# An Automatic Author Profiling from Non-Normative Lithuanian Texts

Monika Briedienė
Vytautas Magnus University
Kaunas, Lithuania
monika.briediene@vdu.lt

Jurgita Kapočiutė - Dzikienė
Vytautas Magnus University
Kaunas, Lithuania
jurgita.kapociute-dzikiene@vdu.lt

*Abstract* - **This paper presents author profiling research done on the Lithuanian texts using automatic machine learning methods. Our research is novel and challenging due to the following reasons: 1) a big number of author profiling dimensions, i.e., gender, age, education, marital status and personality type; 2) very short (avg. ~ 24 tokens) non-normative texts; 3) vocabulary rich highly inflective Lithuanian language. We have performed experimental investigation that resulted in choosing automatic author profiling methods (in particular, classifiers and feature types) that have reached the highest accuracy on the pure texts without any meta-information about their authors. Out of a number of experimentally investigated classifiers using lexical or symbolic features the Naïve Bayes Multinomial method with character n-grams feature type yielded the best performance reaching 84.3%, 52.7%, 79.6%, 76.6%, 79.1% of accuracy in gender, age, education, marital status and personality type detection tasks, respectively.**

***Keywords—gender detection, age detection, education detection, marital status detection, personality type detection, author profiling, the non-normative Lithuanian language, supervised machine learning***

## I. Introduction

In today's world, numbers of electronic texts have exceeded paper texts by several times. However, the vast majority of these texts are written anonymously or pseudonymously. For this reason, court analysts, web forum administrators, social networks supervisors are increasingly facing impersonation, bullying or harassment, discloser of confidential information, dissemination of disinformation, and other issues. Uncovering the exact identity of the person is very complicated and sometimes unsolvable task, whereas to reveal his/her meta-information (i.e., demographic features: age, gender, etc.) is easier, but still very useful. The revealed meta-information that, e.g., *a 50-year-old man is impersonating a 10-year-old girl* may encourage the police to dive more detailed into the data or even take decisive actions for the criminal offense. The manual Internet space monitoring and manual text analysis is hardly possible, because it requires enormous amounts of human resources. Thus, natural language processing technologies become the only solution for tacking similar problems.

The author profiling experimental investigations confirm that the authors' characteristics can be determined by analyzing the style of the text. It is possible due to a phenomenon of the existing human stylome (an analogue of a genome) which allows each person to formulate sentences and express their thoughts in his/her special and unique ways [1]. Similarly, in many research studies, it is claimed that this phenomenon occurs not only in the style of individual, but also in the style of their groups, sharing the same demographic characteristics (as age, gender, education or marital status) or the personality type.

In general, the identification of an authorship has the long history dating back to 1887 [2], but with the Internet era its popularity gained dramatically. Therefore the author profiling – responsible for the automatic extraction of the meta-information about some author (as, e.g., age [3], gender [4], psychological status [5], etc.) – nowadays is an active and important research area. The author profiling research is mainly focused on the English language, whereas for the Lithuanian language it is rather a new subject. The age, gender and political views profiling tasks are solved using parliamentary transcripts [6]; age and gender profiling tasks are solved using the Lithuanian literary texts [17]. However, these research works are done on rather long (having ~ 217 tokens on average) and normative Lithuanian texts. The non-normative Lithuanian language (which is the object of research in this paper) is much more complicated: it is full of out-of-vocabulary words, jargon, foreign language insertions and neologisms. Besides, it faces an important problem of diacritics ignorance (where ą, č, ę, ė, į, š, ų, ū, ž are often replaced with the appropriate ASCII equivalents). However, the author profiling task on the non-normative Lithuanian texts is issued using the gender dimension only [7]. Moreover, some sub-tasks of the author's profiling on the education, marital status, and personality type dimensions have never even been solved before using any types of Lithuanian texts. Consequently, the purpose of this paper is to fill in the above mentioned gap: i.e., to offer the methods (classifiers, their parameters, and features types) able to create the automatic author profiles from the short non-normative Lithuanian texts (Facebook posts, comments and messages).

