Author Gender Prediction in Russian Social Media Texts

Tatiana Litvinova[1[0000-0002-6019-3700]], Dmitry Gudovskikh[1], Alexandr Sboev[1], Pavel Seredin[1[0000-0002-6724-0063]], Olga Litvinova[1[0000-0001-7700-3275]], Dina Pisarevskaya[2], Paolo Rosso[3]

[1] The Kurchatov Institute, Russia
centr_rus_yaz@mail.ru
[2] Independent researcher
dinabpr@gmail.com
[3] PRHLT Research Center, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. Presently natural language processing for social media, in particular in the field of sentiment analysis and topic modeling, is gaining momentum for Russian texts. However, Slavic languages including Russian are still insufficiently explored in terms of computational sociolinguistics and authorship profiling (i.e. automatic identification of latent demographic features of online users such as gender, age, personality based on their texts). Being able to predict these features with a high degree of accuracy would certainly benefit marketing, psychological studies and security. In this paper we are attempting to build classifiers to predict gender of the author in Russian Twitter and Facebook texts and explore the effect of the cross-genre evaluation. We used the most common lemmas, a set of morphological and syntactic parameters as well as the part-of-speech (POS) trigrams as features and multiple classifiers to train and test models. Twitter corpus was used for training, Facebook and test set of Twitter corpus were used for testing. The best models for Twitter were ExtraTreesClassifier and RandomForestClassifier with accuracy 0.72 and linearSVM for Facebook (0.71). The obtained results are comparable with state-of-the art results for Russian language for the texts of different genres.

Keywords: Computational sociolinguistics, Russian social media texts, Gender attribution, POS n-grams, Text categorization, Twitter, Facebook.

1 Problem statement

The rapid growth of social media in recent years, exemplified by Facebook and Twitter, has led to a massive volume of user-generated informal text. This in turn has sparked a great deal of research interest in aspects of social media, including automatically identifying latent demographic features of online users. Many latent features have been explored, but gender and age have generated great interest. Accurate prediction of these features would be useful for marketing and personalization concerns, as well as for security. The majority of recent work in this area has focused on Twitter users [2; 3; 10]. Gender inference accuracy has been reported between 80% and 85%.

Gender prediction is one of the tasks which PAN (a series of scientific events and shared tasks on digital text forensics) participants solve[1] [9; 10]. However advances in latent attribute inference on social media have been largely confined to English content. Gender profiling for social media texts has not been sufficiently researched using Slavic languages except a recent study of the Slovenian language [15]. As for the Russian language, there has been a lot of progress made in social media sentiment analysis [8; 14], topic modeling [4], but to the best of our knowledge, there have been no attempts made to extract social media text author demographics by analyzing their texts with except for work Korshunov et al. [5] who used only token N-grams (N= 1-3) as features and tweets of rather small number of users (450 in total) for training. Besides, they do not describe their corpus in detail.

Russian written texts were investigated in the context of author profiling, including gender prediction. Using RusPersonality corpus which consists of the texts of different genres (e.g. description of a picture, essays on different topics, etc.) labeled with information on their authors (gender, age, results of psychological tests, and so on) authors of [7] obtained models with different F1, 0.74 being the best (ReLU, 1 Hidden Layer with 26 neurons).

The objective of the current paper is to build classifiers for identifying gender of authors of Twitter and Facebook texts in Russian using different algorithms and compare their accuracies. Following PAN 2016 framework [10], we aim at investigating the effect of the cross-genre evaluation, so that models are trained on one genre, which is Twitter here, and evaluated on another genre different from Twitter (Facebook) since it is well-known that "models trained on one genre may not give the same pattern of performance if they are evaluated on a data set, which contains author profiles from a different genre" [1].

