Proceedings of the Federated Conference on Computer Science and Information Systems pp. 347–351
DOI: 10.15439/2017F110
ISSN 2300-5963 ACSIS, Vol. 11

A Comparison of Authorship Attribution Approaches Applied on the Lithuanian Language

Jurgita Kapočiūtė-Dzikienė
Faculty of Informatics,
Vytautas Magnus University,
Vileikos str. 8, LT-44404 Kaunas, Lithuania
Email: jurgita.kapociute-dzikiene@vdu.lt

Algimantas Venčkauskas, Robertas Damaševičius
Faculty of Informatics,
Kaunas University of Technology,
Studentų str. 50, LT-51368 Kaunas, Lithuania
Email: {algimantas.venckauskas, robertas.damasevicius}@ktu.lt

***Abstract*—This paper reports comparative authorship attribution results obtained on the Internet comments of the morphologically complex Lithuanian language. We have explored the impact of machine learning and similarity-based approaches on the different author set sizes (containing 10, 100, and 1,000 candidate authors), feature types (lexical, morphological, and character), and feature selection techniques (feature ranking, random selection). The authorship attribution task was complicated due to the used Lithuanian language characteristics, non-normative texts, an extreme shortness of these texts, and a large number of candidate authors. The best results were achieved with the machine learning approaches. On the larger author sets the entire feature set composed of word-level character tetra-grams demonstrated the best performance.**

I. Introduction

MOST comments or forum posts on the Internet are written anonymously. Due to anonymity people can freely share their thoughts, but cannot feel protected from the negative behavior or cybercrimes. Protective mechanisms (monitoring IP addresses or requesting to register and submit personal data) are not always reliable enough: perpetrators change their IP addresses, use different pseudonyms, or route WebPages through proxy servers. However, even in such complicated situations, the identity can still be disclosed from the existing "stilometric fingerprint" unique to each individual [1].

Apart from the handwriting analysis [2], the textual authorship analysis covers very different applications: author profiling, authorship verification, plagiarism detection, etc. However, in this research we are focusing on the Authorship Attribution (AA) problem which has to detect who of the candidate authors is a real author of some anonymous text document. AA is one of the earliest problems in Computational Linguistic: the oldest attempts were restricted to attributing of the disputed long and homogeneous literary texts to one of few known authors. In the recent decades AA drifted towards practical applications: it copes with the huge number of candidate authors, extremely short texts, limited training data and for all these reasons AA is often called "needle-in-a-haystack" problem [3].

In this research we are solving the AA task using a corpus composed of the Lithuanian Internet comments. Although the corpus does not contain texts produced by convicted cyber criminals, it can perfectly serve for various experiments aimed at detecting authors' style characteristics. The aim of the paper is to determine the best approaches (in terms of the attribution paradigm, the feature type, and the feature selection technique) for the different author set sizes, containing 10, 100, or 1,000 candidate authors. The problem is complicated due to several reasons: 1) very short texts, covering a wide range of topics; 2) the morphologically and vocabulary rich Lithuanian language; 3) non-normative texts; 4) there are no recommendations what could work the best for our solving task. Consequently this research aims at finding the best solutions for the Lithuanian language. Moreover, we anticipate that these solutions could be also useful for the other Baltic or Slavic languages, sharing similar characteristics.

II. Related Work

The statistical methods used to tackle AA tasks can be grouped into two main paradigms: machine learning and similarity-based[1]. The comprehensive review of these methods and various feature types is presented in [4].

The majority of AA research works are carried out on a small number of candidate authors (up to few dozens) and even findings obtained from comparative experiments are very controversial due to the different experimental conditions (languages, datasets, author sets, etc.). Whereas, the comparative experiments tackling "needle-in-the-haystack" problems often claim the superiority of similarity-based approaches (e.g., [5], [6]). However, such experiments are rather rare: i.e., most often methods are chosen and applied without any considerations. Further we will focus on the influential research works dealing with at least one thousand candidate authors.

