A Call for Clarity in Contemporary Authorship Attribution Evaluation

Allen Riddell
Indiana University Bloomington
Bloomington, Indiana, USA
riddella@indiana.edu

Haining Wang
Indiana University Bloomington
Bloomington, Indiana, USA
hw56@indiana.edu

Patrick Juola
EVL Lab, Duquesne University
Pittsburgh, Pennsylvania, USA
juola@mathcs.duq.edu

Abstract

Recent research has documented that results reported in frequently-cited authorship attribution papers are difficult to reproduce. Inaccessible code and data are often proposed as factors which block successful reproductions. Even when original materials are available, problems remain which prevent researchers from comparing the effectiveness of different methods. To solve the remaining problems—the lack of fixed test sets and the use of inappropriately homogeneous corpora—our paper contributes materials for five closed-set authorship identification experiments. The five experiments feature texts from 106 distinct authors. Experiments involve a range of contemporary non-fiction American English prose. These experiments provide the foundation for comparable and reproducible authorship attribution research involving contemporary writing.

1 Introduction

Closed-set authorship attribution picks out the likely author of an unsigned document from a pool of candidate authors. Decades of research show that authors leave conspicuous "fingerprints" in their writing (Juola, 2006). A small amount of preexisting prose (ca. 2,500 words) is often enough to learn enough about the writing "styles" of a set of candidate authors to correctly identify the author of an unsigned document. Authorship attribution techniques have found application in numerous domains. They have been used to resolve uncertainty about authorship in historical research (Mosteller and Wallace, 1964). For writers living today, widespread use of authorship attribution techniques—and related author profiling techniques (Argamon et al., 2009)—poses a privacy risk (Brennan et al., 2012). A better understanding of how authorship attribution techniques work can inform efforts to improve privacy-enhancing language technologies.

While there is no doubt that authorship attribution methods have improved over the past century, recent progress is harder to measure. Some of this difficulty is due to the field's success. In recent decades, new methods improve on old ones by small amounts. Given small improvements, assessing whether or not the advance may be due to aleatory factors such as preprocessing or a particular dataset becomes difficult. Another reason recent progress is difficult to measure is the lack of standard benchmark tasks. Of 15 frequently-cited authorship attribution studies examined by Potthast et al. (2016), original corpora could be found for only 4 (27%) and code could be located for 0 (0%). While other fields, notably machine translation and language modeling, excel at organizing research activity around publicly-accessible benchmark tasks, contemporary authorship attribution research has no such tasks.

Recent experience suggests that without standard benchmarks—and evidence that researchers can consistently reproduce results using them—a field's ability to self-assess progress on well-defined tasks can go astray. The field of recommender systems offers a cautionary tale. Rendle et al. (2019) documents a series of papers being published in prestigious journals over a five year period which do not, in fact, improve on earlier results. Notably, these papers used a standard dataset for their evaluations (Movielens 10M). Where these papers fell short was in their reproduction of previous results—to which their new methods were compared. The papers reported improvements on earlier results

Proceedings of Recent Advances in Natural Language Processing, pages 1174–1179
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_132

which were illusory; models used in earlier research, upon closer examination, outperformed the new methods. Analogous cases exist in other fields. In machine translation, although standard datasets were used, inconsistency in applying a key metric (BLEU) prevented researchers from easily reproducing or comparing results (Post, 2018).

Our paper supports reproducible research in authorship identification by introducing five standard benchmark tasks. Each task features fixed train and test sets. Four of the five tasks have a test set consisting of writing samples on fixed topics, guaranteeing that test set examples do not overlap with training set examples in terms of subject matter. Data for all tasks is available for download without any restrictions.

2 Problem Description

2.1 Problem: Models Cannot be Compared due to Unavailable or Under-specified Test Sets

Comparing the effectiveness of a new model with that of an existing model requires, at minimum, evaluating models on the same data. Because different models may perform differently when applied to texts by different authors or to texts in different genres by the same authors, comparing the performance of two models on a new dataset is often uninformative. Even when the new dataset resembles the original, researchers should worry that the poor performance of an earlier model may be due to accidental errors in re-implementation. Reliable comparisons of new models with previous baselines require that the original data be available.

Having the original data is not enough. The test set, the set of documents whose authorship a model must predict, must also be specified (Bouthillier et al., 2019). If cross-validation is used, the train/test splits must be known. Authorship attribution datasets typically feature a small number of authors (8-100) and much of the variability in model performance can be due to the idiosyncratic composition of cross-validation "folds."

