Influence of Features Discretization on Accuracy of Random Forest Classifier for Web User Identification

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

***Abstract*—Web user identification based on linguistic or stylometric features helps to solve several tasks in computer forensics and cybersecurity, and can be used to prevent and investigate high-tech crimes and crimes where computer is used as a tool. In this paper we present research results on influence of features discretization on accuracy of Random Forest classifier. To evaluate the influence were carried out series of experiments on text corpus, contains Russian online texts of different genres and topics. Was used data sets with various level of class imbalance and amount of training texts per user. The experiments showed that the discretization of features improves the accuracy of identification for all data sets. We obtained positive results for extremely low amount of online messages per one user, and for maximum imbalance level.**

I. Introduction

Today everyone has wide range of technics to hide identity in the Internet. The person, who wants to hide, could use different types of anonymizer tools, anonymous mobile phone sim-cards and other technics. In recent years a large number of studies were carried in field of searching for effective methods in computer forensics, investigation of cybercrimes and evidence collection and analysis [1], [2].

Criminals exploring new ways to communicate even without sending messages via traditional networks. One common thing remains constant in all types and all ways of communication between people – text of the message sent from one to another. We can use it to attribute or identify the author, the web user who wrote it [3]-[7].

As identifier are used various distinctive, measurable characteristics of web user texts, they are used to describe individual. From this point of view linguistic identification is some sort of biometric identification based on behavioral characteristics.

Identification on linguistic or stylometric features helps to solve several tasks in computer forensics and cybersecurity, and can be used to prevent and investigate high-tech crimes and crimes where computer is used as a "tool" [8]-[14].

A. Previous researches

Today for web author identification are used various features types (lexical [15], syntactic [10], [16], [17], structural, context-specific [11], [18], semantic features, function words [11], [19], n-gramm frequencies [20]-[25] e.t.c.) and their combinations [9], [10], [18], [19], [22], [26]-[28].

Structural characteristics of texts (favorite words, specific terms and expressions) were the first used for author attribution. Later it was proved that syntactic features (such as the dividing the text into paragraphs, using of direct speech, prepositional phrase structure, complexity and length of sentences) could be used for discovering the authorship.

Another approach to the author identification based on the analysis of the words used (frequency of words of different lengths, frequencies of individual characters and their sequences, frequencies of function words).

In [11], [18] it was proved that usage of specific features of email messages improves accuracy of author identification.

Several works focus on important problems in web author identification:

- how number of authors influence on identification accuracy? [29]-[30]
- what is the minimum and the optimal text length suitable for web author attribution? [31]
- how class imbalance influence on quality of identification? [22], [32]-[34]

Most of works focus on author attribution of English texts, and only few studies author identification for texts on other languages [27], [36]-[38], and for short Russian messages or texts [18], [31], [34], [39].

Linguistic web user identification is a multi-class text classification task [2], [40]. A lot of works study the question on determination the best classification algorithm; good results were obtained with:

- Support Vector Machines [9]-[11], [15], [17]-[19], [22]-[23], [35], [41], [42];
- Naïve Bayes [21], [25];
- Decision Trees [9], [11], [19];
- Random Forest [27], [34], [37], [38], [43].

ISSN 2305-7254

In our previous researches were carried out experiments showed that Random Forest classifier has the highest accuracy for web author identification on Russian texts [34]. But we opened the question on improving accuracy of Random Forest with different technics.

We have found that dynamic features selection for each identification task and for each set of candidate authors (users) is better than selecting some static feature subset for all existing texts and authors. We tested this approach with some modern feature selection algorithms and the best results showed Relief-f, it algorithm has positive influence on accuracy [44].

In this work we study the question on influence of features discretization on efficiency of Random Forest classifier used in web user identification.

Discretization often is used on data pre-processing step for machine learning algorithms. But it also could be used as feature selection method that can impact the classification accuracy. There were several works on improving classification accuracy of Random Forest with discretization of features [45]-[47]. We focus on linguistic identification based on some features of short Russian-language online texts.

