NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Anastasiya B. Khazova

AUTOMATIC DETECTION OF GENDER IDENTITY: THE PHENOMENON OF RUSSIAN WOMEN'S PROSE

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 71/LNG/2018

This Working Paper is an output of a research project implemented at the National Research University Higher School of Economics (HSE). Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

Anastasiya B. Khazova[1]

AUTOMATIC DETECTION OF GENDER IDENTITY: THE PHENOMENON OF RUSSIAN WOMEN'S PROSE

The article deals with the method of automatic detection of authors ' gender identity on the material of fiction prose of 1980-2000. During this period, there is a special construct, called "women's prose", which is characterized by a special genre and stylistic originality. We set ourselves the task to find out whether the concept of “women's prose” refers only to the non-text reality or is clearly reflected at the level of language.

We have collected corpus of texts 1980-2000 and conducted that identified the most effective machine learning algorithms for the classification of male and female prose.

JEL Classification: Z.

Keywords: women’s prose, fiction, classification, gender, identity.

[1] National Research University Higher School of Economics. Faculty of Humanities, School of Linguistics. E-mail: ahazova@hse.ru

Introduction

A significant part of the modern literary space are works written by female authors. Researchers consider the female emancipation of the end of the XX century an important prerequisite for the active inclusion of women in the literary process. But in reality, the first texts appeared at the end of the XVIII century (eg. A. I. Balasheva-Volyntseva, Catherine II, E. E. Golitsyna, E. R. Dashkova, and others).

The concept of "women's literature" began to form in 30-40 years of the XIX century. But for a long time, these authors were not taken seriously. Changes began to occur in the XX century, which contributed to the publication of collections that have become a kind of Manifesto: "New Amazon", "Champagne Spray"," Not remembering evil."

But it is worth noting that the terms “women's literature” and “women's prose” are not synonymous. If until the 1980s, researchers called women's prose all fiction texts written by women writers, in the early 80's there is the concept of women's prose as a separate construct applied to texts created in 1980-2000. It has its own thematic and stylistic features, as well as outlines a very specific circle of authors: T. Tolstaya, L. Ulitskaya, V. Tokareva, L. Petrushevskaya, G. Shcherbakova (Vorobyova 2006).

The first domestic works about the phenomenon of women's prose appeared under the influence of social changes and the arrival of gender studies in Russia in 1980-1990. Pioneers in this area were N. Gabrielian, I. Trofimova and I. Zherebkina, who described the place of this phenomenon in the context of contemporary culture and society (Titareva 2015).

Later, the researchers began to consider various aspects of this phenomenon. Some pay attention to features of works of individual authors (Y. Vysocina “Intertextuality in the prose by Tatyana Tolstaya: on the material of the novel “Kys””, E. Mercaton “the Poetics of the one-act drama by L. Petrushevskaya”, E. Streltsova “Language features of fiction texts of V. Tokareva, L. Petrushevskaya, L. Ulitskaya”). Others tried to establish the General features of the socio-cultural phenomenon (N. Vorobyova "Women's prose of 1980-2000: dynamics, problems, poetics", G. Pushkar "On the category of gender and gender studies. Artistic opposition femininity / masculinity in modern women's prose. The artistic specificity of the conflict and of the chronotope in women's fiction. Levels of gender artistic conflict.”).

Many authors also pay attention to the factors of emergence and formation of women's prose, and also try to define signs on the basis of which it would be possible to call it as a class of fiction. Part of the researchers focuses on the gender of the author. The other part considers the basis for the allocation of this class genre-style characteristics of the text.

Task

One of the key characteristics for attributing the text to women's prose is the peculiarities of the author's choice of vocabulary, morphological and syntactic constructions.

In research works we can find the statement about existence of male and female variants of language. A. Piperski notes that although there is a difference between the used constructions, but it is "not absolute differences of the sexes, but differences of styles: the female language is usually neutral, and the male language is more rough" (Piperski 2006)

. In addition, we should not exclude the impact of the author's gender identity, which is not synonymous with the concept of biological sex.

Differences in morphology and syntax help to solve the problem of automatic identification of the author of any language product. The features of female speech are hyperbolized expressiveness, frequent use of interjections, positive assessments, colloquial speech, diminutives (Streltsova 2014). In turn, men are more likely to use negative evaluation and reduced vocabulary (Zemskaya, Kitaygorodskaya, Rozanova 1993).

