Transparent Authorship Verification with Machine Learning Models

Stephanie Wan*
Lexington High School
Lexington, MA 02421
stephaniewan07@gmail.com

Abstract

Authorship Verification (AV) is the task of determining if two given documents were written by the same person. AV is critical in addressing issues such as misinformation and impersonation, though it holds risks in violating privacy rights. This paper presents a publicly accessible website hosting transparent AV machine learning models. We aggregate and pre-process diverse datasets to train a lexical model based on embeddings and a stylometric model leveraging feature vectors. To enhance model transparency, we incorporate attention-based highlighting and output important features. The code and website for this paper are available at https://github.com/swan-07/authorship-verification and https://same-writer-detector.streamlit.app/.

1 Introduction

Authorship Verification (AV) is the task of determining if two given documents were written by the same person. AV can have a positive social impact in determining the identity of individuals to "counter misinformation, plagiarism, and inappropriate aggregation", "harassment, impersonation, or criminal activities", as well as potentially distinguishing falsified or AI-generated text (Nguyen et al., 2023). This is especially important as the internet, which facilitates anonymity, is playing an ever-growing role in our lives and societies.

AV approaches commonly focus on linguistics or stylometry, and developments in Transformers and other large language models have been shown to provide competitive results on AV tasks (Nguyen et al., 2023). According to Nguyen et al. (2023), training data diversity can increase the quality and robustness of AV models, as well as their ability to generalize in open environments (environments where testing authors may not appear in the training dataset, such as the real world).

In this paper, we compile and process an array of AV and Authorship Attribution (AA, the task of identifying the author of a given document) datasets, and develop and deploy a website for accessibility using two AV models: a linguistic model in the form of the Embedding model and a stylometric model in the form of the Feature Vector model from Weerasinghe et al. (2021). We increase model transparency with the usage of attention-based highlighting and outputting important features. See Figure I below for an overview of our pipeline.

2 Data

In order to improve our model's ability to generalize, we train on a multitude of different AV and AA datasets (12 in total) of varying lengths, contexts, formalities, and topics. Some authors have shown

*swan-07.github.io

Preprint. Under review.

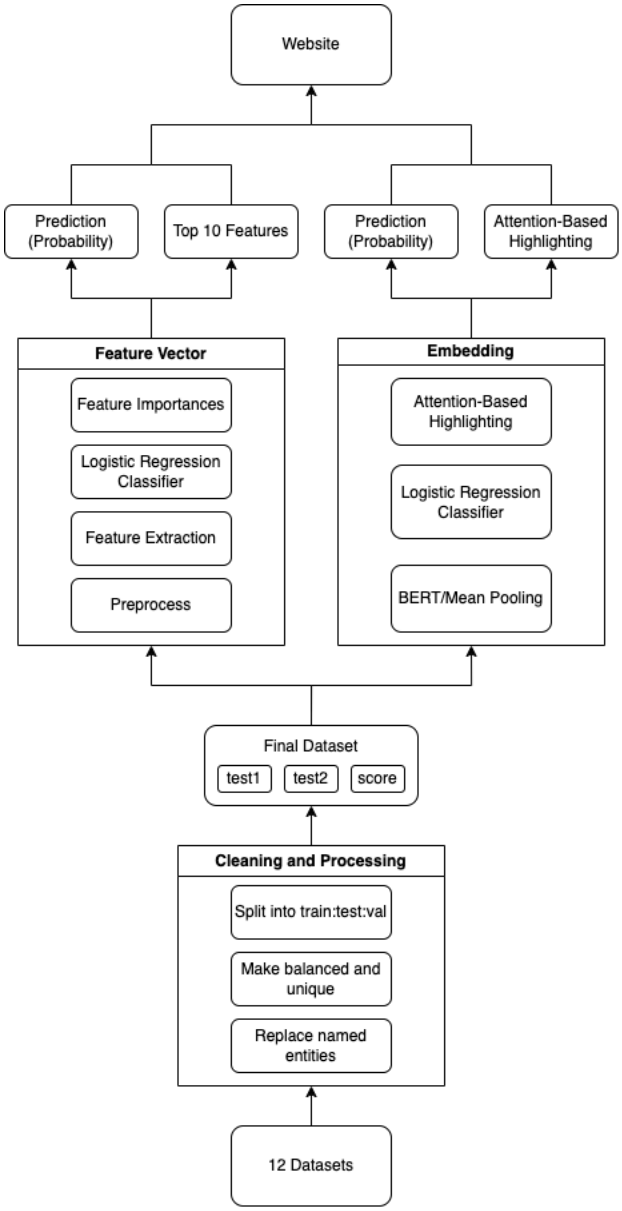

Figure 1: Overall pipeline

that AV models inferred results based on topical information rather than authorship characteristics, so for many of the datasets, we replaced named entities with their corresponding type (Alice -> PERSON) to improve model generalization and mitigate decisions based on topical information using spaCy's EntityRecognizer (Brad et al., 2022).[2] Some datasets' named entities were not replaced to allow our model to see training data with named entities, as shown in Table 3 below.