B. *Main steps of web user identification on stylistic and linguistic features of online texts*

The task of web user identification based on linguistic and stylistic features of online texts or messages could be formulated as follows. Given t_j – some text or message, $U = \{u_1, ..., u_k\}$ – a set of candidate authors, $T = \{t_1, ..., t_m\}$ – set of their messages, where m - number of messages and k – is number of users.

The user u_k is presented as subset of texts $T_k \in T$. It is assumed that the author of t_j is one from the U.

Each text is presented as set of features $t_i = F_i = \{f_1, ..., f_n\}$ and each user is collection of his texts $u_k = T_k$.

We have to calculate the probability for each user to be an author of t_j. The task is solved building the effective classifier.

We split the data for training T_{tr} and test T_{test} subsets. Then we train and test the Random Forest classifier. After that we have validated model, and it can be used to identify author of text t_j.

Identification process includes several important steps.

1) Pre-stage

- Collecting of web users and their messages.
- Features extraction from collected messages.
- Storing the data to the database.

2) Main stage: Web user identification on features of new text t_j.

- Dynamic feature selection to find the best set F' with the most informative features for each set of candidate authors. In this step Relief-f feature selection algorithm is used, that maximizes accuracy of web author identification [44].
- Discretization of continuous features.
- Building and validating of web user identification model: Training the Random Forest classifier on subset of texts, then test it to validate the prediction power of classifier.
- Features extraction and selection of F' subset from new text of uncertain authorship t_j.
- Validated web user identification model can be used to identify web user, the author of text t_j.
- Output results: list of web users sorted by probabilities of their authorship in descending order.

C. *Feature set*

There are a number of possible characteristics of the online messages that can be used for web user identification.

1) Lexical features. In this work are used lexical features of two levels: symbols and words. This group includes the frequencies of function words, frequencies of various groups of symbols (e.g. characters, digits, uppercases), frequencies of abbreviations and acronyms, frequencies of length of words and sentences, message length, average sentense and word length and some others.

2) Syntactic and structural group includes frequencies of text emphasis (bold, italic, etc.), and the logical structure of the text (dividing to blocks and paragraphs), frequencies of punctuations, frequencies of links, quotes and images and others.

3) Meta-text characteristics includes context-specific information, not directly related to the text: time and day of the week when author posted his message.

In this work we use combination of all this features types; are used both qualitative and quantitative features.

Full feature set contains 498 features $t_i = F_i = (F_il + F_is + F_im)$, where F_il - lexical, F_is - syntactic-structural and F_im – metadata (or context specific) features of text t_i. Full list was previously described in [44].

II. Influence of Features Discretization on Accuracy of Random Forest Classifier for Web User Identification

A. *Identification based on Random Forest algorithm*

In our previous researches we have found that Random Forest algorithm have some advantages for web user identification task:

1) tolerance to noisy data, can handle continuous and discrete data [48];
2) is able to handle high-dimensional data;
3) has high accuracy on balanced and imbalanced data sets;
4) is able to handle cases with low amount of training examples;

5) do not demand high processing power and is suitable for practical use in real author identification task.

The basic idea of Random Forest is to construct an ensemble (or forest) of random decision trees. The classification is made by majority vote of decision trees: every tree in the forest classify the instance to one of the classes, i.e., votes for a certain class. Further, the instance belongs to the class, which was voted for the largest number of trees (Fig. 1).

The Random Forest is based on the idea of bagging - a combination of independent classifiers could improve the accuracy. It is assumed that most of the generated trees correctly predict the user, and trees, that are mistaken, classify instance to different classes.

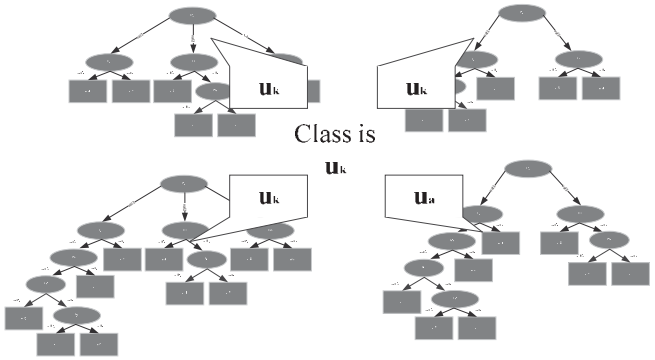

Fig. 1. Random Forest for web user identification visualization

Each tree consists from three types of structures: leaves, interior nodes and branches. The leaves of the tree contain the class or the user, in interior nodes features are stored, and branches are labeled with the values of this features. To classify a text or message, we need to come down from the tree root to a leaf, and get the class that is contained in it.

