Understanding the impact of three derived text formats on authorship classification with Delta

Du, Keli
duk@uni-trier.de
Universität Trier, Deutschland

Introduction

Due to copyright law, Text and Data Mining (TDM) with copyrighted texts faces a lot of restrictions in terms of storage, publication and follow-up use of the resulting corpora, which, however, is against the spirit of open data in digital humanities (DH). As a solution to the problem, the concept of derived text formats (DTFs) has been suggested and discussed (see e.g., Lin et al. 2012, Bhattacharyya et al. 2015, Jett et al. 2020, Schöch et al. 2020). In DTFs, although some information (primarily copyright-relevant features) has been removed from the texts, the texts can still be used for various relevant TDM tasks in DH, such as authorship attribution or topic modeling. Schöch et al. (2020) also provides a very detailed examination of several DTFs from the perspectives of Computational Literary Studies, Computer Science, memory institutions and law. DTFs are extremely meaningful for the DH community, because they match the spirit of open data and make it possible for researchers and libraries to provide more text data for DH research. It also supports the pursuit of open science by encouraging researchers to publish their research data without worrying about violating copyright laws.

However, as far as I know, there is not much research dedicated to the question, how much the loss of information caused by DTFs affects the TDM results. Eder (2013) presented an empirical study of verifying the impact of unwanted noise in texts on authorship attribution and emphasizing that the usefulness of damaged texts should not be underestimated in stylometric studies. He brought noise into texts by (a) randomly replacing a portion of characters, (b) increasing standard deviation of word counts and (c) randomly replacing original words with other words in the same corpus, in order to show the correlation between a dirty corpus and the attribution accuracy. The presented paper did a similar empirical study by transforming texts into token-based DTFs and provide a review on the correlation between information loss caused by these DTFs and the loss of accuracy in authorship classification.

Token-based DTFs

In Schöch et al. (2020), three kinds of token-based DTFs are introduced, to enable the free reusability of text data:

- Simple document-term-matrix: The idea is to transform a corpus into a matrix, which only saves the frequency of each term in each text in the corpus.
- Sequence randomization in segments: The idea is to split a text into segments, randomize the sequence of words in each segment, and reassemble all the segments into a text.
- Selectively reduced information on individual tokens: The idea is to replace a portion of the words (e.g., all the function words) in text with their POS-tags.

Applying the first and the second DTFs to frequency-based authorship attribution does not present any challenge. Take the most well-known method in authorship attribution Burrows's Delta (Burrows, 2022) as an example: Delta test follows the "bag of words" model for representing documents and only requires the frequency of each word in each text to distinguish between authors. The sequence information of words in texts is not necessary. Therefore, the first and the second DTFs keep all the information needed for the Delta test and the transformation does not affect the test results. As a matter of fact, if one only wants to publish a corpus so that the reported classification results of authorship could be verified, all one must do is to publish the document-term-matrix and the metadata table of the corpus.

In comparison, if the texts are transformed into the third DTF, although the frequency information of some words in the text will not match the original situation, the sequence information of words could be kept. This opens the possibility of using the data in this form for other TDM tasks such as sentiment analysis or named entity recognition that require the sequence information of words. If a corpus is published with the expectation that it can be applied to multiple TDM tasks, it makes more sense to prepare the corpus in this format. And of course, it is important to understand how much this format will affect the outcome of different TDM tasks. Therefore, this paper evaluates the usefulness of the third token-based DTF on authorship attribution as a start. In the next sections, the method and the results of the evaluation are reported.

Method

For the evaluation, three corpora representing different languages and text types have been constructed: deu_DraCor (German plays), fra_ELTeC (French novels) and eng_RSC (English journal articles). The relevant information about the corpora is shown in Table 1.

