Proceedings of the Federated Conference on Computer Science and Information Systems pp. 169–172
DOI: 10.15439/2018F22
ISSN 2300-5963 ACSIS, Vol. 15

Lithuanian Author Profiling with the Deep Learning

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

Robertas Damaševičius
Kaunas University of Technology
K. Donelaičio 73, LT-44029,
Kaunas, Lithuania
Email: robertas.damasevicius@ktu.lt

***Abstract*—We address the Lithuanian author profiling task in two dimensions (AGE and GENDER) using two deep learning methods (i.e., Long Short-Term Memory – LSTM) and Convolutional Neural Network – CNN) applied on the top of Lithuanian neural word embeddings. We also investigate an impact of the training dataset size on the author profiling accuracy. The best results are achieved with the largest datasets, containing 5,000 instances in each class. Besides, LSTM was more effective on the smaller datasets, and CNN – on the larger ones. We compare the deep learning methods with the traditional machine learning methods (in particular, Naive Bayes Multinomial and Support Vector Machine), and frequencies of elements as the feature representation). The comparison revealed that the deep learning is not the best solution for our author profiling task.**

I. Introduction and Related Work

AUTHOR Profiling (AP) is a specific subfield of Authorship Identification that aims at revealing characteristics of authors (e.g., age, gender, psychometric traits, etc.) from their writing style: synonymy and sentence structures used, grammatical or syntax errors made, etc. Thus, the AP task is solvable due to the stylometric "fingerprint" (so-called human stylome [1]): a phenomenon of individuals to express their thoughts in the written text in the specific unique ways. The stylome is also valid for the groups of individuals sharing the same demographic or psychometric characteristics. In some cases, the stylome is even attributed to the other human biometrics, as handwriting, gait or voice, and it tends to develop over time [2], depending on the age, education, social status of a person. Due to a number of potential applications in such fields as forensics, security or e-commerce, the importance of AP is constantly growing. These tasks are tackled with the automatic methods and continuous improvements of these methods contribute to the increase of the AP accuracy.

The majority of AP tasks are solved with the traditional machine learning methods and the weight vectors of features [3], [4]. The most influential examples of this field refer to Support Vector Machines (SVMs) [5], Multi-Class Real Winnow [6], Mean Proximity Clustering [7] and Holomorphic Transforms [8]. While a range of explored feature types usually covers stylistic (e.g., average sentence length, standardized type/token ratio), lexical (e.g., bag-of-words, function words), character (e.g., document or word-level character n-grams), morphological (e.g., part-of-speech tags) levels of feature representation types. The detailed description of these techniques can be found in [9].

Since methods are usually tested under different experimental conditions (various languages, profiling dimensions or datasets) it is difficult to determine, which one is the best. It is the reason why the scientific PAN competition of shared tasks plays an important role in the AP research field.[1] The comprehensive comparative analysis on the benchmark datasets reveals potential of tested methods and the new trends.

In 2013 [10], 2014 [11] and 2015 [12] PAN competition age and gender profiling was done on the English and Spanish datasets with the traditional supervised machine learning approaches: Logistic Regression, Random Forest, SVMs, etc. In 2016 PAN competition [13] the goal was to test the robustness of methods from the cross-genre perspective and SVMs were the dominant paradigm. In 2017 [14] two more languages (i.e., Arabic and Portuguese) were added to the dataset. Despite SVMs were still chosen by many participants, deep neural networks (in particular, Windowed Recurrent Convolutional Neural Network as an extension of Recurrent Convolutional Neural Network) achieved state-of-the-art performance on the gender dimension.

In the whole area of authorship identification, authorship attribution is the most explored topic for the morphologically complex Lithuanian language (the recent research work is described in [15], [16]). Unfortunately, the deep learning methods have never been applied on the Lithuanian language in any of these tasks, including AP. The aim of this research is: 1) to test their robustness on the AGE and GENDER dimensions; 2) to compare obtained results with the results produced by the traditional machine learning methods, described in [17].

II. Deep Learning Methods

Our solving task can be formulated as the supervised machine learning, where classifiers are the deep learning methods:

- *Long Short Term Memory* (LSTM) [18]. This method is a modification of Recurrent Neural Network (RNN) having a memory unit and able to learn long-term dependencies. The memory unit with input, output and forget gates is used to remember the values over arbitrary time intervals. The output with 256 nodes in the LSTM layer is an input to the fully connected softmax layer which output is the probability distribution over classes.