The experiments described in [7] are carried out on the Twitter corpus: the introduced "flexible patterns" (taking into account the surrounding information around function words) significantly outperform other feature types based solely on word or character n-grams with SVM. The work in [8] is addressing the open-class issue and deals with the blog dataset of 10,000 authors. It tests a combined similarity-based and machine learning technique on 3 text representation types: tf-idf on content words, tf-idf on various stylistic features, and

[1]Despite by the nature similarity-based approaches are the part of machine learning, they are distinguished and discussed separately in many AA works.

idf on content words. The similarity-based part of this hybrid approach ranks authors according to the cosine values and afterwards the top-rank pair (composed of the anonymous text and the most likely author) is tested on the meta-learning SVM classifier. The high precision in [9], [10], [11] is achieved using the cosine similarity-based technique aggregating several attribution decisions, taken on the different randomly selected subsets of character tetra-grams. These researchers, experimenting with 10,000 blog writers, are also addressing the open-class issue. The research in [12] solves the AA task on the Japanese microblogs of 10,000 authors with the cosine similarity-based approach and character-level n-grams (with n equal to 1, 2, and 3). Adopted three new techniques –in particular, the combined selection for the training dataset, the biased weighting scheme for n-grams, and the part-of-speech tag combined n-grams– assure both the relatively high precision and the short execution time. Another task of 19,000 blog writers is successfully tackled with the Latent Dirichlet Allocation (LDA) technique by measuring the distances between the LDA-based representations (as mixtures of topics) in the anonymous text and in training text samples. The authors of this research [13] claim that offered similarity-based technique applied on the author profiles with enough training data even yields state-of-the-art performance. The authors in [6] are dealing with 100,000 blog writers. They explored 3 different classifiers (SVMs, Naïve Bayes, and Regularized Least Squares Classification) and, in addition, estimated the confidence of their outputs – in particular, measured the difference between the best two matching classes, ran several classifiers, and presented the final AA decision only if they agreed.

Unfortunately the surveyed research works offer no research-based recommendations for the morphologically rich, highly inflective, derivationally complex non-normative Lithuanian language. Despite for the Lithuanian language there are done: 1) lots of descriptive research works (e.g., [14], [15]); 2) some experiments with machine learning (carried out on parliamentary transcripts or forum posts of only 100 candidate authors) [16] or similarity-based approaches (using very limited training data) [17]; these findings do not guarantee the best results for our solving AA task. Our aim is at performing the comparative investigation and at finding the best method, feature type, and feature selection technique for our AA task (with 10, 100, and 1,000 candidate authors) on the corpus of the Lithuanian Internet comments.

III. The Corpus

The created corpus[2] is composed of the Lithuanian Internet comments.

The texts of authors were selected in the way not to get the topic-per-author distribution. Some author was included into the corpus only if all his/her comments were written under the same unique pseudonym and IP address (both considered as a single unit), but not included if 1) his/her pseudonym was used under several IP addresses; 2) more pseudonyms were used under the same IP address. The aim was to reduce the risk of disputed authorship and to get as clean corpus as possible. Although some exceptions (when the same author is writing under several separate IP addresses using different pseudonyms) may still occur, we anticipate they are rare enough to have the significant impact.

During pre-processing all recognized non-Lithuanian characters and reply messages were filtered out, meta information about the author and his/her posts was also eliminated, comments shorter than 30 symbols were excluded.

The most important characteristics about the composed corpus, depending on the different author set sizes (experimentally investigated in this paper) are given in Table I. The authors with the largest number of texts were selected to form the author sets of 10 and 100 candidate authors. The average texts/per author distribution is $\sim$155, but the corpus is unbalanced: i.e., text samples per author varies from only 39 to 2,837. 13 authors have more than 1,000 texts, 575 authors have less than 100, and only 12 authors have the least number of texts. The random ($\sum_j P^2(c_j)$) and majority ($\max(P(c_j))$) baselines (where $P(c_j)$ is the probability of some author c_j obtained by dividing a number of texts written by particular c_j from all number of texts in the corpus) must be exceeded that the AA method could be considered appropriate.

