Examining the Performance of Classification Algorithms for Imbalanced Data Sets in Web Author Identification

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

Abstract—**Individuals, criminals or even terrorist organizations can use web-communication for criminal purposes; to avoid the prosecution they try to hide their identity. To increase level of safety in Web we have to improve the author (or web-user) identification and authentication procedures. In field of web author identification the situation of imbalanced data sets appears rather frequent, when number of one author's texts significantly exceeds the number of other's. This is common situation for the modern web: social networks, blogs, emails etc. Author identification task is some sort of classification task. To develop methods, technics and tools for web author identification we have to examine the performance of classification algorithms for imbalanced data sets. In this work several modern classification algorithms were tested on data sets with various levels of class imbalance and different number of available web-post. The best accuracy in all experiments was achieved with Random Forest algorithm.**

I. Introduction

In modern Web millions of new electronic texts appears every day. Web-users identification and authentication procedures are not enough effective; users could produce numerous of identities, present a fake identity or even hide it [1]. Author identification can be used in various tasks in field of intelligence, computer forensics and in cyber (or on-line) security to determine the identity of the author based on analyses of his texts [2], [3], [4].

Author identification can be seen as a multi-class text classification task [4]. So, there are two main questions to be solved: What are the most suitable features and what classification algorithm to use? In this work is examined the performance of several classification algorithms and identification accuracy on imbalanced data sets.

In field of web author identification the situation of imbalanced data sets appears rather frequent. Almost each time we want to apply some machine learning technique to solve "real-world" author identification problem.

A. Previous researches

Mainly all author identification methods can be split into two groups: profile-based and instance-based [5]. In profile-based approaches author is represented as one text, obtained by concatenating all his texts in one.

In instance-based approaches author is represented as collection of his texts, and each text is training instance.

The profile-based approaches are useful when only few texts (even only one) are available and if texts are extremely short. In instance-based approaches it is easier to combine different features types and it performs better when there is wide range of candidate authors [6].

In most previous works are used balanced data sets, but this differs from reality. Commonly in experiments is used training set, that contains large amount of texts per one author, and often texts are significantly longer then real web-posts.

There were only few works on author attribution on imbalanced data [7], [8], [9].

In [7], [8] various sampling and re-sampling methods were studied to handle the problem of class imbalance in both profile-based and instance-based author identification approaches. It was proposed to produce many short text samples for the minority classes and less but longer text samples for the majority classes. This method could be applied to improve the accuracy of identification.

The key idea of [8] is to represent each text in three views, and then perform Tri-Training. Experiment results show that the accuracy of proposed approach outperforms accuracy achieved in [7], [8] and other baselines.

In [9] was proposed approach called Document Author Representation, it builds document vectors in a space of authors, calculating the relationship between textual features and authors. The best accuracy was achieved with 1-Nearest Neighbors algorithm, with using the 2500 most frequent character 3-grams.

B. Classification task in author identification

Task of author identification can be formulated in word of classical classification task. Given a set of texts $T = \{t_1, ..., t_m\}$ and set of authors $U = \{u_1, ..., u_k\}$, where m - number of texts and k – is number of authors. In this work is used instance-based approach, the author u_i can be presented as subset $T_j \in T$.

There is some subset of text $T' \in T$ of known author – the samples data $\{(t_1,u_1),\ldots, (t_m,u_k)\}$. We have to find effective algorithm or classifier a: $t_j \rightarrow U$, that calculates the probability of authorship for each author to be an author of text t_j and output probabilities sorted in descending order Pr (u_i author t_j).

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

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

In this work is used instance-based approach. We split the samples data for two subsets: T_{tr} and T_{test}. At first, we train the classifier on subset T_{tr}, and then test it on T_{test} to validate the prediction power of model. After the classifier is trained and tested, we have validated model for author identification, and it can be used to identify author of text t_j. This is classical approach of supervised learning.

C. *The problem of imbalanced data sets in web author identification*

The problem of imbalances data set appears when number of texts of one author significantly exceeds the number of another author's texts. This is extremely common situation for the modern web: social networks, blogs, emails etc. This can be illustrated with an example for two authors.

