Bias Analysis and Mitigation in the Evaluation of Authorship Verification

Janek Bevendorff * **Benno Stein*** **Matthias Hagen**† **Martin Potthast**‡

*Bauhaus-Universität Weimar
†Martin-Luther-Universität Halle-Wittenberg
‡Leipzig University

<first>.<last>@uni-{weimar, leipzig}.de
<first>.<last>@informatik.uni-halle.de

Abstract

The PAN series of shared tasks is well known for its continuous and high quality research in the field of digital text forensics. Among others, PAN contributions include original corpora, tailored benchmarks, and standardized experimentation platforms. In this paper we review, theoretically and practically, the authorship verification task and conclude that the underlying experiment design cannot guarantee pushing forward the state of the art—in fact, it allows for top benchmarking with a surprisingly straightforward approach. In this regard, we present a "Basic and Fairly Flawed" (BAFF) authorship verifier that is on a par with the best approaches submitted so far, and that illustrates sources of bias that should be eliminated. We pinpoint these sources in the evaluation chain and present a refined authorship corpus as effective countermeasure.

1 Introduction

When tackling a problem in empirical research, a sound and reliable evaluation of competing solution approaches is a prerequisite to achieve agreement on the state-of-the-art performance. For authorship verification, the PAN series of shared tasks caters for the most important benchmarks to which new approaches refer and compare against. The fundamental problem in authorship verification is to decide whether two given texts were written by the same author. When experimenting within the PAN setting, we learned that one can quickly achieve a competitive performance for this task—with one of the most basic approaches: a TFIDF-weighted character 3-gram model. By extending this model with a few additional features, such as the Kullback-Leibler divergence and related measures, we were able to reach the performance of the best verifiers submitted so far.[1] However, reality caught up with us when we applied our verifier to other authorship verification problems with little success. To get to the bottom of this rather baffling outcome, we carried out a systematic analysis of the entire evaluation chain, its problem definition, its corpora, its evaluation procedure, and of course our model, in search of any sources of bias that may have artificially inflated the performance of our approach. The paper in hand introduces our "Basic and Fairly Flawed" (BAFF) model and reports on our bias analysis. Moreover, in an attempt to improve the situation and call for better data, we not only contribute a new and carefully curated authorship verification corpus,[2] but also collect a few best practices for the creation of such corpora. The outlined situation calls into question a lot of what we believed to know about the state of the art, and future PAN tasks on verification will have to rectify these issues in order to provide for a more valid assessment of the state of the art.

2 Related Work

Authorship verification is a young task in the field of authorship analysis. Proposed by Koppel and Schler (2004), and mostly solved on book-sized texts right away, it remains a challenging task on short texts. The numerous verification approaches developed over the years employ a wide array of features, methods, and corpora (Stamatatos, 2009), rendering a comparison between approaches difficult. A dedicated shared task series at PAN (Stamatatos et al., 2015, 2014; Juola and Stamatatos, 2013; Argamon and Juola, 2011) was a key enabler for comparability and reproducibility. The verifiers submitted by Bagnall (2015), Fréry et al. (2014), and Modaresi and Gross (2014) form the state of the art. While new verifiers are run against the shared task's data to assess their performance against these baselines (e.g., Halvani et al., 2017; Kocher and Savoy, 2017), PAN continues to develop new benchmarks on closely related tasks.[3]

[1] https://www.tira.io/task/authorship-verification/

[2] Code and corpus: https://github.com/webis-de/acl-19

[3] See http://pan.webis.de for an overview of these tasks.

3 BAFF: A Baffling Authorship Verifier

In authorship verification, the most basic question to answer is whether two given texts p and q have been written by the same author.[4] Key to solving the task is finding a good representation $\mathbf{r}$ of the style difference between p and q. We resort to seven well-known measures for this purpose.