For an example, consider the task of reproducing the work of Abbasi and Chen (2008) with the Enron email corpus. Abbasi and Chen (2008) evaluate different techniques using ten-fold cross-validation with varying number of candidate authors. Comparing the performance of a new model with their result requires knowledge of the composition of the folds they used. Small improvements in classification accuracy could be due to different partitions of the set of authors into cross-validation folds. A different partition could, by chance, end up with folds featuring writers who have distinct writing styles, making achieving higher accuracy easier.

2.2 Problem: Inappropriately Homogeneous Training and Test Corpora

Evaluations of authorship attribution techniques often use corpora consisting of homogeneous texts. Corpora consisting of texts in a single genre (e.g., newspaper article, blog post, email message) are common. This method of evaluation is not ideal. It is at odds with traditional presentations of authorship attribution, which typically claim that methods work in a variety of settings (Koppel et al., 2009; Juola, 2006). To eliminate any doubt that methods are, in fact, picking up on content-independent authorial fingerprints, test set texts should not resemble training set texts.

For an illustration of the problem, consider the use of a corpus of 100 newspaper articles written by 10 different authors. Using such a corpus to evaluate the performance of an authorship attribution method may not yield the expected information: an estimate of how well the method will perform on similar authors in a different setting. The risk of a model using topical information is clear. Newspaper writers tend to have distinct areas of expertise ("beats") which influence the types of subjects they write about. Writers from the same generation or similar social backgrounds may tend to write about certain topics. Senior writers may be more likely than junior writers to receive certain topics as assignments. Methods which appear to be using content-independent features may, in fact, be picking up on subtle signals of topic.

Unfortunately this kind of homogeneity in evaluation corpora is common. It features in all the corpora considered by Abbasi and Chen (2008) as well as the "C10" corpus drawn from Reuters (RCV1) (Potthast et al., 2016).

One method of addressing this problem is to use test set documents which are distinct from training set documents. Test set documents might be written in a different setting or different document genre. If, say, training set documents are work e-mails, then test set documents might be personal essays. Using test documents from a different time period would also help address the concern of topical homogeneity. Koppel et al. (2009) illustrate such a

division in a dataset involving two authors by using e-mails written before a fixed date as training and e-mails written after the date as testing.

Another method involves conducting a field experiment and eliciting prose on a fixed topic from writers. The elicited writing samples form the test set. This method is expensive but guarantees that models will not perform better by leveraging information about the topics specific authors tend to write about. Both Juola (2004) and Brennan et al. (2012) use this approach.

Authorship attribution methods are consistently presented as relying on the identification of topic-independent fingerprints. Evaluation tasks should be aligned with this presentation.

2.3 Problem: Unavailable or Restricted Corpora

The practice of restricting access to corpora appears to be more common in authorship attribution research than in the machine translation and language modeling communities. We considered including the C10, PAN12, and PAN13 authorship attribution tasks in our suite of benchmark tasks but found that all three are restricted and cannot be downloaded without permission.[1] We know of no cases in current machine translation or language modeling research where performing a standard evaluation requires access to a restricted dataset. Data for the news translation tasks distributed by the Conference on Machine Translation are available for immediate download.[2] Data for the widely-used language modeling benchmarks (GLUE, SQuAD) are publicly available (Wang et al., 2018; Rajpurkar et al., 2018). Of the 81 language modeling tasks cataloged by the NYU-based team developing the Jiant evaluation tool, 69 tasks (85%) can be downloaded *automatically*, that is, by the evaluation software itself.[3]

Making a dataset publicly available increases the likelihood that other researchers will reproduce results. Recent experience has shown that the probability that a result may not be reproducible or replicable is higher than previously appreciated (less than 70% according to Baker (2016)). The problem of non-reproducible results is sufficiently serious that certain conferences are exploring adopting additional measures—beyond submission of code and data—which will alleviate the problem.[4]

There is no reason to suspect that the reproducibility rate of authorship attribution research is conspicuously different from the rate in other areas of computational linguistics. Indeed, in the study of 15 frequently-cited authorship attribution papers, Potthast et al. (2016) document one failure to replicate results (Potthast et al., 2016, 403). If reproducing or replicating results is difficult in as many as 6% (1 in 15) of papers, then reproduction (or replication) should be a regular practice. And reproducing results requires that the original code and data be easy to access.

