Indonesian Journal of Electrical Engineering and Computer Science
Vol. 17, No. 1, January 2020, pp. 248~255
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v17.i1.pp248-255

Author identification for under-resourced language Kadazandusun

Nursyahirah Tarmizi, Suhaila Saee, Dayang Hanani Abang Ibrahim
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Malaysia

Article Info

Article history:

Received Apr 28, 2019
Revised Jun 29, 2019
Accepted Jul 12, 2019

Keywords:

Author identification
Kadazan dusun
Machine learning
Stylometry
Under-resourced language

ABSTRACT

This paper presents the task of Author Identification for KadazanDusun language by using tweets as the source of data to perform Author Identification task of short text on KadazanDusun, which is considered as one the under-resourced language in Malaysia. The aim of this paper is to demonstrate Author Identification of short text on KadazanDusun. Besides, this paper also examines the performance of two machine learning algorithms on the KadazanDusun data set by analyzing the stylometric features. Stylometric features are used to quantify the writing styles of the authors which includes character n-grams and word n-grams. The workflow of Author Identification implements the machine learning approach to solve the single-labelled multi-class problem and predict the author of a given message in KadazanDusun. Two classifiers are used to compare the accuracy including Naïve Bayes and Support Vector Machine (SVM). The results show that the combination of n-grams which is word-level unigram and {1-5}-grams with character 3-grams are the most relevant stylometric features in identifying the author of KadazanDusun message with an accuracy of 80.17%. The results also show that SVM classifier has outperformed Naive Bayes in this Author Identification task with the accuracy of 80.17%.

Corresponding Author:

Nursyahirah Binti Tarmizi,
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarwak (UNIMAS),
Kota Samarahan, 94300, Sarawak, Malaysia.
Email: syahirahmizi93@gmail.com

1. INTRODUCTION

Author Identification is a process of identifying the author of an anonymous text given the predefined set of candidate authors and corresponding samples of their texts. Author Identification task analyses the writing style of each author by extracting out the stylometric features from the text and the features will represent as the writing style of each author [1]. From a machine learning perspective, approaches in Author Identification can be viewed as a single-labelled, multi-class classification problem, in which a set of class labels is known as a priori. The challenges in Author Identification task exist in modelling the classification task so that the automatic methods will be able to assign class labels (authors) to the objects (text samples). Previous studies show that Author Identification task has been used in a small but diverse number of application areas such as identifying authors in literature [2], in program code, and in forensic analysis for criminal cases [3].

In recent years, the vast popularity of social media has created a special interest in authorship attribution area, both theoretically and computational in short text [4]. The circumstances have led to the development of authorship attribution projects that experimented with web data i.e. web forum, e-mails [5], blogs [6] and social media i.e. Twitter, Facebook, Instagram [7]. Due to the increasing number of available documents in digital form in social networks, Author Identification has become decisive task in analysing the

digital document to solve cybercrime issues such as cyberbully. One of the social media platforms, Twitter comprised of messages that are posted by users which are called tweets that are strictly limited to 280 characters. The character restriction imposes major difficulties on Author Identification systems since authorship attribution methods often work well on long text or messages [8, 9] not as useful applied to short text [10, 11]. While, to the best of our knowledge, none of the previous studies has focused on Author Identification task for under-resourced languages like KadazanDusun as Under-Resourced Language (U-RL) in Malaysia. This paper works on identifying the author of short text for U-RL using KadazanDusun tweets as the source of data. The objectives of this paper is to demonstrate author identification of short text on KadazanDusun and to examine the performance of several machine learning algorithms such as Naïve Bayes and Support Vector Machine on KadazanDusun language data set.

The rest of the paper is organized as follows. Section 2 describes the literature review. Section 3 describes the architecture of AI workflow and the implementation in details. Section 4 reports the results of the experiments. The last section of this paper states our conclusion and future works.

2. LITERATURE REVIEW

This section discusses the definition of indigenous language and U-RL besides reviewing the issue and gap regarding U-RL based on previous papers. Also, we review the state-of-art of Author Identification of short text (tweets) including the performance of the systems.