2 Materials and methods

2.1 Material

Twitter. Annotating social media texts is what makes designing such corpora particularly challenging. Some researchers automatically built Twitter corpora while others have solved this problem by using labor-intensive methods. For example, Rao et al. [11] use a focused search methodology followed by manual annotation to produce a dataset of 500 English users labeled with gender. The gender tag was ascribed based on the screen name, profile picture, self-description ('bio') and – in the few cases that this was not sufficient – the use of gender markings when referring to themselves. For this research we used the same approach with manual labeling for tweet author gen-

[1] http://pan.webis.de/clef17/pan17-web/author-profiling.html, http://pan.webis.de/clef16/pan16-web/author-profiling.html

der. For those cases where the gender information was not clear, we discarded the user. API-query was made for 200 posts, i.e. over 1000 words from each user. Retweets were removed[2].

Facebook. 228 users of different age groups (20+, 30+, 40+) from different cities and occupations were randomly chosen (to get minimum mutual friendships) with no less than 1000 words per user.

Corpus statistics is presented in Table 1.

The general principles of processing the text corpora are as follows:

1. Non-Russian texts were removed;
2. Citations were removed;
3. Accounts of public people were not used as they might have someone else writing for them;
4. http references were removed.
5. All hashtags (marked with #) were replaced with the «hashtag» tag;
6. Named entities (marked with @) are replaced with the NER tag.

Table 1. Dataset

Source	Male users	Female users
Twitter	543	543
Facebook	114	114

Given that the most frequent terms tend to select the most discriminative features when applied to stylistic studies [12], first of all we identified the most frequent lemmas for each class. As was shown in [7], morphological and syntactic features are important for gender prediction. Part-of-speech (POS) N-grams were shown to be especially effective since they can efficiently encode syntactical information. Below we discuss our feature sets in more detail.

1) top 250 lemmas. We used 250 most frequent lemmas in the corpus. Note that we chose to employ lemmas as Russian is a morphologically rich language where gender is expressed explicitly in a range of grammatical structures. However, grammatical gender markings are easily falsified. As we have in mind as a general goal to build a system for gender prediction efficient, even in the case of gender imitation (valuable for security reasons), we made a decision to avoid using the token-based approach. For lemmatization we use Mystem [13], a freely available stemmer for Russian;

2) we used a set of morphological (the frequencies of POS), syntactical features (frequencies of different types of syntactic relationships between heads and dependents[3]) and psycholinguistic markers (derivative coefficients which reflect different ratios of POS), 56 in total;

---

[2] Twitter corpus is available at http://en.rusprofilinglab.ru/rusprofiling-at-pan/korpus/ Last visited 04/07/2017

[3] We used SynTagRus corpus tag system, http://www.ruscorpora.ru/instruction-syntax.html, https://github.com/UniversalDependencies/UD_Russian-SynTagRus

3) we have chosen top 15 frequent POS trigrams which averaged values are different in males and female texts. At the first stage we chose POS trigrams which occurred in 75 % of the documents of the class and then calculated the difference between the average values of the frequencies of POS trigrams in texts by males and females and as a result, 15 POS trigrams with the largest difference in the average values were selected.

The values of all the properties were normalized either by the number of words in the documents or that of POS trigrams.

The majority of prior work in gender inference (and latent inference in general) has used support vector machine (SVM). We explored multiple classifiers from the scikit-learn Python library to train and test our models. The training was conducted using cross-validation for 30 folds by means of StratifiedKFold strategies. The Twitter corpus was divided into the training (90 %) and test (10 %) sets. Twitter was used for training; Twitter test set and the whole Facebook corpus were used for testing.

3 Results and discussion

The classification results are presented in Table 2. The results listed in the tables are average values of each "training-testing" cycle.