There is no consensus about the minimal text length appropriate for the AA tasks: some researchers claim 2,500 words is optimal [18], others achieve reasonable results with $\sim$60 [19]. In our task we have to deal with extremely short texts where an average length ranges from $\sim$20 to $\sim$26 tokens. Besides we are dealing with the sparse non-normative texts full of out-of-vocabulary words, abbreviations, missing diacritics (where Lithuanian letters having the diacritic marks are replaced with the corresponding Latin equivalents), diminutives, etc.

TABLE I
Characteristics of the Lithuanian Internet comment corpus.

Number of authors	**10**	**100**	**1,000**
Number of texts	14,443	63,131	155,078
Number of tokens (letters & digits)	289,462	1,511,823	4,068,231
Average text length (in tokens)	20.042	23.947	26.233
Classification accuracy baselines			
Random baseline	0.001	0.002	0.003
Majority baseline	0.018	0.018	0.018

IV. Classification Approaches

In this research we have explored the following approaches:

- *Support Vector Machine* (SVM) (introduced in [20]), which efficiently handles the high dimensional feature spaces, the sparseness of the feature vectors, and does not perform an aggressive feature selection. In our experiments we selected Sequential Minimal Optimization (SMO) algorithm with the polynomial kernel implementation in WEKA, version 3.8 [21] and all remaining parameters were set to their default values.

[2]The corpus can be downloaded from http://dangus.vdu.lt/~jkd/wp-content/uploads/2015/04/INT_KOMENTARAI_INDV2.7z.

- *Naïve Bayes Multinomial* (NBM) (introduced in [22]) which is often selected due to simplicity, low data storage resources, the fast processing, robustness to cope with the large number of features having equal significance. We used implementation in WEKA with the default parameter values.
- *Similarity-based approach* (SB) with cosine measure [23]. In this paper we explore a simple similarity-based approach with the top N ranked features (SB-TopN) and the approach based on the randomized feature sets (introduced in [9]) (SB-RFS). The SB-RFS technique is adjusted to cope with very concise texts; performs especially well on a small number of features, because the final attribution decision incorporates the generalized results of several decisions obtained during a few iterations. In our experiments we used SB-TopN and SB-RFS implementations presented in [17].

V. Feature Extraction

In our research we have investigated the impact of the most popular and the most accurate feature types (for the statistics see Table II):

- *lex* – a bag-of-words. In our corpus we do not have topic-per-author distribution, therefore this feature type can be used without any risk to get topic classification instead of AA.
- *lem* – a morphological feature type based on the word lemmas. This type is usually recommended for the highly inflective languages. The texts were lemmatized using "Lemuoklis" [24].
- *chr4* – a character feature type based on the word-level tetra-grams. This type was superior to the other types in the topic classification task for the Lithuanian language [25].

Lemmas and character features decrease the sparseness of the feature vectors (see Table II): the lower the sparseness is, the more robust classifier is created. The sparseness can also be reduced with the selection of the most relevant features, therefore in our experiments we investigated the following feature selection techniques:

- *Whole set of features* – i.e., we used the entire set of all *N* available features (presented in Table II). This technique was tested with SVM, NBM, and SB-TopN methods.
- *Feature ranking and selection of top N.* All features were ranked according to their chi-square values and afterwards the top *N* were chosen to form the new set. In our experiments we have investigated $N = 30,000$, because this value was proved to be minimum but optimal in the similar AA experiments [17]. We have explored this technique with SVM, NBM, and SB-TopN.
- *Random selection of features with a fixed size N.* The *N* features (with $N = 30,000$) were randomly selected from the whole feature set. The random selection was done in $K = 20$ iterations with SB-RFS method. The final attribution decision was based on the majority vote of attribution decisions obtained in all K iterations.

