A Preliminary Performance Comparison of Machine Learning Algorithms for Web Author Identification of Vietnamese Online Messages

Bui Khanh
ITMO University
Saint Petersburg, Russia
khahu132@gmail.com

Alisa Vorobeva
ITMO University
Saint Petersburg, Russia
alice_w@mail.ru

***Abstract*— With the rapid development of the Internet and accompanying technologies, communication between people has become easier than ever. Email, news sites, social networking applications become an indispensable connection tool. However, the Internet is also a favorable environment for cybercriminals with malicious activities. Therefore, it is necessary to develop a method to determine which user is the author of the online message. There has been a lot of researches with different corpora and various languages. In this article, we propose an approach to identify the authors of online messages in Vietnamese based on machine learning algorithms. Algorithms used include Naïve Bayes, SVM, Random Forest, and Logistic Regression. The algorithm that has yielded the best results in most cases is Random Forest.**

I. Introduction

More than 50 years have passed since the internet was born. While information security has become one of the main concerns of users, governments and businesses, the cyber world provides an anonymous environment for cybercriminals to operate. Cybercriminals can use the internet for malicious purposes, such as spamming, trolls or phishing users. Ensuring the confidentiality of user information is extremely important. Besides that, it is required to build up innovative tools and techniques to appropriately analyze large volumes of suspicious online messages.

Identification of the Internet user, who wrote some text, is part of authorship attribution or authorship identification. Authorship identification is the task of identifying who wrote a given piece of text from a given set of candidate authors [6].

In fact, with very large numbers of users, the use of algorithms and applying models is rather inefficient. However, in reality, there is only a limited number of suspicious users. These people may be extremists, terrorists using social networking tools and internet to serve the purpose of propaganda, threats, or claims of violence. Research offers valuable techniques in digital forensics for supporting crime investigation and security as in [1], [3], [11], [14].

This paper compares the performance of various classifiers in terms of accuracy for the user identification task of online messages. Naïve Bayes, Support Vector Machines, Random Forest, and Logistic Regressions classifiers are used for performing experimentation. This paper also investigates the appropriate classifier for solving the authorship of anonymous online messages.

After introducing previous researches, Vietnamese language characteristics in section I, we show the model design and research's steps in section II. The experimental results will be presented in section III while the last section are conclusions and future works.

A. Previous researches

In the last 20 years, there has been a great deal of research regarding authorship attribution based on various methods and data sets.

In previous studies, the methods used were also very diverse, focusing on:

- SVM [4], [10], [21];
- Naïve Bayes [23], [24];
- Decision trees [4], [22];
- Random Forest [25]-[29], etc.

Approaches can be divided into two main directions:

- Handcrafted features: Approaches for authorship attribution base on features that are manually engineered like average length of words, the frequency of digits used, the frequency of letters used and use them to classify texts.
- Learned features: Approaches base on features which are automatically obtained from text with tools (example word2vec, glove, etc.).

Traditionally, research in this field has focused on formal texts, such as essays and novels, but recently more attention has been given to texts generated by online users, such as e-mails and blogs. For long text, to make it easier to find the user's stylometry, several studies on long text have been presented in [18], [19]. Authorship attribution of such online texts is a more challenging task than traditional authorship attribution, because the number of candidate authors is often large, and texts usually in short length. The problem we must face is that the represented characteristics will become less and the limitation in text length makes some computational constraints on some measures such as vocabulary richness. However, there have been numerous research studies in the short text, with diverse datasets from email [21], tweet [7], [13], to forum posts [4], [17], blog post [9].

The language of texts used in research is also diverse. Most of them are English data [16], [19], [22], besides there are studies using Arabic [11], Russian [15], Chinese [22]. Some studies focus on a few local languages like Portuguese [32], Spanish [20]. For each different language, there is a different unique feature. Several studies with different languages are shown in the Table I according to [15],[35],[20].