If there is a data set consisting of texts T of two authors $U = \{u_1, u_2\}$. u_1 is the regular author, and u_2 is some author-criminal.

$$u_1 = T'_1 = \{ t_1, \ldots, t_n \}, \text{ where } n = 100.$$

$$u_2 = T'_2 = \{ t_1, \ldots, t_m \}, \text{ where } m = 10.$$

Most of texts of author u_2 will be classified as texts of author u_1. Suppose that some machine learning algorithm builds two possibly models:

1) Model 1 classified 7 out of 10 texts of author-criminal as texts of regular author and 12 out of 100 texts of regular author as texts of author-criminal.
2) Model 2 classified 2 out of 10 texts of author-criminal as texts of author regular author and 97 out of 100 texts of regular author as texts of author author-criminal.

To determine the performance of classifier can be used approach based on the number of mistakes. Then Model 1 is much better as it mistakes only on 19 texts, instead of 99 incorrectly identified authors by Model 2. If we want to improve identification performance, e.g. minimize the mistakes rate, we have to choose Model 1.

However, our main goal is to identify texts written by author-criminal, and in this case, we have to choose Model 2, because it incorrectly identified author only of two criminal texts.

If machine-learning algorithm uses approach, described earlier, to determine the performance of built model it will choose Model 1. This leads to a situation, when high amount of criminal texts could appear on some web site, although we could block the criminal-author using Model 2.

Different classification algorithms use various approaches to solve the problem of imbalanced data sets classification accuracy. Therefore, selection of classification algorithm with the best performance in imbalance data sets in web author identification need experimental improvement.

II. PERFORMANCE OF CLASSIFICATION ALGORITHMS ON IMBALANCED DATA SETS IN WEB AUTHOR IDENTIFICATION

D. *Feature set*

In this work is used two types of features: qualitative and quantitative. Traditional approach of using only stylometric or writing-style features is extended with web-post metadata features and with "emphasis" features.

All of them existing on three levels of web-posts: lexical (Fl), syntactic-structural (Fs) and metadata (Fm).

Full feature set is $F=(Fl+Fs+Fm)$ and contains 490 different features.

1) *Lexical group* includes frequencies of different groups of symbols (e.g. characters, digits, punctuations), functional words and specific expressions, frequencies of certain language constructs, frequencies of abbreviations and acronyms, words in foreign languages, links, images, frequencies of words and sentences with different length, and some other.
2) *Syntactic-structural group* includes frequencies of different punctuation symbols, text emphasis (bold, italic), and the logical structure of the text (blocks, paragraphs).
3) *Meta-features group* contains features withsome additional information about web-posts, not directly obtained from the texts. Here are the time and day of the week when author posted his text. User activities are tracked and stored on web site, and usually posts has information about the publication time.

E. *Classification algorithms*

Therefore, we have to find some classification algorithm that:

1) is able to handle high-dimensional data;
2) has a high accuracy on imbalanced data sets;
3) is able to handle cases with low amount of training examples;
4) is suitable for practical use in real author identification task: do not demand high processing power.

In most previous works, described in this paper earlier, is used several classification algorithms. To examine performance for imbalanced data sets were chosen algorithms that satisfy above-mentioned requirements, and showed rather good accuracy in experimental results reported in previous researches. These algorithms are:

1) Support vector machine (SVM)[10, 11];

2) Multilayer perceptron (MLP);
3) Decision trees (DR);
4) Random forest (RF) [12, 13, 14];
5) Logistic regression (LR);
6) Naïve Bayes (NB).

1) Support vector machine is favored for its ability to handle high-dimensional data. In addition, SVMs shows good performance for text classification tasks (along with LR). It uses polynominal kernel that allow to build non-linear classification models in high-dimensional and implicit feature space. Instead of linear classifiers SVM uses a flexible representation of the class boundaries. However, training SVM is one of its main weakness. To fix this problem in this work was used sequential minimum optimization[15].