[1]More information about the PAN competition is in http://pan.webis.de/.

- *Convolutional Neural Network* (CNN) [19]. The convolution is performed on the sequentially connected word vectors (the detailed description is in [20]). The feature map is produced when the filters (in particular, of 3, 4, and 5 widths) are applied on each possible window of words in the text. The max-over-pooling operation on the feature map generates a single maximum value for each filter. Values from different filters are passed to a fully connected layer which outputs the probability distribution over classes.

The LSTM and CNN methods were tested using *deeplearning4j*[2] – the open-source distributed deep learning library for the Java Virtual Machine. Original method implementations were adjusted to solve only binary classification problems, therefore necessary adjustments to multi-class classification were done by the authors of this paper. All parameters were set to their default values, except for the maximum text length: i.e., it was set to 300 tokens (i.e., words or other text elements separated by spaces or punctuation) to match the maximum possible length of the input text (described in Section III-A).

Both deep learning methods were applied on the top of Lithuanian neural word embeddings (the description is in [21]), in particular, continuous bag-of-words of 300 dimensions generated with the negative sampling as the training algorithm. Since Seimas transcripts of ∼23.9 million tokens (described in Section III-A) are also the part of word embeddings corpora, our deep learning methods are protected from the out-of-vocabulary problem in all AP tasks. Despite we analyze the spoken edited language, the vocabulary of each speaker remains untouched. Since the vocabulary itself becomes one of the strongest evidence of the authorship, word embeddings should be the proper feature type for our solving task.

III. Experimental Set-Up and Results

A. Datasets

The datasets for our AP tasks are composed of the Lithuanian parliamentary text transcripts, representing speeches and debates by the Lithuanian Seimas members produced at regular parliamentary sessions and cover the period of 7 parliamentary terms from 1990 till 2013.

All texts perfectly represent formal spoken Lithuanian language, because: 1) the language of transcripts is unedited (texts match soundtracks), 2) words are grammatically correct. Only texts of the length between 100 and 300 tokens are considered, because: 1) very short texts are less informative; 2) too long texts might have the unclear authorship, i.e., long parliamentary speeches for parliamentarians might be written by someone else.

The experiments are carried out on the datasets for these dimensions:

- *AGE* dimension was composed of 6 classes (25,439 texts, 5,395,677 tokens, 161,010 types, ∼212.10 tokens/per text) related with the age intervals: *to-29* (inclusive), *30-39*, *40-49*, *50-59*, *60-69*, and *from-70* (inclusive).[3]
- *GENDER* dimension was composed of 2 classes (10,000 texts, 2,168,664 tokens, 101,951 types, ∼216.87 tokens/per text): *male* and *female*.

Each dimension was tested with 6 balanced datasets of 100, 300, 500, 1,000, 2,000, and 5,000 texts (i.e., instances) in each class. Except for the AGE dimension: the *to-29* class contained 707 and *from-70* class contained 4,732 instances at most. All datasets were composed by randomly selecting the determined number of text documents from the whole set of texts.[4]

The experiments with AGE and GENDER dimensions were performed with relevant datasets (described in Section III-A) of different sizes, containing 100, 300, 500, 1,000, 2,000, and 5,000 instances in each class.

B. Evaluation

We have tested two deep learning methods (in particular, LSTM and CNN) with the Lithuanian neural word embeddings (described in Section II) on the dataset described in Section III-A. Rough texts (without any normalization and dimensionality reduction) were given as the input. The stratified 10-fold cross-validation was used in all our experiments. The effectiveness of methods was evaluated with the *macro-accuracy* and *macro-f-score* measures (explanation is in [22]) averaged over classes and folds.

To determine if 1) obtained results are reasonable, and 2) differences between results are statistically significant, we have 1) calculated random and majority baselines, and 2) performed McNemar [23] test with one degree of freedom, respectively. The random ($\sum(P(c_j)^2)$) and majority ($max(P(c_j))$) baselines are the same in all datasets except for the AGE dimension with 1,000, 2,000 and 5,000 instances in each class (because it's classes *to-29* and *from-70* contained 707 and 4,732 instances, respectively). For the McNemar test, we have set the significance level equal to 95%, which means that the differences are considered statistically significant, if the calculated *p-value* is lower than 0.05.