Table 1: Overview of the corpora.

corpus	corpus size (million words)	average text length (words)	no. of texts	no. of authors	period	language	text type
deu_DraCor (Fischer et al., 2019)	5.69	18237	312	55	1650 - 1928	German	play
fra_ELTeC (Odebrecht et al., 2021)	11.33	80370	141	30	1840 - 1912	French	novel
eng_RSC (Kermes et al., 2016)	7.92	6206	1276	69	1665 - 1869	English	journal article

The test is designed as follows: First, for each document in a corpus, a certain percentage of words (0%, 10%, ..., 100%) were randomly selected and replaced by their corresponding POS-tags. Since function words are crucial to authorship attribution, instead of only replacing function words as suggested in Schöch et al. (2020), any kind of word (including punctuation) may be replaced. The next step is the standard procedure for Delta test: creating a document-term-matrix, computing the z score of each value in it and then select the most frequent word types as feature to classify documents. The classification was done by a linear SVM classifier with 5-fold cross-validation. Following the results in Evert et al. (2017), the 2000 most frequent word types in each corpus were taken as feature for all the classification tasks. A further test was also performed: All POS-tags that replaced their corresponding words were removed from the text and the Delta test were then conducted again. By comparing the classification results of these two tests, we can also determine the contribution of POS-tags to the results of the classification. Since the words in texts are randomly replaced or removed which could introduce some random variation into the results, each of the above-described tests is repeated 10 times.

Before presenting the classification results, three text passages are prepared to give an impression of readability of the texts in DTF. The original text, the texts with 10% and 50% of words replaced by their corresponding POS-tags, are listed in Table 2.

Table 2. Text passages in original format and DTF (The percentage value indicates the proportion of words replaced or removed.).

percentage	text
0% (original)	The members of this new group of alkaloids are so numerous, their deportment is so singular, and their derivatives raimify in so many directions, that I have not as yet been able to complete the study of these substances in all their bearings; nor is it my intention to go fully into the chemistry of the subject in the present communication, my object being merely to establish the existence of these bodies, and to give a general outline of their connection with the volatile bases, and of their most plominent chemical and physical properties, reserving a detailed description of their salts and derivatives to a future memoir .
10%	The members of this new group of alkaloids are so numerous, their deportment is so singular, and their derivatives raimify in so many directions, that I have not as RB VBN able to complete the study of these substances in all their bearings ; nor is it my intention to go fully into DT chemistry of DT subject in the present communication, PP$ object being merely TO establish DT existence of these bodies, and to give a general outline of their connection IN the volatile bases, and of their most plominent chemical and physical properties, reserving a detailed description of their salts and derivatives to a JJ memoir.
50%	DT members IN DT JJ NN IN NNS VBP so JJ, PP $ deportment is RB JJ, CC their NNS raimify in RB many NNS, WDT PP VHP RB RB yet been JJ TO VV DT study IN these NNS IN all PP$ NNS : CC is it my NN TO go RB into DT chemistry IN the subject in DT JJ communication, my object NN merely to VV DT NN of DT NNS, and TO VV DT JJ outline IN PP$ connection IN DT volatile bases, CC IN their RBS plominent chemical and physical properties, reserving DT JJ NN IN their NNS and derivatives TO a future NN.

Results

The classification results on the German play collection, the English article collection, and the French novel collection are presented in Figures 1, 2 and 3, respectively. The y-axis is the F1(macro)-score, and the x-axis shows the portion of words that are replaced or removed. The blue boxplots and the yellow boxplots represent the classification results, when the words in texts are replaced with POS-tags or removed, respectively. As the reference value, the classification results for the original data are also shown in the figures. The Welch's t-test is also performed to determine the difference in classification results. The "ns" in the figures means non-significance.

In all the three figures, the same trend can be observed: Step by step, the median of F1-score distributions get worse as the percentage increases. Especially when more than half of the words were replaced or removed, the tendency for the classification results to become worse became particularly obvious. In addition, the variance of the F1-scores always becomes larger, if a certain percentage of words in texts are replaced or removed. According to the Welch's t-test, in all cases, whether the words are replaced or removed does not affect the classification. This observation indicates that the POS-tags do not contribute to the distinction of authorships.

