Rethinking the Authorship Verification Experimental Setups

Florin Brad[*] Bitdefender

Andrei Manolache[*] Bitdefender,
University of Stuttgart

Elena Burceanu Bitdefender

{fbrad,amanolache,eburceanu}@bitdefender.com

Antonio Barbalau University of Bucharest
abarbalau@fmi.unibuc.ro

Radu Tudor Ionescu University of Bucharest
raducu.ionescu@gmail.com

Marius Popescu University of Bucharest
popescunmarius@gmail.com

Abstract

One of the main drivers of the recent advances in authorship verification is the PAN large-scale authorship dataset. Despite generating significant progress in the field, inconsistent performance differences between the closed and open test sets have been reported. To this end, we improve the experimental setup by proposing five new public splits over the PAN dataset, specifically designed to isolate and identify biases related to the text topic and to the author's writing style. We evaluate several BERT-like baselines on these splits, showing that such models are competitive with authorship verification state-of-the-art methods. Furthermore, using explainable AI, we find that these baselines are biased towards named entities. We show that models trained without the named entities obtain better results and generalize better when tested on DarkReddit, our new dataset for authorship verification.

1 Introduction

Identifying the author of a text is one of the most versatile NLP tasks, with applications ranging from plagiarism detection to forensics and monitoring the activity of cyber-criminals. The task spans several decades and was tackled using statistical linguistics (Mendenhall, 1887; Zipf, 1932; Mosteller and Wallace, 1964), and, more recently, machine learning (de Vel et al., 2001; Zhao and Zobel, 2005; Koppel et al., 2007; Stamatatos, 2009). Due to the typically small data setup of authorship analysis tasks, deep learning methods had a slow start in this domain. Nevertheless, inspired by the impressive performance of pre-trained language models, such as BERT (Devlin et al., 2019), these methods gained traction in authorship analysis as well. Saedi and Dras (2021) showed that Convolutional Siamese Networks are more robust than a BERT-based method over large-scale authorship attribution tasks.

Barlas and Stamatatos (2020) investigated pre-trained language models for cross-topic and cross-domain authorship attribution and showed that BERT and ELMo (Peters et al., 2018) achieve the best results while being the most stable approaches. Fabien et al. (2020) introduced BERT for Authorship Attribution (BertAA) in which they combine BERT with stylometric features for authorship attribution. The authors remarked that their model is unable to perform text similarity evaluation in the context of the more difficult authorship verification problem, which we tackle.

One of the main contributors to the active developments in authorship analysis is the PAN organizing team, who proposed annual shared tasks since 2009. While the recent PAN 2020 and 2021 contests increased the difficulty of the authorship verification task and enabled large-scale model training (Kestemont et al., 2020, 2021), there are still possible generalization issues due to the dataset splits. For instance, models from 2020 trained on the *closed-set* data surprisingly performed better on the *open-set* test data (which is arguably more difficult) than on the *closed-set* test data (Kestemont et al., 2021). We therefore argue that in order to better assess the generalization capabilities of authorship verification systems, a more fine-grained approach to dataset splitting may be needed.

To address these issues, we introduce a set of five carefully designed splits of the publicly available PAN dataset, ranging from the easiest setup (*closed-set*) to the most difficult (*open-set*). Our splits progressively alleviate information leaks in the test data, enabling a more confident evaluation.

Furthermore, we release our splits publicly[1] to allow other members of the community to evaluate models on any computing infrastructure, enabling the evaluation of large-scale models. Along with the new splits, we introduce a set of BERT-based models (Devlin et al., 2019) to serve as baselines

[*] Equal contribution.

[1] https://github.com/bit-ml/Dupin/tree/main

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5634–5643
December 7-11, 2022 ©2022 Association for Computational Linguistics

Test split	*O2D2**	*O2D2*	*BERT*	*Naive*†	*Comp.*†
Closed	93.5	**96.4**	95.6	75.6	72.2
Clopen	94.0	96.0	**97.4**	74.1	71.1
Open UA	92.6	**92.6**	90.2	78.6	68.5
Open UF	91.4	**95.1**	91.6	79.9	79.0
Open All	80.6	67.5	**88.7**	75.6	76.9
PAN Closed	93.3	**93.5**	-	74.7	74.2
PAN Open	93.3	**94.4**	-	75.3	74.5

Table 1: Overall scores of several models evaluated on our public test splits. We also list the reported results of the models on the private PAN splits. BERT is competitive with the top-scoring O2D2 model of the PAN 2021 competition and both methods greatly outperform the PAN baselines (Naive and Compression). O2D2 performs poorly on our most difficult split Open All. However, performance on the development set is much closer to the BERT results on the test set. The neural models were trained on the large training splits. †Models trained on the small datasets. *Models evaluated on the validation set.

for future research. We show that these language models are competitive with the top scoring O2D2 (*out-of-distribution detector*) system at PAN 2021 (Boenninghoff et al., 2021).