At each step of tree construction are selected the feature and its value that gives the best split or minimizes entropy. To choose which feature to split on is used information gain. Is selected the "purest" split that results in the purest nodes.

B. *Improving accuracy of web user identification with discretization of continuous features*

Discretization of continuous features is an operation of transforming the continuous-valued features to discrete or nominal features (by creating a set of intervals). It is supposed that this improves accuracy of web user identification.

The entire range of feature values can be discretized into a number of partitions or intervals. After discretization every value of nominal feature represents some interval of originals values.

For example, the post publication time can be transformed in three discrete values representing three intervals: 08:00-09:00, 13:00-14:00 and 01:00-02:00.

The discretization is a part of the data preprocessing for some important reasons: building and validating of web user identification model goes faster, discretization can provide some non-linear relations and it can harmonize heterogeneous data: some features are numerical and some are binary.

During the process of building tree any continues-valued feature f is discretized by partitioning all its values into two intervals. Is defined the value of the feature to split on or f^t - threshold value. The values $f \le f^t$ are assigned to the left brunch and $f^t > f$ are assigned to the right.

As for decision trees, most of existing algorithms use binary discretization. It seems to be rather useful to use the multi-interval discretization before constructing the tree, it helps to reduce the tree size and to improve classification accuracy.

The discretization is made is to find a set of cut points d to split the continuous range of values into numerous of informative intervals or zones.

Continuous range of values $R=(x_1), \ldots, (x_o)$ of some feature f_i is partitioned to numerous of intervals, where o – is the number of existing values.

Some continues feature f_i produces a set of intervals $\beta(x)$ (each is subset of R). Example of such intervals are:

$\beta(x)= [f_i(x) \le d], d \in R,$

$\beta(x)= [d \le f_i(x) \le d'], d,d' \in R, d \le d'.$

On Fig. 2 is presented example of values of some feature f_i and cut points d.

Fig. 2. Values of continues feature $f_i(x)$ and cut points d

In this work, we use supervised discretization, where the criterion of minimum description length (MDL) is used to stop splitting the intervals. To define the best bins algorithm finds cut-points that minimizes information entropy.

Empirical results presented in work [49] showed that this approach (MDL) allows constructing better decision trees for the same data. RF algorithm could benefit from this type of discretization as it uses information entropy minimization heuristic for selecting cut-points.

It is necessary to perform experiments to verify that discretization of continuous features can improve author identification accuracy and provides a better identification results.

C. *Text corpus and data sets*

To examine the influence of features discretization on accuracy of web user identification were carried out series of experiments.

In experiments we used text corpus, contains Russian online texts of different genres and topics; previously it was used in [34], [47], [50].

Texts have variable length; distribution is shown on Fig. 3; most of them are from 142 to 699 characters length.

Except the problem that online texts are quite short, there is the another problem - the variation in the number of messages per users. We have to consider this fact, thus number of training samples should not reduce the probability-of the fact that the user is correctly identified.

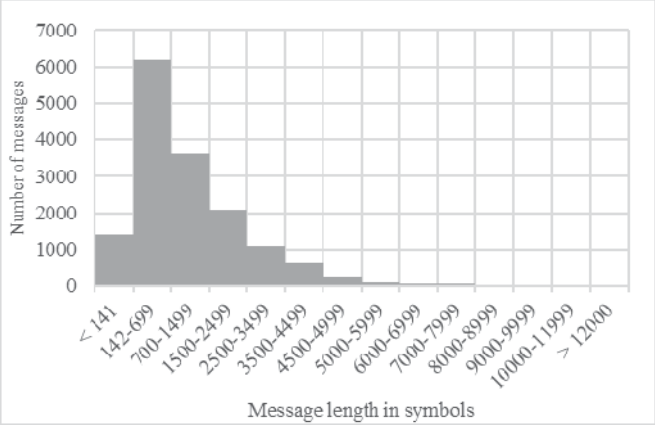

Fig. 3. Message texts length

In light of the above-mentioned conditions we generate two types of data sets:

a) with normal (or medium) amount of training samples;

b) with relatively low amount of training samples.