Another interesting observation is, when all words in texts are replaced by POS-tags, the classification results improve, relative to a reduction of 90%, in the case of the German and English data, but not for the French data. To understand this situation, the change in the number of word types in each corpus was checked. As presented in Figure 4, when 90% of the words are replaced or removed, there are still around 20,000 word types in each text collection. But when all the words are replaced, only a few dozen types remain. Their number becomes so small

that it looks like it is reduced to zero in the Figure 4. Since the classification is based on the most frequent 2000 types, although 90% of the words are replaced or removed, the 2000 features used for classification are still mostly from the remaining 10% of words. In the German and English collections, these words bring apparently noise to the classification task. In contrast, the remaining 10% of words in the French corpus are still able to guarantee a relatively good classification result. From the data in Table 1 we can see that number of authors in the French corpus is smaller, which indicates the classification task on the French corpus is easier. More importantly, as presented in previous studies (e.g., Eder 2015, Romanov & Grallert 2022), that pulling random samples of at least 5000 words length out of texts will be sufficient for ensuring reliable authorship attribution. Considering the average length of the French novels is over 80,000 words, when 90% of the words are replaced or removed, the remaining 10% (that is, about 8000 words) is still sufficient to guarantee a good classification result. To clarify this issue, further research would certainly be of great interest.