The final goal of this research can be achieved after performing the following intermediate tasks: (1) a related work analysis (see Section II), (2) a construction of the representative corpus containing non-normative Lithuanian texts (see Section III), (3) an analytical selection of the most promising methods (see Section IV), (4) a precise experimental evaluation of selected methods (see Section V). The conclusions (recommendations) and future research plans for the author profiling tasks when using short non-normative Lithuanian texts are in Section VI.

## II. Related Works

There are many methods used to deal with the author profiling task. All existing approaches can be grouped according to the following criteria: the percentage of training instances in the dataset, an amount of information they provide, (i.e., a recognition-training feedback) and the nature of knowledge. Based on these criteria, the approaches are [8]: Rule-based, Unsupervised Machine Learning, Supervised Machine Learning, and Similarity-Based.

The obsolete rule-based methods use rules that have been constructed by human-experts. The development process itself is very difficult and requires linguistic competence. In addition, rules are created for the specific solution, therefore are hardly transferable to the new areas.

Unsupervised machine learning (or clustering methods) is chosen when no meta-information (i.e., no training instances) is provided. Examples of the text are grouped according to their similarity. The main disadvantage of these methods is that their grouping does not necessarily correspond an imaginary grouping of a human. Usually because of their low accuracy, these methods are not popular in author profiling tasks.

If texts are supplemented with the necessary meta-information about the particular author characteristic (so-called class) the supervised machine learning is one of two best choices. The stylistic, lexical or symbolic text characteristics (i.e., so-called features) are presented as the input. The classifier summarizes training information and creates a model as its output. This model afterwards can be used for the author profiling of unseen texts. A main disadvantage of all supervised machine learning methods is that they require a comprehensive and representative training data to create a reliable and comprehensive model. The advantage of supervised methods is that they can be flexibly adjusted to the new tasks or areas by adding new text samples and retraining the classifier. The deep learning methods [9] [10] (that became extremely popular recently for many text classification tasks) are also representatives of this group. The popularity of the Neural Networks (Convolutional [10], Recurrent [9], etc.) is also growing recently. Such popularity has also been driven by the technical progress: it has led to the faster computing and processing huge amounts of data. The deep learning is used for the author profiling [10] and authorship attribution [9] tasks. Despite the deep learning methods are successfully applied in many natural language processing tasks, on the smaller datasets (as in our paper) they underperform the other supervised machine learning approaches, such as Support Vector Machines or Naïve Bayes Multinomial [11]. The similarity-based approaches (often researched and discussed separately) are very similar to the supervised machine learning approaches by their nature. The only difference is that instead of creating a model, they preserve all training instances and use similarity measures to determine to which of available classes some incoming unseen instance is the most similar. An advantage of similarity-based methods is that they keep the entire training set; so the information is not lost during generalization.

The majority of research done for solving the author profiling tasks involve these popular supervised approaches (e.g., Naïve Bayes [12], Naïve Bayes Multinomial [13], Support Vector Machines [14]) and similarity-based (e.g., k-Nearest Neighbor) or the comparative experiments proving the superiority of Naïve Bayes Multinomial and Support Vector Machines (as in [15]). Since, it is proved that these approaches are not only the most popular, but the most accurate for the author profiling tasks, further we will focus only on these types of methods.

When analyzing the Lithuanian non-normative texts, we follow the recommendations formulated for the other languages. However, a language factor itself should also be taken into account. The Lithuanian language (used in our research) has rich vocabulary, morphology, word derivation system and relatively free-word order in a sentence. Despite the Lithuanian language (especially non-normative) is rather complicated, some of previously mentioned language characteristics do not necessary have to complicate our solving tasks, i.e., it might occur that our investigated groups of individuals are bind to the very different, but very representative non-normative sentence structures or vocabularies.

## III. Corpus

Unfortunately, the author profiling benchmark corpora are not available on the Internet for the non-normative Lithuanian language, therefore in this research we are using the corpus that was specifically created for our tasks. The corpus is composed of unprocessed posts (without any appearance of the third party texts) manually harvested from the Facebook social network in the period of 2016-2017. The author profiling research for the other languages mostly focuses on the Twitter [15], but not Facebook [16] texts. It is due to the convenient APIs that help crawling tweets; besides, in some countries Twitter is more popular than Facebook. In our work we have chosen Facebook social network due to its popularity in Lithuania and opportunity to store more demographic characteristics such as education, marital status (not only age or gender) reported by the users themselves.