TABLE I. RESEARCH ON AUTHORSHIP IDENTIFICATION

Author	Year	Language	Algorithm
Kjell et al. [38]	1995	English	k-NN
Hoom et al	1999	Dutch	NN, NB
Kukushina [39]	2001	Russian	Distance Markov
Stamatatos [40]	2001	Greek	LDA
Benedetto et al [41]	2002	Italian	Distance (compression)
Diederic et al [42]	2003	German	SVM
Khmelev and Teahan [43]	2003	Russian	Distance Markov
Abbasi [2]	2005	Arabic	SVM
Zheng and Li [22]	2006	English Chinese	NN, SVM
Pavelec [32]	2007	Portuguese	SVM
Vishnu	2013	Telugu	NB, SVM
Mario Crespo [20]	2016	Spanish	SVM
Vorobeva A.A. [26]	2016	Russian	SVM, NB, DC, RF
Kale et al. [36]	2018	Marathi	k-NN,NB, N-gram

For this study, we used Vietnamese because besides being one of the 15 most popular languages on the Internet [33] as shown in Fig. 1. With the number of people with access to the internet of up to 50 million, the Vietnamese language is a huge data source for research. There are some studies on Vietnamese language related to author profiling [2], [5], text summarization [30], topic modeling [12], but they have not been considered for web author identification before.

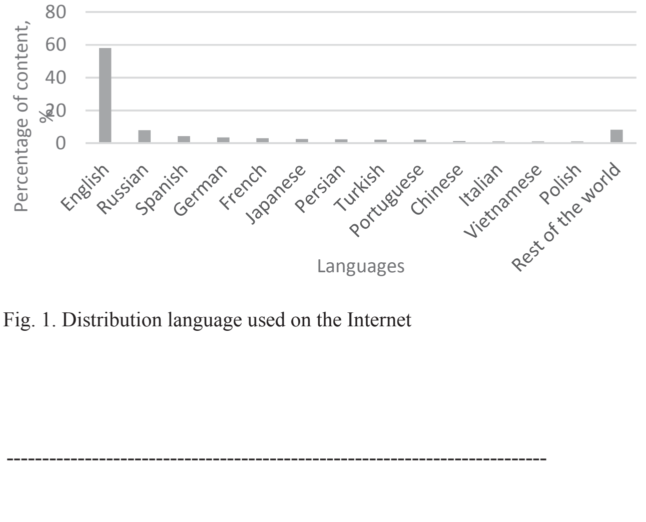

Fig. 1. Distribution language used on the Internet

Among the 10 websites with the largest traffic in Vietnam in 2019, there are 6 social networking sites and news, which is an extremely large environment for criminal activity. That leads us to develop an approach for web author identification (or user identification) of Vietnamese online messages.

B. Classification task in user identification

The task of user identification can be formulated in word of classical classification task.

Given a set of messages $M = \{m_1, ..., m_t\}$ and set of users $U = \{u_1, ..., u_k\}$, where t - number of messages and k – is number of authors. In this work is used instance-based approach, the user u_i can be presented as subset $M_j \in M$.

There is some subset of message $M' \in M$ of known user – the samples data $\{(m_1, u_1), \ldots, (m_t,u_k)\}$. We must find effective algorithm or classifier a: $m_j \rightarrow U$, which calculates the probability of authorship for each user to be an author of message m_j and output probabilities sorted in descending order Pr (u_i author m_j).

Most of classification algorithms assign a sample m_j a class label (or user) u_i with maximum probability and outputs only one user with Pr ($u_{\max}$ author of m_j).

However, having full list of probabilities for each user instead of only one most probable user is rather useful for manual authorship analysis in forensic linguistics to narrow the set of candidate users.

In the actual problem of investigating cybercrime, terrorist tracing, the number of candidate users will be narrowed down to a certain group with patterns so that the results can be more accurate. While absolute accuracy cannot be achieved, the application of a classification model can assist in forensic linguistics. We select the number of candidates for each test from 10 to 30 users, the number of messages will depend on the purpose of the task to be tested.

Also, we split the samples data for two subsets: M_{tr} and M_{test}. At first, we train the classifier on subset M_{tr}, and then test it on M_{test} to validate the prediction power of model. After the classifier is trained and tested, we have validated model for user identification, and it can be used to identify author of message m_j. This is classical approach of supervised learning.

C. Vietnamese language characteristics

The Vietnamese language goes through a period of complex development. Historically, Vietnam has long used variant Chinese characters (chữ Nôm). Today Vietnam uses the alphabet based on Latin alphabet, but unlike the English alphabet of 26 letters, the Vietnamese alphabet includes 29 letters as shown below (see Table II).

TABLE II. VIETNAMESE LANGUAGE ALPHABET

a	â	ă	b	c	d	đ	e	ê	g
h	i	k	l	m	n	o	ô	ơ	p
q	r	s	t	u	ư	v	x	y	

Vietnamese is an isolating and monosyllabic language. Vietnamese word forms never change, contrary to occidental languages that make use of morphological variations (plural form, conjugation, etc.); Therefore, in Vietnamese does not exist lemmatization and stemming.

