok okeeeeeee

After normalization: **REFTAG** okay ba ok kak ok okeeeeeee

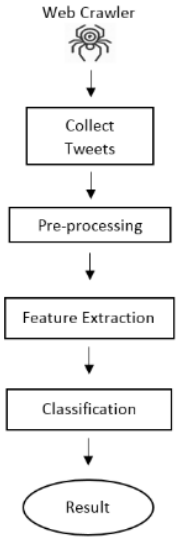


Figure 1. The workflow of Author Identification task

During tokenisation, the text is tokenised using a tokenizer which is called TweetTokenizer. TweetTokenizer is used because this tokeniser is designed to be flexible and easy to adapt to new domains.

3.2. Feature Extraction

During feature extraction process language-independent stylometric features are extracted. The stylometric features involved different levels of n-grams which are the character-level and word-level n-grams. The features are represented in Bag-of-n-grams models.

Character-level n-grams. In this study, character-level 3 and 4-grams are used as feature sets. They can capture the punctuation and capitalization information related to the authors' personal writing style [3][7]. Below shows the example of character 3-grams feature set:

Text: "Sik rindok naaaa"
Char 3-grams: "Sik", "ik_", "k_r", "_ri", "rin", "ind", "ndo", "dok", "ok_", "k_n", "_na", "naa", "aaa", "aaa"

Word-level n-grams. In this study, word n-grams ($n \in \{1\text{-}5\}$) are used as the feature sets. The feature sets can capture Internet idiosyncratic language usage such as emoticons, onomatopoeia (words that formed from sound e.g. achoo) and abbreviations [1]. The following example shows a simple phrase with a complete list of word unigrams and bigrams extracted from it.

Text: Sik rindok naaaa
Unigrams: "Sik", "rindok", "naaaa"
Bigrams: "Sik_rindok", "rindok_naaaa"

3.3. Classification

For the classification process, the experiments were conducted based on the instance-based approach. It is easier to combine all different text representation feature sets by using the instance-based approach. Also, this approach is robust when the class of candidate authors is large [15]. To perform AI on tweets using the instance-based approach, a set of tweets collection $T = \{t_1, \ldots, t_n\}$ and set of

candidate authors $A = \{a_1, ..., a_h\}$, where n – number of tweets and h – number of candidate authors. The candidate authors, a_i are presented as subset $T_k \in T$. The subset of tweets $T_k \in T$ of candidate authors – the samples data $\{t_1\ a_{1,\ ...,}\ t_n\ a_h\}$. During classification, the classifier assigns a sample data t_n a class label (author) a_h with maximum probability and outputs with most plausible author with Pr (a_{max} author t t_n). In this work, Random forest and Support Vector Machine are used as the classifiers.

Random Forest (RF). RF was chosen as the classifier is that RF is known to handle noisy data fairly well which in this study, a highly diverse set of noise features is used. RF creates a set of decision tree classifiers from randomly sub-samples of the training set [8]. Then, the votes from different decision trees are aggregated to improve the predictive model accuracy. Tree with a high error rate is given low weight value and vice versa. The random subsets created in different decision tree may overlap to reduce the effect of noise albeit this will increase the running cost for the classifier to produce the result.

Support Vector Machines (SVM). Support Vector Classification (SVC) is employed in this study. SVC is a type of SVM with an RBF kernel where the implementation is based on libSVM [9]. For the classifier optimization, kernel functions were specified. In this paper the kernel was set as linear when conducting the experiments. Linear kernel utilized a one-vs-rest strategy for a multi-class problem which is faster and can be scaled a lot better in high dimensionality of feature vectors.

4. Experimental Setup

The experimental settings are explained in this section in detailed. All experiments were performed to test the accuracy of the AI task using three different datasets which are English, KadazanDusun and Iban.The experiments were carried out in a controlled environment, in which all data are analysed on a single machine. This is to assure that the results of the performed test are consistent.

4.1. Input Dataset

For each dataset, 10 different authors with varies number of tweets were collected. The preparation of UR-L datasets is a bit different from English. During pre-processing, an additional cleaning process was done on the U-RL tweets. U-RL tweets that contain any English words inside the tweet itself were removed and the remaining words contain only the native languages itself. The purpose is to conserve the originality of the U-RL languages in the tweets. NLTK English text corpora is used to eliminate English words in the tweets. An example below shows a KadazanDusun tweet that has been pre-processed:

Original text: "Hello babe... Idup dlm reality bh!"
Pre-processed text: "… Idup dlm bh!"