Our corpus contains posts, comments and messages of 200 individuals (for statistics see Figure 1), one text per person (to avoid the authorship attribution impact on the author profiling results). 102 and 98 texts belong to women and men, respectively (see *Gender* column in Figure 2). The youngest participant is 18 years old, the oldest – 78, the mean age of respondents is ~ 36.9. Respondents are divided into six age groups (see *Age* column in Figure 2). The selected grouping is used in surveys of psychologists, in the social studies, in the largest European and Lithuanian data archives. Besides, it is also used in the similar research works [17], making our results more comparable to the previously reported for the Lithuanian language.

The education level of 105 and 95 respondents is higher and secondary, respectively (see *Education* column in Figure 2). 114 and 86 individuals claimed they are married and single, respectively (see *Marital status* column in Figure 2). 112 and 88 people attributed themselves as extrovert and introvert, respectively (see *Personality type* column in Figure 2).

The corpus consists of 4.830 tokens (including in-the-vocabulary and out-the-vocabulary words, numbers, and non-normative “words” with embedded digits or punctuation) in total. The shortest text (without symbols and emoticons) is only 2 tokens length, the longest – 161, the average length per text is only ~ 24 tokens.

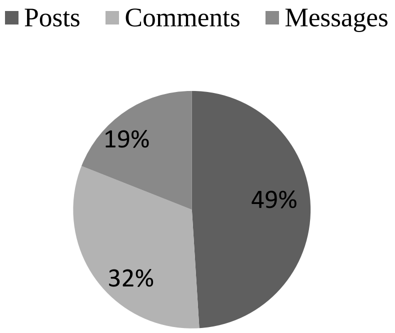

Fig. 1 A percentage of posts, comments and messages in our corpus

## IV. METHODOLOGY

The methodological part covers two main directions: 1) the proper selection of the classifier and 2) the proper selection of the feature type.

To come up with the very best, we have analyzed the following classifiers of these groups:

- *Supervised machine learning*. A representative of this type is the Support Vector Machine (SVM) method (introduced by Cortes C. and Vapnik V. in 1995 [18]). It is a discriminatory instance-based approach, currently considered as one of the most popular text classification techniques. The method effectively copes with the huge number of features, sparse feature vectors and does not perform an aggressive feature selection, which may result in the loss of valuable information and accuracy [19]. Another representatives are Naïve Bayes (NB) and its modification Naïve Bayes Multinomial (NBM) (introduced by Lewis D. D. and Gale W. A. in 1994 [20]). These techniques are generative profile-based approaches, often chosen due to their simplicity and sufficiently high accuracy. The NB assumption about the feature independence allows each parameter to be learned separately; these methods work especially well when a number of features having equal significance is high; they are fast and do not require large data storage resources. Moreover, Bayesian methods often play a baseline role in the evaluation.
- *Similarity-based*. A representative of this type is the IBK method (introduced by Aha D. and Kibler D. in 1991 [21]). This nearest neighbors’ classifier chooses the appropriate *k* value, based on the *k*-time cross-check after the distance evaluation (between a testing instance and all samples in the training set). Another representative is Kstar method (introduced by Cleary J. G. and Trigg L. E. in 1995 [22]). On the contrary to IBK, Kstar calculates not a distance measure, but a similarity function. It differs from the other approaches of this type, because uses the entropy-based distance function. These two last-mentioned methods store