Unless otherwise specified, the train:test:val split for each dataset was 0.7:0.15:0.15 and datasets were processed to be "balanced" and "unique", with equal amounts of same and different text pairs of unique text pairs (where every text only appears in one pair).

An overview of the datasets can be seen in Table 3. A compiled list of our processed datasets, as well as links to access the original datasets, can be accessed on HuggingFace.[3]

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

[3]https://huggingface.co/datasets/swan07/authorship-verification

Table 1: Summary of AV and AA datasets used

Dataset	Text Form	Pairs Used*	Average Chars**	Length***	Formality	Named Entities
Reuters	News Articles	1202	2770	Medium	Formal	Replaced
Blogs	Blog Posts	58930	1086	Short	Informal	Replaced
Victorian	Book Excerpts	10718	4923	Long	Formal	Replaced
arXiv	Paper Abstracts	704	803	Short	Formal	Replaced
DarkReddit	Reddit Comments	1028	2751	Medium	Informal	Replaced
BAWE	Student Writing	1150	14702	Long	Formal	Replaced
IMDB62	Movie Reviews	30982	1668	Short	Informal	Replaced
PAN11	Enron Emails	4650	300	Short	Mix	Replaced
PAN13	Various	120	7143	Long	Mix	Replaced
PAN14	Novels and Essays	900	15843	Long	Formal	Not Replaced
PAN15	Various	1265	3167	Medium	Mix	Not Replaced
PAN20	Fanfiction	275409	21473	Long	Informal	Not Replaced

*The amount of pairs of text from this dataset that were in the final compiled dataset
**Average amount of characters in each text used in the final compiled dataset
***General length of each text in the dataset

2.1 Datasets

Reuters50 contains Reuters articles with at least one subtopic of corporate/industrial to minimize topic differences (Liu, 2016).

The Blog Authorship Corpus (Blogs) consists of blog posts collected from blogger.com in August 2004 (Schler et al., 2006). We get rid of rows with only whitespace and duplicates.

Victorian comprises excerpts from books of prolific English language 19th century authors (Gungor, 2016).

arXiv includes single-author paper abstracts from authors with at least 10 papers, focusing on machine learning and computer science (Moreo, 2022).

DarkReddit (open) is a dataset of samples from /r/darknet3 (Brad et al., 2022). This is an open-set dataset, meaning that the authors in test and validation are unseen (not present) in the train split. This dataset can thus test the generalizability of our model. We keep the original splits.

British Academic Written English (BAWE) is composed of proficient British student writing, varying in length and fairly evenly distributed across "Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences... [and] undergraduate and taught masters level" (Nesi et al., 2008).

IMDB62 is a dataset of IMDB reviews from 2009 by reviewers with more than 500 submissions (Seroussi et al., 2014).

We also have a variety of datasets from PAN throughout the years. PAN is a series of scientific events and shared tasks on digital text forensics and stylometry, including authorship verification (PAN, n.d.).

PAN11 is a dataset based on the Enron email corpus (Argamon & Juola, 2011). We use texts from the LargeTrain split.

PAN13 is a dataset with texts of a variety of themes, formality, and lengths and languages (Juola & Stamatatos, 2013). We use the given train and test splits (swapped, so we use their test split as our train split and vice versa), comparing the first known text to the unknown text, and split our test split in half to make test and evaluation splits.

PAN14 consists of novels and essays in several languages/genres (Stamatatos et al., 2014). We take the English texts. To provide variety, entity names were not removed from this dataset.

PAN15 consists of cross-topic and cross-genre documents from a variety of languages: Dutch, English, Greek, and Spanish (Stamatatos et al., 2015). We take all documents from all languages, swap their test and train split, and split our test split in half to make test and validation splits.

PAN20 contains English documents from fanfiction.net (Bevendorff et al., 2020). We use the XL open-all split from Brad et al. (2022) and Kestemont et al. (2020), where "authors and fandoms in the test set have not been seen in the training data" and "authors in validation set have not been seen in the training set, but validation fandoms are similar to the training fandoms".

3 Models

Our website hosts modified versions of two AV models: a stylometric "feature vector" approach, as detailed in Weerasinghe et al. (2021), and an embedding model. Models were trained on RunPod with an A100 SXM.

3.1 Feature Vector

We implemented a Feature Vector model, as described in Weerasinghe et al. (2021). This model was chosen for its high performance in the PAN2021 AV shared task. They "modeled [AV] as a binary classification problem, in which the input to our classifier is a feature vector encoding the two documents and the target variable, indicating whether or not the two documents were written by the same author" (Weerasinghe et al., 2021).