We also qualitatively inspect the models' predictions and find that they often rely on named entities to verify authorship. We show that by replacing the named entities in the dataset with placeholders, we are able to obtain significant performance gains and better generalization capabilities.

In summary, **our contributions** are threefold:

1. We **introduce five splits**, based on the PAN dataset, with a decreasing degree of shared information between train and test sets. These configurations enable benchmarking large models, providing a robust evaluation environment, on which we run several BERT-based baselines.

2. Using explainable AI (XAI) methods, we find that **BERT-like models focus on named entities** to determine authorship. We replace them with placeholders and retrain our models, which brings a significant performance boost.

3. We introduce the DarkReddit dataset for authorship verification, which is significantly different in style to the fanfictions in PAN. We test the **generalization capabilities** of the models trained on PAN, by evaluating them on DarkReddit. Our previous finding is further confirmed by our model trained without named entities, which generalizes better and improves the overall metric by 5.6%.

Split	*authors in val*	*fandoms in val*	*authors in test*	*fandoms in test*
Closed	✓	✓	✓	✓
Clopen$^\diamond$	✓	✓	✓	✓
Open Unseen Authors	✗	✓	✗	✓
Open Unseen Fandoms	✓	✗	✓	✗
Open All	✗	✓	✗	✗

Table 2: Dataset splits sorted from the easiest (train authors and fandoms are seen in the validation and test sets) to the most difficult (train authors and fandoms are not found in the test set). $^\diamond$Some of the authors of the Different Authors train pairs in Clopen may be unknown at test time, making it a mix between Closed and Open.

2 Datasets

We use the PAN 2020 authorship verification dataset[2]. A document d_i belongs to a fandom (topic) f_i and is written by an author a_i. Author verification is a classification task which asks whether documents d_i and d_j are written by the same author (SA) or by different authors (DA). The dataset comes in two sizes: small (52k examples) and large (275k examples). The latter one is better suited for deep learning models.

2.1 New PAN 2020 splits

The PAN 2020 competition is a *closed-set* verification setup, meaning that the unseen test set contains documents whose authors and fandoms were seen at training time. The PAN 2021 competition has a more difficult *open-set* setup, in which the training data is the same as in 2020, but the submitted solutions are privately tested against document pairs from previously unseen authors and fandoms. The PAN testing infrastructure makes it difficult to evaluate large models quickly. To this end, we release several dataset splits, ranging from the easier *closed-set* setup to the more difficult *open-set* variants. We summarize the splits in Tab. 2 and provide a more detailed description in the Supplementary Material B. For each split, we propose a small (XS) and a large (XL) version.

[2] https://pan.webis.de/clef21/pan21-web/author-identification.html

	Closed$_{XL}$			Clopen$_{XL}$			Open UA$_{XL}$			Open UF$_{XL}$			Open All$_{XL}$		
Metric	**O2D2**	**cB**	**B**	**O2D2**	**cB**	**B**	**O2D2**	**cB**	**B**	**O2D2**	**cB**	**B**	**O2D2**	**cB**	**B**
F1	**96.6**	93.8	95.0	96.1	95.6	**96.8**	**93.6**	85.4	89.2	**95.2**	88.6	90.8	45.0	74.8	**86.9**
F0.5	94.3	93.4	**94.5**	94.0	96.5	**97.0**	**89.6**	87.0	88.1	**94.5**	92.3	88.8	70.1	84.6	**87.2**
c@1	**95.9**	93.3	94.5	95.4	95.4	**96.6**	**91.5**	84.8	88.2	**93.1**	88.8	90.0	65.2	78.9	**87.0**
AUC	**98.7**	98.0	98.6	98.4	99.1	**99.4**	**95.8**	92.4	95.3	**97.6**	96.5	96.9	89.5	91.1	**94.0**
overall	**96.4**	94.7	95.6	96.0	96.7	**97.4**	**92.6**	87.4	90.2	**95.1**	91.5	91.6	67.5	82.3	**88.7**

Table 3: Comparison of neural models on the PAN 2020 XL splits. O2D2 outperforms BERT (**B**) on three out of five splits, while BERT outperforms charBERT (**cB**) on all the splits. Note how the *closed-set* results (left side) are considerably better than the *open-set* ones (right side), indicating that models overfit the styles of the known authors from the closed splits. We report all the PAN 2020 metrics for the test split. The best result per split is in bold.