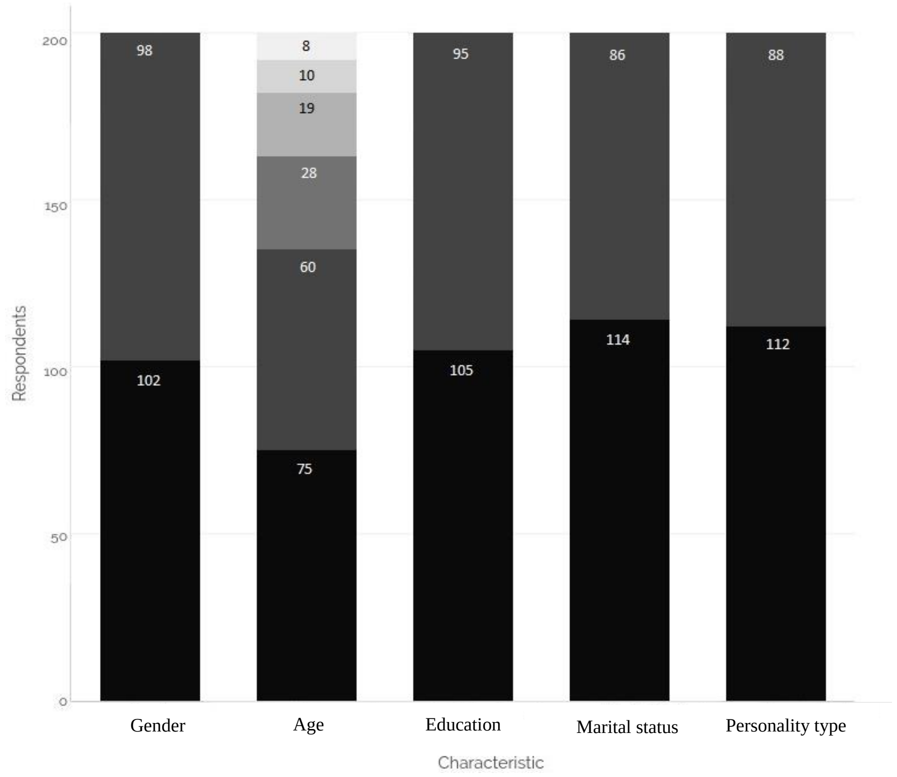

*Fig. 2 Distribution of respondents according to characteristics*

all available instances; therefore, are prevented from the information loss during training.

Our second research direction involved the proper selection of the feature type. In our experiments we have explored:

- Lexical feature types: token uni-grams ($n$=1) (individual tokens) and token tetra-grams ($n$=4) (sequences of 4 tokens in a window sliding one token at the time). For instance, from the phrase "author profiling from the Lithuanian texts" it would be generated 6 unigrams: "author", "profiling", "from", "the", "Lithuanian", "texts" and 3 tetra-grams "author profiling from the", "profiling from the Lithuanian", "from the Lithuanian texts".
- Character features, in particular, character n-grams similarly to token n-grams are sequences of items, but instead of tokens they contain characters. For instance, from the phrase "author profiling" it would be generated the following document-level character 4-grams: "auth", "utho", "thor", "hor_", "or_p", "r_pr", etc. (where "_" marks the whitespace). It is important to mention that a value of $n$ not necessary has to be fixed. E.g., with the interval $n$ = [2,4] all bi-grams ($n$=2), trigrams ($n$=3), and tetra-grams ($n$=4) would be generated and used as features.

## V. EXPERIMENTS AND RESULTS

Our experiments were carried out on the corpus described in Section III using the methods and feature types described in Section IV.

We used the implementations of the methods integrated into the WEKA 3.8 machine learning toolkit[1]. WEKA [23] allowed both: the extraction of features and selection of the classifier.

All experiments were performed using stratified10-fold cross validation and evaluated with the accuracy (1) and f-score (2) metrics. The results are considered acceptable and reasonable if the achieved author profiling accuracy is above majority (3) and random (4) baselines.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \qquad (1)$$

$$F\_score = \frac{2 * tp}{2 * tp + fp + fn} \qquad (2)$$

here $tp$ (true positives), $tn$ (true negatives), $fp$ (false positives), $fn$ (false negatives) denote a number of correctly classified instances $c_i$ with $c_i$ and $c_j$ with any other $c_j$, incorrectly classified instances $c_i$ with any other $c_j$ and any other $c_j$ with $c_i$, respectively

$$\max(p_i) \qquad (3)$$

$$\sum_i p_i^2 \qquad (4)$$

here $p_i$denote the probability of the class

Our preliminary experiments have involved the selection of the most accurate classification technique when using word tokenizer with unigrams ($n$=1) (denoted as *word1*), n-gram tokenizer with unigrams ($n$=1) (*lex1*) and tetra-grams ($n$=4) (*lex4*), alphabetic tokenizer with unigrams (n=1) (*alph1*), character n-gram tokenizer with unigrams ($n$=1) (*char1*) and tetra-grams ($n$=4) (*char4*) (the best results are presented in Figures 3-7). The overall best results were achieved with SVM and NBM methods and character n-grams[2]. These methods also demonstrated the best performance in the author profiling tasks on the morphologically complex Arabic language [15].