2.1. Indigenous Language in Malaysia and U-RL

In an article by [12], Malaysia has a high density of indigenous languages. Indigenous language is defined as a language that has a stable community of speakers with a considerable time-depth, a genetic relationship with other native languages in the same geo-linguistic region and recognized as a native language by the community themselves. In Malaysia, there are slightly about 100 of indigenous languages covering the east and west Malaysia. Besides Malay language, Iban of Sarawak and KadazanDusun of Sabah can be said have a large number of native speakers as Malaysian indigenous languages. Nowadays, the use of these three native languages in Malaysia have been widely used not only as of the mother tongue of the community but rather actively been used in social online communication as well. Although the native languages (Malay, Iban and KadazanDusun) have a wide range of usage in their communities, yet, there are still no writing system has been ascribed to them which appertain these native languages as under-resourced languages [12].

According to [13], under-resourced language (U-RL) is referred to as language with some (if not all) lack of unique writing system or stable orthography, the limited presence of the web, lack of linguistic expertise, and lack of electronic resources for both speech and language processing. However, the inflation of these U-RLs in online communication has become an important factor for natural language processing (NLP) tasks to be able to analyze these texts for the purpose of cybersecurity and cybercrime for instance if the text implicates the usage of the indigenous languages.

2.2. State-of-art Author Identification

Author identification is an important task to detect or reveal the culprit in terms of cybercrime and cyber-attacks [14]. As stated by [15], Author Identification task involves techniques in performing forensics of online messages to collect practical evidence by automatically analyses a large collection of suspicious online messages from a number of suspects. The task involves the classification of authors and the accuracy of the system is influenced by the text length, the combination of stylometric features and the algorithm used to classify the authors.A study done by [16] explores the stylometric likability of tweets data. They used character n-grams as stylometric features and Naïve Bayes as the classifier. With a subset of 300-2000 tweets data of a varying number of users, they obtained 92% of accuracy ac hived for unigram model while 100% achieved through the bi-grams model. According to [17], Naïve Bayes is simple yet effective method in designing a text classifier with high accuracy rate and fast speed given a large number of training data. On the other hand, [7] studied the use of character n-grams using Convolutional Neural Network (CNN) as the classifier to identify the author. Using 1000 tweets per user with 9000 users, the best character n-grams model they obtained was character unigram model with 76.1% of accuracy. While [18] considered verifying compromised Twitter account using Author Verification by using a range of 50-100 words per user with 10,000 users. In this study, they used profile-based approach where they implement Simplified Profile Intersection (SPI) method to verify the author. Using word n-grams (N=6) as their best stylometric features with 100 words from each user, they obtained an accuracy of 95.8%.

3. METHODOLOGY

In this paper, we proposed the workflow of Author Identification for short text to identify the author of KadazanDusun tweets from Twitter. We used two supervised machine learning algorithms, Support Vector Machines (SVM) and Naïve Bayes, to learn the stylometric features from each user. The stylometric features are extracted from the training data that has been collected from different Twitter users. Figure 1 shows the workflow of the proposed approach. The explanation of each phase in the workflow is described in the following subsections respectively.