Table 2. Classification results (w – women, m – men)

Twitter test

	Precision		Recall		F1-score		Accuracy
	w	m	w	m	w	m	
SVM (linear)	*0.445*	*0.761*	*0.654*	*0.581*	*0.52*	*0.656*	*0.603*
SVM (rbf)	*0.442*	*0.768*	*0.67*	*0.579*	*0.523*	*0.658*	***0.605***
SVM (poly)	*0.319*	*0.867*	*0.705*	*0.563*	*0.429*	*0.681*	*0.593*
ExtraTreesClassifier (n_estimators = 50)	0.762	0.666	0.698	0.744	0.726	0.699	0.714
ExtraTreesClassifier (n_estimators = 100)	0.757	0.689	0.711	0.743	**0.73**	**0.71**	**0.722**
AdaBoosting	0.656	0.689	0.683	0.671	0.666	0.676	0.673
RandomForestClassifier(n_estimators=10)	0.733	0.632	0.672	0.705	0.697	0.661	0.682
RandomForestClassifier(n_estimators=100)	0.74	0.707	0.72	0.739	**0.726**	**0.717**	**0.724**
DecisionTreeClassifier(criterion='entropy')	0.661	0.713	0.707	0.68	0.678	0.692	0.687

Facebook

	Precision		Recall		F1-score		Accuracy
	w	m	w	m	w	m	
SVM (linear)	*0.599*	*0.8*	*0.716*	*0.705*	0.652	0.749	***0.709***
SVM (rbf)	*0.523*	*0.75*	*0.637*	*0.652*	0.574	0.697	*0.646*
SVM (poly)	*0.569*	*0.719*	*0.631*	*0.666*	0.598	0.591	*0.651*
ExtraTreesClassifier (n_estimators = 50)	*0.889*	*0.457*	*0.579*	*0.832*	0.701	0.588	*0.654*
ExtraTreesClassifier (n_estimators = 100)	0.917	0.484	0.599	0.875	0.724	0.622	0.681
AdaBoosting	0.676	0.635	0.61	0.702	0.639	0.666	0.654
RandomForestClassifier(n_estimators=10)	0.823	0.434	0.551	0.744	0.659	0.545	0.611
RandomForestClassifier(n_estimators=100)	0.899	0.513	0.608	0.86	0.725	0.642	0.689
DecisionTreeClassifier (criterion='entropy')	0.637	0.495	0.514	0.621	0.568	0.549	0.56

ExtraTreesClassifier (estimators=100) and RandomForest (estimators=100) performed the best for Twitter and outperformed baseline (50 %). They performed slightly worse on Facebook texts but still beating the baseline (0.68 and 0.69 respectively). It is interesting to note that linearSVM performed best on Facebook data than on Twitter (0.7 and 0.6 respectively). Positive effect of cross-genre was found in PAN 2016 evaluations where models trained on Twitter performed better on blogs: "Learning with Twitter where people share their comments without censorship, in a sponta-

neous way, and where researchers can obtain a high number of texts per author, could be a good manner to improve the performance of author profiling tasks in other genres (such as blogs) for which it is more difficult to obtain sufficient training data" [10].

The analysis of the frequencies of linguistic parameters enabled us to make some interesting observations (cf. [5; 8]) (t-test was used, $p < 0.05$). More frequent character flooding (greaaattt ("отлиичноо"), oomph ("уффффф"), hurray ("уррааааа"), theyyy "онииии"), diminutives (referring to men and food), named entity mentions (marked with @) are typical for female Twitter texts. As for Facebook, woman tend to use more specific geographic and proper names, as well as mention clothes, family and religious more often than males. In male Facebook texts there are more military, drinking, computer, food, car vocabulary.

For Twitter and Facebook texts the following are common: females tend to use the conjunctions "and" («и»/«а»), negations "not" («не»), significantly more pronouns "I" («я»), "my" («мой»), "own" («свой») as well as relative pronouns "the whole"(«весь»), "this" («этот»), "such" («такой»), preposition "near" («у»); slightly more – "we" («мы»), "you" («вы»), "she" («она»), adverb "very" («очень»). Males prefer the prepositions "in" («в»), "on" («на»), "around" («по»), "about" («о»), "from" («из»), "for" («для»), conjunction "but" («но»), pronoun "they" («они»). The fact that pronouns are more common in female texts and prepositions in those by males is in agreement with the observations made for other languages (for different text genres), as well as for Russian texts of different genres (see [6] for details). Note that these differences are more distinct in Facebook texts.