Each text in the dataset went through a pre-processing and feature extraction. Pre-processing outputs were stored with the original text when passed to feature extraction.

Pre-processing included tokenization, Part-of-Speech (POS) Tagging, and POS Tag Chunking and features extracted consisted of common stylometric features such as Character n-grams, POS-Tag n-grams, Special Characters, Frequency of Function Words, Average number of characters per word, Distribution of word-lengths, Vocabulary Richness, POS-Tag Chunks, POS chunk construction, Stop-word and POS tag hybrid tri-grams, Part-of-Speech tag ratios, and Unique spellings (Weerasinghe et al., 2021).

Features were standardized by removing the mean and scaling to unit variance. The absolute vector difference between the feature vectors for each text pair were also scaled, and the standardized result was fed to the classifier. The classifier used was a Logistic Regression classifier using a Stochastic Gradient Descent training algorithm with a logarithmic loss function, implemented through SKLearn's SGDClassifier.

We extracted the top 10 features by importance (multiplying differences between two texts' feature vectors with the coefficients of the features) for each comparison between two texts and outputted the classifier's probability of outputting a value of 1 (determining the two texts had the same author) as our probability. Our metrics are shown in Table II below.

3.2 Embedding

BERT was chosen for its high performance in AV tasks, as seen in Barlas and Efstathios Stamatatos (2020) and Prasad and Chakkaravarthy (2022). We followed the implementation by Prasad and Chakkaravarthy (2022), using BERT as the base for a Siamese network trained with contrastive loss. However, after training on 18000 texts we found the loss had plateaued (Figure II). We found that performance was poor and similar to mean pooling (Figure III, Table III), so we used mean pooling in future steps.

We trained a Logistic Regression Classifier using SKLearn to calibrate the cosine similarity score between embeddings to a probability. We used attention weights for attention-based highlighting on our texts, as visualized attention can provide transparency behind the predictions of the network.

Table 2: Feature Vector model metrics

Dataset	AUC	c@1	$F_{0.5}$	F1	Brier	Overall
Evaluation	0.646	0.599	0.606	0.653	0.627	0.626
Arxiv	0.658	0.660	0.696	0.710	0.678	0.680
Blogs	0.594	0.574	0.568	0.550	0.594	0.576
British	0.744	0.705	0.719	0.771	0.732	0.734
DarkReddit	0.772	0.716	0.716	0.717	0.738	0.732
IMDB	0.732	0.672	0.680	0.630	0.700	0.683
PAN11	0.479	0.501	0.528	0.536	0.517	0.512
PAN13	0.569	0.444	0.486	0.583	0.479	0.512
PAN14	0.605	0.598	0.594	0.566	0.627	0.598
PAN15	0.629	0.565	0.596	0.636	0.601	0.605
PAN20	0.693	0.620	0.621	0.698	0.653	0.657
Reuters	0.565	0.525	0.548	0.613	0.557	0.562
Victorian	0.549	0.508	0.548	0.640	0.523	0.554

For individual datasets, we use their test split.

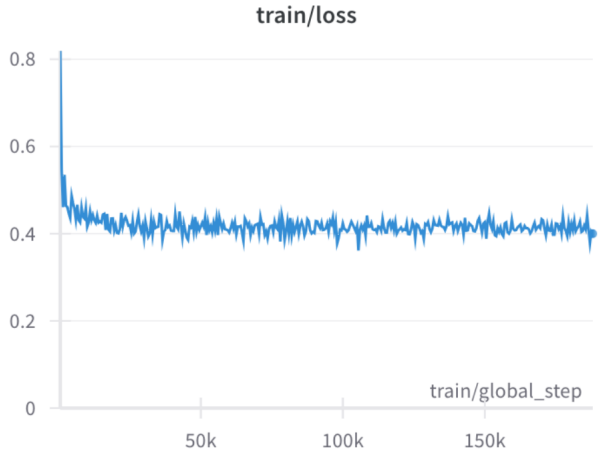

Figure 2: BERT finetuned model training loss

3.3 Website

We deployed our models on a Streamlit website accessible to the public where users can input two texts and receive predictions from the embedding model with attention based highlighting and predictions from the Feature Vector model with a list of important features.[4]

4 Discussion and Conclusion

In this paper we presented the approach for transparent AV models and deployed a website for interacting with our models. We aggregated and processed datasets in the AA and AV fields, and implemented AV models and transparency strategies. In creating our website, we hope to increase accessibility and transparency behind AV methods.