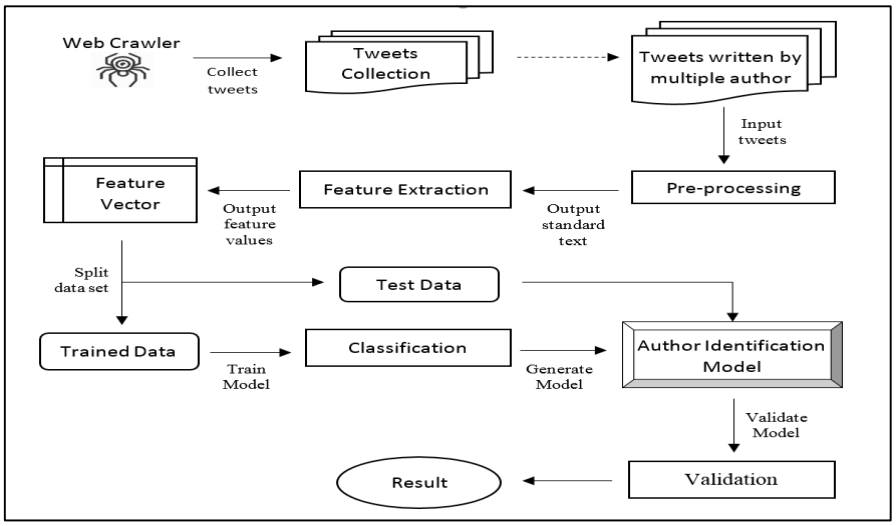

Figure 1. The workflow of author Identification of short text

3.1. Tweets Collection

According to [3], there are no public data set exist for authorship attribution. Therefore, we need to crawl out own set of data for the experiment purpose from Twitter using Twitter API namely *tweepy*. Since Twitter is a very popular platform where cybercrime activities frequently happened [1], the data set is built of a list of tweets that are extracted based on a list of vulgar words in KadazanDusun language which is obtained from a web blog [19]. A list of 13 vulgar words was inserted in the crawler to crawl a list of tweets containing given vulgar words. Below examples show tweets in KadazanDusun with vulgar words in them:

a) "ni bkn stkt paluiii.. ***basug***! kimbet! ***toburus***!"
b) "Nda lama ko akan post suruh org pgi mati sbb jaga jodoh. ***Mulau***. Dunia2."

3.2. Pre-processing

The tweets are collected in JSON format using crawler provided by Twitter API. After a list of tweets is collected, additional tweets are collected from each user based on the 'user_id' and are saved in CSV format for data pre-processing. After collecting the tweets from each author, the tweets are pre-processed by removing the meta-data and noises. Some information is excluded during the extraction including the retweets messages and also tweets that contain meme. The only information that is kept are the text and author columns. Tweets with less than four words are removed.

In this paper, the pre-processing stage will be focusing more on text normalization where the tweets are normalized to standard text. Hence, the pre-processor takes away the original text and replace the text with standard tags that represent the replaced content. This process is important as it would greatly reduce the number of features to be analysed, for example, long numbers or web links and repeated date and time.
The following examples show tweets before and after the aforementioned pre-processing procedure:

Before pre-processing: "**@QcLyn** hahahhaa okay sya study satu pun tiada **#finalexam #stress**
After pre-processing: "**@REF** hahahhaa okay sya study satu pun tiada **HTAG HTAG**

3.2.1. Native features of Twitter

In this paper, instead of depending solely on the stylometric features, native features of Twitter such as hashtags, user references, web links and Emoji are utilized in all the experiments. The employment of native features of Twitter in Author Identification is to further refine the accuracy of attribution by pursuing a

hybrid approach that extends beyond just stylometry. As stated in [3] hashtag is referred as keywords used in tweets to identify messages on a specific topic and they are preceded by a '#' character. While user references are the users of Twitter that are being mentioned by other users in tweets using an '@' followed by their username. Emoji are cartoon figures that used to express ideas or emotion in text.

3.3. Feature Extraction

To build a feature vector, we need to tokenise the text beforehand. The process of tokenisation will use a suitable tokenizer that will be able to tokenise the KadazanDusun text in social media. The extraction of features includes the lexical and syntactical stylometry features. Different level of n-grams will be extracted include the character-level and word-level n-grams will be extracted and represented in Bag-of-Word (BoW) models.

3.4. Classification

Each text is represented as the vector and each text is labelled with its respective author or class. This is based on the instance-based approach where each training text samples act as a unit that contributes separately to the attribution model. Then each feature sets will be classified using different types of classifiers to build the classification model. The classifiers involved are Naïve Bayes (NB) and Support Vector Machine (SVM). These classifiers will yield different accuracy results and running time-based on their capacity to handle the high dimensionality of features. The evaluation of the model is based on the *k*-fold cross-validation. This validation will return the mean accuracy of the model based on the k-fold of training and testing.

4. EXPERIMENTAL SETUP

In this section, the experimental settings are laid out as follows. First, the experimental setup is briefly described for the purpose of presenting the criteria adopted in building a balanced data set. Next, the stylometry features used in this experiment are described in detail as well. Lastly, two machine learning algorithms implemented in this experiment are discussed followed by the standard evaluation used in the experiments.