2.2 DarkReddit

To test an even more difficult scenario than our *open-set* splits, we created a small authorship verification dataset. This dataset could be used to benchmark the generalization capabilities of AV models, while also being useful for cybersecurity applications. The dataset was constructed by crawling 1026 samples from `/r/darknet`[3], a subreddit dedicated to discussions about the Darknet. There is an equal number of same author and different author pairs, resulting in a balanced dataset. A document has 2,500 words on average, 9 times less than the PAN 2020 splits. The two datasets also differ in other aspects (*e.g.* topics, authors, text purpose, self-contained message). We illustrate the differences between PAN and DarkReddit examples in Figure 1.

3 Experiments

Training. We fine-tune **BERT** (**B**) (Devlin et al., 2019) and **Character BERT** (**cB, charBERT**) (Boukkouri et al., 2020) as binary classifiers for authorship verification. Given two documents d_i and d_j, we concatenate and feed them to the Transformer encoder. When a document is longer than 256 tokens, we sample a random chunk of length 256. The chunks are resampled at every epoch, hence increasing the variety of the training set. To make predictions, we add a linear layer on top of the $h_{[CLS]}$ vector and optimize the entire model via the binary cross entropy loss. We use the same set of hyperparameters across all of the experiments. For the other models (**O2D2**, **Naive** and **Compression**) we used the provided code and default hyperparameters.

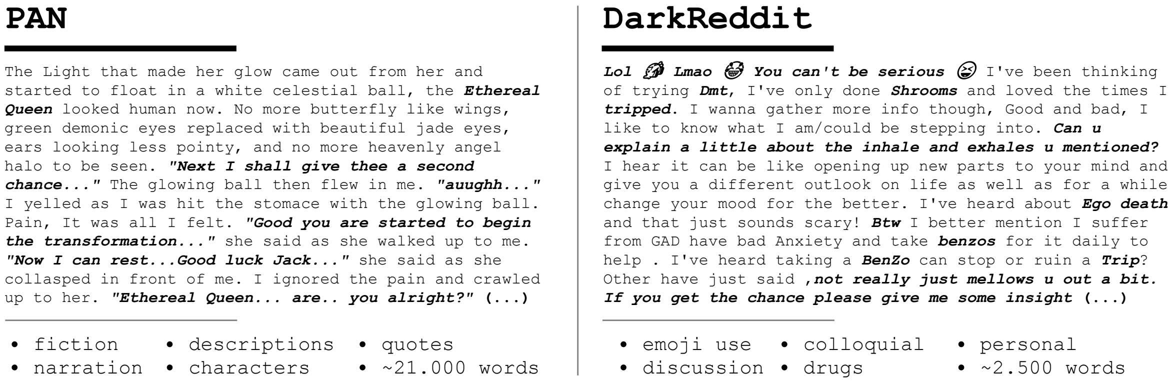
Evaluation. We report the $overall$ metric from PAN 2020 (the mean over $F1$, $F0.5$, $c@1$ and AUC). To use information from all document pairs (d_i, d_j), we split each of them into 256-length non-overlapping chunks. We then feed each chunk pair to the model, obtaining the class probabilities. Finally, we average the probabilities of all the chunk pairs to obtain the prediction for the document pair. Unsurprisingly, using multiple chunks outperforms randomly picking only one chunk from each document, leading to up to 10% improvements in the overall score.

3.1 Model comparison

Comparison to PAN models. As can be seen in Tables 1 and 3, BERT is competitive with the PAN 2021 winner on our public test splits. Both models greatly outperform the PAN baselines, a naive distance-based approach (Kestemont et al., 2016) and a compression-based approach (Halvani and Graner, 2018). BERT performs worse on the more difficult open splits. O2D2 performs surprisingly poor on the Open All test split, which may be due to its calibration step, since the performance on the development split is much larger (80.6 vs 67.5). Evaluating BERT on the private PAN sets is slow due to access to CPU-only machines (≈1200h on a machine powered by Intel Xeon E5 CPU with 8GB RAM memory). However, based on the scores of the O2D2 and baseline approaches, we expect it to perform similarly to the open test sets.