4.2. Random Sampling

Random sampling is done by selecting tweets randomly and standardize the number of training sets. All experiments are run using 10 authors where each author with 100, 200, 300 and 400 samples. This selection process excludes any duplication of tweets in preventing any bias which later will affect the accuracy of the AI system. Table 3 below depicts the statistical description for all three datasets.

The decision to perform this study on a small number of 10 authors as the candidate authors are explained by the real-world application of authorship identification. In most civil case, the case involved a small number of candidate authors for a piece of anonymous text [4]. Therefore, the settings of the experiments are established on the basis of a realistic situation to ensure that the results reported are realistic and that the accuracy is not overly estimated.

Table 1. Statistical description of three datasets.

Dataset	Num. of tweets	word_length	character_length
English	100	15087	81778

	200	17584	106803
	300	26118	157530
	400	59201	319210
KadazanDusun	100	8793	53260
	200	17584	106803
	300	26118	157530
	400	35260	213131
Iban	100	9194	53133
	200	17831	105318
	300	26856	158360
	400	36127	211492

4.3. Stylometric Features

In this study, the character-level and word-level n-grams were evaluated as feature sets (see Sec. 3.2). These feature sets are analysed based on their performance using SVM as a classifier in order to access the adequacy of the feature sets in English and U-RL datasets. The feature sets that are analysed in this study include word unigrams and word {1-5}-grams for word-level n-grams. While for character-level n-grams, character 3 and 4 grams are analysed separately. According to [1], character n-gram are capable in capturing unusual features in tweets, but character 4-grams could generally include many of unigrams and significantly increases the length of the feature vectors. Therefore, character 3-grams are analysed too in this study to truncate feature vectors. Apart from that, the combination of word-level and character-level are analysed to offer better accuracy for short text. There are 2 different combinations of feature sets include word unigram and word {1-5}-grams combined with character 3 and 4-grams separately. Due to insufficient information of short text in tweets, combining different features sets could contribute in increasing the accuracy of AI [14].

4.4. Evaluation

In this study, the 10-fold cross-validation was used to validate the performance of the feature sets and classifiers in terms of accuracy. The validation process will divide the datasets into training and validation sets. Subsequently, 10 iterations of training and validation are performed where, within each iteration, a different fold of data was held-out for validation, while the remaining folds were used for learning. This validation approach will obtain an aggregate accuracy from the iteration. 10-fold cross-validation was used as it is the most commonly used in data mining and machine learning field.

5. Results and Discussion

In this section, the experimental results are presented on three different datasets. The results include the performance of different feature sets and classifiers using a different number of samples using three different datasets.

5.1. Accuracy comparison of different feature sets

To access the usefulness of feature sets, the experiments were performed using different feature sets as mentioned in Sec. 4.3. SVM is used as the base classifier. The experiments were conducted using a fixed number of 10 authors with 400 tweets for each author. With 400 tweets per author, sufficient stylometric information is prepared to analyse the data. The following table 2 below reports the accuracy results of different feature sets.

Table 2. Accuracy comparison of different feature sets between English and U-RL datasets.

Feature sets	English (%)	KadazanDusun (%)	Iban (%)
Word unigram	**70.7**	**67.9**	69.8
Word {1-5}-grams	67.9	66.0	66.7
Character 3-grams	69.7	67.3	69.3

Character 4-grams	68.8	67.9	**71.0**
Combination 1	74.9	71.5	73.6
Combination 2	74.3	71.2	73.4

As can be observed in table 2, the results reveal that word unigrams feature set gains the highest accuracy as individual feature set for English and KadazanDusun dataset with 70.7% and 67.9% respectively. This is relevant as word unigrams can capture the choices of particular words that are unique for each author. It appears that authors may have varying primary modes or motif in using Twitter For instance, there are authors that appear to tweet on their social life, by updating their personal status and sharing location updates. While, some authors may use Twitter as a medium to blog on certain issues, which majority of the tweets they posted will focus on specific opinion and information.