The smallest unit of Vietnamese language is called "tiếng" (syllabic). In linguistics, a word is a basic unit. English words are syllables separated by space. However, a word in Vietnamese can be simple words (monosyllable) or complex words (including spaces, polysyllable, can be reduplicative or compound words). For example, "học sinh" – a compound word meaning "student" in English. Word segmentation is one of the very important jobs, even can be said that its the most important work in Vietnamese natural language processing. There have been many studies with different approaches and accuracies in the Vietnamese word segmentation task. Some research results are presented in Table III below according to [31], and pyvi tool from Viet-Trung Tran [8].

TABLE III. VIETNAMESE WORD SEGMENTATION RESULTS

Approach	Accuracy, %
VnTokenizer	97.33
DongDu	97
UETsegmenter	97.87
JVnEgmenter-Maxent	97.00
RDRsegmenter	97.9
Pyvi	97.9

Vietnamese is made up of tones (includes 6 tones) that are written with letters. Different words with different tones will have different pronunciation and have a different meaning (e.g. mau – fast, màu – color while máu – blood, 3 words that differ only in tone).

In this study, we used pyvi as a process to calculate the density of features and accept the errors caused by this tool.

To be able to accurately assess the grammatical differences of 2 different users when using Vietnamese is really an extremely complex job. However, we will consider the example taken from the data set presented later in the study. For example, we have 2 users.

User A:

Mess 1: Thế mà HLV ko cho vào từ đầu, Công Phượng quẩy cho chắc ko cần phải cứu =))

Mess 2: cải thiện chất lượng đường xá, mở rộng trước khi nghĩ đến phương tiện công cộng. Đường HN tôi thấy quá tệ

Mess 3: dẹp mà dẹp bỏ đc cả ôtô và xe máy luôn thì tuyệt vời nhỉ. đi bộ, xe đạp và tàu điện thôi. quá ngon ! mà nói vậy chứ hệ thống chất lượng giao thông VN làm sao đáp ứng dc việc đó. chắc còn lâu lắm =))

User B:

Mess 1: Về nhà rồi...thật là nhẹ nhõm cả người...xin chúc mừng

Mess 2: Thật là thất đức, nếu còn mà không bán...

Mess 3: Việt Nam vô địch sau bao nhiêu năm trời!

Through 2 examples above, if you pay attention to the signs, you can see the style of 2 users are different. If the first user has a high frequency of using abbreviations and special characters, the second user often uses string "..." to express emotions. Therefore, for this study, we chose a statistic-based calculation approach. Details of the implementation process are shown in section below.

II. FRAMEWORK ARCHITECTURE FOR AUTHOR IDENTIFICATION OF VIETNAMESE ONLINE MESSAGE

In this section we will describe the architecture of the framework, the data preparation process, and the features used.

D. Model design

A model for user identification is made up of:

- Build data scraping tool to build corpus from users;
- Preprocessing (delete blank or duplicate messages);
- Extract features from messages;
- Features selection and Algorithm selections;
- Training and classification.

We propose a framework for web author identification of online messages of Vietnamese online messages as described in Fig. 2.

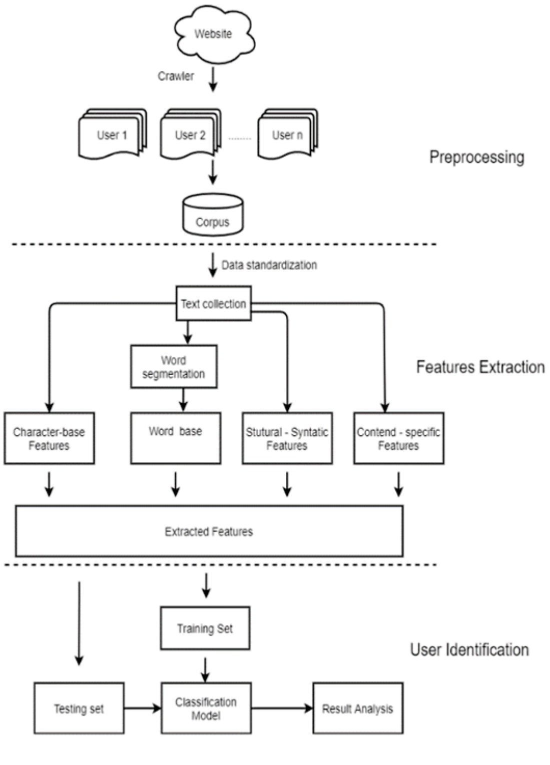

Fig. 2. Framework for web author identification of Vietnamese online messages

The main steps include data preparation, data processing, features selection and application of algorithms.