Comparison on our splits. We fine-tune and evaluate the BERT-based models on the larger XL splits introduced in Sec. 2 and compare them to the PAN 2021 winner, O2D2. We also report performances of two other models and their ensem-

[3] `https://www.reddit.com/r/darknet/`



PAN

The Light that made her glow came out from her and started to float in a white celestial ball, the ***Ethereal Queen*** looked human now. No more butterfly like wings, green demonic eyes replaced with beautiful jade eyes, ears looking less pointy, and no more heavenly angel halo to be seen. ***"Next I shall give thee a second chance..."*** The glowing ball then flew in me. ***"auughh..."*** I yelled as I was hit the stomace with the glowing ball. Pain, It was all I felt. ***"Good you are started to begin the transformation..."*** she said as she walked up to me. ***"Now I can rest...Good luck Jack..."*** she said as she collasped in front of me. I ignored the pain and crawled up to her. ***"Ethereal Queen... are.. you alright?"*** (...)

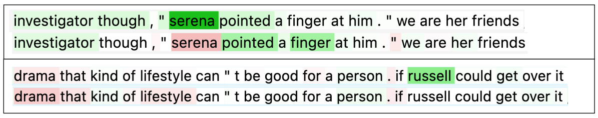
- fiction
- narration
- descriptions
- characters
- quotes
- ~21.000 words

DarkReddit

Lol 🤣 ***Lmao*** 😂 ***You can't be serious*** 😆 I've been thinking of trying ***Dmt,*** I've only done ***Shrooms*** and loved the times I ***tripped***. I wanna gather more info though, Good and bad, I like to know what I am/could be stepping into. ***Can u explain a little about the inhale and exhales u mentioned?*** I hear it can be like opening up new parts to your mind and give you a different outlook on life as well as for a while change your mood for the better. I've heard about ***Ego death*** and that just sounds scary! ***Btw*** I better mention I suffer from GAD have bad Anxiety and take ***benzos*** for it daily to help . I've heard taking a ***BenZo*** can stop or ruin a ***Trip***? Other have just said ***,not really just mellows u out a bit. If you get the chance please give me some insight*** (...)

- emoji use
- discussion
- colloquial
- drugs
- personal
- ~2.500 words

Figure 1: A PAN-2020 sample compared to a DarkReddit one. Note the contrasting style, topics, vocabulary and size between the two samples.



investigator though , " serena pointed a finger at him . " we are her friends
investigator though , " serena pointed a finger at him . " we are her friends

drama that kind of lifestyle can " t be good for a person . if russell could get over it
drama that kind of lifestyle can " t be good for a person . if russell could get over it

Figure 2: Explainability analysis using Integrated Gradients. Words highlighted in green help the correct prediction, while those in red distract from it. In each pair, the rows' attributions are from BERT fine-tuned on Closed$_{XL}$ and Open UF$_{XL}$ respectively. Fine-tuning on the latter split changes the words' attribution scores. Specifically, the focus on named entities (*e.g.* serena, russell) in the 1st row of each pair, which should not be relevant in author detection, diminishes in the 2nd row.

bles in the Supplementary Material A.1: a Siamese model (siamBERT) and a domain-adapted BERT pretrained on the PAN 2020 corpus with the MLM objective, then fine-tuned on each split. In Tab. 3 we notice that BERT outperforms charBERT on all the splits over almost all the metrics. We expected charBERT to provide better contextual embeddings for rare words (like named entities), since they incorporate character n-grams into the embeddings. Though BERT may represent rare words noisily, it is sufficiently robust for the PAN 2020 corpus.

3.2 Qualitative examples reveal biases

We next focus on better understanding the models' predictions through explainable AI (Tjoa and Guan, 2019) techniques. Inspecting the attention scores is a common method of explaining a model's prediction that has been called into question in recent years (Pruthi et al., 2020; Serrano and Smith, 2019). We therefore follow recent explainability results (Bastings and Filippova, 2020) and use the Integrated Gradients (IG) (Sundararajan et al., 2017) method from the Captum library (Kokhlikyan et al., 2020) to reveal the individual importance of words.

We analyze BERT models fine-tuned on a closed and an open set, checking for potential biases arising from the dataset splitting process. In Fig. 2, we show how important each word is in the authorship verification decision. For BERT trained on the Closed$_{XL}$ split (1st row per pair), the most important ones are the named entities. This initial focus is reduced when fine-tuning the model on the Open UF$_{XL}$ split (which keeps training and testing fandoms disjoint). This suggests that fandom-specific named entities encountered at test time are less likely to be exploited for the prediction, since they were not seen during training. Furthermore, the open validation splits help with generalization at the *model selection* step. This is due to measuring the model's performance against fandom and author-specific information unseen at training time.