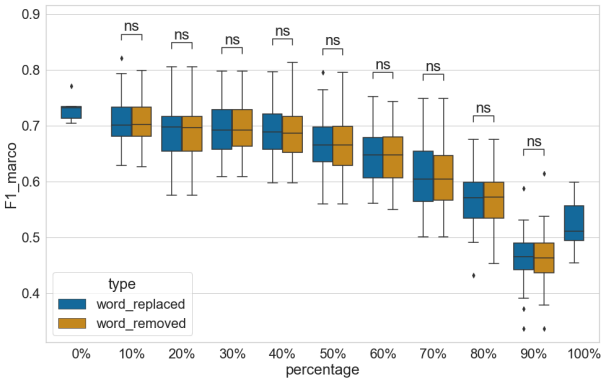

Figure 1. Authorship classification on the German play collection

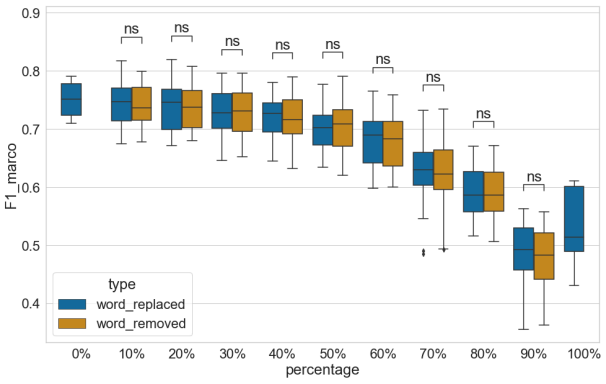

Figure 2. Authorship classification on the English article collection

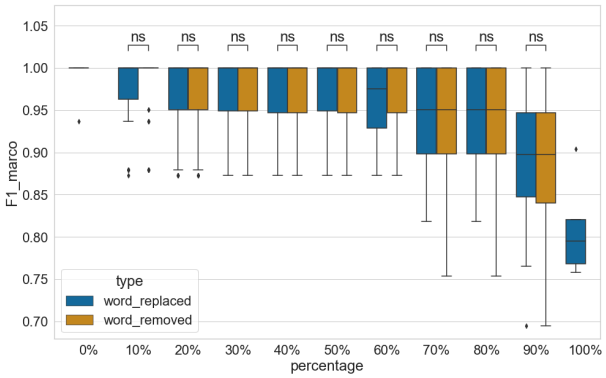

Figure 3. Authorship classification on the French novel collection

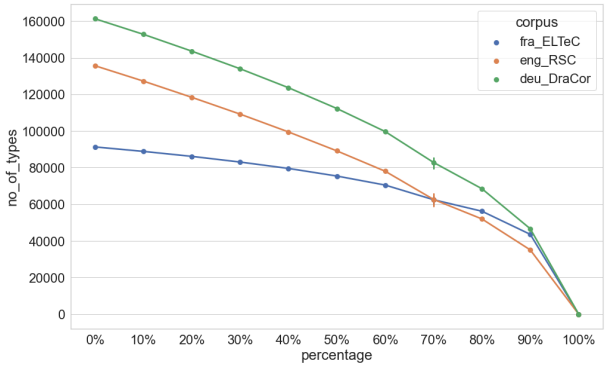

Figure 4. Change of the number of word types in three text collections

Conclusion

This paper provides an exploration of the usefulness of three token-based DTFs for frequency-based authorship classification with Delta. As presented, selectively reducing information on individual tokens could ensure, to a certain extent, that the authorship classification results are not affected too much. The impact of token-based DTFs on the results of Delta test can be reduced by considering only replacing or removing content words, while all function words remain unchanged. But this limits the application of the texts on other TDM tasks such as topic modeling. For the future work, a series of tests are planned on evaluating the usefulness of token-based DTFs on other TDM tasks. The goal is to find DTFs that could balance various factors (e.g. word frequency, sequence information, content vs. function words, copyright) so that texts could be published and used for as many TDM tasks as possible without violating copyright law.

Bibliographie

Bhattacharyya, Sayan, Peter Organisciak, und J. Stephen Downie. 2015. „A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features“. *Interdi-*

sciplinary Science Reviews 40 (1): 61–77. https://doi.org/10.1179/0308018814Z.000000000105 .

Burrows, John. 2002. „'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing* 17 (3): 267–87. https://doi.org/10.1093/llc/17.3.267 .

Eder, Maciej. 2013. „Mind your corpus: systematic errors in authorship attribution". *Literary and Linguistic Computing* 28 (4): 603–14. https://doi.org/10.1093/llc/fqt039 .

Eder, Maciej . 2015. Does size matter? Authorship attribution, small samples, big problem, *Digital Scholarship in the Humanities* , Volume 30, Issue 2, Pages 167–182. https://doi.org/10.1093/llc/fqt066 .

Evert, Stefan, Fotis Jannidis, Thomas Proisl, Steffen Pielström, Thorsten Vitt, Christof Schöch, und Isabella Reger. 2017. „Understanding and Explaining Distance Measures for Authorship Attribution". *Digital Scholarship in the Humanities* . https://academic.oup.com/dsh/article-pdf/32/suppl_2/ii4/21298943/fqx023.pdf .

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C. and Trilcke, P. 2019. Programmable corpora: Introducing DraCor, an infrastructure for the research on European drama. *Digital Humanities 2019* : 5 doi: doi:10.5281/zenodo.4284002 .

Jett, Jacob, Capitanu Boris, Kudeki Deren, Cole Timothy, Hu Yuerong, Organisciak Peter, Underwood Ted, Koehl Eleanor Dickson, Dubnicek Ryan, Downie J. Stephen. 2020. " The HathiTrust Research Center Extracted Features Dataset (2.0)" . HathiTrust Research Center. https://doi.org/10.13012/R2TE-C227 .

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and Teich, E. 2016. The royal society corpus: From uncharted data to corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* . pp. 1928–31.

Lin, Yuri, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, und Slav Petrov. 2012. „Syntactic Annotations for the Google Books NGram Corpus". In *Proceedings of the ACL 2012 System Demonstrations* , 169–74. Jeju Island, Korea: Association for Computational Linguistics. https://aclanthology.org/P12-3029 .

Odebrecht, C., Burnard, L. and Schöch, C. 2021. European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels. Zenodo doi: 10.5281/ZENODO.4662444 . https://zenodo.org/record/4662444 (accessed 9 December 2022).

Romanov, Maxim, Grallert, Till . 2022. 'Establishing Parameters for Stylometric Authorship Attribution of 19th-Century Arabic Books and Periodicals'. *DH2022* , Tokyo, 23 July 2022. https://dh-abstracts.library.virginia.edu/works/11858 .

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen". *Zeitschrift für digitale Geisteswissenschaften (ZfdG)* 5. http://dx.doi.org/10.17175/2020_006 .