TABLE II
Feature types in the corpus of the Lithuanian Internet comments.

Feature type	Number of features		
	10 authors	100 authors	1,000 authors
lex	56,064	172,257	315,590
lem	39,498	109,935	201,469
chr4	40,855	78,773	119,008

VI. Experimental Set-up and Results

The experiments were carried out on the stratified corpus (described in Section III). Instances were selected for the training (of 80%) and testing (of 20%) sets. The same training/testing sets were used in all our experiments exploring different methods (see Section IV), feature types, and feature selection techniques (see Section V).

The experiments were evaluated using *accuracy* and *f-score* (averaged over different classes) performance measures. We also performed the McNemar test (with the significance level of 95%) to check if the differences between observed results are statistically significant.

The results obtained on the datasets with 10, 100, 1000 candidate authors are presented in Fig. 1, Fig. 2, Fig. 3, respectively. *Accuracy* and *f-score* values are presented in white and gray columns, respectively. The first two columns of the *accuracy* and the *f-score* present the results achieved on the entire feature set, the second two – on 30,000 features (with SB-RFS all results are obtained with 30,000 features). The dashed line indicates the higher one of the random and majority baselines.

VII. Discussion

Not all results are reasonable: i.e., the *accuracy* on the token lemmas (*lem*) with SB-RFS is below the majority baseline (equal to 0.018).

The similarity-based approaches are outperformed with machine learning in all our datasets of 10, 100, 1,000 candidate authors. The differences between *lex* + SB-RFS and 30,000 *lem* + NBM on 10 candidate authors and the differences between entire *chr4* + SB-TopN and 30,000 *lem* + NBM on 100 candidate authors are statistically significant with the probability density function $p \ll 0.05$. Whereas, the difference between entire *chr4* + SB-TopN and 30,000 *lem* + NBM on 1,000 candidate authors is not statistically significant with marginal $p = 0.05$. Since the NBM method and the similarity-based approaches maintain similar performance levels on the largest dataset, SVM is obviously superior to any similarity-based technique in any dataset ($p \ll 0.05$). However, similarity-based approaches can still be suitable with the larger author sets. The superiority of SVM compared to the similarity-based methods with the increase of the candidate authors is declining and probably some breaking point can be reached. These investigations are already in our future plans.

If SVM is obviously superior to NBM, the similarity-based approaches produce very controversial results: *lex* + SB-RFS, entire *lex* + SB-TopN, entire *chr4* + SB-TopN are the best