3 Improving Authorship Attribution Evaluation

The problems described in the previous section complicate a range of authorship attribution research (e.g., identification, verification, profiling). We propose a suite of five tasks which address the problems for one area of authorship attribution research: closed-set author identification involving contemporary English-language non-fiction prose. Lessons learned developing standard benchmark tasks in this area will, we hope, inform the development of analogous tasks in other areas.

Two arguments support our focus on contemporary non-fiction texts. First, collecting redistributable non-fiction prose from a diverse set of writers is relatively easy. A considerable share of the English-using population writes non-fiction prose. Demonstrating (some) competency in English composition is a requirement in secondary education across the English-speaking world. Second, many researchers are interested in the efficacy of authorship attribution methods applied to contemporary non-fiction English prose. English is, for the moment, the *lingua franca* of diplomacy, science, and international business. Authorship attribution methods which work on English therefore enjoy broad applicability. The stakes of author profiling research—research informed by authorship attribution research—are also significantly higher for research involving living writers than for writ-

[1] The restricted-download datasets may be found at the following URLs: https://zenodo.org/record/3759064 (C10), https://zenodo.org/record/3713273 (PAN12), https://zenodo.org/record/3715864 (PAN13).

[2] For example, http://www.statmt.org/wmt18/translation-task.html

[3] See https://github.com/nyu-mll/jiant/blob/master/guides/tasks/supported_tasks.md for a list of the tasks.

[4] "ML Reproducibility Challenge," https://paperswithcode.com/rc2020

Task	Number of authors (training set)	Words per author (training set)	Number of authors (test set)	Words per author (test set)	Fixed topic
AAAC–fixed-topic	13	2,563	13	843	Yes
AAAC–free-topic	13	2,563	13	2,008	No
EBG–obfuscation	45	8,866	45	555	Yes
RJ–fixed-topic	48	7,492	21	575	Yes
RJ–obfuscation	48	7,492	27	565	Yes

Table 1: Summary statistics of documents used in the five tasks.

ers active in previous centuries. Only in the former case is, say, an individual's privacy at risk.

3.1 Reproducible Authorship Attribution Benchmark Tasks (RAABT)

Five closed-set authorship identification tasks make up the Reproducible Authorship Attribution Benchmark Tasks (RAABT). Table 1 summarizes the tasks. All tasks feature a fixed test set. Test set documents do not overlap with training set documents. In four out of five of the tasks, authors write test set documents on a fixed topic. Three of the tasks involve writing from a diverse set of adults living in North America. In aggregate, the tasks feature 106 different authors.

The tasks are published at `https://zenodo.org/record/5213898`.

Task	Algorithm	Training LOO-CV	Testing
AAAC–fixed-topic	Baseline (chance)	7.7%	7.7%
	Logistic regression	18%	**31%**
	Linear SVM	16%	23%
AAAC–free-topic	Baseline (chance)	7.7%	7.7%
	Logistic regression	18%	38%
	Linear SVM	16%	**46%**
EBG–obfuscation	Baseline (chance)	2.2%	2.2%
	Logistic regression	67%	4.4%
	Linear SVM	67%	**8.9%**
RJ–fixed-topic	Baseline (chance)	2.1%	2.1%
	Logistic regression	60%	**9.5%**
	Linear SVM	60%	4.8%
RJ–obfuscation	Baseline (chance)	2.1%	2.1%
	Logistic regression	60%	**7.4%**
	Linear SVM	60%	**7.4%**

Table 2: Performance of two simple models on the five tasks. Table shows classification accuracy for multi-class logistic regression and linear SVM. Both models use the same feature set consisting of frequencies of 512 function words. This set of 512 function words has been used extensively in previous research (Koppel et al., 2009; McDonald et al., 2012).

3.2 Task descriptions

1. **Ad-hoc Authorship Attribution Competition, fixed topic (AAAC–fixed-topic).** The first task is "Problem A" from the 2004 Ad-hoc Authorship Attribution Competition (AAAC) (Juola, 2004, 2006). Texts were gathered from 13 authors in a 2013 undergraduate writing course at a university in the United States. For the test set documents, participants were asked to write on the topic of "work."

2. **Ad-hoc Authorship Attribution Competition, free topic (AAAC–free-topic)** The second task is "Problem B" from the AAAC. Test documents are additional course essays on other topics. Test set documents do not overlap with training documents. Training documents are the same as in the first task.