4.1. Dataset

In this paper, the experiment is conducted using a data set consist of a collection of KadazanDusun tweets. These tweets are crawled based on a list of KadazanDusun vulgar words as mentioned in Section 3. There are a total of 14,284 tweets collected from 15 different Twitter users. The number of words that are posted in a tweet varies in terms of the number of characters. Twitter allows 280 characters as the maximum characters that are able to post including the alphanumeric, web link, emoticons etc. The preparation to build a training data set involves random sampling. Users that posted more than 400 tweets are selected. There are 10 users and their tweets are randomly selected up to 400 tweets so that each author has a balanced distribution of tweets. The purpose of random sampling is to avoid imbalanced data set and bias towards certain authors that have a higher number of tweets during the classification process later. After random sampling of the data takes place, the process of pre-processing and normalization are continued so that the data is cleaned from noises and in a standard form.

4.2. Stylometric Features Used

In the proposed workflow, the Bag-of-Word (BoW) approach is implemented which consists of language-independent stylometric features i.e. word and character n-grams are used [20].

4.2.1. Word n-grams

As [21] points out, word-level n-grams are used to take advantage of contextual information. Word-level n-grams are a continuous sequence of *n* words of a longer portion of a text. Such pattern of choices of particular word sequences are unique and different for each author as an individual's cognitive representation of language which is influenced by the socio-historical linguistic background of that author [22]. As an example, let us consider the following sample tweet from a user with a list of word-level unigrams and bigrams from the tweet:

Text: "bagus lagi ko doa malam diam2 dari post2."

Unigrams: ("bagus", "lagi", "ko", "doa", "malam", "diam2", "dari", "post2")

Bigrams: ("bagus lagi", "lagi ko", "ko doa", doa malam", "malam diam2", "diam2 dari", "dari post2")

4.2.2. Character n-grams

Character-level n-grams are able to capture the nuances of style including the lexical information and syntactic information [20] such as the alphabetical, digit characters, uppercase and lowercase counts as well as the letter frequencies and punctuation marks counts. Besides, this feature type is able to capture lexical and even grammatical and orthographic preferences without the need for linguistic background [23]. Apart from that character n-grams are also tolerant to noise, the use of this feature is compatible with social media text such as tweets. Tweets usually contain high grammatical error and high usage of punctuation which this type of feature can handle. Below shows an example of character 3-grams:

Text: "paluiii!!!"
3-grams: ("pal", "alu", "lui", "uii", "iii", "ii!", "i!!", "!!!")
4-grams: ("palu", "alui", "luii", "uiii", "iii!", "ii!!", "i!!!")

4.3. Classifiers Used

Authorship identification is a single-labelled and multi-class text classification problem. Selection of classifier is appropriate in performing the identification and should be done carefully. The purpose of conducting the experiment using a different type of classifiers is to evaluate the performance of each classifier with the selected feature sets used. The evaluation is done using *k*-fold cross-validation to measure the performance of each classifier.

4.3.1. Naïve Bayes

In our work, we use multinomial Naïve Bayes (NB) that manipulates discrete features i.e. word counts, word frequencies etc. which is suitable for our case. This probabilistic classifier has secure independent assumptions based on the application of Bayes Theorem. Let the set of classes be denoted by *C*. Let *N* be the size of the vocabulary. Then Multinomial NB will assign the test document t_i to the class that has the highest probability $\Pr(c|t_i)$. The class prior $\Pr(c)$ can be estimated by dividing the number of documents that belong to the class *c* by the total number of documents. In a study done by [24], $\Pr(t_i|c)$ is the probability of obtaining a document like t_i in class *c* is calculated as 1:

$$\widehat{Pr}(t_i|c) = \alpha \prod_n \Pr(w_n|c)^{f_{ni}} \quad (1)$$

Where α is a constant and f_{ni} is the count of word *n* in the test document t_i while $\Pr(w_n|c)$ is the probability of word *n* given class *c*.