E. Prepare Dataset

This data set is obtained from Vnexpress.com, which is the most visited newspaper website in Vietnam. All posts and comments are public. We have written a data scraping tool from Selenium and python libraries to process HTML / XML files. After collecting data, we have built the dataset only includes text classified by author number, all other information such as username, biography, etc. is removed for research purposes.

After the data collection process, the dataset consisted of 99431 messages from 100 users. Because Vnexpress.com does not allow inserting images, icons or tables into the article, so we retain the original data for processing.

Then with each research task, we selected different numbers of users and messages. The number of users per test will be from 10 to 30.

The characteristics of online text messages are usually short, as can be clearly seen in Fig. 3.

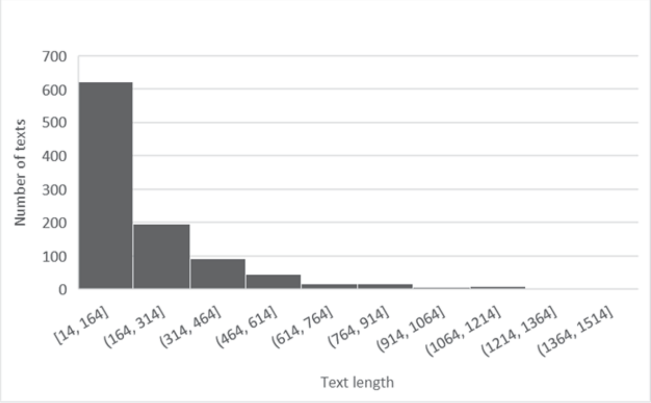

Fig. 3. Distribution of texts length

On the chart is the number of messages from 10 random users taken from the dataset, it can be seen that the message length is concentrated from 14 -914 characters.

However, each user will have a different distribution, which is shown in Fig. 4 below.

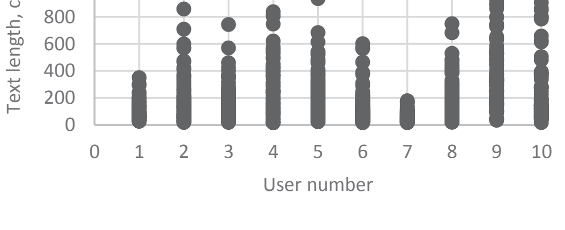

Fig. 4. Distribution of texts length per user

While the first user and 7th user have a rather low dispersion of the message length, the 10th user is the opposite.

F. Features Extraction

We extract character-based features and word-based features from all messages using scripts written by Python.

Character-based features include features based on frequency of characters such as:

- Total number of characters;
- Frequency of Vietnamese alphabetic character;
- Frequency of non-Vietnamese alphabetic character;
- Frequency of special characters;
- Frequency of capital characters etc.

Word-based features include features based on word frequency:

- Total number of words;
- Total number of complex words;
- Total number of single words;
- Total unique words/ Total words;
- Measures defined by Yule, Simpsom, Sichele, Brune, Honore;
- Frequency of capital words, etc.

As mentioned above, we use the pyvi tool to serve the purpose of word segmentation.

2) Structural - Syntactic Features: Total number of sentences and paragraph, features related to paragraph.

- Frequency of punctuations
- Frequency of Vietnamese functions words
- Texts end with special characters or has quote content etc.

3) Content-specific Features Group:

- Frequency of some Vietnamese personal pronouns (tôi, bạn, tớ, cậu, tau, mi, etc.).
- Time posted
- Frequency of abbreviations, algorithm counts the number acronym developed by us.

Total features set includes 215 features.

G. User identification

After calculating the features, we have implemented algorithms to classify users. Many issues can affect the accuracy of user classification algorithms such as:

- Number of candidate users;
- Corpus size (texts length or number of texts);
- Distribution of training corpus over user (balance or imbalance) [26].

And to evaluate the effectiveness of algorithms, we have set out several tasks to answer the following problems:

1) The influence of the number of users on the algorithm accuracy.

2) Effect of the number of messages on algorithm accuracy.

3) Performance of the algorithm for different users group and different texts length.