3.1 Features: Style Difference Measures

To compute the style difference measures listed below, we first represent p and q as character trigram vectors $\mathbf{p}$ and $\mathbf{q}$; character n-grams are considered robust style indicators across many authorship analysis tasks (Stamatatos, 2013). Given $\mathbf{p}$ and $\mathbf{q}$, we calculate the following well-known measures:[5]

1. Cosine similarity (TF-weighted)
2. Cosine similarity (TFIDF-weighted)
3. Kullback-Leibler divergence (KLD)
4. Skew divergence (skew-balanced KLD)
5. Jensen-Shannon divergence
6. Hellinger distance
7. Avg. logarithmic sentence length difference
   (a feature frequently used by PAN participants)

After assembling $\mathbf{r}$ as a 7-dimensional vector from these difference measures, we rescale all computed features to the interval $[0, 1]$ with respect to the dataset so as to align the diverse value ranges. We fully expect the divergence measures to be correlated to a greater or lesser extent; the learning algorithm will select the best-performing ones.

3.2 Performance Results

Table 1a shows the performance of four WEKA classifiers based on our model on the PAN15 test dataset. The decision tree performs best, beating Bagnall's winning deep learning approach in terms of accuracy by one percentage point for an overall second place (Table 1e). We can produce similar results on the PAN14 novels dataset (Table 1f), and, switching to a random forest, even claim first place on the essays dataset (Table 1g). Altogether, with very little effort, our model outperforms the 31 approaches submitted to PAN in 2014 and 2015, competing with much more elaborate solutions.

[4] In forensic applications, a text of unknown authorship and one or more texts known to be written by a given author are considered (van Halteren, 2004). If solved, other authorship-related tasks, such as authorship attribution, would be solved as well, since they can be reduced to a series of verifications.

[5] Except for the cosine similarity and the average sentence length difference, the other statistical difference measures we use have rarely been considered for verification to date.

4 Bias Analysis

Unable to reproduce these outstanding results on other verification problems, our ensuing analysis of the evaluation chain revealed several *interdependent* sources of bias in all its components, namely our model, the data, and the evaluation procedure. In what follows, we discuss these biases, outline their underlying flaws, and ways to mitigate them.

4.1 Model Bias

In an attempt to pinpoint which feature contributes how much to the overall performance, we ran an ablation test. While the removal of each feature causes some performance loss, the removal of Feature 2, the TFIDF-weighted cosine similarity, resulted in the loss of 19 percentage points, by far the largest among all features. What makes TFIDF special is its IDF factor, which was the key to identify two sources of bias in our model:

(B1) Corpus-relative features. TFIDF is used so matter-of-factly throughout machine learning that hardly anyone discusses the origin of its document frequency (DF) values. In the absence of any explanation, one may assume that they are computed from the currently processed dataset. This is perfectly alright for most tasks, but crucially not for authorship verification where computing DF from the evaluation datasets at runtime is both unrealistic and prone to overfit. The rather small number of test cases in the PAN datasets combined with Bias B4 allows the learning algorithm to "reverse-engineer" part of the ground-truth from the DF values, while in practice, a forensic linguist analyzes only one case at a time, not many (see Bias B6). Table 1c ("scaled" rows) shows BAFF's performance when computing DF from the processed corpus, and when using the Brown corpus instead, revealing a severe drop of performance. Hence, corpus-relative features should be avoided.

(B2) Feature scaling. Another machine learning technique that is often applied without second thought is scale normalization of all features. However, applying the same reasoning as for the (I)DF calculation, scale normalization biases our features towards corpus specifics. Table 1c shows BAFF's performance with and without scale normalization. We experience a massive performance drop in combination with corpus-relative IDF, but much less so with "external" IDF from the Brown corpus. This aggravation of Bias B1 through feature scaling is most likely influenced by Biases B3–B6.

(a) *BAFF on PAN15 corpus*

	Acc.	Prec.	F_1	ROC
Naive Bayes	0.674	0.675	0.674	**0.771**
SVM	0.700	0.700	0.700	0.700
Decision Tree	**0.768**	**0.773**	**0.767**	0.746
Random Forest	0.660	0.661	0.660	0.717

(b) *BAFF on Gutenberg corpus (unscaled, w/o TFIDF)*

Naive Bayes	0.934	0.634	0.634	**0.756**
SVM	**0.695**	0.701	**0.693**	0.695
Decision Tree	**0.695**	**0.765**	0.674	0.695
Random Forest	0.683	0.687	0.681	0.741

(c) *Corpus-relative IDF against external IDF*

Corpus IDF (scaled)	**0.768**	**0.773**	**0.767**	**0.746**
Corpus IDF (unscaled)	0.622	0.684	0.651	0.639
Brown IDF (scaled)	0.598	0.611	0.586	0.598
Brown IDF (unscaled)	0.590	0.605	0.575	0.590