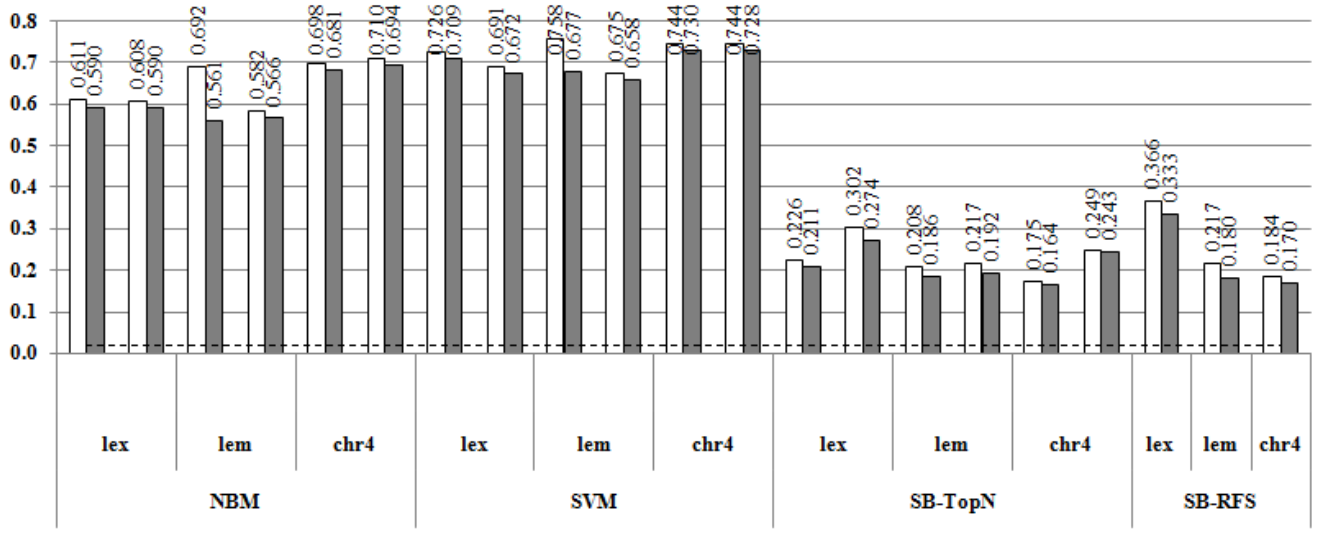

Fig. 1. The influence of the selected approach on the results with 10 candidate authors.

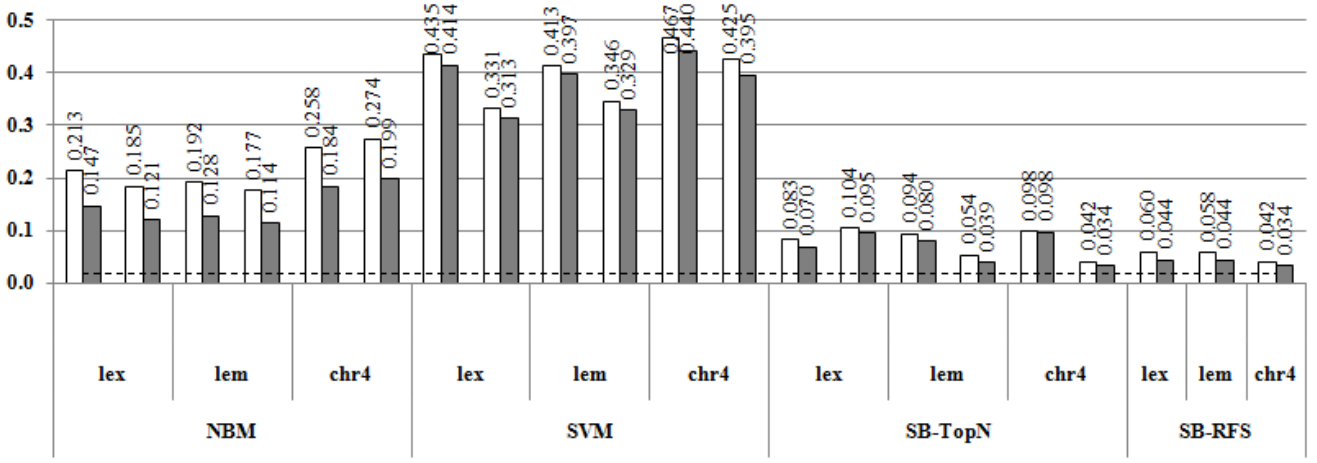

Fig. 2. The influence of the selected approach on the results with 100 candidate authors.