Identification accuracy is calculated by the formula (1):

$$\text{Accuracy} = \frac{\text{number of users identified correctly}}{\text{number identified users}} \quad (1)$$

The algorithm used to classify:

- SVM,
- Naïve Bayes,
- Random Forest,
- Logistic Regression.

1) SVM. An auxiliary vector machine (SVM) was introduced by Cortess and Vapnik in 1995.

SVM is an effective method for data classification problems. It is a specialized tool for text classification and opinion analysis problems with the advantage of processing on space with its height and flexibility. Classification is usually non-linear, the ability to apply a new kernel allows flexibility between linear and non-linear methods, thereby making the classification performance larger.

2) Naïve Bayes (NB) is a popular classification technique in supervised machine learning. The main idea of this technique is to rely on conditional probabilities between words or phrases and classification labels to predict which new text belongs to. Naïve Bayes has many applications of text classification problems, building automatic spam filters; or in point-of-view problems because of its easy-to-understand, easy-to-deploy as well as accuracy.

The basic idea of the Naïve Bayes approach is to use conditional probabilities between characteristics and labels to predict the probability of the label of a text to be classified.

Due to simplicity, NB models are often beaten by models properly trained and tuned using the other algorithms.

3) Random Forest (RF) is a member of the Decision Tree algorithm family and is one of the most accurate classification algorithms available today. Its effectiveness with high dimensional data is also one of the advantages of Random Forest.

4) Logistic Regression (LR) is a common algorithm, especially used in classification problems. Its advantages are outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting.

We also used Relief – f algorithm [37] to support features selection as we realize that selecting certain features will yield greater results if applied to all features. All performances of classification algorithms will be shown in next section.

III. Performance of classification algorithms

We, in turn, perform experiments to solve the tasks mentioned above.

1) The influence of the number of users on the accuracy of algorithms.

To solve the first task, we selected 30 random users out of a set of 100 users. Then took 200 messages each user, from this dataset we divided into 2 subsets. One subset called training set, which is used to train the classification, and the other called testing set, it's used to validate the power of the model. The ratio between them is 8:2 and the larger is training set. Detailed results algorithms are shown in Fig. 5.

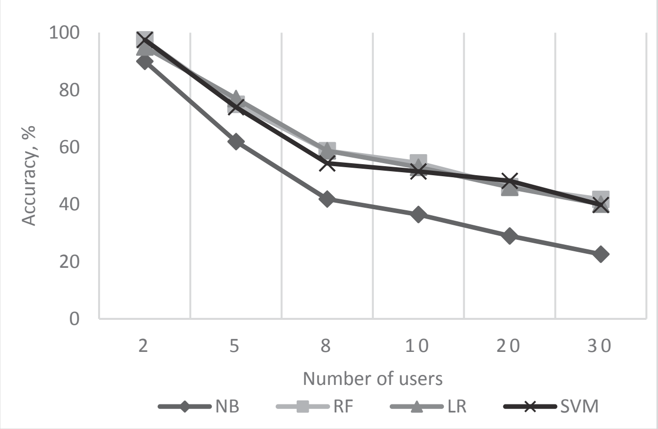

Fig. 5. Performance of algorithms with different number of users

On the figure we can see that all algorithms reduce accuracy when increasing the number of users. This is understandable for identifications task. Until the number of users reached 30, the algorithm accuracy only ranged from 20-40%. Table IV below describes the accuracy of each algorithm.

TABLE IV. Identification accuracy of selected algorithm on datasets withs different number of users

	Identification accuracy, %					
Number of users	**2**	**5**	**8**	**10**	**20**	**30**
NB	90	62	41.875	36.5	29	22.6
RF	97.5	75	58.75	54.5	46.25	41.8
LR	95	77	58.75	53	46	40.1
SVM	97.5	74	54.375	51.5	48.25	39.83

The results of algorithms are rather high when using only 2-5 users.

RF keeps the best accuracy in most attempts. While the RF, LR, SVM algorithms keep the accuracy quite good and approximately the same, NB has the lowest accuracy, down from 90% (applied to 2 users) to only about 20% when applied to 30 users.

2) The influence of the number of messages on the accuracy of algorithms.