In this way from text corpus were formed eight groups of data sets, varying levels of class imbalance and number of training messages per user.

Each data set group contains 20 data sets, including 10 web users (U) and their texts. The total number of test data sets used in experiments is 160.

Ratio between minimum and maximum numbers of texts are shown in Table I (min:max). In imbalanced data sets number of texts has normal distribution.

Therefore, were formed two groups of balanced and six groups of imbalanced data sets, as it is shown in Table I.

TABLE I. DATA SETS WITH DIFFERENT LEVELS OF CLASS IMBALANCE

Data set group	Number of texts per web user (min:max)	
Low imbalance	20:25	8:10
Medium imbalance	10:20	5:10
High imbalance	5:25	2:10
Balanced	24:25	10:10

D. *Experiments and results*

Series of experiments were carried to study influence of the features discretization on efficiency of Random Forest classifier for web user identification. In the following subsections are presented descriptions of performed experiments.

For each data set classification was processed on data before discretization likewise the classification on discretized data.

1) Accuracy estimation

The accuracy of identification (A) is the ratio of the number of correctly identified users $IdentU_{corr}$ to the total number of the test samples (text messages) $|T_{tr}|$ (1).

$$A = \frac{IdentU_{corr}}{|T_{tr}|} \times 100\% \quad (1)$$

Accuracy estimation was carried out by 10-fold cross validation, the ratio of the training and test samples - 90% and 10%. The results are shown below.

2) Influence of the discretization on the accuracy of Random Forest classifier on imbalanced data

We have performed two series of experiments to estimate how features discretization effects on accuracy of Random Forest classifier. Classification process was executed of two types of data sets, as described above.

The first series of experiments was executed on medium amount of training text samples.

Table II indicated that Random Forest performs better on discretized feature set. Experiments results proved the hypothesis that discretization of continues features has positive influence on accuracy of Random Forest.

TABLE II. IDENTIFICATION ACCURACY IN EXPERIMENTS ON INFLUENCE OF FEATURES DISCRETIZATION ON DIFFERENT DATA SETS GROUPS WITH MEDIUM AMOUNT OF TRAINING SAMPLES

	Classification accuracy, %		
Data set group	**Number of texts per web user (min:max)**	**Before discretization**	**After discretization**
Low level of class imbalance	**20:25**	78.03	81.15
Medium level of class imbalance	**10:20**	75.01	82.2
High level of class imbalance	**5:25**	73.76	81.62
Balanced data set	**24:25**	78.17	81.23

In all experiments accuracy after discretization is much higher; maximum accuracy increase was achieved on high imbalanced data sets – 5.31%.

In experiments on non-discretized data accuracy decreased with increasing level of imbalance, but after the features discretization accuracy is stable high.

The influence of features discretization on different data sets groups is shown in Fig. 4.

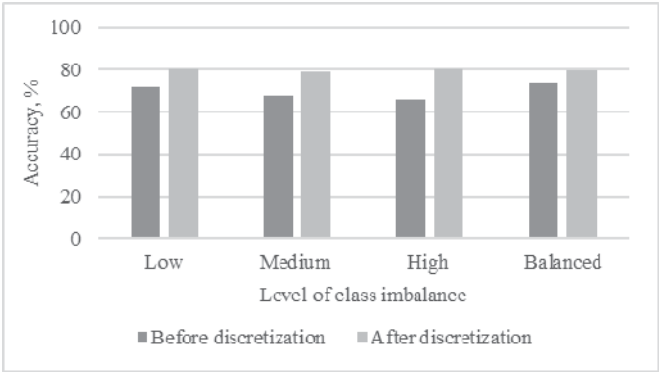

Fig. 4. Influence of features discretization on identification accuracy in experiments with medium number of training texts

The second series of experiments was executed to find how the discretization effects on the accuracy of Random Forest classifier on imbalanced data with low amount of training samples.