2) Multilayer perceptron or Neural networks has remarkable ability to find hidden relations by itself in complicated, unstructured and noisy data, that are too complex to be noticed by other machine learning technics. Therefore, MLP operation can be unpredictable; the training examples must be carefully selected otherwise the computing time increases significantly or the MLP might be functioning incorrectly. In addition, MLP demand high processing power.

3) Decision trees DT have some advantages for author identification task. They are tolerance to noisy data, can handle both continuous and discrete data. Also useful is their ability to select features with the most discriminatory power.

4) Random forest as SVM also has great ability to handle high-dimensional data, has good accuracy in texts classification tasks. As RF is just an ensamble of decision trees it has all their advantages.

5) Logistic regression

Logistic regression is one of the best discriminative probabilistic classification models used in a wide number of tasks. It is robust to noise and can be used as a noise tolerant classification method. The LR output can be interpreted as a probability; this is important, as we had mentioned above. However, the three main drawback of LR is its training time, problems with processing of categorical features and its nature – it is a linear classifier, it assumes is that there is one linear decision boundary.

6) Naïve Bayes, which is a simple probabilistic classification algorithm that often performs well in many domains.

F. Corpus and Data sets

Full text corpus contains 23546 web-post and 1004 authors. It was formed by collecting posts in Russian language from blog-hosting www.livejournal.com. The corpus contains texts of different genres and topics. For experiments carried out in this work was randomly selected 165 authors and 3853 their texts.

Texts have variable length; most of them are from 142 to 699 characters length. Distribution of texts length is shown on Fig. 1.

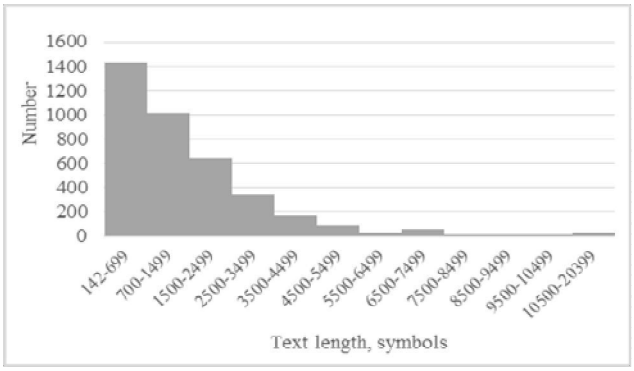

Fig. 1. Texts length

From this corpus were produced eight versions of data sets, varying levels of class imbalance and number of training texts per author. Therefore, were formed two balanced and six imbalanced data sets, as it is shown in Table I.

Each data set contains 20 sets of candidate authors (*U*), including 10 authors and their texts. Therefore, total number of test data sets is 160.

Two types of data sets were formed: with low amount of training texts and with medium amount. The first one simulates real situation when we have only few texts of one author. The second one simulates more optimistic situation.

Data sets with low amount of texts:

1) Balanced (10 texts per author);
2) Low class imbalance (at least 8 training texts are available for all the authors and 10 is the maximum amount of training texts per author);
3) Medium class imbalance (at least 5 training texts are available for all the authors and 10 is the maximum amount of training texts per author);
4) High class imbalance (at least 2 training texts are available for all the authors and 10 is the maximum amount of training texts per author);

Data sets with medium amount of texts:

1) Balanced (25 texts per author);
2) Low class imbalance (at least 20 training texts are available for all the authors and 25 is the maximum amount of training texts per author);
3) Medium class imbalance (at least 10 training texts are available for all the authors and 20 is the maximum amount of training texts per author);
4) High class imbalance (at least 5 training texts are available for all the authors and 25 is the maximum amount of training texts per author).

TABLE I. DATA SETS WITH DIFFERENT LEVELS OF CLASS IMBALANCE

Level of class imbalance	Number of texts per author	
Low	min. 20 max. 25	min. 8 max. 10
Medium	min. 10 max. 20	min. 5 max. 10
High	min. 5 max. 25	min. 2 max. 10
Balanced	min. 24 max. 25	min. 10 max. 10

In imbalanced data sets number of texts per authors has normal distribution. The distribution of training texts per author, it is visualized in Fig. 2-4.