4 Conclusions and Future Work

The results for prediction of Twitter user gender using most common lemmas and morphosyntactical parameters as features are comparable to those for Russian offline texts from the corpus RusPersonality [7]. Although we trained on Twitter, our models showed similar accuracies on Facebook texts which means that our feature set is useful for gender identification in different genres of social media.

Note however that we consciously excluded the gender-marked parameters, which made the task more challenging. In future we are planning to expand our feature sets with special attention on content-independent parameters including readability measures, etc. which are more useful in gender detection designed for real-world application where users can imitate their writing pretending to be a people of opposite sex (for example, a pedophile can imitate the writing style of a young girl, to groom a child, etc.)

Acknowledgment. This research is financially supported by the Russian Science Foundation, project No 16-18-10050 "Identifying the Gender and Age of Online Chatters Using Formal Parameters of their Texts".

The work of the last author was in the framework of the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

References

1. Ashraf S., Iqbal H. R., Muhammad R., Nawab A. Cross-Genre Author Profile Prediction Using Stylometry-Based Approach. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR-WS.org. Évora, Portugal (2016).

2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309. Edinburgh, United Kingdom. ACM (2011).

3. Ciot, M., Sonderegger, M., Ruths, D. Gender inference of Twitter users in non-English contexts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1136–1145 (2013).

4. Koltsova, O., Koltcov S., Nikolenko S. Communities of co-commenting in the Russian LiveJournal and their topical coherence. Internet Research 26 (3) (2016).

5. Korshunov, A., Beloborodov, I., Gomzin, A., Chuprina, K., Astrakhantsev, N., Nedumov, J., Turdakov, D. Definition of demographic attributes of users of microblogging, In: Proceedings of the Institute of System Programming, Russian Academy of Sciences 25 179-194 (2013).

6. Litvinova, T., Seredin, P., Litvinova, O. Zagorovskaya, O. Gender Identification in Russian Written Texts. XLinguae 3, 176–183 (2017).

7. Litvinova, T., Seredin, P., Litvinova, O., Sboev, A., Zagorovskaya, O., Gudovskikh, D., Moloshnikov, I., Rybka, R. Gender Prediction for Authors of Russian Texts Using Regression and Classification Techniques. Proc. of The Third International Workshop on Concept Discovery in Unstructured Data (CDUD 2016): CEUR Workshop Proceedings. Vol-1625. Moscow, Russia, 44-53 (2016).

8. Loukachevitch, N., Blinov, P., Kotelnikov, E., Rubtsova, Ju., Ivanov, V., Tutubalina, H. Sentirueval: Testing Object-Oriented Sentiment Analysis Systems in Russian. In: Proceedings of International Conference Dialog, pp. 3-13. Rossiiskii Gosudarstvennyi Gumanitarnyi Universitet, Moscow, Russia (2015).

9. Rangel, F., Celli, F., Rosso P., Potthast, M. Stein, B. Daelemans, W. Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Labs and Workshops, Notebook Papers, CEUR-WS.org. Toulouse, France (2015).

10. Rangel, F., Rosso P., Verhoeven, B., Daelemans, W., Potthast, M. Stein B. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR-WS.org. Évora, Portugal (2016).

11. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M. Classifying latent user attributes in twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, pp. 37–44. ACM (2010).

12. Savoy, J. Comparative Evaluation of Term Selection Functions for Authorship Attribution. Digital Scholarship in the Humanities 30(2), 246-261 (2015).

13. Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA. (2003)

14. Vasilyev, V. G., Denisenko, A. A., Solovyev, D. A. Aspect Extraction and Twitter Sentiment Classification by Fragment Rules. In Proceedings of International Conference Dialog. Rossiiskii Gosudarstvennyi Gumanitarnyi Universitet, Moscow, Russia, pp. 100–110 (2015).

15. Verhoeven, B., Krjanec, I., Pollak, S. Gender Profiling for Slovene Twitter Communication: The Influence of Gender Marking, Content and Style. Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pp. 119-125. Valencia, Spain. ACM (2017).