The results are presented in Table III.

TABLE III. IDENTIFICATION ACCURACY IN EXPERIMENTS ON INFLUENCE OF FEATURES DISCRETIZATION ON DIFFERENT DATA SETS GROUPS WITH LOW AMOUNT OF TRAINING SAMPLES

	Classification accuracy. %		
Data set group	**Number of texts per web user (min:max)**	**Before discretization**	**After discretization**
Low level of class imbalance	**8:10**	71.72	80.07
Medium level of class imbalance	**5:10**	67.66	79.01
High level of class imbalance	**2:10**	65.79	80.37
Balanced data set	**10:10**	74	79.52

As in previous series, highest influence on accuracy was achieved on data sets with high imbalance level; accuracy increased on 14.58%. The average 79.7% of correctly classified instances is on about 9.95% better than for the experiment on the data before discretization (average accuracy 69.8%) (Table III).

On Fig. 5 this results are visualized.

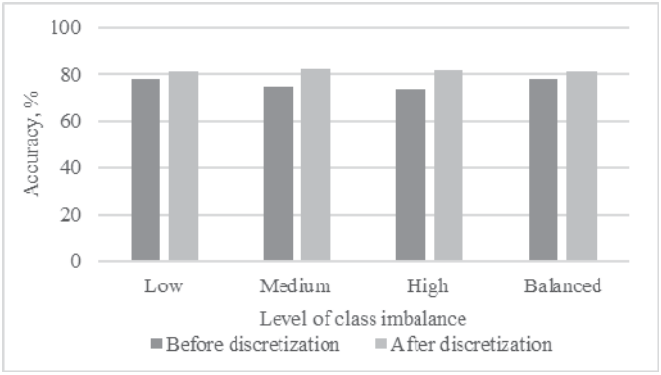

Fig. 5. Influence of features discretization on identification accuracy in experiments with low number of training texts

Fig.6 presents the accuracies increase of Random Forest classification for two types of data sets. Influence of discretization increases with increasing level of class imbalance.

Comparison of the all experiments results showed that the greatest effect on the accuracy discretization has on data sets with low amount of training texts.

All experiments showed that the application of this approach – discretization of features, improves the identification accuracy on an average of 7.63%.

We obtained positive results for extremely low amount of online messages per one user, and for maximum class imbalance level – accuracy increased on 14.58% (Fig. 5). These results are interesting and rather valuable for practical use.

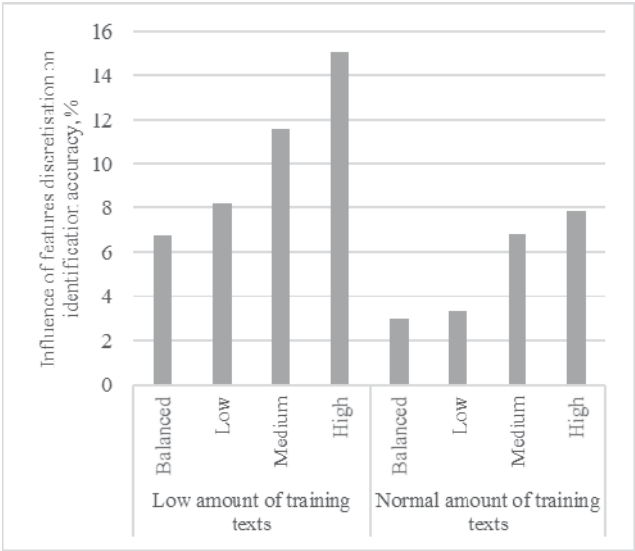

Fig. 6. Influence of features discretization on identification accuracy for different data sets groups and various number of available training texts

It is important that in all experiments discretization has positive influence on Random Forest accuracy.

For real web user identification tasks (with only few available online messages and large spread of their amount between users), combination of Random Forest and features discretization would be good practice solution.

VII. CONCLUSION

In this work we study the question on influence of multi-interval discretization of continuous-valued features on efficiency of Random Forest classifier, used for web user identification.