Several methods and approaches have been proposed for solving the problem of web author identification of online messages, it is still not clear what the volume of message needs to be for reliable identification. This time we used 10 authors from the dataset, each author has the same number of messages each attempt, ranging from 25 to 200 messages. As in task 1, we divided dataset into two subsets with ratio 8: 2. The accuracy of identification is shown in Fig. 6

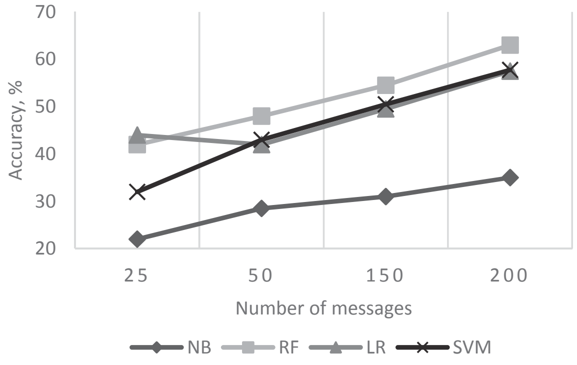

Fig. 6. Performance of algorithms for difference number of messages

Accuracy tends to increase with increasing quantity. When the number of messages reached was 200 per user, the best accuracy was achieved when using RF.

3) *Performance of algorithms for different user groups and influence of message length*

To be able to simultaneously solve the problems posed in task 3, we have chosen a dataset with special criteria. We selected 30 users from a list of 100 users which satisfy the conditions for the number and length of messages.

Split user randomly into 3 different groups of authors $U_i = \{u_1, ..., u_{10}\}, 1 \le i \le 3$. This is also the number of candidate authors that are often chosen for identification in researches.

Each user has 200 messages, including 100 messages of short length (<500 characters), and 100 messages of medium to long length (>500 characters). For each group we have conducted 2 experiments, one for short messages, and the other for messages with larger teams. The main context is still to determine who is the author of any message, but the randomization of groups allows us to have an overview of the accuracy of the algorithms when applied. Does an algorithm have good accuracy for all different groups? From there, we can choose appropriate algorithms for real-life problems. Experiments also show the influence of the length of the text on the accuracy of the classification. Results are shown in Fig. 7.

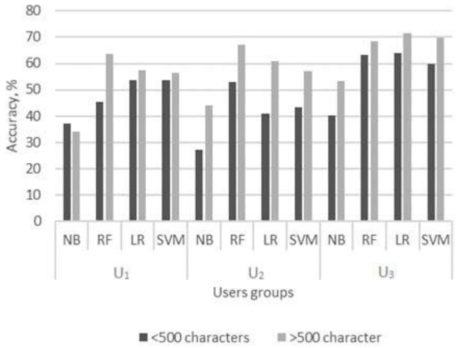

Fig. 7. Performance of algorithms for different user groups and different type of messages

Look at the figure we can clearly see, generally, in most cases, messages with longer lengths will result in more accurate classification. This is understandable when long messages bring more "knowledge" about style and personal aspects of the user.

Once again, RF achieved the highest classification result in almost tasks (20% higher than NB when applying for both messages> 500 characters and <500 characters), however in some cases, LR reached the highest results.

Next, we perform the calculation to have the average result of identifying 3 user groups as shown in the following table (see Table V).

TABLE V. AVERAGE ACCURACY OF SELECTED ALGORITHM FOR VARIOUS TEXTS LENGTH

Algorithm	Average accuracy, %	
	> 500 characters	< 500 characters
NB	34.84	43.8
RF	53.91	66.35
LR	52.72	63.4
SVM	52.21	61.1

Compare the results in table V when performing algorithms in task 3 and the results obtained in task 1 can see that when the message length is stable at higher than 500 characters, the result will be much higher than when applying algorithms with variable length.

IV. CONCLUSION

In this article, we have proposed an approach for web author identification for the new language (Vietnamese) based on machine learning methods. We have set up three tasks, and in most cases, Random Forest provided the best results. First experiment has shown that the best identification accuracy is achieved for 2-5 users (97,5% with Random Forest and SVM for 2 users, 77% with Logistic Regression for 5 users) and decreases with an increasing number of users (41.5% with Random Forest for 30 users). For second experiments with 10 users, best accuracy is 63% (with Random Forest for 200 messages per user). Experiments demonstrate that for longer messages, the classification accuracy is better than low-length messages. The results are quite good, though accuracy is not so high as with the state of art for messages in English. Partly most of the data we used have short length, another reason because Vietnamese language still has many difficulties in handling, so there are many things we cannot use to research such as Part-of-speech tagging, Named-entity recognition, etc. The research also has limitations when the data used is online messages from online newspapers, leading to similarity of the content of the message. But the main purpose of the research is not only to provide an approach to processing online messages, but also to apply those techniques to identify other messages, which can be used by cybercriminals to spam, defame or harass. Besides that, the Vietnamese language also has many interesting linguistic features such as tones, spells or local characteristics words. So in the future, when the techniques of natural language processing with Vietnamese language reach the allowed accuracy, we will try to apply it as well as intensifying experiments to find threshold about length, number of appropriate messages per author to increase the accuracy of web author identification.