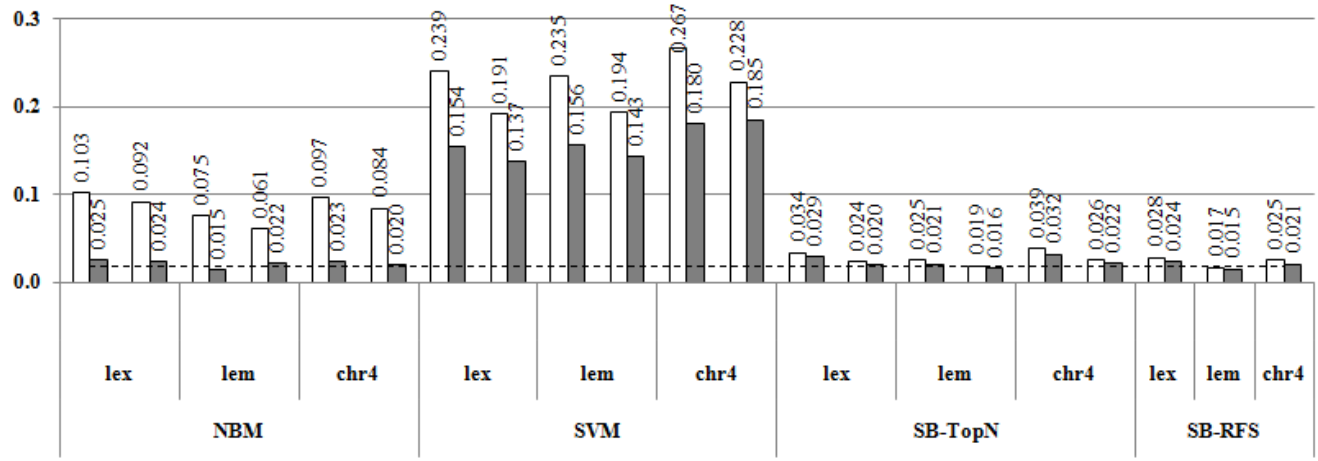

Fig. 3. The influence of the selected approach on the results with 1,000 candidate authors.

approaches on the datasets containing 10, 100, and 1,000 candidate authors, respectively. Due to these findings is hard to say which similarity-based method is actually the best.

The important findings lie in the analysis of the various feature representation types. *lem* and *chr4* types give the best results with NBM; with SVM is difficult to determine the best type (because differences in the accuracies between different feature types are not statistically significant); *lex* is the best type with the similarity-based methods on the dataset of 10 candidate authors. *chr4* type gives the best results with NBM and SVM; with the similarity-based methods is difficult to determine the best type on the dataset of 100 candidate authors. *lex* and *chr4* types are the best with NBM; *chr4* type is the best with SVM; and marginally the best with the similarity-based approaches on the dataset with 1,000 candidate authors. Thus, summarizing all these findings it can be concluded that the best feature type (especially on the larger author sets) is character tetra-grams (*chr4*). Morphological tools are helpless on the non-normative texts, but character features are robust to deal with the morphologically complex languages by capturing the patterns of complex inflection morphology intrinsically.

The restriction of the feature set size to 30,000 features speeds up the calculation time, but, statistically significant degrades the accuracy, except with the SB-RFS method.

VIII. Conclusions and Future Work

The main contribution of this research is a comparative study of AA approaches (machine learning, similarity-based), feature types (lexical, morphological, character), feature selection techniques (whole set, feature ranking, random selection) and the author set sizes (of 10, 100, and 1,000 candidate authors) on non-normative Internet comments using the morphologically complex Lithuanian language.

The best results were achieved with the machine learning approaches; on the larger author sets the word-level character tetra-grams with the whole set of features demonstrated the best performance.

The obtained authorship attribution results are low enough to encourage us to continue seeking for the better solutions. In the future research we also plan to experiment with the larger authors sets and with the other types of non-normative texts.

Acknowledgment

The authors acknowledge the contribution of the project "Lithuanian Cybercrime Centre of Excellence for Training, Research and Education", Grant agreement No. HOME/2013/ISEC/AG/INT/400005176, co-funded by the Prevention of the Fight against Crime Programme of the EU.

References

[1] H. Van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistics*, vol. 12, 2005, pp. 65–77.

[2] D. Połap, and M. Woźniak. Flexible Neural Network Architecture for Handwritten Signatures Recognition. *International Journal of Electronics and Telecommunications*, vol. 62, no. 2, 2016, pp. 197–202.