Generally, discretization is used on the step of data pre-processing for classification algorithms that could work only with discrete data. In this work, we use supervised discretization, where the criterion of minimum description length is used to stop splitting the intervals. Despite the fact that Random Forest handle both discrete and continues features, discretization can also be useful for improving its accuracy. Empirical results showed that this approach allows

constructing better decision trees for the same data. Random Forest algorithm benefits from this discretization as it uses information entropy minimization heuristic for selecting cut-points.

We focus on linguistic identification based on lexical, structural, syntactic and some context-specific features of short Russian-language online texts.

To evaluate the influence of features discretization on accuracy of Random Forest a series of experiments were carried out on text corpus, contains Russian online texts of different genres and topics. To simulate the real-world situation was used data sets with various level of class imbalance and amount of training texts per user.

The experiments showed that the discretization of features, improves the accuracy of identification on an average of 7.63%. We obtained positive results for extremely low amount of online messages per one user, and for maximum imbalance level – accuracy increased on 14.58%.

Obtained results and conclusions give us the direction for further work, and for deeper learning of discretization process and technics for improving web user identification accuracy.

REFERENCES

[1] I. Zikratov, I. Pantiukhin, A. Sizykh "The method of classification of user and system data based on the attributes", *18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT)*, 2016, pp. 404-409.

[2] R. Mostovoy, P.S. Borisenko, A.B. Levina "Mobile Phone Security: Side Channel Point of View", *18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT)*, 2016, pp. 567-569.

[3] E. Stamatatos. "A survey of modern authorship attribution methods", *Journal of the American Society for Information Science and Technology*, vol. 60(3), 2009, pp. 538–556.

[4] E. Stamatatos, W. Daelemans, B. Verhoeven, P. Juola, A. L´opez-L´opez, M. Potthast, B. Stein, "Overview of the author identification task at PAN 2015". *In CLEF 2015 - Conference and Labs of the Evaluation forum*, 2015.

[5] F. Iqbal, H. Binsalleeh, B.C.M. Fung, M. Debbabi, "A unified data mining solution for authorship analysis in anonymous textual communications", *Information Sciences*, vol. 231, 2013, pp. 98-112.

[6] L. van der Knaap, F.A. Grootjen, "Author identification in chatlogs using formal concept analysis", *19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC2007)*, 2007, pp. 181-188.

[7] M. Eder, M. Kestemont, J. Rybicki, "Stylometry with R: a suite of tools", *Digital Humanities 2013: Conference Abstracts*, 2013, pp. 487-489.

[8] M. Corney, A. Anderson, G. Mohay, O. de Vel, "Identifying the authors of suspect email", 2001. Web: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.

[9] A. Abbasi, H. Chen, "Applying Authorship Analysis to Extremist-Group Web Forum Messages", *IEEE Intelligent Systems*, 2005, vol. 20, no.5, pp. 67-75.

[10] 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.

[11] 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.

[12] I. Frommholz, H.M. al-Khateeb, M. Potthast, Z. Ghasem, M. Shukla, E. Short, "On Textual Analysis and Machine Learning for Cyberstalking Detection", *Datenbank-Spektrum*, vol. 16, No. 2, 2016. pp. 127-135.

[13] Rosenblum N., Zhu X., Miller B.P. "Who Wrote This Code? Identifying the Authors of Program Binaries", *Computer Security – ESORICS 2011: Lecture Notes in Computer Science*, vol. 6879, 2011, pp. 172-189.

[14] Juola P. "An overview of the traditional authorship attribution subtask", *In CLEF 2012 Notebooks*, Web: http://ims-sites.dei.unipd.it/documents/71612/155385/CLEF2012wn-PAN-Juola2012.pdf.

[15] M. Bhargava, P. Mehndiratta, K. Asawa, "Stylometric analysis for authorship attribution on Twitter", *International Conference on Big Data Analytics*, 2013, pp. 37-47.

[16] J. Albadarneh, B. Talafha, M. Al-Ayyoub, B. Zaqaibeh, M. Al-Smadi, Y. Jararweh, E. Benkhelifa, "Using big data analytics for authorship authentication of arabic tweets". *IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, 2015.

[17] A. Anderson, M. Corney, O. de Vel, and G. Mohay, "Multi-topic email authorship attribution forensics", *ACM Conference on Computer Security*, 2001.