In their study, A. Piperski notes that "male texts typical of nouns, adjectives and prepositions, and female – pronouns, verbs, adverbs and interjections" (Piperski 2006). These deep features (Gomon 1990) provide ample opportunities for the use of the identified features of writing in the automatic analysis of texts of different directions.

Literary studies are often subjective. In his "Why Literary Time Is Measured in Minutes" study, Underwood T. writes that "Models provide a perspective on the world designed to address a specific question; like textual interpretations, they can acknowledge assumptions, use subjective evidence, and simplify some patterns in order to bring others forward" (Underwood 2018). The use of machine learning is an actual method for the application to this task. Firstly, it avoids bias in the evaluation of data. Secondly, it makes it possible to use quantitative indicators, which become decisive in the process of identifying the authorship of the text.

There are many experiments on the classification of various texts, including those related to fiction (Diederich, Kindermann, Leopold, Paass 2003), methods of machine learning (Argamon 2009). Many researchers use gender and age as classes, focusing on texts from blogs and social networks. The leading experts in this field are the following: M. Koppel, S. Argam, E. Stamatatos, M. Liwicki, and various methods of determining the gender of a number of algorithms, such as the balanced winnow algorithm, naive bayes, decision trees, etc.

The variant of the exponential gradient algorithm, which in turn is a generalization of the Balanced algorithm, is used by M. Koppel and S. Argamon in their works (Koppel, Argamon, Shimoni 2002).

Also popular in the problem of automatic profiling are using SVM and K-nearest neighbor algorithms. O. De Vel (Vel, Corney, Anderson, Mohay 2002) and T. Kucukyilmaz (Kucukyilmaz, Cambazoglu, Can, Aykanat 2008) gave a description of the applications for text classification.

Despite the fact that the semantic features of the text can potentially be a marker of the author's style, they are difficult to measure, so there is not much research in this area. The most significant is the article of S. Argamon (Argamon, Dawhle, Koppel, Pennebaker 2005), in which the author uses the semantic features of the text, namely, each set of parts of speech in the form of graphs capable of expanding, deepening or limiting the meaning of the text.

In most cases, researchers use stable characteristics of the text: parts of speech, service words, length of sentences and words, syntactic relations, the using of different characters, because they give good results in the field of attribution of texts, and, in particular, the gender identification of their author.

One of the leading associations in this direction is PAN. There are also many annual competitions to determine the gender of the author. Russian researchers are actively involved in the study of the author's profile, the important work of T. A. Litvinova and Zagorovskaya O.V., the authors of the corpus RusGendAttr, created on the basis of Internet texts.

The joint laboratory of corpus sociolinguistics and author studies (Voronezh) conducts research in the field of profiling of Internet texts and text corpora with marked neuropsychological characteristics of their authors.

Despite the urgency of the problem of author's profiling, the problem of gender classification of texts was not solved by the material of Russian women's prose.

But these studies were not conducted on the material of Russian women's prose. Unfortunately, modern algorithms do not allow one hundred percent probability to determine the sex of the writer. But often enough to identify and 90%, which are provided by the compilation of a statistical model based on a number of formal features.

The task of gender identification by machine learning methods becomes easier due to the variety of texts written by women authors in 1980-2000. The researchers distinguish them in a separate construct, therefore, obtaining these data allows us to trace whether the concept of “female prose” refers only to non-text reality or is clearly reflected at the level of language. In addition, they help to determine whether the differences in female and male writing during this period contribute to a more accurate identification of the author's gender.

Method

Gender, often ignored in earlier linguistic studies, is of particular importance with the development of gender linguistics. The gender category is a product of culture and captures perceptions of the differences between male and female. It is reflected both in folk art and traditions, and in the language (Holmes, Meyerhoff 2008).

1. Speech and communicative behavior of people implies the presence of typical strategies for the selection of certain lexical units and syntactic structures, "the specifics of male and female speaking."

2. The reflection of gender in language. This approach consists in the description of functioning of people of different sex in semantic areas, "nominative system, lexicon, syntax, category of a sort".

The concept of gender identity, i.e. categorization by the author himself, includes a lot of categories such as agender, ternary gender models, and many others. We use a binary gender model based on the biological field of the author. It is dictated by the specifics of the corpus, since the authors are women whose works are included in the corpus, deliberately distinguish themselves in the literary environment and do not try to hide their biological affiliation.

D. Tannen introduced the concept of "genderlect", which includes a set of permanent language features for speech of both sexes. According to the hypothesis of the researcher, the choice of certain speech structures is dictated not only by different communication goals, but also by the specifics of socialization that affects people in the process of their growing up.