[3] M. Koppel, J. Schler, and Sh. Argamon. Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 60, no. 1, 2009, pp. 9–26.

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

[5] K. Luyckx, and W. Daelemans. Authorship Attribution and Verification with Many Authors and Limited Data. *Proceedings of the 22Nd International Conference on Computational Linguistics*, vol. 1, 2008, pp. 513–520.

[6] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. Ch. R. Shin, and D. Song. On the Feasibility of Internet-Scale Author Identification. *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, 2012, pp. 300–314.

[7] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel. Authorship Attribution of Micro-Messages. *Empirical Methods in Natural Language Processing*, 2013, pp. 1880–1891.

[8] M. Koppel, J. Schler, Sh. Argamon, and E. Messeri. Authorship Attribution with Thousands of Candidate Authors. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 659–660.

[9] M. Koppel, J. Schler, and Sh. Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, vol. 45, no. 1, 2011, pp. 83–94.

[10] M. Koppel, J. Schler, and Sh. Argamon. Authorship Attribution: What's Easy and What's Hard? *Journal of Law & Policy*, vol. 21, 2013, pp. 317–331.

[11] M. Koppel, J. Schler, Sh. Argamon, and Y. Winter. The "Fundamental Problem" of Authorship Attribution. *English Studies*, vol. 93, no. 3, 2012, pp. 284–291.

[12] S. Okuno, H. Asai, and H. Yamana. A Challenge of Authorship Identification for Ten-Thousand-scale Microblog Users. *IEEE International Conference on Big Data*, 2014, pp. 52–54.

[13] Y. Seroussi, I. Zukerman, and F. Bohnert. Authorship Attribution with Latent Dirichlet Allocation. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 181–189.

[14] G. Žalkauskaitė. *Idiolekto požymiai elektroniniuose laiškuose. [Idiolect signs in e-mails]*, Vilnius University, Lithuania. PhD thesis, 2012 (in Lithuanian).

[15] A. Venčkauskas, R. Damaševičius, R. Marcinkevičius, and A. Karpavičius. Problems of Authorship Identification of the National Language Electronic Discourse. *ICIST 2015: 21st International Conference on Information and Software Technologies*, 2015, pp. 415–432.

[16] J. Kapočiūtė-Dzikienė, L. Šarkutė, and A. Utka. The Effect of Author Set Size in Authorship Attribution for Lithuanian. *NODALIDA: 20th Nordic Conference of Computational Linguistics*, 2015, pp. 87–96.

[17] J. Kapočiūtė-Dzikienė, A. Utka, and L. Šarkutė. Authorship Attribution of Internet Comments with Thousand Candidate Authors. *ICIST 2015: 21st International Conference on Information and Software Technologies*, 2015, pp. 433–448.

[18] E. Maciej. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, vol. 30, no. 1, 2013, pp. 167–182.

[19] K. Luyckx. Authorship Attribution of E-mail as a Multi-Class Task. *CLEF 2011 Labs and Workshop, Notebook Papers*, (eds.) V. Petras and P. Forner and P. Clough, 2011.

[20] C. Cortes, and V. Vapnik. Support-Vector Networks. *Machine Learning*, vol. 20, no. 3, 1995, pp. 273–297.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, vol. 11, no. 1, 2009, pp. 10–18.

[22] D. D. Lewis, and W. A. Gale. A Sequential Algorithm for Training Text Classifiers. *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 3–12.

[23] G. Salton, and Ch. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, vol. 24, no. 5, 1988, pp. 513–523.

[24] V. Daudaravičius, E. Rimkutė, and A. Utka. Morphological annotation of the Lithuanian corpus. *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies (ACL'07)*, 2007, pp. 94–99.

[25] J. Kapočiūtė-Dzikienė, F. Vaassen, W. Daelemans, and A. Krupavičius. Improving Topic Classification for Highly Inflective Languages. *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, 2012, pp. 1393–1410.