4.3.2. Support Vector Machines

In this paper, we implement Support Vector Classification (SVC) as the classifier which the implementation is based on libSVM. For optimization, Kernel functions can be specified for the decision function. In the experiment, the kernel is set as linear. This algorithm uses a *one-vs-rest* strategy for multi-class problem which is faster and can be scaled a lot better. In this algorithm, the data item (text) is plotted as a point in *n*-dimensional space (where *n*=num. of features) with the value of each feature being the value of a particular coordinate. Then the classification is performed by finding the hyper-plane that differentiate the classes very well by making the distance interval between each category maximize each other. As reported by [25], the calculation of minimum distance of hyper-plane is defined in 2:

$$\min \varphi(\omega, \xi) = \frac{1}{2}(\omega.\omega) + C\sum_{i=1}^{1} \xi_i \quad (2)$$

5. RESULTS AND DISCUSSION

The experiment setup was run using a fixed pool of 10 authors and implements *k*-fold (*k*=10) cross-validation to validate the classification model. In this section, the results of the experiments that we had performed to examine the approach introduced in Section 3 are as follows:

a) Accuracy comparison for different stylometric feature sets using a fixed pool of 400 tweets per author using SVM as the base classifier
b) Performance comparison between two classifiers, Naïve Bayes and SVM, in terms of accuracy and time taken by varying the number of tweets

5.1. Comparison of Different Feature Sets

In order to access the usefulness of the feature types, the experiment are conducted using different sets of features using SVM as the base classifier. In this paper, the feature sets used are:

a) Word Unigram
b) Word {1-5}-grams
c) Character 3-grams
d) Character 4-grams
e) Combination 1: Word Unigram, Word {1-5}-grams, Character 3-grams
f) Combination 2: Word Unigram, Word {1-5}-grams, Character 4-grams

At this point, one could wonder what would be the impact of using only character 3-grams and 4-grams as well as word n-grams in the identification task. Our choice for the features sets as listed above was motivated by previous work in the area [4], [3] and [22]. Data obtained in the previous study by [3] using PMSVM as the base classifier yields a result of character-level 4-grams as the most relevant independent feature set which has the highest accuracy using English data set. According to [4], word-level n-grams features substantially improve over character n-gram features. In this paper, word-level and character-level n-grams features together with their combinations are used to analyse the accuracy of identification of anonymous author for a given text in KadazanDusun. Figure 2 depicts the accuracy of different feature sets using 400 tweets for each author using SVM as the base classifier.

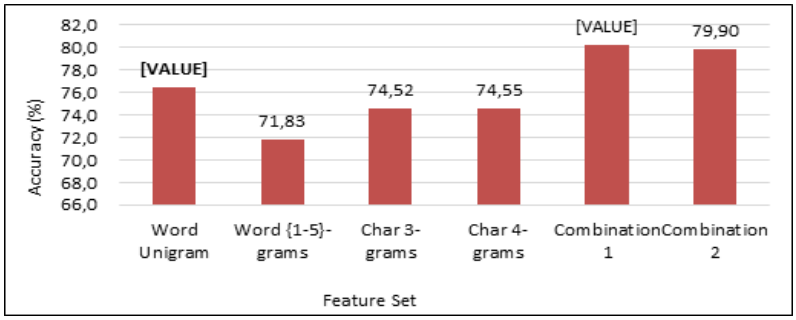

Figure 2. The accuracy of different feature sets

The accuracy results were compared between the feature sets to analyse which feature sets is more relevant in identifying the author of an anonymous text in KadazanDusun. Figure 2 shows that word unigram feature set is the best individual feature set with the highest accuracy of 76.42% followed by character-level 4-grams and 3-grams with a slight difference of 0.03%. As can be seen, this is relevant as the word unigram feature set is able to capture the choices of particular words which are unique and different for each author. Although character-level 4-grams combination a bit below than character-level 3-grams, it still offers competitive accuracy with only a slight difference. The identification accuracy improved even higher (fairly by 4%) as both word-level and character-level feature sets are combined. According to Figure 2 above, the highest accuracy is obtained by Combination 1 feature sets with more than 80% accuracy followed by Combination 2 feature sets, which slightly below by 0.3%. It can be observed, both Combination 1 & 2 by far gain higher accuracy which suggest that character 3-grams and 4-grams feature sets are able to capture the different choice of emoticons, abbreviations, and creative punctuation used by each author.

