An automatic gender detection from non-normative Lithuanian texts

Monika Briedienė
Vytautas Magnus University
K. Donelaičio 58,
LT-44248, Kaunas, Lithuania
e-mail: monika.briediene@fc.vdu.lt

Jurgita Kapočiūtė-Dzikienė
Vytautas Magnus University
K. Donelaičio 58,
LT-44248, Kaunas, Lithuania
e-mail: jurgita.kapociute-dzikiene@vdu.lt

***Abstract*—This paper describes the gender detection research done on Lithuanian texts using automatic machine learning methods. The main contribution of our work is investigations done namely on the very short (avg. ~ 39 tokens) non-normative texts. With this paper we analyze a fundamental problem: how to choose automatic methods (in particular, classifiers and feature types) that could achieve the highest accuracy in our solving author profiling task (when the short pure text itself is the only evidence used for determining the author's meta-information). The related research analysis helped us to select the methods which demonstrated encouraging results on the other languages and to apply them on the Lithuanian dataset. Out of a number of experimentally investigated classifiers with lexical or symbolic features the Naïve Bayes Multinomial method with character n-grams (of n = [1, 5]) feature type yielded the best performance reaching 83.6% of the accuracy.**

Keywords—gender detection; author profiling; non-normative Lithuanian language; supervised machine learning

I. INTRODUCTION

Due to the constant increase of electronic texts the various natural language processing works become especially relevant. However, lots of these texts are written anonymously or pseudonymously, therefore court linguistic analysts, administrators of Internet forums, supervisors of social networks more and more often face such problems as impersonation, bullying or harassment, discloser of confidential information, dissemination of disinformation, etc. Although to disclose an identity of a particular person sometimes is rather difficult, the meta-information (i.e., demographic characteristics: age, gender, etc.) also may provide some clues: e.g. a system detects that a 50 years old man is impersonating a 12 years old girl and encourages the police to dive more into details or even to take the decisive actions in finding a criminal.

Researchers confirm that the authors' characteristics can be determined during an analysis of the text style. It is possible due to an existing phenomenon of a human stylome (an analogue of a genome) which enables each person to formulate sentences and to express their thoughts in special unique ways [*1*]. Similarly, a number of studies prove this phenomenon occurs not only in the style of individuals, but also in the style of their groups, sharing the same demographic characteristics (as age, gender or social status) or psychological state.

In general, the authorship identification has a long history dating back to 1887 [*2*], but the Internet era opened a gate to even greater popularity for it. Due to it 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 research area. The author profiling research is mainly focused on the English language, whereas for the Lithuanian it is rather a new topic. Moreover, some author profiling sub-tasks (as, e.g., the gender detection using non-normative Lithuanian texts) have never been solved before. Consequently, an aim of this paper is to fill the previously mentioned gap: i.e., to explore the methods on short non-normative Lithuanian texts (Facebook posts, comments and messages) and to formulate the recommendations (about classifiers, their parameters and features types) for the automatic gender detection task.

The ultimate goal of this research can be achieved by 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) (5) conclusions (recommendations) for the gender detection when using short non-normative Lithuanian texts and our further research plans (see Section VI).

II. RELATED WORKS

All existing author profiling approaches can be grouped according to the following criteria: a percentage of training instances in the dataset, an amount of information they provide, (i.e., a recognition-training feedback) and the nature of knowledge. According to these criteria the approaches are [*6*]: Rule-based, Unsupervised Machine Learning, Supervised Machine Learning and Similarity-Based.

The obsolete rule-based approaches use rules manually constructed by humans. The development process itself is very laborious and requires linguistic expertise. Moreover, created rules are tied to that specific solving problem, therefore are hardly transferable to new domains.

Unsupervised machine learning (or clustering methods) is selected when no meta-information is provided. The text samples are grouped according to the similarity between them. A main drawback of these methods is that their grouping does not necessarily correspond an imaginary grouping by a human. Mostly due to the very low accuracy these methods are not among the most popular choices in any author profiling tasks.