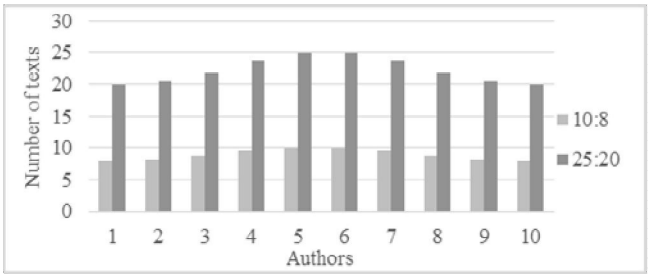

Fig. 2. Distribution of training texts per author in data sets with low class imbalance and various number of available texts per author

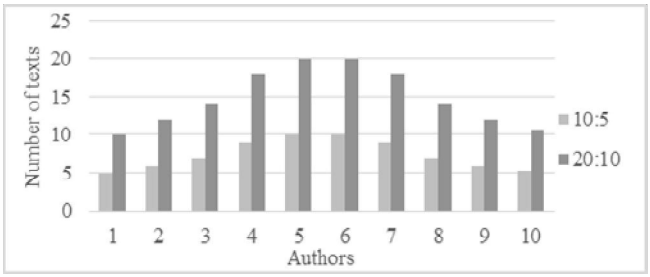

Fig. 3. Distribution of training texts per author in data sets with medium class imbalance and various number of available texts per author

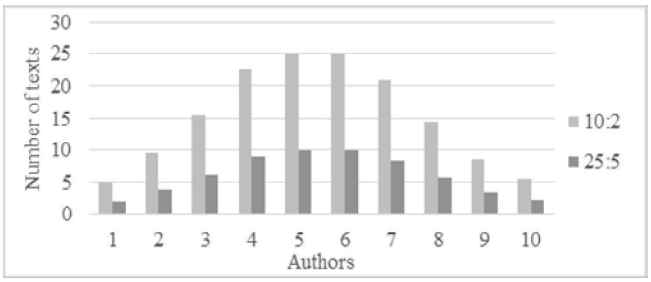

Fig. 4. Distribution of training texts per author in data sets with high class imbalance and various number of available texts per author

G. *Experiments and results*

Various experiments were held to compare author identification accuracy achieved with previously selected classification algorithms. All experiments were carried out on a 10-fold cross validation.

Accuracy (A) is ratio of correctly identified authors of texts T_c, to total number of test texts T_t (1).

$$A = T_c / T_t \quad (1)$$

1) Investigate classification algorithms training time to approve possibility of practical use in author identification tasks

Large number of features makes training time extremely long for some classification algorithms, this leads to impossibility to use them in practice.

To study the effect of the features number on the training time were made seven data sets with various features number (form 30 to 90 features) and one set with 490 features.

In Fig. 5 are shown the results of this short experiment.

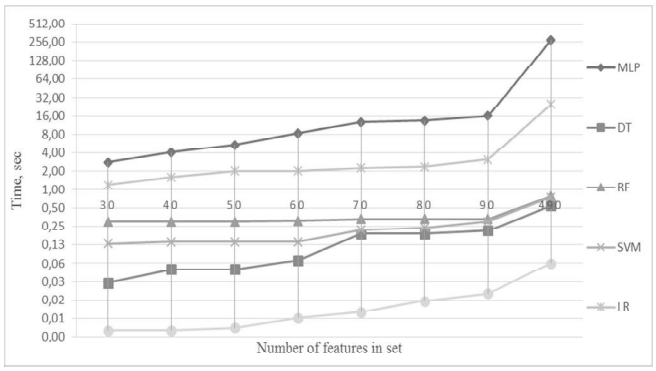

Fig. 5. Training time of classification algorithms on feature sets of various length

Experiment results and comparison of training time showed that MLP and LR has the highest training time; with increasing the number of features in set the training time grows significantly. The results coincides with the results of [16] obtained earlier, and also are confirmed by other researchers [17].

Accuracy of MLP and LR is comparable with the accuracy obtained by using other methods, that allow excluding the MLP and LR from further study.

2) Classification accuracy on imbalanced data with medium amount of training samples

First series of experiments simulates the situation of normal amount of available web-posts; maximum 25 web-posts for one author.