ACKNOWLEDGMENT

This paper is supported by Government of Russian Federation (grant 08-08).

REFERENCES

[1] Ingo F., Haider M., Martin P., Zinnar G., Mitul S. & Emma "On Textual Analysis and Machine Learning for Cyberstalking Detection", *Datenbank Spektrum* 16, 2016, 127–135

[2] Pham D.D., Tran G.B., Pham S.B "Author profiling for Vietnamese blogs", *International Conference on Asian Language Processing*, 2009, pp.190-194.

[3] Zheng, R., Qin, Y., Huang, Z., Chen, H., "Authorship analysis in Cybercrime Investigation", *Intelligence and Security Informatics*, 2003, pp. 59-73.

[4] Abbasi, A. and H. Chen, "Applying authorship analysis to extremist-group Web forum messages," *IEEE Intelligent Systems*, vol. 20, no. 5 (Special issue on artificial intelligence for national and homeland security), 2005, pp. 67–75.

[5] Duong Tran Duc , Pham Bao Son , Tan Hanh, "Author Profiling of Vietnamese Forum Posts - An Investigation on Content-based Features", *VNU Journal of Science: Computer Science and Communication Engineering*, [S.l.], v. 33, n. 1, aug. 2017. ISSN 2588-1086.

[6] Ahmed M. Mohsen, Nagwa M. El-Makky, Nagia Ghanem, "Author Identification Using Deep Learning", *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 898-903.

[7] Roy Schwartz, Oren Tsur, Ari Rappoport Moshe Koppel "Authorship Attribution of Micro Messages", *Empirical Methods in Natural Langauge Processing*, 2013, pp.1880–1891.

[8] Python Vietnamese Toolkit. Viet T.T Web: https:/pypi.org/project/pyvi.

[9] A.A. Vorobeva, "Analiz vozmozhnosti primeneniya razlichnih lingvisticheskih harakteristik dlja identificacii avtora anonimnih korotkih soobshenij v globalnoj seti Internet", *Informaciya i kosmos*, 2013, no. 4, pp. 42-47.

[10] J. Diederich, J. Kindermann, E. Leopold, G. Paass. Leibniz, "Authorship Attribution with Support Vector Machines", *Applied Intelligence*, 2000, vol. 19, issue 1, pp. 109-123

[11] Bedoor Y. AlHarbi, Mashael S. AlHarbi, Nouf J. AlZahrani, Meshaiel M. Alsheail,Jowharah F. Alshobaili, Dina M. Ibrahim, "Automatic Cyber Bullying Detection in Arabic Social Media", *International Journal of Engineering Research and Technology*. ISSN 0974-3154, Volume 12, Number 12 (2019), pp. 2330-2335

[12] Ha Nguyen Thi Thu, Tinh Dao Thanh, Thanh Nguyen Hai, Vinh Ho Ngoc, "'Building Vietnamese Topic Modeling Based on Core Terms and Applying in Text Classification", *Proc. of Fifth IEEE International Conference on Communication Systems and Network Technologies*, 2015, pp. 1284-1288, DOI 10.1109/CSNT.2015.22,

[13] Nilan Saha Pratyush Das Himadri Nath , "Authorship Attribution of Short Texts using Multi Layer Perceptron", *International Journal of Applied Pattern Recognition*, 2018 Vol.5 No.3, pp.251 - 259

[14] Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace." *ACM Transactions on Information Systems (TOIS),* 2008, pp.1-29

[15] Vorobeva A.A. "Forensic linguistics: automatic web author identification", *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 2, pp. 295–302, doi:10.17586/2226-1494-2016-16-2-295-302

[16] A. Caliskan-Islam Stylometric Fingerprints and Privacy Behavior in Textual Data, PhD thesis, Drexel University, 2015.

[17] Pillay S. R., & Solorio T. "Authorship attribution of web forum posts", *eCrime Researchers Summit*, 2010, pp.1-7.

[18] Monaco, J. V., Stewart, J. C., Cha, S. H., and Tappert, C. C. (2013, September). "Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works", *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS),* 2013*,* pp. 1-8.