If texts are supplied with the necessary meta-information about the certain author characteristic (so-called class) the supervised machine learning is one of two best choices. The stylistic, lexical or symbolic text characteristics (extracted from the training instances) are provided as an input for a classifier. It generalizes all input information and produces a model as an output. This created model afterwards can be used for the author profiling of unseen texts. A main drawback of all supervised machine learning methods is that they require a comprehensive and representative dataset to create an exhaustive and robust model. An advantage is that the method can be flexibly adjusted to new tasks or domains: after adding new text samples the classifier can be easily retrained. The similarity-based approaches are very similar to the supervised machine learning by their nature. An only difference is that instead of creating the model they memorize and store all training instances and use similarity measures to determine to which of available classes some incoming instance is the most similar. An advantage of similarity-based methods is that they store the entire training set; therefore no information is lost during its generalization. Since both supervised machine learning and similarity-based approaches are the most accurate, they are the most popular for the various author profiling tasks. This important observation narrows down our research area to these approaches only.

The research done on various languages usually involves the investigation of these popular approaches for supervised machine learning (e.g., Naïve Bayes [*7*], Naïve Bayes Multinomial [*8*], Support Vector Machines [*9*]) 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 [*10*]).

When investigating the Lithuanian non-normative texts we considered recommendations formulated for the other languages. However, a language factor itself is also very important, therefore must be taken into account as well. The Lithuanian language (that we are coping in this research) is rich in the vocabulary and morphology, has the rich word derivation system and the relatively free-word order in a sentence. Despite the Lithuanian language is rather complicated, some of previously mentioned language characteristics do not necessary complete the solving problem, i.e., it might occur that our investigated groups of individuals are bind to the very different sentence structures or vocabulary.

In fact the gender detection task for the Lithuanian language is not absolutely new: it has been solved using the supervised machine learning methods [*11*]. However, these authors used rather long normative texts (having averagely ~217 tokens in each). Whereas the non-normative Lithuanian language (which is the object of research in this paper) is notably different: it is full of out-of-vocabulary words, jargon, foreign language insertions and neologisms. Moreover, the non-normative Lithuanian 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). Hence, in this research we are planning to check how much the accuracy is affected by a shortness of texts and a type of the language.

III. Corpus

A gender detection task was solved using the specifically prepared corpus of non-normative Lithuanian language texts. The corpus was composed of original posts (without any appearance of third party texts) harvested from the Facebook social network in October, 2016. It contains posts, comments and messages of 70 persons (for statistics see Figure 1) (32 and 38 texts belong to women and men, respectively (see Figure 2)). The youngest participant is 18 years old, the oldest – 77, the mean age of respondents is ~33.8. 43 and 27 people indicated that their level of education higher and secondary, respectively. 33 and 37 individuals claimed they are married and unmarried, respectively.

The corpus consists of 2.729 tokens in total[1] (of which 1.433 are written by men and 1.296 by women) (see Figure 3)). The shortest text (without symbols and emoticons) is only 4 tokens length, the longest – 161, the average length of texts is ~39 tokens.

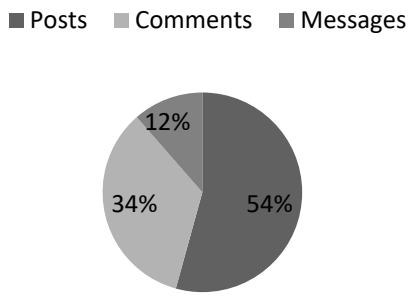

Fig. 1 A percentage of posts, comments and messages in our corpus

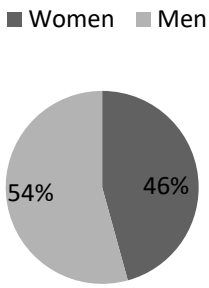

Fig. 2 A percentage of texts in our corpus written by men and women

[1] It is important to notice, that instead of words we focus on tokens in this work. Besides regular words, tokens also include out-of-vocabulary words, numbers, and non-normative "words" with embedded digits or punctuation marks.

