Cross-genre Gender Identification in Russian Texts Using Topic Modeling Working Note: Team DUBL

Gabriella Skitalinskaya
Institute of Technology Tallaght
Dublin, Ireland
gabriellasky@icloud.com

Liliya Akhtyamova
Institute of Technology Tallaght
Dublin, Ireland
liliya.akhtyamova@postgrad.ittdublin.ie

John Cardiff
Institute of Technology Tallaght
Dublin, Ireland
john.cardiff@it-tallaght.ie

ABSTRACT

In this paper, we describe the results of gender identification from Team DUBL. We used a topic modeling approach for identifying the author's gender based on his/her written texts. The model was trained on the RusProfiling PAN 2017 Twitter Corpus that contains data in the Russian language. The model has been evaluated on texts of other genres, including texts such as letters to a friend, online reviews, Facebook posts and etc. Our model has obtained competitive results and has been shown to outperform more sophisticated algorithms on gender identification.

CCS CONCEPTS

• **Computing methodologies → Information extraction**;

KEYWORDS

gender identification, natural language processing, topic modeling

1 INTRODUCTION

Author profiling is a broad field which focuses on revealing the author's demographics, psychological characteristics and mental health attributes from his(her) written texts. These attributes include age, gender, social status, personality traits, native language, writing style, etc. The performed analysis could be used later for plagiarism detection, authorship attribution detection, forensics, etc.

A series of shared tasks on digital text forensics called PAN[1] (Plagiarism, Authorship and Social Software Misuse) has been conducted since 2013. However, Slavic languages, in particular, the Russian language is less investigated from the author identifiÑĄation standpoint and has never been presented at PAN. This year to solve the mentioned problem the Rusprofiling[2] shared task has been organized [7]. The focus of the RusProfiling shared task is Cross-genre Gender Identification in Russian texts, meaning investigating the effect of the cross-genre evaluation. The models are trained on one genre, in our case the Twitter corpus, and evaluated on other genres, such as texts describing images, letters to a friend, motivation letters, Facebook posts, online reviews, etc.

The rest of this paper is organized as follows. We discuss relevant literature in Section 2. Section 3 gives details on the training dataset and the description of the proposed approach. Section 4 provides experimental evaluation, and important insights gained during our work. We conclude in Section 5, outlining our contributions and directions for future research.

[1] http://pan.webis.de/clef17/pan17-web/author-profiling.html
[2] http://en.rusprofilinglab.ru/rusprofiling-at-pan/

2 RELATED WORK

One of the standard and quite successful techniques of analyzing texts involves analyzing the writing style of the author using stylometric features. These writing style patterns are used to identify different attributes of an author. For example, in [15] over 1,000 stylometric features were proposed: word- and character-based stylometric features, function words, profanities, punctuation, etc. Many different approaches to performing analysis of such features exist. For example, in the early work [3] the authors investigated the authorship gender and language background cohort attribution from e-mail text documents. They used an SVM classifier to perform analysis on over 800 e-mails. The classifier was fed 222 stylometric-, structural-, and gender-specific features, obtaining F-score about 80%.

In the shared task organized by PAN one of subtasks includes author profiling of a Twitter post corpus. The latest track focused on language variety identification with 4 languages and 19 varieties included, consisting of 11400 tweets in total [12]. The total number of teams, participating in the track was 22. Overall, the best result was obtained using an SVM classifier with tf-idf n-grams, outperforming more sophisticated methods [1].

For the Russian language the research conducted in this area is quite limited. The following papers are worth mentioning. In [10] the authors use statistical methods to calculate the correlation of frequencies of parts-of-speech (POS) bigrams and traits of the author. A simple regression model was used to calculate the accuracy of the model. The training dataset consisted of students' essays written in the Russian language. The obtained results have shown 65%, 79%, and 88% of accuracy for the gender, neuroticism, and openness identification accordingly, proving the usefulness of POS bigrams. In another paper [8] the authors gathered a corpus of essays from 60 respondents on the following topics: letter to a friend and motivation letter to the employee. Then, these texts were analyzed to predict the self-destructive behaviour of the authors. Using a statistical approach based on the presence and frequency of different stylometric features, the authors achieved an accuracy of about 80% on the mentioned dataset.