3.3 Replacing named entities improves generalization

We hypothesize that replacing the named entities may further help with generalization in a *data-centric* fashion, prohibiting the model to exploit them at train time. To this end, we replace the named entities from the Open All XS dataset with their corresponding type (*e.g. Wolverine→person*)[4]. We notice in Tab. 4 that this replacement step improves the *overall* score for both models, strengthening our hypothesis about the role of named entities in authorship verification. Our results are in line with the previous works of Layton et al. (2010) and Ding et al. (2015), which show that removing entities such as mentions, hashtags

[4] https://spacy.io/api/entityrecognizer

	BERT		charBERT	
Metric	**w/ NE**	**w/o NE**	**w/ NE**	**w/o NE**
F1	**73.5**	**73.5**	54.1	**75.2**
AUC	91.1	**94.3**	**89.0**	84.4
F0.5	84.1	**85.9**	**72.9**	66.5
C@1	78.1	**78.7**	68.0	**68.3**
overall	81.7	**83.1**	71.0	**73.6**

Table 4: Performance of BERT and charBERT on the PAN Open All XS test split when using the raw dataset and hiding the named entities (w/o NE).

training set	*F1*	*AUC*	*F0.5*	*C@1*	*overall*
Open All w/ NE	69.5	83.0	58.9	56.4	67.0
Open All w/o NE	**74.1**	**86.4**	**64.4**	**65.4**	**72.6**

Table 5: Cross-corpus evaluation on DarkReddit. We compare the BERT models trained on the Open All XS dataset with and without named entities. Removing the named entities from the training set significantly improves the model's generalization across corpora.

and topic information improves performance of authorship attribution.

Our results are further confirmed in a zero-shot scenario, under a significant distribution shift, when testing on the DarkReddit corpus introduced in Sec. 2.2. Specifically, we demonstrate in Tab. 5 a significant performance gain when training without named entities. This suggests that the initial model was focusing on named entities in a spurious way.

4 Conclusions

We introduced and published five splits of the PAN dataset ranging from the easiest *closed* setup to increasingly more challenging settings. This enables a fine-grained evaluation and model selection. We showed that BERT-based baselines are competitive with top-scoring authorship verification methods and significantly outperform non-neural baselines.

Using Integrated Gradients, we showed that, distinctly from the closed split, the open splits help generalization at the *model selection* step by preventing the model from overfitting on named entities of specific train authors or fandoms. We further improved generalization by replacing the named entities, making the models more robust to spurious features. This claim also holds under a strong distribution shift, when cross-evaluating the models on the significantly different DarkReddit dataset.

Acknowledgements

This work has been supported in part by UEFISCDI, under Project PN-III-P2-2.1-PTE-2019-0532. Andrei Manolache was also supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

Limitations

Closed vs. open splits. While our paper focuses on building more difficult open set splits, deploying authorship verification systems is application specific. This means that having methods trained on closed splits may be desirable in certain scenarios, such as when we are guaranteed that the test authors are known.

Noisy examples. Collecting texts for building corpora for authorship verification can suffer from noisy data. Concretely, in both cases of PAN and DarkReddit, one user can write under multiple pseudonyms, leading to some different author examples to actually have the incorrect label. Moreover, multiple users can share the same account, leading to another issue where same author pairs are wrongly labeled. However, large-scale authorship verification models should be robust to this issue due to the large dataset size.

Long documents. Our BERT-based baselines are capped at sequences of 512 tokens at most. This means that we can process at most 256 tokens from each text in a pair at a time. During training, we overcame this issues by selecting random chunks of texts. During evaluation, we aggregated predictions from all the chunks to obtain a prediction for the documents pair. This limits the representation power during both training and evaluating, due to encoding smaller contexts. Moreover, it slows down inference on longer examples, making it even more difficult to evaluate models on limited infrastructure. Further works should also include models that accommodate longer sequences.

Ethics Statement

Authorship Verification systems may be deployed in non-ethical ways, by different organizations and parties, in order to track down vulnerable categories of people, such as journalists, dissidents, whistleblowers, etc. However, we believe that opening up research regarding authorship verification can help these vulnerable categories by raising awareness of the possibilities and limitations of state-of-the-art techniques and by mitigating their misuse.

Our datasets are based on publicly available data and do not contain sensitive information.

References

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations - 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5-7, 2020, Proceedings, Part I*, volume 583 of *IFIP Advances in Information and Communication Technology*, pages 255–266. Springer.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pages 149–155. Association for Computational Linguistics.