In our later experiments we have performed the tuning of the character n-gram parameter $n$ by keeping the classifier parameter stable and equal to SVM or NBM (because these classifiers demonstrated the best performance in the classifier selection experiments). The obtained results with the different author profiling dimensions are reported in Figure 8.

The overall best results (reaching 0.843 of the accuracy and 0.843 of f-score) on the short non-normative Lithuanian texts in the gender detection task were achieved with the NBM and character n-grams of $n$ = [6, 7] as the feature type. The best results reaching 0.527 of accuracy and 0,473 of f-score on the age dimension were achieved with the NBM and character n-grams (of $n$ = [5, 5]). In the education detection NBM and character n-grams (of $n$ = [5, 5]) demonstrated the best performance reaching 0.796 of accuracy and 0.796 of f-score. Experiments with the marital status showed the best results reaching 0.766 of accuracy and 0.767 of f-score with the NBM and character n-grams (of $n$ = [6, 6]). Tests with the personality type proved the superiority of NBM again: the highest 0.791 accuracy and 0.792 f-score was achieved with the character n-grams (of $n$ = [6, 6]). Thus, the Naïve Bayes Multinomial classifier and previously reported feature types would be recommended for the similar tasks and languages.

On the contrary, the best previously reported age and gender profiling results on the normative Lithuanian language were achieved with the SVM classifier and lemma bi-grams as the feature type [17]. It is not surprising having in mind that morphological tools (dealing with the normative texts) were maximally helpful. Besides, the second best feature type was also based on the character n-grams. Despite our best method achieved slightly higher accuracy compared to the previously reported, the direct comparison is hardly possible due to the very different experimental conditions (datasets and their sizes, language types, text lengths, etc.).

In general, the gender detection task is solved for a rather big group of languages, reaching ~ 80% and ~ 56.53% of accuracy on the normative English in [4] and [24], respectively; 64.73% on the Spanish blogs in [24] and ~ 82.60% on the Greek blogs [25]. On the non-normative tweet texts the obtained accuracies are still surprisingly high reaching, e.g., ~ 98% on Arabic in [15] and ~ 99% on English in [26]. However, the reported results, especially for the English language, are very controversial (from ~ 56.53% in [24] to even ~ 99% in [26]). The age detection task is also thoroughly researched for many

[1] Download from: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[2] Since the f-score values demonstrate the same trend compared to the accuracies, we do not present them in the figures.

languages, reaching 64.0%, 43.80%, 19.09% on the English texts [24], [3], [27]; 64.30%, 37.50% on the Spanish [24] [27]; 71.3% on the Dutch [28]; 80% on the Chinese [29]. Research on the personality type is mostly done on the normative English language [5] and reaches ~ 58.2% of accuracy.

Hence, the observed results are very different, due to the different test samples, methods, or chosen languages.

Due to the very different experimental conditions (different datasets, used methods and language types) these results are hardly comparable between; as well as they are hardly comparable with the results obtained in our research work.