With the goal of gender identification, in [9] the authors fed different morphological and syntactic features to machine learning algorithms and were able to obtain an F-score of 74% using ReLu. They used the RusPersonality dataset [6], consisting of 1867 texts in different genres (including descriptions of pictures, essays on different topics, letters, etc.). In another paper [13] Sboev et al. achieved even better results on the same dataset using deep learning algorithms with an F-score of 86%.

3 APPROACH

In this section, we give an overview of the proposed approach and describe its main components.

3.1 Dataset

The training dataset is a Twitter corpus and contains tweets from 600 users in the Russian language. For each tweet, there is information about its gender. The task organizers provided five different datasets for testing, each dataset belonging to a particular genre:

1. Offline texts (picture descriptions, letters to a friend, motivation letters to employees) from the RusPersonality Corpus (370 texts) [6].
2. Facebook (228 texts)
3. Twitter (400 texts)
4. Product and service online reviews (776 texts)
5. Gender imitation corpus (women imitating men and the other way around) (94 texts)

3.2 Data Preprocessing

Every text from the dataset went through the following preprocessing procedures:

- removal of stopwords
- removal of short words (less than 2 characters)
- lemmatization.
- removal of links, hashtags and mentions (optional)

3.3 Method Based on Topic Modeling

Topic modeling is a rapidly developing technique capable of revealing hidden topics in text collections. Originating from the text analysis, topic modeling found its implications in many other areas, which include signal, image and video processing and network analysis [2, 11].

Regarding text analysis, practical implications of topic modeling include information retrieval, summarization, segmentation and classification of texts, as well as regression analysis. However, most topic modeling based approaches capable of solving the mentioned problems are too difficult for practitioners to understand and apply. These approaches are based on Bayesian learning and require sophisticated tuning and good theoretical knowledge of Bayesian algorithms[14]. As a consequence only basic models are in common practice, for example, such algorithms as *Probabilistic Latent Semantic Analysis* (PLSA) [4] and *Latent Dirichlet Allocation* (LDA) [5], which are ineffective in many cases.

In this sense, *Additive Regularization of Topic Models* (ARTM) proposed in [16] is free of redundant probabilistic assumptions and provides a simple inference for many combined and multi-objective topic models. This method is based on classical, non-Bayesian regularization, using a semi-probabilistic approach. Moreover, besides the simplicity of the approach, another advantage is the ability to take into account different data or "modalities" accompanying texts to build a model, which could be images, audio and video attachments, user log data, different metadata (for example, user's age, gender), etc.

In ARTM the construction of the topic model is based on an iterative two-step expectation-maximization (EM) algorithm, where at the first step (E-step) the expected value of the likelihood function is calculated, followed by the maximum likelihood estimation (M-step). The steps are repeated until convergence. At this point, the adding of regularizers (or constraints) helps to prevent the likelihood function from the problem of non-uniqueness and instability. The detailed explanation of the ARTM algorithm could be found in [16].

Table 1: Parameters of model for each Run

	Run 2	Run 3	Run 4
Number of topics	50	50	70
Φ sparsity regularizer	-	-0.1	-5*1e5
Θ sparsity regularizer	-	-0.15	-
Φ decorrelator	-	-	10*1e5
Number of iterations	20	20	10

All experiments were carried out in Python using the open-sourced realization of an ARTM algorithm – BigARTM tool[3].