(d) *10-fold cross-val. naive Bayes on corpus-rel. IDF*

PAN15 Test (scaled)	**0.742**	**0.749**	**0.740**	**0.796**
Gutenberg (scaled)	0.570	0.628	0.515	0.599

(e) *PAN15 submissions*

	C@1	ROC	Final
Bagnall	0.757	**0.811**	**0.614**
BAFF	***0.768***	*0.746*	*0.573*
Castro et al.	0.694	0.750	0.520
Gutierrez et al.	0.694	0.740	0.513
Kocher and Savoy	0.690	0.738	0.508
Halvani and Winter	0.601	0.762	0.458

(f) *PAN14 novels submissions*

Modaresi and Gross	**0.715**	0.711	**0.508**
Zamani et al.	0.650	0.733	0.476
BAFF	*0.651*	*0.715*	*0.465*
Khonji and Iraqi	0.610	**0.750**	0.458
Mayor et al.	0.614	0.664	0.407
Castillo et al.	0.615	0.628	0.386

(g) *PAN14 essays submissions*

BAFF	***0.722***	***0.761***	***0.549***
Fréry et al.	0.710	0.723	0.513
Satyam et al.	0.657	0.699	0.459
Moreau et al.	0.600	0.620	0.372
Layton	0.610	0.595	0.363
Modaresi and Gross	0.580	0.603	0.350

(h) *PAN15/14 and our Gutenberg corpus statistics*

	Num. Cases		Avg. Words / Text	
	Training	Test	Training	Test
PAN15	100	500	340	510
PAN14 Novels	100	200	1,540	6,000
PAN14 Essays	200	200	830	820
PAN14 Essays[a]	200	200	3,040	2,940
Gutenberg	192	82	3,900	3,930

(i) *Gutenberg corpus subsets (genre and time period)*

Corpus subset	Num. Cases	Unique Authors
19th cent. adventures	118	177
19th cent. sci-fi	60	90
20th cent. sci-fi	96	144
Total	**274**	**390**[b]

[a] Counting "known" texts as a single large text. A case in the essays corpus has one "unknown" and up to five "known" texts.
[b] Not all authors are unique across subsets.

Table 1: Column 1 shows the results of different classifiers on the PAN15 (*a*) and our Gutenberg corpus (*b*), an analysis of BAFF on the PAN15 corpus with different IDF values (*c*), and a comparison of 10-fold cross-validation naive Bayes with corpus-relative TFIDF as the only feature between the two corpora (*d*). Column 2 ranks BAFF against the top-5 PAN15 (*e*) and PAN14 (*f*/*g*) submissions (final score = C@1 · ROC). Column 3 lists general statistics for all corpora (*h*) and genres and time periods covered by our Gutenberg corpus (*i*).

4.2 Data Bias

Just as the creators of a verification model should mitigate bias by avoiding unsuitable features and techniques, so should the creators of an evaluation dataset take precautions not to make it readily exploitable. The reason why Biases B1 and B2 inflated the performance of our model is largely due to the fact that the data is biased, too, or else the model's biased features would not have had such a significant positive effect. Reviewing PAN's datasets, we identify three sources of bias.

(B3) Plain text heterogeneity. Inspecting the plain text files of the datasets, many of them carry artifacts that are unlikely to signal authorial style, but rather originate from the plain text converter used or the human transcriber. Examples we observed include mixed use of ASCII and Unicode ellipsis markers (some as iconic as ". . . ."), a wide variety of quotation marks and em dashes (also mixed encodings), and curly braces for parentheses. Moreover, the texts are formatted to be human-readable by preserving white space, including indentations and line breaks, which vary greatly across authors, but were not necessarily introduced by them. Given that many verification models use character n-grams as basic style representation, n-grams covering these artifacts may indicate authorship even *across* cases. To mitigate this bias, the texts in a dataset should be fully homogenized (particularly in the presence of Bias B4).