As for Iban dataset, it appears that the character 4-grams is the best individual performance with the highest accuracy of 71%. The result is slightly different from English and KadazanDusun, yet it is relevant as character 4-grams can capture the extensive usage of punctuation and emoticons. It has been proved by previous study [1] that this feature set reasonably captures idiosyncrasies commonly used in micro-messages such as Twitter which leads to identifiable patterns characteristic to each author. Authors may use words in full capital letters to bold their expression towards something whereas other authors may prefer to express their tweets using emoticons or exclamation marks.

The results also show that the combination of word unigrams with other word n-grams and character n-grams help to boost up the accuracy. It appears that the highest accuracy achieved by the feature sets combination that consists of word unigrams, word {1-5}-grams and character 3-grams for all the datasets. The finding suggests that each author have certain preferences for some words and punctuations that they are more comfortable with. The selection of words is impossible to be captured by using only character n-grams due to their size of appearance with other word affixes. Yet, character n-grams have been used widely as the supportive feature in solving attribution problem since they are relatively tolerant to spelling errors and non-standard use of punctuation

The combination of character 3-grams with the word n-grams shows a competitive performance with character 4-grams combination with a difference that not more than 1%. Character n-grams are very effective in boosting the classification accuracy, but they used to generate high dimensionality of vectors to be processed. A very compatible multi classifier is needed to be able in handling high-dimensional feature representation and large-scale classification. Therefore, it seems that the combination with character 3-grams will be more relevant by offering higher accuracy with significantly less cost in terms of space and time.

5.2. Accuracy comparison of different classifiers

The experiments were conducted using two different classifiers which are SVM and RF. The accuracy of each classifier was tested using a various size of training sets ranging from 100 to 400 tweets with a fixed number of 10 authors. For this experiment, we purposely use the Combination 1 feature sets consist of word unigram, word {1-5}-grams and character 3-grams as in the previous experiment in Sec. 5.1 reveals that Combination 1 obtained the best accuracy for all three datasets. Figure 2 below depicts the result for English dataset.

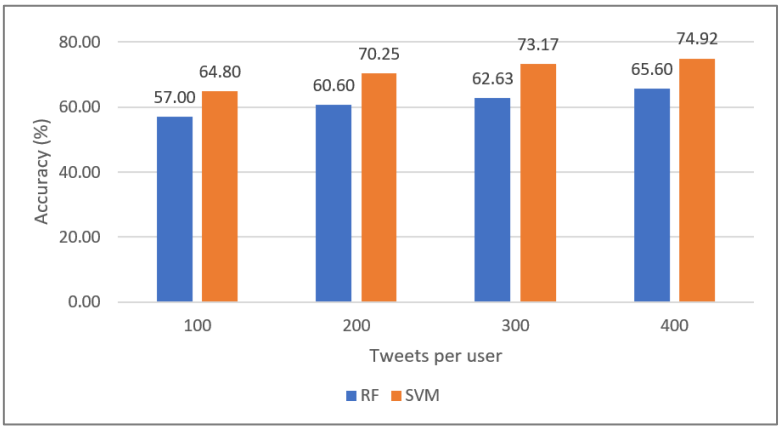
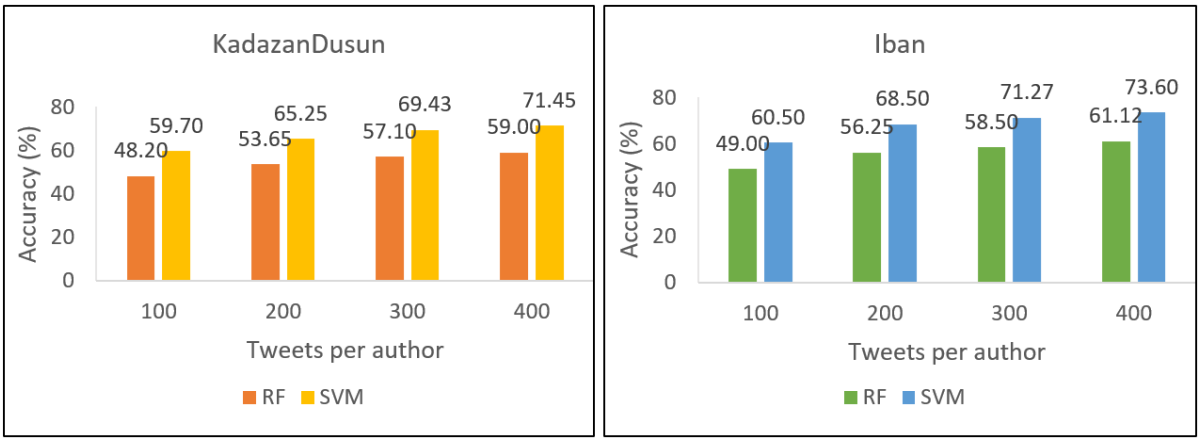


Figure 2. Accuracy of different feature sets for English dataset.