3.4 Other Approaches

In our initial experiments, we have tried a number of different methods to solve the considered author profiling problem. Among them were the bag-of-words models and models that were based on Russian grammar rules. In the Russian language, the gender influences the formation of past tense of verbs, which allows identification of the genders of the subject of the verb. We have tried looking for sentences containing verbs in the first person singular past tense and analyzing them.

Moreover, we have tried various classifiers: Random Forest, Linear Regression, Naive Bayes, SVMs, topic modeling based on Non-negative Matrix Factorization, Latent Dirichlet Allocation, Latent Semantic Analysis etc. However, our experiments revealed that these solutions demonstrate the same or worse performance compared to the one proposed in this work, therefore the topic modeling based on ARTM was chosen for submission and final evaluation.

4 EXPERIMENTS AND RESULTS

We have submitted 4 runs experimenting with different model settings, including different preprocessing of texts, different numbers of topics, regularizers. A detailed description of the differences between each run is described in Table 1[4]. Additionally there are slight differences in the preprocessing of texts for Runs 2,3 and Run 4. The preprocessing for Run 4 include removing stop words, short words(<3 characters), hashtags, links and mentions, whereas for Runs 2,3 only conjunctions, special characters, numbers, short words(<3 characters) and stop words have been removed.

The results obtained by our runs for each dataset as well as the best result in the track are presented in Table 2. It can be seen that the addition of regularizers in Runs 3, 4 allows to increase the performance. Additionally, the increase in the number of topics leads to better performance for some test datasets. Our Run 4 with the accuracy of 63% placed third on the Test 1 dataset. This could be due to the fact, that since the dataset consisted of essays, such as

[3] http://bigartm.org

[4] The results of Run 1 are not presented as the experiments setup is identical to the setup of Run 2.

Table 2: Accuracy of the results obtained for different runs on the test data

	Test 1 (Offline Texts)		Test 2 (Facebook)		Test 3 (Twitter)		Test 4 (Online Reviews)		Test 5 (Gender Imitation)	
Run	Rank	Accuracy	Rank	Accuracy	Rank	Accuracy	Rank	Accuracy	Rank	Accuracy
Best Result	1	0,7838	1	0,9342	1	0,6825	1	0,618	1	0,659
Run 2	9	0,5486	9	0,7543	9	0,6125	18	0,475	9-16	0,5
Run 3	10	0,5486	**8**	**0,7587**	**5**	**0,63**	**17**	**0,4793**	9-16	0,5
Run 4	**3**	**0,6297**	10	0,75	7	0,6275	19	0,463	9-16	0,5

letters to a friend, motivation letters, descriptions of pictures and etc, the average length of texts was longer and more topics were covered. Thus, using a higher number of topics leads to capturing topics of higher granularity, resulting in higher accuracy. It should be mentioned, that the results obtained for the Test 3 dataset with the accuracy of 63% are not far off from the best result achieved in the task.

Overall, it can be seen that the dataset containing online reviews was the hardest for the gender identification task. The reason for this may be the difference in the nature of the train and test datasets. The Test 3 dataset with online reviews contains specific corpora that may not be covered adequately in the training dataset (Twitter). The highest accuracy of 76% has been achieved on the Facebook dataset (Test 2), which is significantly higher than the accuracy of 63% obtained for the Twitter dataset(Test 3). In a way, this is surprising since the nature of the Test 3 dataset is the same as of the training dataset. Such results may be explained by the Facebook posts being longer and richer in information, in addition to containing fewer misspellings, syntactic errors, abbreviations and etc.

The results obtained for the Gender imitation corpus (Test 5) are not very high. In our opinion, this could be mostly due to the chosen approach. In the case of topic modelling, it would have been more appropriate and interesting to train on a dataset, such as a Gender imitation corpus, and not only build a classifier to predict people imitating the opposite gender, but also learn which topics are more frequently discussed by such people.

5 CONCLUSION AND FUTURE WORK

In this paper, we reported the approach of the DUBL solution submitted to the RusProfiling Shared Task. All four runs performed competitively with one of our runs achieving high results in identifying the author's gender based on offline texts.