There are a large number of theories about socialization and its role in the formation of gender: the theory of socialization (R. Lakoff, E. Ox), the principle of male domination and female subordination (Thorne, Henley, etc.), opposite examples of language innovations in the female gender perspective, extending to the male gender perspective, the hypothesis of "gender subcultures". The researchers criticized her for ignoring the context and hyperbolizing the assimilation of communication tactics compared to other important factors.

In addition to the characteristics of women's prose, to which it is impossible to apply automatic methods (conceptual and genre originality), there are language features that allow you to apply machine learning methods to the problem of classification of texts related to women's prose (Tannen 1996).

The most popular for solving this problem are several programs: Weka - the software for data analysis, package Gensim for Python, used for thematic modeling and library Scikit-Learn, which contains a large number of machine learning algorithms.

To carry out an experiment to identify the best algorithm for classifying fiction texts 1980-2000 on the basis of gender, we have chosen the program Weka, because it, along with a convenient mechanism for pre-processing of the material and the basic algorithms for its processing, presents a flexible set of properties, filters and additions.

Our task is to determine the most effective classification algorithms from among those traditionally adopted to solve the problem of classification of texts that can correctly classify the document as male or female works. Such experiments are actively conducted on the material of English, Italian, Danish and other languages, but this issue has not been studied on the Russian material.

There are several machine learning algorithms for solving the problem of text classification. Consider some of them:

1. Support vector machine or SVM (support vector machines).

This method implements a binary classification (Borisov 2013). It divides the input vectors in two specified groups. For example, it divides reviews into positive and negative. The best data type to use is the set with the maximum difference between the two categories.

When it is implemented, the points in space expressed by vectors are separated by hyperplanes. The task of the algorithm is to find a plane that divides the vectors into two or more groups. In the case when two categories-classification is called binary, in all other cases the term “multi-class”is used.

The algorithm can work with already designated classes, then it will be “learning with the teacher", and without: in this case, we have a clustering problem.

To train the classifier, the SMO (sequential minimal optimization) method is used, which changes the SVM parameters to match the training data.

In our case, we are faced with the problem of binary classification with a teacher, because we have a training sample with pre-marked classes: male and female texts.

2. The Naive Bayesian classifier is a probabilistic classifier. On the basis of a number of variables and their dependencies, the dependence is calculated by Bayes ' theorem, which expresses the degree of confidence in the truth of the judgment

The peculiarity of the algorithm is that all classes are considered independently of each other.

Hence, the method is called "naive", because in most cases the data are somehow still connected with each other.

The Naive Bayes classifier requires learning from labeled data. For all its simplicity, the use of the algorithm allows to achieve high accuracy in tasks related to text analysis.

BayesNet is a graph probabilistic model consisting of a set of variables and their probabilistic Bayes dependencies (Getoor, Taskar 2007).

3. The k-nearest neighbor method is a classification algorithm that evaluates the similarity of the nearest neighbors and, based on this similarity, classifies an object into one of the groups (Jain 2010).

To train the algorithm, a sample is used, in which the ratio of objects to one of the classes is indicated. All objects are in the n-dimensional space, where n is the number of features on which the classification is based.

4. Random forest (Random Forest) is the set of critical trees, which while solving the problem of classification based on voting by selected characteristics. The algorithm successfully shows itself in solving problems of clustering, classification and regression, because it is able to include a large amount of data and analyzed classes (Liaw 2002).

The decision tree is a method with the following structure: there is a root node from which nodes depart. They contain the attributes on which the input data is distributed. From nodes depart sheets, which are classified objects.

To make a decision about the classification of an object, you must go from the root node through all the child nodes on which the decision is based to the sheet.

Corpus

To solve the problem of classification, we have assembled a body consisting of works by the authors of the second half of the XX - beginning of the XXI century: L. Petrushevskaya, T. Tolstaya, V. Tokareva, L. Ulitskaya, G. Shcherbakova, V. Pelevin, D. Bykov, E. Limonov, V. Erofeev. We chose those authors on the basis of differences in style and subject of texts. It's allowed to take into account the possible differences imposed by these characteristics.

The analysis of fiction is always complicated by the individual stylistic features of the author. But at the same time, in writing there are various features at the level of morphology and syntax, allowing to distinguish male from female writing.