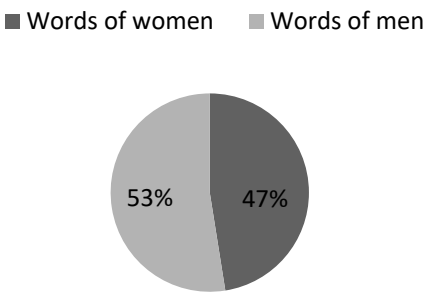

Fig. 3 A percentage of tokens in our corpus written by men and women

IV. METHODOLOGY

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

To come up with the very best, we investigated the following classifiers for:

- *Supervised machine learning*. A representative of this type is the Support Vector Machine (SVM) method (introduced by Cortes C. and Vapnik V. in 1995 [*12*]). It is a discriminatory case-based approach, currently considered as the most popular text classification technique. 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 [*13*]. Another representatives are Naïve Bayes (NB) and its modification Naïve Bayes Multinomial (NBM) (introduced by Lewis D. D. and Gale W. A. in 1994 [*14*]). These techniques are generative profile-based approaches, often chosen due to their simplicity. 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 [*15*]). 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 [*16*]). 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 all available instances; therefore are prevented from the information loss during training.

The second direction involved the proper selection of the feature type. In our experiments we investigated:

- Lexical features: 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 "gender detection from the Lithuanian texts" it would be generated 6 unigrams: "gender", "detection", "from", "the", "Lithuanian", "texts" and 3 tetra-grams "gender detection from the", "detection 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 "gender detection" it would be generated the following 4-grams: "gend", "ende", "nder", "der_", "er_d", "r_de", etc. (where "_" denotes the whitespace). It is important to mention that a value of n not necessary has to be fixed: i.e., ranges are also possible. With range, e.g., n = [2,4] it would be generated bi-grams (n=2), plus trigrams (n=3), plus tetra-grams (n=4).

V. EXPERIMENTS AND RESULTS

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

We used the implementations of the methods incorporated into the WEKA 3.8 machine learning toolkit[2]. WEKA [*17*] allowed both: the extraction of features and selection of the classifier.

In all our experiments we used 10 fold cross validation and evaluated accuracy (1) and f-score (2). The results are considered acceptable and reasonable if the accuracy is above random (3) and majority (4) baselines equal to 0.502 and 0.540, respectively.

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

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

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

$$\max(p_i) \quad (3)$$

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

Our preliminary experiments involved the selection of the most accurate classification technique when using token unigrams (n=1), token tetra-grams (n=4) and character tetra-grams (n=4) (the results are presented in Figure 4). The best results were achieved with SVM and NBM and character tetra-grams[3]. These methods also demonstrated the best

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

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

performance in gender detection tasks on the morphologically complex Arabic language [*10*].

Later on we used only SVM and NBM (because they demonstrated the best performance in the preliminary experiments) by tuning a parameter *n* in the character n-grams (see obtained results in Figure 5).

The overall best results (reaching 0.836 of the accuracy) on the short non-normative Lithuanian texts for the gender detection task were achieved with the NBM and character n-grams of *n*=[1, 5] as the feature type (see Figure 6); therefore we would recommend them for the other similar tasks and languages.

By the way, the best of the only previously reported results for the Lithuanian language in the gender detection task were achieved with the SVM and lemma bi-grams as the feature type [*18*]. 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 based on the character n-grams, too. Despite our best method achieved slightly higher accuracy (by 0.089) 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.).

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 [*19*], respectively; 64.73% on the Spanish blogs in [*19*] and ~82.6% on the Greek blogs [*20*]. On the non-normative tweet texts the obtained accuracies are surprisingly high reaching, e.g., ~98% on Arabic in [*10*] and ~99% on English in [*21*]. As we can see, the reported results, especially for the English language, are very controversial (~56.53% in [*19*] and even ~99% in [*21*]). Due to very different experimental conditions (different datasets, used methods and language types) these results become hardly comparable. They are also hardly comparable with the results obtained in our research work.

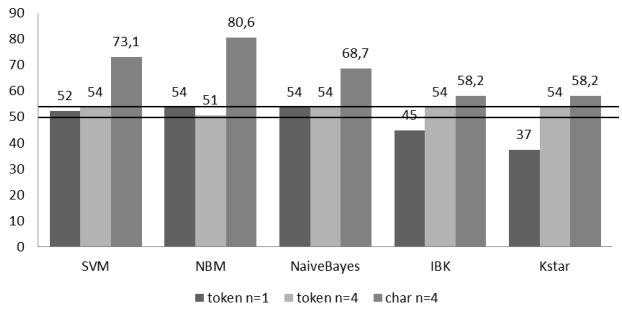

Fig. 4 Accuracies (in percentage) obtained with token unigrams (token *n*=1), token tetra-grams (token *n*=4) and character tetra-grams (char *n*=4) (an upper horizontal line represents a majority baseline, lower – a random baseline)