5.2. Comparison of Classifiers

The performance of these two classifiers is compared in terms of the accuracy and the cost of running time using Combination 1 feature set. The reason for using Combination 1 feature set is because this feature set has the highest accuracy compared to other feature sets from the previous section. Thus, it is relevant to choose this feature set for the identification task. For this experiment, we considered using 10 authors with a variable number of tweets that range between 100 to 400 tweets. Table 1 below shows the performance of two different classifiers, Naive Bayes and SVM.

Table 1. The Performance of Naive Bayes and SVM with Different Number of Tweets

Number of tweets	Naïve Bayes		SVM	
	Accuracy (%)	Time Taken (s)	Accuracy (%)	Time taken (s)
100	69.20	7.8	72.10	25.1
200	73.05	15.3	74.85	84.8
300	76.73	22.0	79.20	171.7
400	**77.58**	17.6	**80.17**	211.8

As can be seen in Table 1, the running time for both classifiers increase as the number of tweets increase. This is relevant as the feature vectors increase with the higher number of tweets used as training data. Result from Table 1 proves that the time taken for Naïve Bayes to yield results is faster compared to SVM. This is because the Naïve Bayes possess more simple method in building up a classification model that resulted in producing fast speed results [17]. Figure 3 shows the comparison for the performance of the classifiers in terms of the accuracy

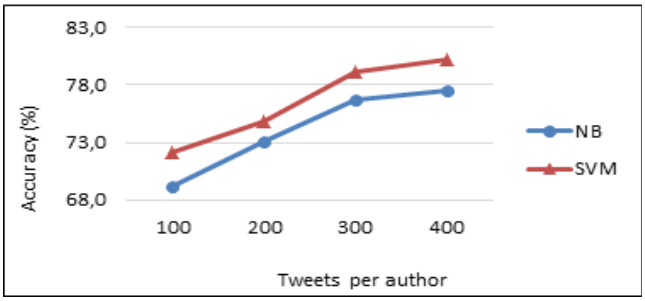

Figure 3. Accuracy comparison between Naïve Bayes and SVM

Figure 3 depicts that SVM gains substantially high accuracy than Naive Bayes. Data from a previous study [8] showed that the accuracy of the classifier will drop gradually with an increase in tweets from 300 to 500 tweets, as it is dependent on the data set. In their paper, the result shows that Linear SVC has outperformed other classifiers with 300 tweets up to 72.66% but then the accuracy drop significantly to 68% when they used 500 tweets. The accuracy of both classifiers keeps increasing with the number of tweets and the highest accuracy achieved by SVM with over 80% accuracy using 400 tweets. It can be observed, the increment of accuracy suggest that the more tweets are gathered as training data, the better the performance of the classification model. Besides, the results obtained prove that as more data is available to capture the author's style and discriminate the writing styles of the authors.

6. CONCLUSION

This paper has implemented the KadazanDusun tweets as an Under-Resourced language data set on Author Identification task for short text. For the purpose of automatic AI for short texts, experiments have been conducted by combining different sets of stylometric features that are independent-language features including word-level and character-level n-grams. The combination of word unigram and character 3-grams and all word n-grams results with high accuracy with 80.17% accuracy. The features set combinations are highly predictive for AI task in KadazanDusun language data set. On the other hand, SVM classifier achieved high performance in this task on the combination of word-level n-grams with character 3-grams that have shorter execution time compared to character 4-grams. In future, other types of features related to language-dependent such as Part-of-Speech (POS) n-grams will be tested. Moreover, different classifiers can be explored to examine the performance of that classifiers on this data set.