Benedikt Boenninghoff, Robert M. Nickel, and Dorothea Kolossa. 2021. O2D2: Out-Of-Distribution Detector to Capture Undecidable Trials in Authorship Verification—Notebook for PAN at CLEF 2021. In *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. Characterbert: Reconciling elmo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6903–6915. International Committee on Computational Linguistics.

O. de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55–64.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Steven H. H. Ding, Benjamin C. M. Fung, and Mourad Debbabi. 2015. A visualizable evidence-driven approach for authorship attribution. *ACM Trans. Inf. Syst. Secur.*, 17(3).

Maël Fabien, Esaú Villatoro-Tello, Petr Motlícek, and Shantipriya Parida. 2020. Bertaa : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing, ICON 2020, Indian Institute of Technology Patna, Patna, India, December 18-21, 2020*, pages 127–137. NLP Association of India (NLPAI).

Oren Halvani and Lukas Graner. 2018. Cross-domain authorship attribution based on compression: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. Overview of the Authorship Verification Task at PAN 2021. In *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Mike Kestemont, Justin Anthony Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of julius caesar. *Expert Syst. Appl.*, 63:86–96.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896.

Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *J. Mach. Learn. Res.*, 8:1261–1276.

Robert Layton, Paul Watters, and Richard Dazeley. 2010. Authorship attribution for twitter in 140 characters or less. In *2010 Second Cybercrime and Trustworthy Computing Workshop*, pages 1–8.

T. C. Mendenhall. 1887. The characteristic curves of composition. *Science*, ns-9(214s):237–246.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1*

(Long Papers), pages 2227–2237. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4782–4793. Association for Computational Linguistics.

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Comput. Speech Lang.*, 70:101241.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2931–2951. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Assoc. Inf. Sci. Technol.*, 60(3):538–556.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Erico Tjoa and Cuntai Guan. 2019. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Information Retrieval Technology, Second Asia Information Retrieval Symposium, AIRS 2005, Jeju Island, Korea, October 13-15, 2005, Proceedings*, volume 3689 of *Lecture Notes in Computer Science*, pages 174–189. Springer.

G.K. Zipf. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press.

A Other quantitative results

A.1 SiamBERT and domain-adapted BERT

We list the performance of two other large pretrained BERT-based models on the PAN XL dataset splits in Tab. 6. The large gap between other models and siamBERT (**sB**) could be due to how the model functions, without learning over both documents simultaneously. BERT processes a pair of sequences, so the word-piece representations interact at every level before making a prediction based on the sequence pair embedding $h_{[CLS]}$. In contrast, siamBERT processes each sequence separately, making the word-pieces 'interact' at the end through the sequence embeddings $h^{(1)}_{[CLS]}$ and $h^{(2)}_{[CLS]}$. The domain-adapted BERT (BERT†) obtains similar results to BERT. Thus, the MLM finetuning step on Closed$_{XL}$ is not warranted, showing that adapting the representations to the domain of the downstream task brings no improvements.

A.2 Ensembling

We measure the performance of various combinations over the previously described models. In Tab. 7, we see how ensembling improves the performance on the Open UF$_{XS}$ set over the best model with over 2%. However, unexpectedly, the ensemble performance is weaker on the Open UA$_{XS}$ set. This hurts the ensemble's robustness and might be a sign of overfitting, explained by having too many similar models that collapse to the same output, failing in the same points and overwriting the better prediction.

B Datasets

B.1 PAN dataset

The PAN-2020 competition featured two datasets, a smaller one (52k pairs), intended for traditional shallow verification methods, and a larger one (275k pairs), intended for deep learning solutions. A document has an average of 21k words.

PAN XL. The large dataset has balanced classes (same vs different authors). Document pairs written by the same author always come from different fandoms (*e.g.* Star Wars vs Harry Potter), while pairs written by different authors can belong to the same fandom or to different fandoms. Same author pairs are constructed from 41k authors, while different author pairs are constructed from 251k authors, with an overlap of 14k authors in

	Closed$_{XL}$			Clopen$_{XL}$			Open UA$_{XL}$			Open UF$_{XL}$		
Metric	**sB**	**B†**	**B**	**sB**	**B†**	**B**	**sB**	**B†**	**B**	**sB**	**B†**	**B**
$F1$	85.0	94.4	**95.0**	79.8	**96.8**	**96.8**	84.1	84.0	**89.2**	83.1	**91.3**	90.8
$F0.5$	84.8	93.5	**94.5**	80.2	**97.3**	97.0	85.3	86.9	**88.1**	84.1	**91.2**	88.8
$c@1$	86.0	93.9	**94.5**	81.4	**96.6**	**96.6**	85.5	83.7	**88.2**	84.2	**90.8**	90.0
AUC	93.1	98.4	**98.6**	89.6	**99.5**	99.4	92.7	92.4	**95.3**	92.0	96.8	**96.9**
$overall$	87.2	95.1	**95.6**	82.7	**97.5**	97.4	86.9	86.7	**90.2**	85.9	**92.5**	91.6

