© Copyright HDSKD
All Rights Reserved

Author Identification Using Different Sizes of Documents: A Summary.

S. Bourib[1], S. Khennouf[2]

[1] Faculty of Electronics and Informatics, University Houari-Boumediene
[2] University of Msila

Abstract_ In the present research work, we deal with the problem of authorship attribution of ancient Arabic text documents, which were written by several ancient philosophers. For that purpose, we conducted several authorship attribution experiments applied with different text sizes. A special dataset, called "A4P" (**A**uthorship **A**ttribution for **A**ncient **A**rabic **P**hilosophers), has been constructed by extracting texts of different sizes from the books of those 5 ancient Arabic philosophers, where the genre and topic are quite similar. The size of the texts varies from 100 words to 3000 words per text. In our approach two types of features are employed; character N-grams and words and several classifiers are used, namely: SMO based SVM, Multi Layer Perceptron, Linear Regression, Stamatatos distance and Manhattan distance. Results show that the minimum required text size (for getting good authorship attribution performances) depends on the used features and classification technique, but in the overall the performances of the proposed techniques are quite interesting.

Keywords*: Authorship attribution, Performances vs size, Pattern recognition, Artificial intelligence.*

1. Introduction and related works

Authorship Attribution (*AA*) is a research field concerned with the automatic classification of text documents with regards to their author(s). It tries to respond to the following question: Who is the author of this document?

© Copyright HDSKD
All Rights Reserved

Many studies have been reported during the last decades as described in [Juola 2006], [Stamatatos 2009] and [Sayoud 2015], where many disputes were reported and several types of features and techniques were proposed too. In most cases the amount of data was big enough to bring significant characteristics to the author.

However, in 2001, de Vel, Anderson, Corney and Mohay [Vel 2001] reported a challenging research work of author identification on small texts such as in emails. The problem in such works is: was the small data provided by an email sufficient enough to make a fair author identification? Even though several works conducted on email documents reported quite good results too, their works were not so convincing since the scientific community agree that an efficient authorship attribution requires a quite big amount of text. In other words: the longer is the text, the higher is the precision of authorship attribution.

Now, what could be the minimum/optimum data size for that purpose? Some recent studies developed by Kim Luyckx and Walter Daelemans in 2011 [Luyckx 2011] tried to investigate the size issue and showed the real effect of author data size in authorship attribution. Afterward, an interesting work was reported by Maciej Eder in 2013 [Eder 2013], giving several responses to the problem by providing some key solutions to the minimum data size required for different cases and different languages, except for the oriental languages such as Arabic for instance (*i.e. those types of languages were not investigated*).
The results of Eder were interesting and useful, but we cannot extend them to all the languages that were not investigated, such as Arabic. Moreover, certain features were not tested by the author, which make his results, even though interesting, not extensible to every feature or classifier either.
That is, by the present investigation, we try to find out the minimum text size required to get a consistent authorship characterization for the Arabic Language. For that purpose, different types of features, distances and classifiers are tested and commented.

2. Dataset

The corpus used in our experiments consists of 45 text documents from 5 different authors, namely: 9 texts per author. These texts were written by five ancient Arabic philosophers (*without translation*) from various regions and date from the 11th, 12th and 13th centuries. Moreover, those texts have somewhat the same genre and topic. The original texts are quite long, but are divided herein into medium and short texts, so that we get several fixed sizes: the shortest text is about only 100 words and the longest one is about 3000 words. Hence, the different text sizes are 100, 500, 1000, 1500, 2000, 2500 and 3000 words per text.

Those different texts (*the 5 philosophers*) are extracted from the Universal Library (*Elwaraq*).

During our experiments, the corpus was arbitrarily divided into two subsets, one subset for the training and another one for the testing. The training corpus, used to train each author model is composed of 10 long texts of 3000 words and the testing corpus is composed of 35 different texts.

© Copyright HDSKD
All Rights Reserved

To our knowledge, most researchers, in the literature, used their own text corpus for the task of author attribution. Although the Greek corpus Stamatatos et al. (1999) [Stamatatos 1999] and Chinese corpus Peng et al. (2003) [Peng 2003] were included in Keselj et al. (2003) [Keselj 2003] for instance, unfortunately they were forced, like us, to assemble their Arabic corpus from classical texts (*eg. Ibn Roched, Elfarabi, etc.*). Finally, for purposes of further comparison, the entire corpus, which we built (*A4P*), has been made freely available on our personal website.

3. Results and discussion

In this paper, we were looking for the minimum text size for a given document able to ensure a good AI task. For that purpose, we have conducted several experiments of Authorship identification applied on multi-size text documents: from 3000 words down to only 100 words per document. For that purpose an Arabic dataset has been constructed and collected from the books of 5 ancient Arabic philosophers.Two types of features were investigated: chacter 5-gram and words. Furthermore, several classifiers were employed: SMO-SVM, MLP, Linear regression, Stamatatos distance and Manhattan distance.

Results have shown that the minimum text size required for performing a fair Authorship Identification, depends on the features and classification method that are employed. But, in general, the size of 2500 words per document seems to be the minimum amount of textual data required for a fair AI, in many cases. Furthermore, concerning the Robustness vs Size-Reduction, an interesting conclusion has been deduced from this survey. For instance, when using character 5-grams the MLP appears to be the most interesting classifier to use in AI; and when using words, Manhattan distance appears to be the most interesting one. In general, the feature character penta-gram seems better than words by showing much better performances than this last one for almost all classifiers. The unique exception is noticed with Manhattan distance, which presents a paradoxical result showing that the word feature is quite better than character n-grams.

References

[Eder 2013] M. Eder, Does size matter? Authorship attribution, small samples, big problem. Literary and Linguistic Computing Advance Access published November 14, 2013, doi:10.1093/llc/fqt066.

[Juola 2006] P. Juola, JGAAP, Authorship Attribution, Foundations and Trendsin Information Retrieval, Vol. 1, No. 3, 2006, pp. 233–334, Now Publisher.

[Keselj 2003] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, N-gram-based author profiles for authorship attribution Authors, Publication date 2003/8, Proceedings of the conference “pacific association for computational linguistics”, Volume 3, pp. 255-264.

[Luyckx 2011] K. Luyckx, and W. Daelemans, The effect of author set size and data size in authorship attribution, Literary and Linguistic Computing, Vol. 26, No. 1, 2011, pp. 35-55.

© Copyright HDSKD
All Rights Reserved

[Peng 2003] F. Peng, X. Huang, D. Schuurmans, and S. Wang, Text classification in Asian languages without word segmentation, Proceedings of the sixth international workshop on Information retrieval with Asian languages, Volume 1, pp. 41-48.

[Sayoud 2015] H. Sayoud, A Visual Analytics based Investigation on the Authorship of the Holy Quran, 6th International Conference on Information Visualization Theory and Applications, March 11-14, 2015, Berlin, pp. 177-181.

[Stamatatos 1999] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, Automatic Authorship Attribution, In Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics, 1999, pp. 158-164.

[Stamatatos 2009] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for Information Science and Technology, 2009, 60(3), pp. 538–56.

[Stamatatos 2013] E. Stamatatos, On the Robustness of Authorship Attribution Based on Character n-gram Features, Journal of Law and Policy, 21(2) , 2013, pp. 421-439.

[Vel 2001] O. de Vel, A. Anderson, M. Corney, and G. Mohay. ACM SIGMOD Record. Volume 30 Issue 4, December 2001, ACM New York, NY, USA, pp. 55 - 64.