(B4) Population homogeneity. Many monographs are required to construct a verification dataset. But the sources tapped so far lack scale, so that three shortcuts are commonly applied to maximize yield:[6] For same-author cases, *more than one* case is constructed for a given author, (1) by systematically pairing *more than two* texts by that author, and/or (2) by splitting long texts (e.g., books) to obtain more text chunks from that author. For different-authors cases, (3) texts from authors for whom same-author cases exist are reused, using different, or even the same chunks also found in same-author cases. Such imbalance causes authors' styles to be over-/underrepresented. Steady use of these shortcuts also gives rise to Bias B5.

(B5) Accidental text overlap. The strong contribution of the TFIDF-weighted cosine similarity points to text overlap in same-author cases that renders them easier-to-discriminate from different-authors cases. Caused by Bias B4, text overlap includes named entities (e.g., speaker names in the plays of PAN15), topic words shared between text chunks taken from the same source text, repeated phrases, and unique character sequences. The fanfiction used for PAN14 contains text reuse from the original books. Accidental overlap between cases may lead a learning algorithm astray, especially in the presence of Biases B1 and B6. For mitigation, a text overlap analysis and correction is necessary.

[6] E.g., the PAN15 dataset consists of hundreds of cases constructed from only 15 stage plays by six different authors.

4.3 Evaluation Bias

Lastly, the evaluation procedure itself is biased.

(B6) Test conflation. At testing time, authorship verifiers can usually access the entire test dataset. This is unrealistic; a forensic linguist works on a case-by-case basis, and cases are independent of one another, or their underlying population is unknown. Emulating this scenario, a verifier should process only one test case at a time, without referring to previously processed cases to solve the next one. Incidentally, this policy would mitigate many of the aforementioned biases. While not enforcible in individual evaluations and shared tasks with run submissions, at PAN, it may indeed be, by adjusting the TIRA platform (Potthast et al., 2019) to handle the software runs accordingly.

5 The Webis Authorship Verification Corpus

With the goal of avoiding all data biases, we constructed a new authorship verification corpus based on books obtained from Project Gutenberg:[7] the Webis Authorship Verification Corpus 2019. We validate the corpus using our BAFF approach.

5.1 Corpus Construction

At Project Gutenberg, transcriptions of many public domain books are provided. Given their diversity, we limit our choice to fiction books from the 19th and 20th century and the two specific genres adventure and science fiction, controlling for respective style variation. Table 1h and i compare the corpus statistics with the three PAN corpora.

To avoid Bias B4, we ensured that each author is unique within, though not necessarily across any combination of time period and genre. Moreover, no texts were reused to construct different-authors cases, but texts from previously unused authors were collected. The same-author cases were created so that both texts are from different books, and where possible, neither book is from the same series of books. Altogether, we created a total of 274 verification cases of which 50 % are same-author and the rest different-authors cases, with a 70/30 split of training and test. The size of each text varies between 3,500 and 4,000 words (21,870 characters on average), with a few individual texts being shorter due to insufficient material. Unlike the PAN datasets, we aimed for a corpus that can also be processed by Koppel and Schler's unmasking, an important state-of-the-art approach.

[7] https://www.gutenberg.org/

To avoid Bias B3, all texts were carefully normalized to remove editorial and non-authorial artifacts. We stripped book and chapter titles, illustration placeholders, ASCII art, repeated character runs, footnotes, and obvious quotations from the texts (to also avoid Bias B5), as well as any Gutenberg-related front pages and additions to the original text. Gutenberg books make use of underscores to signify italic text; we removed those as well. Special characters like ellipses and quotation marks were manually replaced by a consistent ASCII representation. We further collapsed all newlines and other white space into a single space character to avoid incidental and inadvertent bias due to formatting.

5.2 Corpus Validation

As per Bias B1, a high performance of TFIDF-weighted cosine similarity hints at a biased dataset. To validate our corpus in this respect, we cross-validated a naive Bayes classifier using only this feature (Table 1d), which achieved merely 57 % accuracy compared to 74 % on PAN15. Excluding cosine similarity, BAFF still gets up to 70 % accuracy (Table 1b), which marks statistical divergence measures as promising features for future verifiers.

6 Conclusion

In shared tasks, sometimes basic approaches outperform more sophisticated ones. This is frequently the case when machine learning meets small data. Inadvertent properties of the data act as confounders that a learning algorithm will gladly fit onto if they are not controlled. In the case of authorship verification as per PAN, this was a major part of the problem. As long as much larger corpora remain out of reach for lack of a sufficient source of monographs, extra care needs to be taken in preparing the data, as exemplified for our corpus.