[19] Koppel, M., and Schler "Authorship verification as a one-class classification problem", *In Proceedings of the twenty-first international conference on Machine learning*, July 2004, pp. 62.

[20] Mario Crespo, "Analysis of parameters on author attribution of Spanish electronic short texts", *Research in Corpus Linguistics*, 2016, pp. 25-32.

[21] O. de Vel, A. Anderson, M. Corney, G. Mohay, "Mining e-mail content for author identification forensics", *ACM Sigmod Record*, vol. 30(4), pp. 55-64.

[22] R. Zheng, J. Li, Z. Huang, H. Chen, "A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques", *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 57, no. 3, 2006, pp. 378-393.

[23] Peng F., Schuurmans D., Wang S., Keselj V, "Language independent authorship attribution using character level language models", Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, vol. 1, 2003, pp. 267-274.

[24] Almishari M., Kaafar D., Oguz E., Tsudik G. "Stylometric linkability of tweets", *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, 2003, pp. 205-208

[25] A. Bartoli, A. Dagri, A.D. Lorenzo, E. Medvet, F. Tarlao "An author verification approach based on differential features", *CLEF 2015 Evaluation Labs*, 2015.

[26] Vorobeva A. A. "Examining the performance of classification algorithms for imbalanced data sets in web author identification", *18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT)*, 2016, pp. 385-390

[27] M. L. Pacheco, K. Fernandes, A. Porco, "Random forest with increased generalization: A universal background approach for authorship verification", *CLEF 2015 Evaluation Labs*, 2015.

[28] P. Maitra, S. Ghosh, D. Das, "Authorship verification: An approach based on random forest", *CLEF 2015 Evaluation Labs*, 2015.

[29] Vorobeva A. A. "Influence of features discretization on accuracy of random forest classifier for web user identification", *20th Conference of Open Innovations Association (FRUCT)*. – IEEE, 2017. – pp. 498-504.

[30] Nguyen-Hoang, Tu-Anh et al. "An Efficient Vietnamese Text Summarization Approach Based on Graph Model", *IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, 2010, pp.1-6.

[31] Dat Quoc Nguyen , Dai Quoc Nguyen , Thanh Vu , Mark Dras , Mark Johnson, "A Fast and Accurate Vietnamese Word Segmenter", *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2018, pp.2582-2587.

[32] Daniel Pavelec, Edson Justino, and Luiz S. "Oliveira Author Identification using Stylometric Features", *of Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*. Vol 11, No 36, 2007, pp. 59-65.

[33] W3Techs - World Wide Web Technology Surveys. Historical trends in the usage statistics of content languages for websites Web: https://w3techs.com/technologies/history_overview/content_language

[34] Luyckx K., & Daelemans W., "The effect of author set size and data size in authorship attribution", *LLC*, 26, 2011, pp. 35-55.

[35] A, Pandian & Sadiq, Abdul Karim., "Innovative Methods in Identifying Authors of Documents*", International Journal of Engineering and Technology*. Vol.6 No.6 ,2014, pp. 2512-2520

[36] Kale, Sunil & Prasad, Rajesh. "Author Identification on Imbalanced Class Dataset of Indian Literature in Marathi". *International Journal of Computer Sciences and Engineering*, Vol 6, 2018, pp. 542-547.

[37] Kononenko I. "Estimating attributes: analysis and extensions of RELIEF", *Lecture Notes in Computer Science*. Vol 784, 1994, pp. 171–182.

[38] Kjell, Bradley, W. Addison Woods, and Ophir Frieder. "Information retrieval using letter tuples with neural network and nearest neighbor classifiers." 1995 *IEEE International Conference on Systems, Man*

and Cybernetics. Intelligent Systems for the 21st Century. Vol. 2. IEEE, 1995.

[39] Kukushkina, Olga V., Anatoly A. Polikarpov, and Dmitry V. Khmelev. "Using literal and grammatical statistics for authorship attribution." *Problems of Information Transmission*, 2001, pp.172-184.

[40] Stamatatos, Efstathios. "A survey of modern authorship attribution methods." *Journal of the American Society for information Science and Technology*, 2009, pp. 538-556.

[41] Basile, Chiara, et al. "An example of mathematical authorship attribution." *Journal of Mathematical Physics*, 2008.

[42] Diederich, Joachim, et al. "Authorship attribution with support vector machines." *Applied intelligence*, 2003, pp. 109-123.

[43] Khmelev, Dmitry V., and William J. Teahan. "A repetition based measure for verification of text collections and for text categorization." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. 2003.