The results demonstrate that the accuracy for English dataset is considerably high for both classifiers, RF and SVM, which over than half. The performance achieved by SVM classifier is comparable with RF. Though, SVM outperformed RF with the best accuracy of 74.9% obtained with 400 training samples. SVM yields better results due to its capacity in handling high dimensionality of features and sparse data in tweets. Although the accuracy of RF appears to be less accurate than SVM, the results suggest that both classifiers increase substantially with the increase in the number of training samples. Apparently, with the increase in the number of samples RF improves 8% while, SVM improves 10% of accuracy.

As for U-RLs, the datasets were prepared slightly different to English dataset with the additional cleaning process as mentioned in previously Sec. 3.1. The accuracy of U-RL datasets are demonstrated in Figure 3 *(a)* and *(b)* below.



(a) (b)

Figure 3. Classification accuracy for *(a)* KadazanDusun and **(b)** Iban datasets.

The accuracy results of both datasets have a similar pattern with English whereby SVM outperformed RF. Fig. 3 above demonstrates the results obtained using KadazanDusun and Iban dataset which represent as the U-RL datasets. The following Figure 3 *(a)* above displays the accuracy of classifiers for KadazanDusun. The result demonstrates that SVM by far achieved the highest accuracy, 71.5%, with 400 training samples. SVM is significantly 12.5% more accurate than RF for KadazanDusun dataset.

As for Iban dataset, the results obtained in Figure 3 *(b)* above shows that the best result was achieved by SVM compared to RF which barely below SVM. Again, SVM particularly achieved the best accuracy, 73.6%. The findings suggest that the accuracy for both classifiers increases with the

increase of samples. Again, SVM classier particularly achieved the best accuracy with 400 tweets per author, where the accuracy reaches 73.6% for Iban dataset. The findings from Figure 3 *(b)* suggest that the accuracy of classifiers increases with the increase of training samples. Particularly, SVM improves with 13.1% (from 49% to 73.6%) and RF improves significantly with 12.12% (from 49% to 61.1%).

It is apparent that all results obtained in Figure 2 and Figure 3 show that SVM gives better results compared to RF because SVM handles sparse data better than RF. The figures also summarise that the number of samples affected the accuracy of the classifiers. The size of training samples plays an important role which can affect the accuracy of AI in identifying the author of an anonymous text. With a large number of training samples, more reliable statistics can be obtained and leads to a significant improvement in the prediction for test data.

Though, it appears that data obtained in [4] refute to the findings gained in this study. The results in [4] demonstrated that RF works better on Arabic dataset. Using 10 authors with 25 tweets for each author, the results from [4] reveals that RF gains better accuracy with 20% divergence in accuracy ahead from SVM. It is obvious that there is a significant divergence between Arabic dataset used in [4] and U-RL datasets used in this study, as the number of tweets is lower compared to U-RL. In their study RF work well with limited data size compared to SVM.

The findings conclude that the performance of AI using English dataset obviously better compared to the U-RL. This is because the additional cleaning process was done to the U-RL datasets. It is done purposely to preserve the originality of the native languages. Still, the performance of AI using the U-RL datasets portray competitive results which can be enhanced ore using a different type of stylistic features that are more suitable to hone the accuracy of AI using the U-RL datasets.

6. Conclusion

In summary, this study presented an approach for AI task using English and U-RL short-text messages. The main goal is to analyse the performance of AI using English and the U-RL datasets in terms of accuracy. The findings in this study have shown that the performance of English dataset yields better accuracy compared to KadazanDusun and Iban datasets. Word unigrams work as the best individual feature sets for English and KadazanDusun datasets, while character 4-grams for Iban dataset. The combination of word n-grams with character 3-grams achieved the highest accuracy with almost 75% using English dataset. As for the classifiers, SVM gained better result for all three datasets compared to RF. Overall the performance of AI yields better results using English dataset compared to KadazanDusun and Iban datasets. Apart from that, word unigrams and character 4-grams independently achieved good accuracy. Yet, the combination of feature sets proved to have better results compared to individual feature sets.