In the future, we plan to improve our model based on topic modeling by augmenting it with more features (stylometric, morphological and syntactic). Moreover, we are planning to make more intricate preprocessing, i.e. adding word and character bigrams to our model, taking into account counts of hashtags, links and mentions found in texts. We believe that with more parameter tuning, we can achieve better results than presented in this paper and be able to advance the state-of-the-art in the task of author profiling in general.

REFERENCES

[1] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland* (Sept. 2017). http://ceur-ws.org/Vol-1866/

[2] Jen-Tzung Chien. 2015. *Topic Modeling for Speech and Language Processing*. Springer Japan, Tokyo, 87–111. https://doi.org/10.1007/978-4-431-55339-7_4

[3] Olivier De Vel, Malcolm Corney, Alison Anderson, and George Mohay. 2002. Digital Forensic Research Conference Language and Gender Author Cohort Analysis of E-mail for Computer Forensics Language and Gender Author Cohort Analysis of E-mail for Computer Forensics. *The Digital Forensic Research Conference* (2002). http:/dfrws.org

[4] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), 50–57.

[5] Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25, 2-3 (1998), 259–284.

[6] Tatiana Litvinova, Olga Litvinlova, Olga Zagorovskaya, Pavel Seredin, Aleksandr Sboev, and Olga Romanchenko. 2016. "Ruspersonality": A Russian corpus for authorship profiling and deception detection. *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, 1–7. https://doi.org/10.1109/FRUCT.2016.7584767

[7] Tatiana Litvinova, Francisco Rangel, Paolo Rosso, Pavel Seredin, and Olga Litvinova. 2017. Overview of the RUSProfiling PAN at FIRE Track on Cross-genre Gender Identification in Russian. *Notebook Papers of FIRE 2017, FIRE-2017* (2017). Bangalore, India, December 8-10, CEUR Workshop Proceedings. CEUR-WS.org.

[8] Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. 2016. Profiling a set of personality traits of text author: what our words reveal about us. *Research in Language* 14, 4 (1 2016). https://doi.org/10.1515/rela-2016-0019

[9] Tatiana Litvinova, Pavel Seredin, Olga Litvinova, Olga Zagorovskaya, Alexandr Sboev, Dmitry Gudovskih, Ivan Moloshnikov, and Roman Rybka. 2016. Gender Prediction for Authors of Russian Texts Using Regression And Classification Techniques. *Proceedings of the Third International Workshop on Concept Discovery in Unstructured Data (CDUD 2016)* (2016), 44–54. http://ceur-ws.org/Vol-1625/paper5.pdf

[10] T A Litvinova, P V Seredin, and O A Litvinova. 2015. Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study. *Indian Journal of Science and Technology ISSN* 8, S9 (2015), 93–97. https://doi.org/10.17485/ijst/2015/v8iS9/51103

[11] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* 5, 1 (2016), 1608.

[12] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. *Working Notes Papers of the CLEF* (2017). http://pan.webis.de

[13] Aleksandr Sboev, Tatiana Litvinova, Irina Voronina, Dmitry Gudovskikh, and Roman Rybka. 2016. Deep Learning Network Models to Categorize Texts According to Author's Gender and to Identify Text Sentiment. *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, 1101–1106. https://doi.org/10.1109/CSCI.2016.0210

[14] Evgeny Sokolov and Lev Bogolubsky. 2015. Topic Models Regularization and Initialization for Regression Problems. *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications* (2015), 21–27. https://doi.org/10.1145/2809936.2809940

[15] Fiona J. Tweedie and R. Harald Baayen. 1998. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32, 5 (1998), 323–352. https://doi.org/10.1023/A:1001749303137

[16] Konstantin Vorontsov and Anna Potapenko. 2014. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *International Conference on Analysis of Images, Social Networks and Texts_x000D_* (2014), 29–46.