The literary process of the XX century was reflected at all levels of the language system. The constant change in the lexical composition of the language, the use of colloquial and obscene vocabulary, the lack of stylistic rigor and much more have formed a new look of works of art and significantly influenced women's writing.

The corpus of texts was collected on the basis of open sources of the Internet. Works of different genres and lengths were selected for the analysis: novels, stories, fairy tales. It was possible to achieve a variety of text material for both groups, which will help to more accurately determine the difference between them.

The collected corpus of women's prose and men's texts were later combined and divided into training and test samples.

The size of the training sample consisted of 89 000 tokens, 12 men's and 12 women's texts. The training sample was expertly labeled. The test sample consisted of 20 male and 22 female texts and 317,000 word uses.

All texts in txt format and utf-8 encoding have been preprocessed-they have been cleared of punctuation marks and extra characters that can cause an error when reading data in Weka. They were then used to create arff files for later processing.

The process of classifying texts

Assembled corpuses of the female and male prose texts were subsequently combined and divided into training and test sets. The training sample was expertly labeled.

When using a collection of 20 (train) and 42 (test) texts, a text classification procedure was performed, taking into account the selected algorithms. We used the 5-fold cross validation method, which divides the body into 5 equal parts, in which 1 part becomes the object of test data, and the remaining 4 - training data. After passing 5 iterations, the results were combined, which allowed to assess the integrity of the formation of the learning model.

For the experiment to identify the most effective algorithm for determining the gender identity of the author, we have identified 5 most effective algorithms for solving this problem: “support vector machine”, “random forest”, “naive Bayes”, “BayesNet”, “k-nearest neighbors”.

To build models, weka used the following classifiers: NaiveBayes, BayesNet, SMO (variation of SVM), Lbk (k-nearest neighbors), LWL RandomForest.

StringToWordVector was used as a filter, which converts strings into a set of numeric attributes. As a stemmer we chose Lovins Stemmer, and for tokenization - WordTokenizer. This combination contributed to better results for models using standard algorithms without additional attributes.

The use of stop words significantly reduced the quality of the model, which confirms the assertion that the service words in the texts of Russian-speaking authors are one of the signs of stylistic originality (Morozov 1915).

TFIDF and n-gram tokenization did not improve the quality of models, so it was decided to exclude their use.

During the application of models trained on the test sample using the above algorithms, the following results were obtained:

Table 1. Results of classification of male and female texts

men 20, women 22, total 42	**SMO**	**Naive Bayes**	**BayesNet**	**KNN**	**LWL** **Random Forest**
Men	15	15	13	15	14
Women	15	15	12	12	14
Total	30	30	25	27	28
Men %	**75**	**75**	65	**75**	70
Women %	**68,18**	**68,18**	54,54	54,54	63,63
Total %	**71,43**	**71,43**	59,52	64,29	66,67

Result

Most accurate to classify literary texts on female and male proved to the SVM algorithm, he was able to correctly classify 68,18% feminine and 75% masculine texts. Similar results showed NaiveBayes.

The worst result in the problem of classification of female texts showed algorithms BayesNet and KNN-methods poorly correlate female prose with the desired group (54.54%).

The average results were shown by RandomForest. They correctly classified 63.63% of female texts and 70% of male texts.

It was found that the most effective classifiers for fiction are such implementations of algorithms as NaiveBayes and SMO. The best result in determining the gender identity of the authors of works of art showed classifiers based on Naive Bayesian classifier and support vector machine (71.43%). Other algorithms showed not such a high result.

It was revealed 7 female (6 of them - L. Petrushevskaya's texts) and 2 male texts which were not classified by any model. Samples of texts of these authors are present in the training sample, all of them are characterized by different volume, but this is not enough for their correct processing. It can be assumed that the obstacle to the correct classification is a weak feminine signal at the lexical and semantic level of these texts.

11 female texts and 8 male texts are accurately determined by standard classification algorithms, and in other cases (4 and 8 texts) the classifiers periodically make a mistake.

Despite the abundance of methods of classification of texts and active research in this area, currently the algorithms have not reached 100% accuracy. In addition to statistical limitations, the artistic style of speech, as well as the author's invariants, make it difficult to determine the gender identity of the authors of works.

Even so, the problem of identifying the authors of women's prose works in a binary model based on the author's gender is satisfactorily solved by using standard machine learning algorithms used to classify non-fiction texts.

The study revealed that the models based on the SMO and NaiveBayes algorithms show the quality of 64-68%. Thus, texts written by women authors between 1980 and 2000, when presented as vectors, can be automatically identified. This becomes possible due to the pronounced semantic and lexical features of the texts written during this period.