The results produced by the deep learning methods with the neural word embeddings were compared to the results of the traditional classification methods (in particular, Naive Bayes Multinomial – NBM and Support Vector Machine - SVM) with the frequencies of elements as the text document feature representation. The results for NBM and SVM were taken from [17]. NBM and SVM were tested with the different feature representation types: ultimate style markers, document-level character n-grams (with *n*=[2,7]), function words, token n-grams (with *n*=[1,3]), token lemmas (with *n*=[1,3]), part-of-speech tag n-grams (with *n*=[1,3]), and n-grams of concatenated lexical and morphological features. There is no single the

[2]The deep learning library is in https://deeplearning4j.org/.

[3]The chosen grouping is also used in the largest European data archive (http://www.gesis.org) and in the Lithuanian Data Archive for Social Science and Humanities (http://www.lidata.eu).

[4]The *AMŽIUS_PROF* and *LYTIS_PROF* datasets of the *AGE* and *GENDER* dimensions, respectively, can be downloaded from http://dangus.vdu.lt/~jkd/eng/?page_id=16.

best feature representation type: it depends on the classification method and the dataset size (for the best types see Table I). Here *lemmorf* denotes lemmas + fine-grained POS information; *lex* – tokens, *lexpos* – tokens + coarse-grained POS; *lem* – lemmas; *lempos* – lemmas + coarse-grained POS; *lexmorf* – tokens + fine-grained POS information; *chr* – characters. The number next to each tag represents *n* of their n-gram.

C. Results

The results of the deep learning on the top of neural word embeddings and traditional machine learning methods with the best feature types (presented in Table I) are summarized in Figure 1. The figures do not present the *f-score* values, demonstrating the same trend as the *accuracy* values.

Figure 1 allow us to make the following claims. All obtained results are reasonable, because exceed random and majority baselines, except for LSTM with the dataset size of 100 in the GENDER dimension.

Marginally the best accuracies of 0.316 and 0.609 with the AGE and GENDER, respectively, were achieved with the CNN method and the largest datasets of 5,000 instances in each class. Besides, CNN method achieves higher profiling accuracy compared to LSTM on the larger datasets for all dimensions (with 1,000-5,000 for AGE and GENDER). Whereas, CNN is often outperformed by LSTM on the smaller datasets: with 100-500 for the AGE dimension; with 300-500 for GENDER.

According to the McNemar test, the differences in accuracies between tested LSTM and CNN methods are significant with $p < 0.05$ for the AGE and GENDER dimensions with 1,000, 2,000 and 5,000 instances in each class. For 100, 300, and 500 instance datasets for the AGE dimension are not statistically significant with $p = 0.32, 0.22, 0.19$, respectively. The p values for 100, 300, and 500 instance datasets for GENDER are 0.37, 0.99, and 0.38, respectively.

The comparison of LSTM or CNN + neural word embeddings with NBM or SVM + element frequencies as the feature representation type revealed that the deep learning methods are not the best choice for our AP tasks. The neural methods are significantly outperformed by the traditional machine learning methods: i.e., except for GENDER with 100 or 300 datasets.

The accuracies improve by increasing a number of instances in each class. In this research the purpose was to equalize the experimental conditions (in terms of dataset sizes) and to compare the effectiveness of deep learning methods with traditional machine learning methods. However, the deep learning results improve with the increase of the dataset size, whereas, e.g., NBM seems already have reached its limits (i.e., the peak on AGE and GENDER, are with 500 instance datasets, respectively). Maybe it is possible to find the breaking point where the deep learning methods reach or even bypass the effectiveness of traditional methods. Thus, the deep learning experiments with the larger datasets could be possible accuracy improvement direction for the future research.

Despite the experiments are performed with the grammatically correct texts and in-the-vocabulary words, for NBM and SVM lexical features (bag-of-words) are not always the best representation. The deep learning methods are applied on the top of neural word embeddings, however, in the future research would be useful to test the other types of embeddings (e.g., based on characters or lemmas). Moreover, the parameter (i.e., the numbers of layers, filters, or neurons in each hidden layer) tuning of LSTM and CNN could result in the higher AP accuracy, therefore this important step is also on the list of our future plans.

IV. Conclusions and Future Work

The main contribution of this research – the Lithuanian author profiling experiments with the AGE and GENDER dimensions, performed using the deep learning methods (applied on the top of neural word embeddings) that have never been tested for this task on the Lithuanian language. During this research the impact of the dataset size (with 100, 300, 500, 1,000, 2,000, 5,000 instances in each class) was also investigated. Moreover, the achieved results were compared with the traditional machine learning methods with element frequencies as the feature representation type.