7. References

[1] Rocha A, Scheirer WJ, Forstall CW, Theophilo A, Shen B, Carvalho ARB, and Stamatatos E 2017 *Transactions on Information Forensics and Security* **12** 5-33
[2] Okuno S, Asai H, and Yamana H 2014 *A challenge of authorship identification for ten-thousand-scale microblog users* (Washington, DC: IEEE) pp 52-54
[3] Markov I, Baptista J and Pichardo-Lagunas O 2017 *Acta Polytechnica Hungarica* **14** 59-78
[4] Altakrori MH, Iqbal F, Fung BCM, Ding SHH and Tubaishat A 2019 *ACM Trans. Asian Low-Resour. Lang. Inf. Process* **18** 51
[5] Dauber E, Overdorf R and Greenstadt R 2017 *Stylometric Authorship Attribution of Collaborative Documentst* ed S Dolev, S Lodha (Cham: Springer) 10332 pp 115-35
[6] Alrabaee S, Shirani P, Debbabi M, and Wang L 2017 *On the Feasibility of Malware Authorship Attribution* ed F Cuppens et al. (Cham: Springer) 10128 pp 256-72
[7] Joshi M and Zincir-Heywood N 2019 *Classification of Micro-Texts Using Sub-Word Embeddings* (Varna: ACL) pp 526-533
[8] Dangeti P 2017 *Statistics for Machine Learning* (Birmingham: Pact Publishing) chapter 4 pp 149
[9] Deng N, Tien Y and Zhang C 2013 S*upport Vector Machines: Optimization Based Theory, Algorithms, and Extensions* (Florida: CRC Press) chapter 8 pp 243

[10] Azarbonyad H, Dehghani M, Marx M, and Kamps J 2015 *Time-Aware Authorship Attribution for Short Text Streams* (New York: ACM) pp 727-30
[11] Shrestha P, Sierra S, González F, Montes M, Rosso P, and Solorio T 2017 *Convolutional Neural Networks for Authorship Attribution of Short Texts* (Valancia: ACL) **2** pp 669-74
[12] Banga R and Mehndiratta P 2017 *Authorship attribution for textual data on online social networks* (Noida: IEEE) pp 1-7
[13] United Nation Educational, Scientific and Cultural Organization 2019 *Strategic Outcome Document of the 2019 International Year of Indigenous Languages* (Retrieved from IYIL2019 https://en.iyil2019.org)
[14] Soler-Company J and Wanner L 2017 *On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification* (Valencia: EACL) pp 681-87
[15] Kourtis I and Stamatatos E 2011 *Author identification using semi-supervised learning* (Amsterdam: CLEF)

17020134 Author Identification: Performance Comparison using English and Under-Resourced Languages

ORIGINALITY REPORT

9%	1%	8%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universiti Malaysia Sarawak Student Paper	3%
2	Anderson Rocha, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante et al. "Authorship Attribution for Social Media Forensics", IEEE Transactions on Information Forensics and Security, 2017 Publication	2%
3	Malik H. Altakrori, Farkhund Iqbal, Benjamin C. M. Fung, Steven H. H. Ding, Abdallah Tubaishat. "Arabic Authorship Attribution", ACM Transactions on Asian and Low-Resource Language Information Processing, 2018 Publication	1%
4	"Artificial Neural Networks and Machine Learning – ICANN 2017", Springer Science and Business Media LLC, 2017 Publication	1%
5	Munish Kumar, M. K. Jindal, R. K. Sharma,	

	Simpel Rani Jindal. "Performance evaluation of classifiers for the recognition of offline handwritten Gurmukhi characters and numerals: a study", Artificial Intelligence Review, 2019 Publication	1 %
6	research.ijcaonline.org Internet Source	1 %
7	Nirkhi, S.M., R. V. Dharaskar, and V.M. Thakre. "Analysis of online messages for identity tracing in cybercrime investigation", Proceedings Title 2012 International Conference on Cyber Security Cyber Warfare and Digital Forensic (CyberSec), 2012. Publication	1 %