Unfortunately, not all texts have a relatively stable stereotype of signals on the lexical and semantic level, therefore, one way to improve models is the consideration of punctuation features, sentiment analysis and syntax.

Thus, we can hypothesize that there is no universal formal code of the construct of female prose at the level of the text. But in many works we can identify signals of different intensity, which allow to carry out the procedure of automatic identification.

Summing up, this quantitative study is one of the stages of understanding Russian women's prose not only as a phenomenon of non-textual reality, but also as a formal construct. The conclusions presented in this paper require testing on a large amount of data, but give an idea of the semantic and lexical features of texts that can be automatically identified and used in text profiling tasks.

References

1. Argamon S. et al. Automatically Profiling the Author of an Anonymous Text. URL: http://u.cs.biu.ac.il/~koppel/papers/AuthorshipProfiling-cacm-final.pdf
2. Getoor L., Taskar B. (ed.). Introduction to statistical relational learning. – Cambridge : MIT press, 2007. – T. 1. c.
3. Jain A. K. Data clustering: 50 years beyond K-means //Pattern recognition letters. – 2010. – T. 31. – №. 8. – C. 651-666.
4. Joachim Diederich J., Kindermann J., Leopold E., Paass G. Authorship Attribution with Support Vector Machines. // Applied Intelligence, V.9, Iss.1, 2003. URL:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.7558&rep=rep1&type=pdf

5. Koppel M., Argamon S., Shimoni A. R. Automatically Categorizing Written Texts by Author Gender. // Literary and Linguistic Computing 17(4), 2002. URL: http://u.cs.biu.ac.il/~koppel/papers/male-female-llc-final.pdf
6. Liaw A. et al. Classification and regression by randomForest //R news. – 2002. – Т. 2. – №. 3. – С. 18-22.
7. Underwood T. Why Literary Time Is Measured in Minutes //ELH. – 2018. – Т. 85. – №. 2. – С. 341-365.
8. Борисов Е. Классификатор на основе машины опорных векторов. SVM/SMO. URL: http://mechanoid.kiev.ua/ml-svm.html
9. Воробьева Н.В. Женская проза 1980-2000 г.: динамика, проблематика, поэтика. Авторефер. канд. дисс. – Пермь, 2006. – С.6.
10. Гомон Т. В. Исследование документов с деформированной внутренней структурой. Дисс. канд. юрид. наук М., 1990. С. 96.
11. Земская Е. А., Китайгородская М. А., Розанова Н. Н. Особенности мужской и женской речи // Русский язык в его функционировании / Под ред. Е. А. Земской и Д. Н. Шмелева. М., 1993. С. 90–136.
12. Морозов Н.А. Лингвистические спектры, как средство для отличения плагиатов от истинных произведений того или другого известного автора и для определения их эпохи. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3
13. Пиперски А. Гендер и язык: есть ли разница между мужской и женской речью? (электронный документ) // Д. Варламова. Проект Theory and Practice. 26 февраля, 2006. URL: https://theoryandpractice.ru/posts/9451-gender-language
14. Родионова Е.С. Методы атрибуции художественных текстов. // Структурная и прикладная лингвистика. Вып. 7: Межвуз. сб. / Под ред. А.С. Герда. - СПб; Изд-во С-Петерб. Унт-та, 2008. с.118-127. URL: http://epir.ru/pragmat!/projects/corneille/files/Metody_atributsii.pdf
15. Стрельцова Е. А. Языковые особенности художественных текстов В. Токаревой, Л. Петрушевской, Л. Улицкой // Вестник Череповецкого государственного университета. 2014. №3 (56). URL: https://cyberleninka.ru/article/n/yazykovye-osobennosti-hudozhestvennyh-tekstov-v-tokarevoy-l-petrushevskoy-l-ulitskoy
16. Таннен Д. Ты меня не понимаешь! Почему мужчины и женщины не понимают друг друга. М.: Вече, Персей, АСТ. URL: https://psychology74.nethouse.ru/static/doc/0000/0000/0019/19727.buk93ifa0n.pdf
17. Титарева Л.Д. Женская проза как феномен современной российской культуры (на примере забайкальского края). Авторефер. канд. дисс. – Чита, 2015. – С.4.

Anastasiya B. Khazova
National Research University Higher School of Economics (Moscow, Russia). Faculty of Humanities, School of Linguistics. Assistant;
E-mail: ahazova@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.