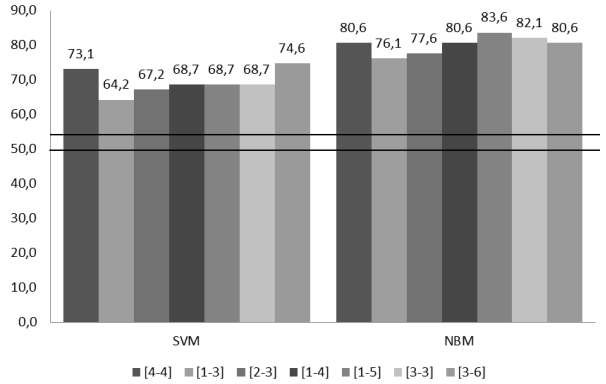

Fig. 5 Accuracies (in percentage) obtained with SVM and NBM methods using different *n* parameters in the character n-grams (an upper horizontal line represents a majority baseline, lower – a random baseline)

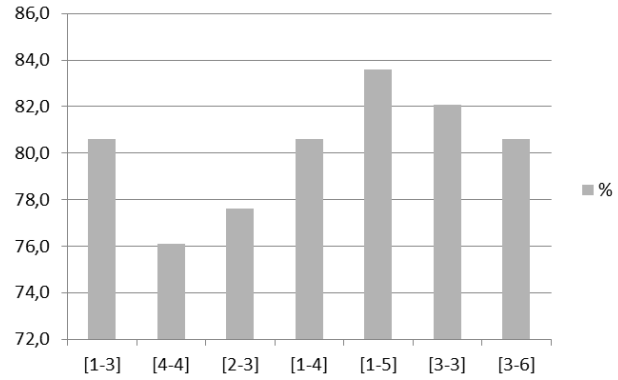

Fig. 6 Accuracies (in percentage) achieved with NBM method and different character n-grams

VI. CONCLUSION AND FUTURE WORKS

In this paper we report the first gender detection results using short (of only avg. ~39 tokens) Lithuanian non-normative texts taken 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 reaching 83.6% of accuracy were achieved with the Naïve Bayes Multinomial method and character n-grams (of *n* = [1, 5]) as features.

Since the majority of the research done for the Lithuanian langue is mostly focused on the normative texts, in the future research we are planning to pay special attention to this problem by increasing the datasets and tackling the other author profiling tasks as age detection, social status detection, etc.

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, vol. 12(1), pp. 65–77, 2005.

[2] T. C. Mendenhall, "The Characteristic Curves of Composition," pp. 37-66, 1851.

[3] J. Schler, M. Koppel, S. Argamon, J. Pennebaker, "Effects of Age and Gender on Blogging," Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp. 191-197, 2006.

[4] M. Koppel, S. Argamon, A. R. Shimoni, "Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing, vol. 17(4), pp. 401-412, 2002.

[5] S. Argamon, S. Dawhle, M. Koppel, J. Pennebaker, "Lexical Predictors of Personality Type," Proceedings of Classification Society of North America, St. Louis MI, 2005.

[6] E. Stomatatos, "A Survey of Modern Author," Journal of the American Society for Information Science and Technology, Wiley, pp. 538-556, 2009.

[7] 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.

[8] 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 & Systems, pp.136-146, 2016.

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

[10] E. AlSukhni, Q. Alequr, "Investigation the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author," International Journal of Advanced Computer Science and Applications, pp. 319-328, 2016.

[11] 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: proceedings of the 6th inter, pp. 99-106, 2014.

[12] C. Cortes, V. Vapnik, "Support-Vector Networks," pp. 273-297, 1995.

[13] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," pp. 137-142, 1998.

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

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

[16] J. G. Cleary, L. E. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure," 12th International Conference on Machine Learning, pp. 108-114, 1995.

[17] (2016) WEKA. [Online]. http://www.cs.waikato.ac.nz/ml/weka/

[18] 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, 2015.

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

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

[21] 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.