Table II indicated that RF has better performance than other studied algorithms.

TABLE II. IDENTIFICATION ACCURACY ON DATA SETS WITH VARIOUS LEVELS OF CLASS IMBALANCE AND MEDIUM AMOUNT OF TRAINING TEXTS

	Classification accuracy			
Texts per author ratio (min:max)	**SVM**	**DT**	**RF**	**NB**
Balanced data set				
24:25	64.65	59.24	**75.42**	54.82
Low level of class imbalance				
20:25	64.21	59.90	**75.33**	53.79
Medium level of class imbalance				
10:20	60.45	56.98	**71.50**	52.44
High level of class imbalance				
5:25	61.74	58.67	**70.21**	53.94

In all experiments, RF achieves the best accuracy. Maximum accuracy was on balanced data set - 75.42%. Accuracy decreased with increasing level of imbalance balance, difference is about 5%.

The effect of class imbalance levels on accuracy of SVM, DT, RF and NB is shown in Fig. 6.

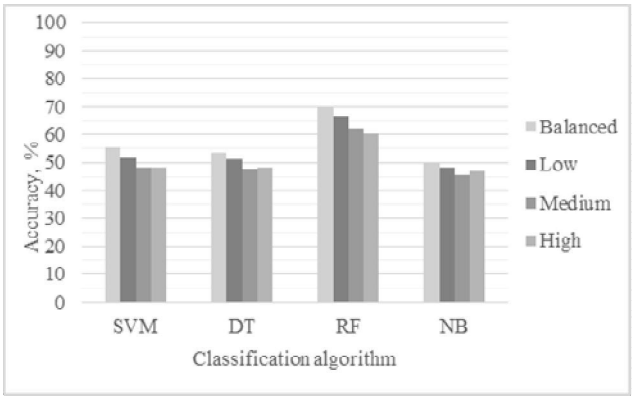

Fig. 6. Identification accuracy of selected algorithm on data sets with different levels of class imbalance and medium number of available training texts

On imbalanced data classification accuracy is little worse – 75.33%, 71.50%, 70.21% (low, medium and high level of imbalance). It is clearly that accuracy decreases with growth of imbalance level. The lowest accuracy is in high imbalanced data, but this was expected.

3) Classification accuracy on imbalanced data with low amount of training samples

Second series of experiments was conducted to estimate the accuracy of the selected algorithms, they were held to simulate the situation where only low amount of web-posts are available; maximum 10 web-posts for one author.

Experiments were carried on a 20 sets of authors $U_i = \{u_1, ..., U_{10}\}$, $1 \le i \le 20$.

The experiments results are summarized in Table III.

TABLE III. IDENTIFICATION ACCURACY ON DATA SETS WITH VARIOUS LEVELS OF CLASS IMBALANCE AND LOW AMOUNT OF TRAINING TEXTS

	Classification accuracy			
Texts per author ratio (min:max)	**SVM**	**DT**	**RF**	**NB**
Balanced data set				
10:10	55.21	53.55	**69.56**	50.09
Low level of class imbalance				
8:10	52.01	51.29	**66.72**	48.18
Medium level of class imbalance				
5:10	48.48	47.90	**62.21**	45.46
High level of class imbalance				
2:10	48.47	48.17	**60.52**	46.97

As expected, maximum accuracy was achieved on balanced data set. Accuracy decreases with increasing level of imbalance, difference is significant - about 9%.

Fig. 7 shows accuracy of author identification for data sets with different level of imbalance.