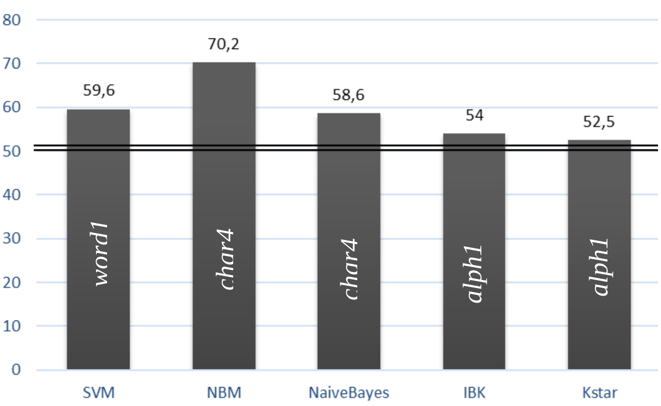

Fig. 3 Accuracies (in percentage) obtained with different classifiers solving gender detection task (an upper horizontal line represents a majority baseline, lower – a random baseline). Every column shows the best result obtained with different feature type: word tokenizer & unigrams denote as word1, alphabetic tokenizer & unigrams - alph1, n-gram tokenizer & unigrams - lex1, n-gram tokenizer & tetra-grams - lex4, character n-gram tokenizer& unigrams - char1, character n-gram tokenizer & tetra-grams - char4.

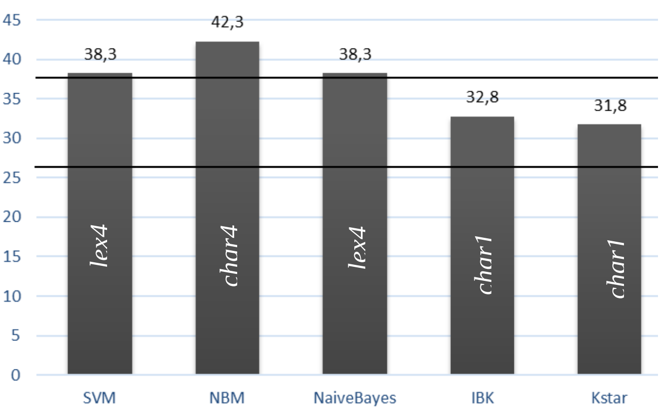

Fig. 4 Accuracies (in percentage) obtained with different classifiers solving age detection task. For the other notations see Fig. 3.

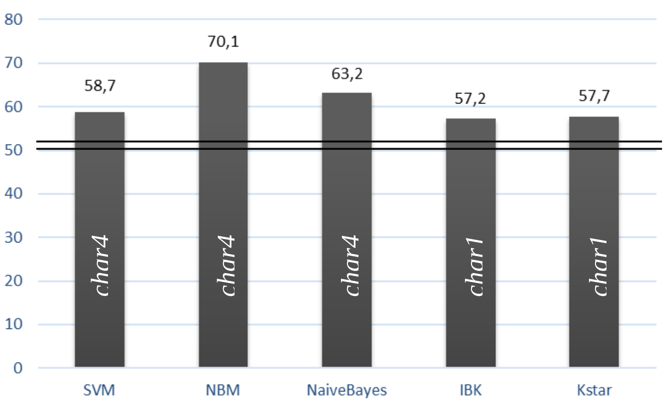

Fig. 5 Accuracies (in percentage) obtained with different classification solving education detection task. For the other notations see Fig. 3.

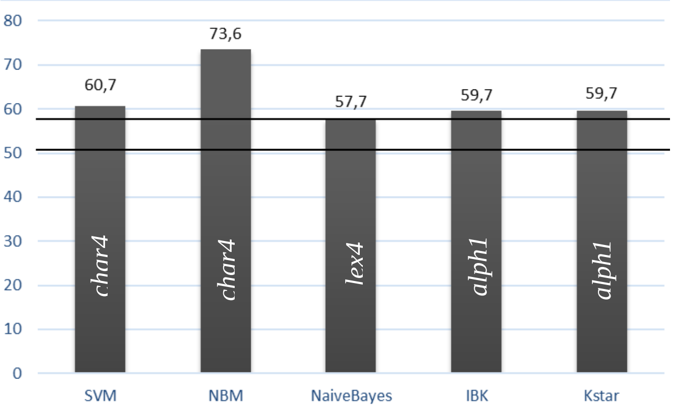

Fig. 6 Accuracies (in percentage) obtained with different classification solving marital status detection task. For the other notations see Fig. 3