The experiments on the grammatically correct texts of the Lithuanian parliamentary transcripts revealed the superiority of the Convolutional Neural Network over the Long Short-Term Memory method with the larger datasets on both profiling dimensions.

The highest accuracies of 0.316 and 0.609 on the AGE and GENDER, respectively, do not exceed the accuracies achieved by the traditional machine learning methods. Summarizing, deep learning methods are not the best choice for our profiling tasks with AGE and GENDER. Despite that in the future research we are planning to continue exploring the deep learning methods (by increasing training set sizes, tuning parameters, selecting different types of word embeddings) for the author profiling.

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". *Quantitative Linguistics*, vol. 12(1), 2005, pp. 65–77.

[2] P. Juola. "Future trends in authorship attribution". *Advances in Digital Forensics III – IFIP International Conference on Digital Forensics*, vol. 242, 2007, pp. 119–132.

[3] H. Gómez-Adorno, G. Sidorov, D. Pinto, D. Vilariño, and A. Gelbukh. "Automatic authorship detection using textual patterns extracted from integrated syntactic graphs". *Sensors*, vol. 16(9), 2016, pp. 1374, https://doi.org/10.3390/s16091374.

[4] V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono, A. E. Nugroho, E. W. Andangsari, and M. N. Suprayogi. "Personality prediction based on Twitter information in Bahasa Indonesia". *Federated Conference on Computer Science and Information Systems, FedCSIS 2017. In the 2nd International Workshop on Language Technologies and Applications (LTA'17)*, 2017, https://doi.org/10.15439/2017F359.

[5] Sh. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker. "Lexical predictors of personality type". *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[6] J. Schler, M. Koppel, Sh. Argamon, and J. W. Pennebaker. "Effects of age and gender on blogging". *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, AAAI, 2006, 199–205.

[7] R. Aljumily. "Hierarchical and non-hierarchical linear and non-linear clustering methods to "Shakespeare authorship question"". *Social Sciences*, MDPI AG, vol. 4(3), 2015, pp. 758–799, https://doi.org/10.3390/socsci4030758.

TABLE I
FEATURE REPRESENTATION TYPES WITH DIFFERENT CLASSIFICATION METHODS AND DATASET SIZES (I.E., A NUMBER OF INSTANCES IN EACH CLASS).

Dataset size	AGE		GENDER	
	NBM	SVM	NBM	SVM
100	lemmorf-2	lex-2	lexmorf-1	chr-7
300	lex-2	lemmorf-2	lempos-1	lempos-3
500	lexpos-1	lempos-1	lem-1	lexmorf-2
1,000	lempos-1	lem-3	lem-1	lem-3
2,000	lemmorf-1	lem-3	lem-1	lem-1
5,000	lem-1	lem-3	lem-1	lem-3

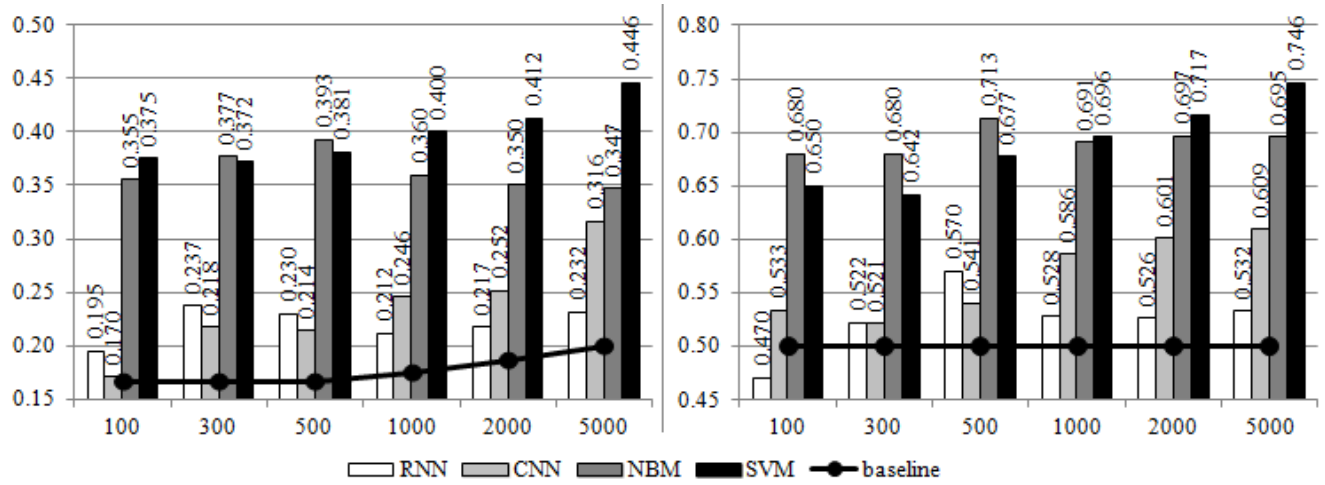

Fig. 1. AGE (left chart) and GENDER (right chart) profiling results: accuracies with different methods and dataset sizes. The *baseline* label denotes the higher value of the random and majority baselines.