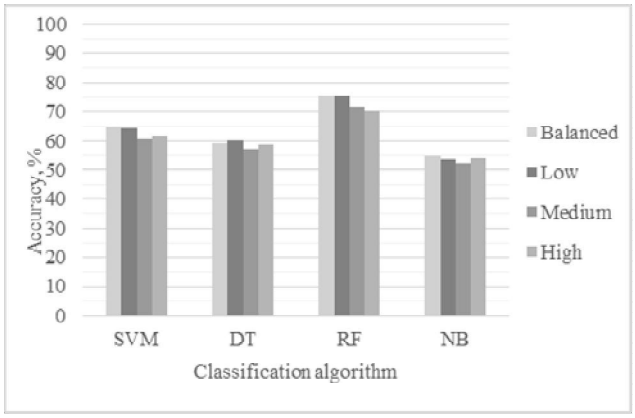

Fig. 7. Identification accuracy of selected algorithm on data sets with different levels of class imbalance and low number of available training texts

As in previous experiments, RF achieves the best accuracy on the balanced data – 69.56%. The accuracy on imbalanced data is much worse – 67.72%, 62.21%, 60.52% (low, medium and high level of imbalance).

Comparison of the all experiments results showed that the highest classification accuracy and the best overall accuracy for all data sets was achieved with RF algorithm, Fig.6 and Fig. 7 visualize it very clearly.

The highest accuracy on balanced data sets with low amount of available web-posts is 69.56%. Experiments with normal amount of training samples showed 75.42% of accuracy. Accuracy is little worse on data sets with low level of imbalance, 66.72% and 75.33% respectively.

However, it is very important that in all experiments Random Forest algorithm outperform others. Therefore, for "real-world" tasks, even if we have only few available texts and huge difference in their number for authors, the RF would be the best choice.

VII. CONCLUSION

Several classification algorithms were selected in theoretical study as the most suitable to solve the problem mentioned above.

To evaluate the performance of these algorithms on data with various class imbalance three series of experiments were carried out. In experiments were used data sets, containing from 2 to 25 web-posts of 10 candidate authors. Length of the most texts varied from 142 to 10500 (99,4%) symbols.

The experiments on various data sets, showed that Random Forest algorithm has better performance than other algorithms that were studied.

It was observed that RF performs better than SVM, DT and NB both on data with low and medium amount of training texts, and on data with all levels of class imbalance.

In future work will be studied effect of texts length on identification accuracy, minimum amount of training texts per one author and some technics to improve classification accuracy on imbalanced data.

REFERENCES

[1] A.V. Gvozdev, I.S. Lebedev, "Model analiza informacionnih vozdeistviy v otkritih computernih sistemah", *Sborik dokladov VII mezhdunarodnoy konferencii Sovremennie problemi prikladnoy informatiki*, 2011, pp. 45-47.

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

[3] Stamatatos E., "A Survey of Modern Authorship Attribution Methods", *Journal of the American Society for Information Science and Technology*, 2009, vol. 60, no.3, pp. 538-556

[4] 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.

[5] Luyckx K. "Scalability Issues in Authorship Attribution", *Ph.D. Thesis*, University of Antwerp, 2010.

[6] 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

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

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

[9] Tieyun Qiana, Bing Liub, Li Chena, Zhiyong Penga, Ming Zhonga, Guoliang Hea, Xuhui Lia, Gang Xuc, "Tri-Training for authorship attribution with limited training data: a comprehensive study", *Neurocomputing*, 2016, vol. 171, pp. 798–806.

[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] Adrián Pastor López-Monroy, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, Jesús Ariel Carrasco-Ochoa, José Fco. Martínez-Trinidad, "A new document author representation for authorship attribution", *Pattern Recognition*, vol. 7329 of the series Lecture Notes in Computer Science pp 283-292.

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

[13] Promita Maitra, Souvick Ghosh, Dipankar Das., "Authorship Verification – An Approach based on Random Forest", *Notebook for PAN at CLEF 2015*, 2015.

[14] María Leonor Pacheco, Kelwin Fernandes, Aldo Porco, "Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification", *Notebook for PAN at CLEF 2015*, 2015.

[15] John C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Technical Report, *Microsoft Research*, 1998.

[16] Alice A. Vorobyeva, Aleksey V. Gvozdev. "Identificaciya anonimnih polzovateley Internet-portalov na osnovanii tehnicheskih i lingvisticheskih harakteristik polzovatelya", *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2014, no. 1, pp. 139-144.

[17] Romanov A.S. "Metodika identifikacii avtora teksta na osnove apparata opornih vektorov", *Dokladi TUSURa*, 2009, vol. 1, no.19, pp. 36-42.