Table 6: Comparison over large pre-trained models on PAN-2020 XL splits. BERT is very competitive and the domain-adapted BERT† does not consistently bring improvements over it. Distinctively from others, siamBERT never sees information from both documents at the same time, which significantly impacts its score. Note how the closed-set results (left side) are considerably higher than the open-set ones (right side), which might indicate the models overfit the styles of the known authors from the closed splits. We report all the PAN-2020 metrics.

	Open UA$_{XS}$					Open UF$_{XS}$				
Metric	**cB**	**sB**	**B†**	**B**	**best ensemble**	**cB**	**sB**	**B†**	**B**	**best ensemble**
$F1$	85.4	85.9	90.4	90.1	92.0	91.3	90.9	93.1	90.9	94.4
AUC	92.6	86.8	96.3	97.3	97.2	94.9	86.3	96.8	97.9	98.0
$overall$	87.0	87.2	92.1	**93.5**	93.4	91.2	88.9	93.2	93.2	**95.3**

Table 7: Ensembling results on the XS Open UA and UF splits. We show that the individual models are complementary for the Open UF$_{XS}$ set, so the $overall$ score can be improved by combining them. However, this is not the case for Open UA$_{XS}$, indicating that the models might overfit on this split, most of them collapsing to a wrong prediction.

both the same and different pairs. The XL dataset has 494k distinct documents that span 1.600 fandoms.

PAN XS. The small dataset is also balanced. Distinctly from the XL dataset, it has only cross-fandom pairs in both class pairs. This split allows fast prototyping through smaller experiments with models that have different components.

B.2 Our splits

We provide the construction details for all our splits below.

Closed split In this setup, authors and fandoms at train time are also found in the validation and test sets (but with different documents). This split can hurt generalization, because it might work only on a subset of authors or even worse, on specific document pairs. Since we have no access to the PAN 2020 test set, we make the train, validation and test sets ourselves, by splitting the original pairs. Each author pair (a_i, a_j) in the DA pairs is unique, so splitting the DA pairs such that both test authors a_i and a_j are seen at train time is impossible. However, we relax this constraint and ensure that at least one of the authors in DA test pairs is seen at train time.

Clopen split The Clopen split is similar to the closed split for the SA pairs. However, we remove the closed set constraint for the DA pairs and assign them randomly into train, validation and test. Thus, authors and fandoms in the Clopen test and validation sets might not be seen in the training set, making it a bit more general (more similar to the open sets).

Open Unseen Authors split In this split, authors from the test set should not appear in the training set. However, this is difficult to achieve strictly, so we split the PAN 2020 dataset into train and validation/test sets such that: i) authors of the SA test pairs do not appear in the SA train pairs; ii) some authors ($< 5\%$) in the DA test pairs may appear in the DA train pairs; iii) most of the fandoms in the test set appear in the training set.

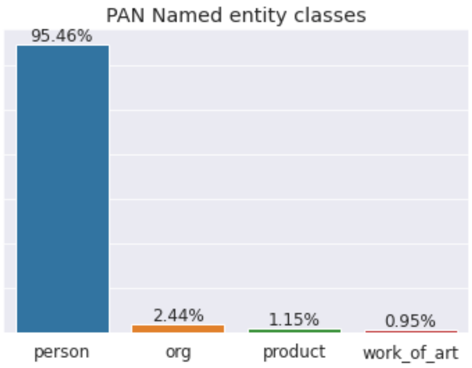


Figure 3: Distribution of PAN named entities

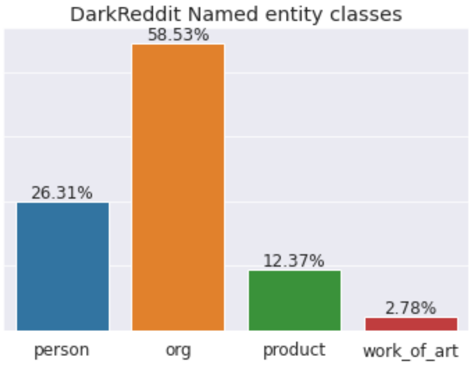


Figure 4: Distribution of DarkReddit named entities.