[8] Ch. Napoli, E. Tramontana, G. Lo Sciuto, M. Woźniak, R. Damaševičius, and G. Borowik. "Authorship semantical identification using holomorphic Chebyshev projectors". *2015 Asia-Pacific Conference on Computer Aided System Engineering*, IEEE, 2015, https://doi.org/10.1109/APCASE.2015.48.

[9] E. Stamatatos. "A survey of modern authorship attribution methods". *Journal of the Association for Information Science and Technology*, John Wiley & Sons, Inc. vol. 60(3), 2009, pp. 538–556, https://doi.org/10.1002/asi.21001.

[10] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. "Overview of the author profiling task at PAN 2013". *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, 2013.

[11] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans. "Overview of the 2nd author profiling task at PAN 2014". *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, 2014.

[12] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. "Overview of the 3rd author profiling task at PAN 2015". *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, 2015.

[13] P. Rangel, M. Francisco, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, Martin, and B. Stein. "Overview of the 4th author profiling task at PAN 2016: Cross-Genre Evaluations". *Working Notes Papers of the CLEF 2016 Evaluation Labs*, 2016.

[14] P. Rangel, M. Francisco, P. Rosso, M. Potthast, and B. Stein. "Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in Twitter". *Working Notes Papers of the CLEF 2017 Evaluation Labs*, 2017.

[15] J. Kapočiūtė-Dzikienė, A. Venčkauskas, and R. Damaševičius. "Comparison of authorship attribution approaches applied on the Lithuanian language". *Federated Conference on Computer Science and Information Systems, FedCSIS 2017. In the 2nd International Workshop on Language Technologies and Applications (LTA'17)*, 2017, pp. 347–351, https://doi.org/10.15439/2017F110.

[16] A. Venčkauskas, A. Karpavičius, R. Damaševičius, R. Marcinkevičius, and J. Kapočiūtė-Dzikienė. "Open class authorship attribution of Lithuanian Internet comments using one-class classifier". *Federated Conference on Computer Science and Information Systems, FedCSIS 2017. In the 2nd International Workshop on Language Technologies and Applications (LTA'17)*, 2017, pp. 373–382, https://doi.org/10.15439/2017F461.

[17] J. Kapočiūtė-Dzikienė, L. Šarkutė, and A. Utka. "Author profiling of Lithuanian parliamentary speeches: exploring the influence of features and dataset sizes". *Human Language Technologies – The Baltic Perspective: Proceedings of the 6th International Conference Baltic HLT*, IOS press, 2014, pp. 99–106, https://doi.org/10.3233/978-1-61499-442-8-99.

[18] S. Hochreiter and J. Schmidhuber. "Long short-term memory". *Neural Computation*, vol. 9(8), 1997, pp. 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE*, 1998, pp. 2278–2324, https://doi.org/10.1109/5.726791.

[20] Y. Kim. "Convolutional neural networks for sentence classification". *Empirical Methods in Natural Language Processing*, EMNLP, 2014, pp. 1746–1751, https://doi.org/10.3115/v1/D14-1181.

[21] J. Kapočiūtė-Dzikienė and R. Damaševičius. "Intrinsic evaluation of Lithuanian word embeddings using WordNet". *CSOC 2018: 7th computer science on-line conference*, 2018, pp. 394-404, https://doi.org/10.1007/978-3-319-91189-2_39.

[22] M. Sokolova and G. Lapalme. "A systematic analysis of performance measures for classification tasks". *Information Processing and Management*, vol. 45(4), 2009, pp. 427–437, https://doi.org/10.1016/j.ipm.2009.03.002.

[23] Q. McNemar. "Note on the sampling error of the difference between correlated proportions or percentages". *Psychometrika*, vol. 12(2), 1947, pp. 153–157, http://doi.org/10.1007/BF02295996.