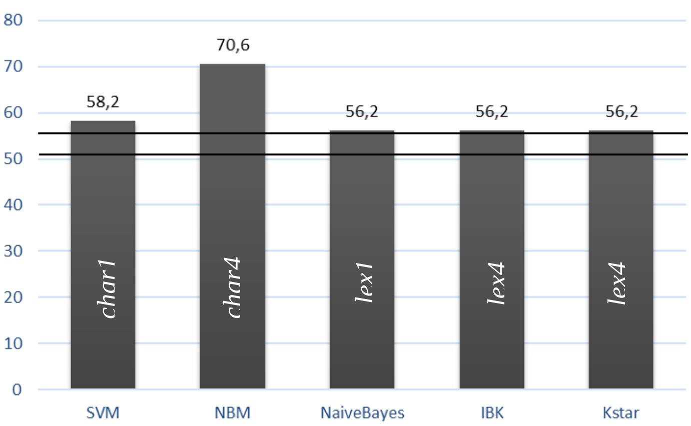

Fig. 7 Accuracies (in percentage) obtained with different classifiers solving personality type detection task. For the other notations see Fig. 3.

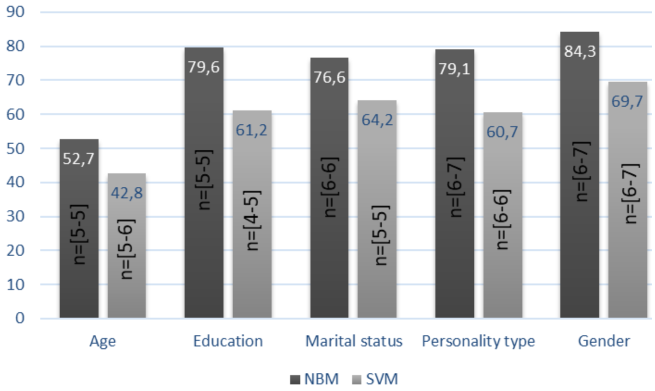

Fig. 8 The best summarized accuracies (in percentage) for the different profiling dimensions.

## VI. CONCLUSION AND FUTURE WORKS

In this paper we report the author profiling task results using short (of only avg. ~ 24 tokens) Lithuanian non-normative texts harvested from the Facebook social network. During our research we investigated the most popular supervised machine learning (Naïve Bayes, Naïve Bayes Multinomial, Support Vector Machine) and similarity-based (IBK, kStart) techniques plus various lexical and character feature types.

The best results on the 1) gender (84.3% of accuracy), 2) age (52.7%), 3) education (79.6%), 4) marital status (76.6%) and 5) personality type (79.1%) dimensions were achieved with 1) Naïve Bayes Multinomial and character n-grams of $n = [6, 7]$; 2) Naïve Bayes Multinomial method and character n-grams of $n = 5$; 3) Naïve Bayes Multinomial and character n-grams of $n = 5$; 4) Naïve Bayes Multinomial and character n-grams of $n = 6$; 5) Naïve Bayes Multinomial method and character n-grams of $n = 6$.

In the future research our focus on the non-normative Lithuanian texts remains. We are planning to increase our author profiling corpus and test it on the different deep learning approaches.

## REFERENCES

[1] H. Van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, A. Neijt, "New Machine Learning Methods Demonstrate the Existence of a Human Stylome," Journal of Quantitative Linguistics, 2005.

[2] T. C. Mendenhall, "The Characteristic Curves of Composition," Science, 1851.

[3] J. Schler, M. Koppel, S. Argamon, J. Pennebaker, "Effects of Age and Gender on Blogging," American Association for Artificial Intelligence , 2006.

[4] M. Koppel, S. Argamon, A. R. Shimoni, "Automatically Categorizing Written Texts by Author Gender," Literary and Linguistic Computing, pp. 401-412, November 2002.

[5] S. Argamon, S. Dawhle, M. Koppel, J. Pennebaker, "Lexical Predictors of Personality Type," Joint Annual Meeting of the Interface and the Classification Society of North America, June 2005.

[6] J. Kapočiūtė-Dzikienė, L. Šarkutė, A. Utka, "Automatic author profiling of Lithuanian parliamentary speeches : exploring the influence of features and dataset sizes.," Human Language Technologies – The Baltic Perspective, pp. 99-106, 2014.

