Proceedings of the Fifth Conference of Mathematical Society of Moldova
IMCS-55, September 28 - October 1, 2019, Chisinau, Republic of Moldova

Can Computers Catch the Authors Style?

Victoria Bobicev, Yulia Hlavcheva, Olga Kanishcheva

Abstract

The paper presents the experiments on authorship attribution for scientific articles written in Russian and Ukrainian. The main aim of this research is to explore the topic influence on author classification.

Keywords: Text classification, authorship attribution, topic classification, machine learning methods, character based learning.

1 Introduction

Authorship Identification is a hot topic in many areas and especially in science and education as a part of plagiarism detection effort. The goal of this study is to verify how strong the topic of the text influences its authorship attribution. We used machine learning methods to classify the scientific articles from different domains by their authors. While the domain specific words could affect the classification, we selected sets of texts from the same domain and performed authorship attribution within the domain. However, even within one domain the papers describe different topics that still affect the classification. We are organizing the experiments in such a way that training and test sets contain different papers of the same author in order to avoid topic influence.

©2019 by Victoria Bobicev, Yulia Hlavcheva, Olga Kanishcheva

2 Related work

The methods used in the domain of authorship attribution vary from purely manual meticulous analysis of the different text elements to the absolutely automate statistical methods such as machine learning. While machine learning is mostly a classification task, classification by the text author is always influenced by the text topic. This phenomena was explored in [1] and the results showed that most of the stylometric variables are actually discriminating topics rather than authors.

Our work is the continuation of the author classification experiments presented in [2] on the scientific papers written in Russian and Ukrainian language. Although there has been made an attempt to avoid the topic influence by experimenting with the papers form one domain (Economics), there was still a significant probability that topics of the papers influenced the classification and helped to obtain such high results (f-measure from 0.85 to 0.95 on different datasets). In this paper we reorganize the experiments in order to avoid topic classification as much as possible.

3 Experiments Description

The Dataset. We used the same dataset described in [2]. There were two data sets: (1) the whole one that included 271 Ukrainian and 77 Russian articles written by 32 and 8 authors respectively (we experimented with them apart) and (2) the part consisting of 175 Economics and 65 IT articles written by 18 and 10 authors respectively (experimented apart as before).

The Method. We worked with classification method on the base of PPM compression algorithm. It demonstrated its ability to classify the short forum posts by the author with impressive accuracy of almost 90% [3]. The method uses as the features all sequences of characters of length 5, 4, 3, 2 and 1 character from texts. We tested two variations of this method: (1) absolutely all characters from texts including upper and lower case letters, numbers, spaces, all kind of punctuation marks and other specific characters which appear in scientific publications; (2) only words composed of alphabet letters converted to lowercase and spaces between them.

Text Organization. The main difference from the previous experiments consist in the text organisation. The main problem with the text we worked with was the differences in their sizes: it differs from 150 words to 50 pages and such differences may influence the classification results. This was the cause for the decision to split all texts in fragments of approximately 150 words and classify these fragments by author. However, we suppose that the fragments of the same article were classified not by their author but by their common topic. Thus, in the current set of experiments we does not divide the texts in fragments and work with them as they are dividing them in training and test sets in such way as every author is presented in both sets. We used 5-fold cross-validation and there were around 2 texts from each author in each test set.

Table 1. Results of the author classification experiments using PPM and two variations of the features: all characters (PPM char) and only lowercase letters (PPM letters) on four sets of texts

	Ukrainian	Russian	Economics	IT
PPM(char)	0.601	0.950	0.604	0.800
PPM(letters)	0.585	0.913	0.580	0.782

Table 1 presents the results of the experiments. In comparison with the previous work we can say that high results remained only for Russian texts. The previous were 0.972 and 0.979; slightly higher. For the rest of the sets the results are much worse. For example, previous results for Ukrainian were 0.854-0.865 and for IT 0.943-0.958. In case of Russian texts, the high results are explained by the differences in topics the 8 authors in question wrote about.

4 Conclusion

In this paper we tested the hypothesis that author classification is highly influenced by the text topics. The results of our experiments confirmed this. There is still possibility that the differences of text sizes influenced the results; we plan to explore the possibilities to even out the text sizes. The other plan is to investigate the influence of the number of the authors on the classification accuracy.

References

[1] George K. Mikros and Eleni K. Argiri. *Investigating Topic Influence in Authorship Attribution.* In Benno Stein, Moshe Koppel, and Efstathios Stamatatos, editors, SIGIR 07 Workshop, PAN 2007.

[2] V. Bobicev, Y. Hlavcheva, O. Kanishcheva, V. Lazu. *Authorship Attribution in Scientific Publications*, Corpora 2019, St-Petersburg, Russia.

[3] V. Bobicev, M. Sokolova, El Emam Khaled, Y. Jafer, B. Dewar, E. Jonker, S. Matwin. *Can Anonymous Posters on Medical Forums be Reidentified?* Journal of Medical Internet Research, 2013, Oct 03; 15(10):e215.

Bobicev Victoria[1], Hlavcheva Yulia[2], Kanishcheva Olga[3]

[1]Technical University of Moldova.
E–mail: `victoria.bobicev@ia.utm.md`

[2]National Technical University “Kharkiv Polytechnic Institute”
E–mail: `glavjul@gmail.com`

[3]National Technical University “Kharkiv Polytechnic Institute”
E–mail: `kanichshevaolga@gmail.com`