3. **Extended Brennan-Greenstadt Corpus, obfuscation condition (EBG–obfuscation)** The Extended Brennan-Greenstadt Corpus (Brennan et al., 2012) (EBG) contains writing from 45 individuals contacted through the Amazon Mechanical Turk platform no later than the year 2012. Participants uploaded examples of their writing. The researchers asked for writing of a "scholarly" nature.

   Participants were then asked to write a short essay on a fixed topic. They were asked to describe their neighborhood to someone unfamiliar with the location. Notably, they were also asked to obscure their writing style. They were, however, not given any instructions on how to accomplish this. These essays form the test set.

   Given prevailing norms on Amazon Mechanical Turk and the monetary incentive to finish quickly (payment did not depend on time spent on the task) we suspect many participants did not devote considerable time to de-

vising strategies for obscuring their writing style. We suggest that this task be treated as, in essence, an additional fixed topic task.

We note that the population of individuals who sell their labor on Amazon Mechanical Turk is quite diverse in terms of age, gender, and region (Coppock and McClellan, 2019).

4. **Riddell-Juola Corpus, control condition (RJ–fixed-topic)**

   The Riddell-Juola Corpus collects texts using essentially the same techniques were used in Brennan et al. (2012). Responses were collected in March and June of 2019. According to self-reported gender and age, participant demographic characteristics are roughly balanced.

   Participants were asked to respond to the same "describe your neighborhood" prompt mentioned earlier. No further instructions were given. (The instruction to obscure one's writing style was not present.)

5. **Riddell-Juola Corpus, obfuscation condition (RJ–obfuscation)** This task is the same as RJ–fixed-topic with one difference. Participants were told to obscure their writing using the same instruction as found in EBG–obfuscation. Again, they were given no instructions on how to accomplish this task.

   Participants were randomly assigned to receive the obfuscation instruction. Therefore the authors of the test set documents in this task do not overlap with the authors of the test set documents in RJ–fixed-topic.

   The training sets for the two tasks involving the Riddell-Juola Corpus are the same.

4 Accuracy of Received Methods

Table 2 reports the performance of two classic methods on the five tasks. We use a familiar 512-word function word feature set with both methods (Koppel et al., 2009). For linear SVM we use the libSVM implementation with default cost parameter ($C = 1$) (Chang and Lin, 2011). For multiclass logistic regression we use L2 regularization ($\lambda = 1$) (Pedregosa et al., 2011).

These baselines are intended to be reference points. They are chosen because they should be particularly easy to reproduce.

5 Discussion

Perceptions of the importance of having reproducible measures of model performance on well-understood tasks have changed over the last decade. Previously regarded as something desirable but by no means essential, reproducible benchmarks are increasingly seen as indispensable. Experience has shown that without such benchmarks, researchers risk overestimating the reliability of existing results or gaining a false sense of a field's progress on particular problems. Our paper contributes a suite of benchmarks which can be used to anchor future authorship attribution research.

These five tasks are a start. Additional tasks would be welcome. Many forms of writing and document types in widespread use today are not featured in the five tasks we introduce here. Short text messages and informal e-mails, in particular, are ubiquitous. Yet many individuals' habits of composition vary dramatically when writing in such genres. Standard benchmarks for cross-register and cross-genre authorship attribution would likely yield new insights into the strengths and weaknesses of existing approaches.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 1814425. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM Trans. Inf. Syst.*, 26(2):7:1–7:29.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452.

Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pages 725–734. PMLR.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy

and Anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3):12:1–12:22.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Alexander Coppock and Oliver A McClellan. 2019. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, 6(1):2053168018822174.

Patrick Juola. 2004. Ad-hoc authorship attribution competition. In *Proceedings of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden.

Patrick Juola. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Andrew WE McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. 2012. Use fewer instances of the letter "i": Toward writing style anonymization. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 299–318. Springer.

Frederick Mosteller and David L Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python (Version 0.24.1). *Journal of Machine Learning Research*, 12:2825–2830.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, and Maike Elisa Müller. 2016. Who wrote the web? Revisiting influential author identification research applicable to information retrieval. In *European Conference on Information Retrieval*, pages 393–407. Springer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Steffen Rendle, Li Zhang, and Yehuda Koren. 2019. On the difficulty of evaluating baselines: A study on recommender systems. *arXiv preprint arXiv:1905.01395*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.