[7] M. Briedienė, J. Kapočiūtė-Dzikienė, "An authomatic gender detection from non-normative Lithuanina texts," Ceur-Ws, Kaunas, 2017.

[8] E. Stomatatos, "A Survey of Modern Author," Journal of the American Society for Information Science and Technology, 2009.

[9] D. Bagnall, "Author identification using multi-headed recurrent," PAN 2015, 2015.

[10] S. Sierra, M. Montes-y-Gómez, T. Solorio, F. A. González, "Convolutional Neural Networks for Author Profiling," Notebook for PAN at CLEF 2017, 2017.

[11] E. A. Zanaty, "Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification," Mathematics Dept., Computer Science Section, Faculty of Science, Sohag University, Sohag, Egypt, 2012.

[12] M. Meina, K. Brodzinska, B. Celmer, M. Czokow, M. Patera, J. Pezacki, M. Wilk, "Ensemble-based classification for author profiling using," Notebook for PAN at CLEF 2013, 2013.

[13] T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy, "Profile specific Document Weighted approach using a New Term Weighting Measure for Author Profiling," International Journal of Intelligent Engineering and Systems, pp. 136-146, december 2016.

[14] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, W. Daelemans, "Overview of the 3rd Author Profiling Task at PAN 2015," 2015.

[15] E. AlSukhni, Q. Alequr, "Investigation the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author," Article Published in International Journal of Advanced Computer Science and Applications, 2016.

[16] M. Fatimaa, K. Hasanb, S. Anwara, R. M. A. Nawab, "Multilingual author profiling on Facebook," Information Processing & Management, pp. 886-904, liepa 2017.

[17] J. Kapočiūtė-Dzikienė, A. Utka, L. Šarkutė, "Authorship Attribution and Author Profiling of Lithuanian Literary Texts," Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pp. 96-105, September 2015.

[18] C. Cortes, V. Vapnik, "Support-Vector Networks," Machine Learning, pp. 273–297, 1995.

[19] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," European Conference on Machine Learning, pp. 137-142, 1998.

[20] D. D. Lewis, W. A. Gale, "A Sequential Algorithm for Training Text Classifiers," SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval , pp. 3-12, July 1994.

[21] D. Aha, D. Kibler, "Instance-based learning algorithms," Machine Learning, pp. 37–66, 1991.

[22] J. G. Cleary, L. E. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure," In Proceedings of the 12th International Conference on Machine Learning, 1995.

[23] 2016. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/.

[24] K. Santosh, R. Bansal, M. Shekhar, V. Varma, "Author Profiling: Predicting Age and Gender from Blogs," Notebook for PAN at CLEF 2013, 2013.

[25] G. K. Mikros, "Authorship Attribution and Gender Identification in Greek Blogs," Methods and Applications of Quantitative Linguistics, pp. 21-32, 2012.

[26] Z. Miller, B. Dickinson, W. Hu, "Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features," International Journal of Intelligence Science, pp. 143-148 , 2012.

[27] J. Marquard, G. Farnadi, G. Vasudevan, M-F. Moens, S. Davalos, A. Teredesai, M. De Cock, "Age and Gender Identification in Social Media," CLEF 2014 working notes; PAN 2014, 2014.

[28] C. Peersman, W. Daelemans, L. Van Vaerenbergh, "Predicting Age and Gender in Online Social Networks," SMUC '11 Proceedings of the 3rd international workshop on Search and mining user-generated contents , pp. 37-44 , 2010.

[29] Li Chen, Tieyun Qian, Fei Wang, Zhenni You, Qingxi Peng, Ming Zhong, "Age Detection for Chinese Users in Weibo," WAIM 2015: Web-Age Information Management, 2015.

[30] A. Venckauskas, A. Karpavicius, R. Damaševičius, R. Marcinkevičius, J. Kapočiūte-Dzikiené, and C. Napoli, “Open class authorship attribution of Lithuanian Internet comments using one-class classifier.” In Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 373-382, 2017..

[31] M. Wróbel, J.T. Starczewski, and C. Napoli, “Handwriting recognition with extraction of letter fragments”. In International Conference on Artificial Intelligence and Soft Computing, pp. 183-192, 2017.