Another important take-away message is that model authors in authorship verification need to be extra careful about their feature selection. Fortunately, this will come naturally to researchers in the field as they are already trained to avoid features that encode topic rather than style. In particular, we strongly suggest that future evaluations should adopt a stateless one-case-at-a-time test policy.

Finally, in a spin-off study on unmasking, we generalized the algorithm to work on short, essay-length texts (Bevendorff et al., 2019): it achieves an accuracy of 0.73, an F_1 of 0.69, and a precision of 0.82, marking the first baseline for our corpus.

References

Shlomo Argamon and Patrick Juola. 2011. Overview of the International Authorship Identification Competition at PAN-2011. In *Notebook Papers of CLEF 2011 Labs and Workshops, pages 19-22*.

Douglas Bagnall. 2015. Author Identification using multi-headed Recurrent Neural Networks — Notebook for PAN at CLEF 2015. In *CLEF 2015 Working Notes Papers*.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Generalizing Unmasking for Short Texts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 654–659. Association for Computational Linguistics.

Esteban Castillo, Ofelia Cervantes, Darnes Vilariño, David Pinto, , and Saul León. 2014. Unsupervised Method for the Authorship Identification Task — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Daniel Castro, Yaritza Adame, María Pelaez, and Rafael Muñoz. 2015. Authorship Verification, Combining Linguistic Features and Different Similarity Functions — Notebook for PAN at CLEF 2015. In *CLEF 2015 Working Notes Papers*.

Jordan Fréry, Christine Largeron, and Mihaela Juganaru-Mathieu. 2014. UJM at CLEF in Author Identification — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Josue Gutierrez, Jose Casillas, Paola Ledesma, Gibran Fuentes, and Ivan Meza. 2015. Homotopy Based Classification for Author Verification Task — Notebook for PAN at CLEF 2015. In *CLEF 2015 Working Notes Papers*.

Oren Halvani and Christian Winter. 2015. A Generic Authorship Verification Scheme Based on Equal Error Rates — Notebook for PAN at CLEF 2015. In *CLEF 2015 Working Notes Papers*.

Oren Halvani, Christian Winter, and Lukas Graner. 2017. Authorship verification based on compression-models. *CoRR*, abs/1706.00516.

Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. In *CLEF 2013 Working Notes Papers*.

Mahmoud Khonji and Youssef Iraqi. 2014. A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF) — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Mirco Kocher and Jacques Savoy. 2015. UniNE at CLEF 2015: Author Identification — Notebook for PAN at CLEF 2015. In *CLEF 2015 Working Notes Papers*.

Mirco Kocher and Jacques Savoy. 2017. A simple and efficient algorithm for authorship verification. *JASIST*, 68(1):259–269.

Moshe Koppel and Jonathan Schler. 2004. Authorship Verification as a One-Class Classification Problem. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 1–7.

Robert Layton. 2014. A simple Local n-gram Ensemble for Authorship Verification — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Cristhian Mayor, Josue Gutierrez, Angel Toledo, Rodrigo Martinez, Paola Ledesma, Gibran Fuentes, and Ivan Meza. 2014. A Single Author Style Representation for the Author Verification Task — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Pashutan Modaresi and Philipp Gross. 2014. A Language Independent Author Verifier Using Fuzzy C-Means Clustering — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Erwan Moreau, Arun Jayapal, , and Carl Vogel. 2014. Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer.

Satyam, Anand, Arnav Kumar Dawn, , and Sujan Kumar Saha. 2014. Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis — Notebook for PAN at CLEF 2014. In *CLEF 2014 Working Notes Papers*.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2):421–439.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López López, Martin Potthast, and Benno Stein. 2015. Overview of the Author Identification Task at PAN 2015. In *CLEF 2015 Working Notes Papers*.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the Author Identification Task at PAN 2014. In *CLEF 2014 Working Notes Papers*.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*.

Hamed Zamani, Samira Abnar, Mostafa Dehghani, Mahsa Forati, and Pariya Babaei. 2014. Submission to the Author Identification Task at PAN 2014. In *CLEF 2014 Working Notes Papers*.