REFERENCES

[1] P. Juola and G. K.Mikros, “Cross-Linguistic Stylometric Features: A Preliminary Investigation”, *in JADT2016: International Conference on Statistical Analysis of Textual Data*, France, 2016.
[2] R. Chen, L. Hong, C. Lu and W. Deng, "Author Identication of Software Source Code with Program Dependence Graphs," *in 2010 IEEE 34th Annual Computer Software and Applications Conference Workshops*, Seoul, 2010.
[3] A. Rocha and W. J. Scheirer et al, “Authorship Attribution for Social Media Forensics”, *IEEE Transactions on Information Forensics and Security*, pp. 5-33, 2017.
[4] R. Schwartz and O. Tsur et al, “Authorship Attribution of Micro-Messages”, *in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, 2013.
[5] O. d. Vel and A. Anderson et al, “Mining E-mail content for Author Identification Forensics”, *ACM SIGMOD Record*, NY, USA, 2001.
[6] H. Mohtasseb and A. Ahmed, “Mining Online Diaries for Blogger Identification”, *in Second International Conference, Big Data Anlaytics (BDA) 2013*, Myosore, India, 2013.

[7] P. Shrestha, S. Sierra and F. Gonzalez, "Convolutional Neural Networks for Authorship Attribution of Short Texts", *in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, 2017.
[8] J. Soler-Company and L. Wanner, "On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification", *in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, 2017.
[9] A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning", *in 2016 15th IEEE 9 International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, 2017.
[10] S. Okuno, H. Asai and H. Yamana, "A challenge of authorship identification for ten-thousand-scale microblog users", *in 2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2015.
[11] R. Banga and P. Mehndiratta, "Authorship attribution for textual data on online social networks", *in 2017 Tenth International Conference on Contemporary Computing (IC3)*, Noida, India, 2018.
[12] A. Omar, "Processing Malaysian Indigenous Languages: A Focus on Phonology and Grammar", *Open Journal of Modern Linguistis*, 4, pp. 728-738, 2014.
[13] S. Krauwer, "The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap", *in Proceedings of the 2003 International Workshop Speech and Computer*, Moscow, Russia, 2003.
[14] I. Frommholz and H. M. al-Khateeb et al, "On Textual Analysis and Machine Learning for Cyberstalking Detection", *Datenbank Spektrum*, vol. 16, no. 2, p. 127–135, 2016.
[15] S. Nirkhi, R. V. Dharaskar and V. Thakare, "An Experimental Study on Authorship Identification for Cyber Forensics", *IJCSN International Journal of Computer Science and Network*, pp. 756-60, 2015.
[16] M. Almishari and D. Kaafar et al, "Stylometric Linkability of Tweets", *in WPES '14 Proceedings of the 13th Workshop on Privacy in the Electronic Society*, Arizona, 2014.
[17] Y. Yu and L. Zhou, "Acoustic Emission Signal Classification based on Support Vector Machine," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 10, no. 5, pp. 1027-1032, 2012.
[18] R. A. Igawa and A. M. G. de Almeida et al, "Recognition of Compromised Accounts on Twitter", *in SBSI 2015 Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective*, Goias, 2015.
[19] C. Jafran, "Perkataan makian dan kutukan (bahasa kasar) dalam bahasa Kadazandusun", available at *http://gagaritabada.blogspot.com/2014/02/perkataan-makian-dan-kutukan-bahasa.html*, 10 February 2014.
[20] M. Koppel and J. Schler, "Computational Methods in Authorship Attribution", *Journal of The American Society for Information Science and Technology*, vol. 60, no. 1, pp. 9-26, 2009.
[21] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods", *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538-556, 2009.
[22] R. Layton, P. Watters and R. Dazeley, "Authorship Attribution for Twitter in 140 Characters or Less", *in 2010 Second Cybercrime and Trustworthy Computing Workshop*, Ballarat, VIC, 2010.
[23] D. Wright, "Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic", *International Journal of Corpus Linguistics*, vol. 22, no. 2, 2017.
[24] D. Li-guo and L. A.-p. Di peng, "A New Naive Bayes Text Classification Algorithm," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 12, no. 2, pp. 947-952, 2014.
[25] A. A. G. Gladwin, M. J. Lavin and D. M. Look, "Stylometry and collaborative uthorship: Eddy, Lovecraft, and 'The Loved Dead'," *DSH*, vol. 32, pp. 123-140, 2017.
[26] A. M. Kibriya and E. Frank et al, "Revisited, Multinomial Naïve Bayes for Text Categorization", *in Australasian Joint Conference on Artificial Intelligence*, Cairns, 2004.