Open Unseen Fandoms split This split type has the following properties: i) fandoms in the validation/test sets are not seen during training; ii) some authors in the validation/test set may appear in the training set. To ensure no overlap between train and validation/test fandoms, training examples (d_1, d_2, f_1, f_2) where either f_1 or f_2 appear in the validation/test fandoms are dropped. This results in approximately *110K fewer train examples*.

Open all split This split is the most difficult and the closest to the true open set setup in PAN 2021. Distinctly from the previous four splits, which were created using the original pairs, this split required sampling new document pairs and has the following properties: i) authors and fandoms in the test set have not been seen in the training data ii) authors in the validation set have not been seen in the training set, but the validation fandoms have been seen in the training set.

B.3 Distribution of named entities

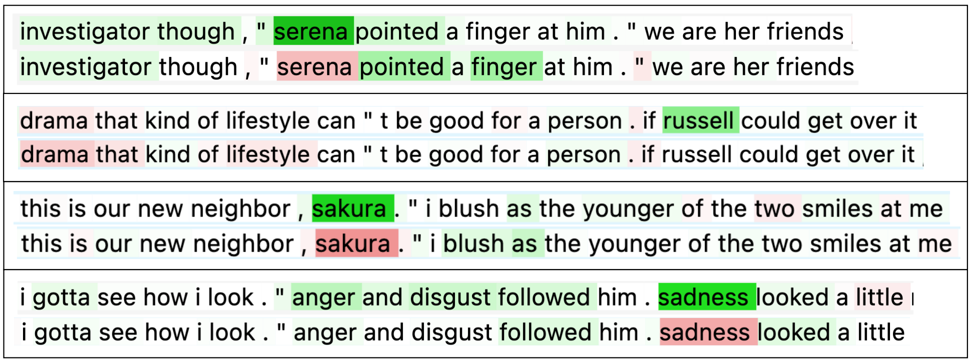
We observe in Figures 3 and 4 that the named entity distributions in the PAN and DarkReddit datasets are very different.

	Closed$_{XL}$ (split size)					Clopen$_{XL}$ (split size)				
	Total	**SA**		**DA**		**Total**	**SA**		**DA**	
Split		SF	CF	SF	CF		SF	CF	SF	CF
Train	248,322	0	133,359	22,064	92,909	248,688	0	133,359	20,945	94,384
Valid	13,449	0	7,024	356	6,069	13,093	0	7,024	1,072	4,997
Test	13,784	0	7,395	355	6,034	13,784	0	7,395	1,114	5,275

Table 8: PAN-2020 XL dataset - Closed-set splits, broken down into Same Author (SA) vs Different Author (DA). Each class is further divided into Same Fandom (SF) and Cross-Fandom (CF) pairs.

	Open UA$_{XL}$ (split size)				Open UF$_{XL}$ (split size)				Open UAll$_{XL}$ (split size)			
	SA		**DA**		**SA**		**DA**		**SA**		**DA**	
Split	SF	CF	SF	CF	SF	CF	SF	CF	SF	CF	SF	CF
Train	0	133,367	18,840	96,492	0	71,826	20,779	41,385	0	124,000	62,286	61,715
Valid	0	7,023	2,230	3,836	0	7,047	1,176	5,232	0	6,852	2,966	3,885
Test	0	7,388	2,061	4,328	0	7,056	1,176	5,233	0	6,853	1,633	5,218

Table 9: PAN-2020 XL - Open-set splits: Unseen Authors (UA$_{XL}$), Unseen Fandoms (UF$_{XL}$) and Unseen All (UAll$_{XL}$), broken down into Same Author (SA) vs Different Author (DA). Each class is further divided into Same Fandom (SF) and Cross-Fandom (CF) pairs.


investigator though , " serena pointed a finger at him . " we are her friends
investigator though , " serena pointed a finger at him . " we are her friends

drama that kind of lifestyle can " t be good for a person . if russell could get over it
drama that kind of lifestyle can " t be good for a person . if russell could get over it

this is our new neighbor , sakura . " i blush as the younger of the two smiles at me
this is our new neighbor , sakura . " i blush as the younger of the two smiles at me

i gotta see how i look . " anger and disgust followed him . sadness looked a little
i gotta see how i look . " anger and disgust followed him . sadness looked a little

Figure 5: Explainability analysis using Integrated Gradients for other samples, when fine-tuning BERT on Closed$_{XL}$ and Open UF$_{XL}$ respectively.