[18] S. Afroz, A. Caliskan-Islam, A. Stolerman, R. Greenstadt, D. McCoy, "Doppelganger finder: Taking stylometry to the underground", *2014 IEEE Symposium on Security and Privacy (SP),* 2014, pp. 212-226.

[19] M. Koppel, J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution", *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, vol. 69, pp. 72–80, 2003.

[20] M. Koppel, Y. Winter, "Determining if two documents are written by the same author", *Journal of the Association for Information Science and Technology*, vol. 65(1), 2014, pp. 178–187.

[21] 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.

[22] T. Qian, B. Liu, L. Chen, Z. Peng, "Tri-training for authorship attribution with limited training data", *ACM*, 2014, pp. 345–351.

[23] C. Sanderson, S. Guenter, "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation", *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 482–491.

[24] U. Sapkota, S. Bethard, M. Montes-y G´omez, T. Solorio, "Not all character n-grams are created equal: A study in authorship attribution", *HLT-NAACL*, 2015, pp. 93–102.

[25] 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.

[26] A. Abbasi, H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace", *ACM Transactions on Information Systems*, vol. 26(2):7, 2008, pp. 7.

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

[28] E. Stamatatos, N. Fakotakis, G. Kokkinakis "Computer-based authorship attribution without lexical measures", *Computers and the Humanities*, vol. 35(2), 2001, pp. 193–214.

[29] Luyckx K., Daelemans W. "Personae, a corpus for author and personality prediction from text", 2008, Web: http://www.academia.edu/2813253/Personae_a_corpus_for_author_and_personality_prediction_from_text.

[30] Yang M., Chow K.P., "Authorship attribution for forensic investigation with thousands of authors", *The 29th IFIP TC 11 International Information Security and Privacy Conference (SEC 2014)*, 2014, v. 428, p. 339-350.

[31] Afroz S. *Deception in Authorship Attribution,* PhD thesis, Drexel University, December 2013.

[32] Stamatatos, E., "Text Sampling and Re-sampling for Imbalanced Author Identification Cases", *Proc. of the 17th European Conference on Artificial Intelligence (ECAI'06)*, 2006.

[33] Stamatatos, E., "Author Identification Using Imbalanced and Limited Training Texts", *Proc. of the 4th International Workshop on Text-based Information Retrieval*, 2007.

[34] 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.

[35] R. S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, B. Maia, "'Twazn me!!! Automatic authorship analysis of microblogging messages", *International Conference on Application of Natural Language to Information Systems*, 2011, pp. 161–168.

[36] Mikros, G. K., Perifanos, K. "Authorship attribution in Greek tweets using authors multilevel n-gram profiles", *AAAI Spring Symposium: Analyzing Microtext*, 2013.

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

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

[39] Sukhoparov M.E., Lebedev I.S. "Methodologies of Internet portals users' short messages texts authorship identification based on the methods of mathematical linguistics", *8th IEEE International Conference on Application of Information and Communication Technologies*, 2014, pp. 1-6.

[40] 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.

[41] M. Koppel, J. Schler, E. Bonchek-Dokow, "Measuring differentiability: Unmasking pseudonymous authors", *Journal of Machine Learning Research*, 2007, vol. 8, pp. 1261–1276.

[42] 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

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

[44] Vorobeva A.A. "Dynamic feature selection for web user identification on linguistic and stylistic features of online texts", *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 1, pp. 117–128.

[45] M. Robnik-Sikonja, "Improving random forests", *ECML*, vol. 3201, 2004, pp. 359-370.

[46] Xu B. et al. "Hybrid weighted random forests for classifying very high-dimensional data", *International Journal of Data Warehousing and Mining*, vol. 8, 2012, pp. 44-63.

[47] Lustgarten J. L. et al. "Improving classification performance with discretization on biomedical datasets", *AMIA*, 2008.

[48] Breiman L., "Random Forests", *Machine Learning*, 2001, vol. 45, pp 5-32.

[49] Usama M. Fayyad, Keki B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning", *Proceedings of the International Joint Conference on Uncertainty in AI*, 1993, pp. 1022-1027.

[50] 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.