Our BERT model performed poorly due to overfitting. The Feature Vector model performed best on the BAWE and DarkReddit datasets and worst on the PAN11 and PAN13 datasets, which was suprising: BAWE is formal and DarkReddit is informal, BAWE and PAN13 are long while DarkReddit is medium and PAN11 is short, and more pairs were used from the PAN13 dataset than from BAWE

[4]https://same-writer-detector.streamlit.app/

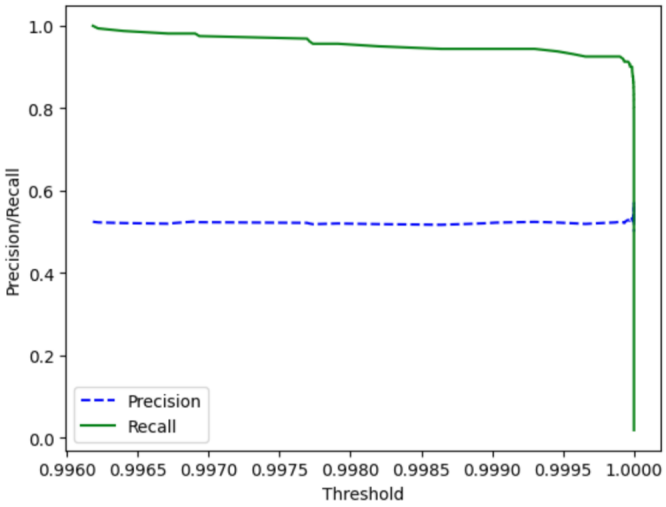

Figure 3: BERT finetuned model precision/recall curve

Table 3: Trained BERT model metrics

Dataset	Accuracy	Precision	Recall	F1 Score
Evaluation	0.524	0.524	1.00	0.688
Reuters	0.492	0.492	1.00	0.659
Blogs	0.498	0.498	1.00	0.665
Victorian	0.492	0.492	1.00	0.659
arXiv	0.566	0.566	1.00	0.723
DarkReddit	0.500	0.500	1.00	0.667
BAWE	0.498	0.498	1.00	0.665
IMDB62	0.524	0.524	1.00	0.688
PAN11	0.444	0.444	1.00	0.615
PAN13	0.500	0.500	1.00	0.667
PAN14	0.530	0.530	1.00	0.693
PAN15	0.500	0.500	1.00	0.667
PAN20	0.500	0.500	1.00	0.667

For individual datasets, we use their test split.
We found the metrics for our test datasets to be the same as the mean pooling metrics.
Numbers are rounded to three significant figures.

or DarkReddit. This potentially demonstrates the effect of entity removal and the feature vector method on the model's ability to generalize.

4.1 Ethics

Although AV's ability to de-anonymize individuals has many benefits, as seen in the Introduction, AV also brings many concerns, such as a lack of privacy and anonymity and potential repression by authorities. Thus, any authorship identification technologies must be handled with care due to their potential for negative social impacts. We chose to study AV rather than AA due to AV being more respectful of overall privacy: given a text with an anonymous author, AV would need to be applied on every possible suspect, which allows it to be useful when the set of suspects is small, but not when the set of suspects is very large, preserving privacy concerns. We believe our website does not pose many societal harms due to the lower performance of our models, and confers benefits by increasing accesibility to transparent AV models.

4.2 Limitations

Some limitations of our project were a lack of compute and time. Given more time and compute, we could train our models for longer and have a more robust dataset. Due to these limits we were unable to train our models fully (as many epochs as the original papers recommended). Our trained BERT model was overfit and performed poorly—future papers could implement methods such as dropout to mitigate overfitting. Future work could also aim to have greater accuracy in AV models as our models had poorer accuracy.

Another limitation is that most texts in our dataset were in English, limiting our model's ability in other languages.

Acknowledgments and Disclosure of Funding

We thank Dr. Gil Alterovitz and Ning Xie from MIT PRIMES and Dr. Manesh Gani and Joanna Gilberti for their feedback and support during the research process. Our work was supported by a grant from Trelis AI Grants. Authors declare no competing interests.

References

Argamon, S., & Juola, P. (2011). PAN11 Author Identification: Attribution [Data set]. In CLEF 2011 Labs and Workshops, Notebook Papers. PAN at Conference and Labs of the Evaluation Forum 2011 (PAN at CLEF 2011). Zenodo. https://doi.org/10.5281/zenodo.3713246

Barlas, G., & Stamatatos, E. (2020). Cross-Domain Authorship Attribution Using Pre-trained Language Models. IFIP Advances in Information and Communication Technology, 255–266. https://doi.org/10.1007/978-3-030-49161-1_22

Bevendorff, J., Kestemont, M., Stamatatos, E., Manjavacas, E., Potthast, M., & Stein, B. (2020). PAN20 Authorship Analysis: Authorship Verification (0.0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5106099

Brad, F., Manolache, A., Burceanu, E., Barbalau, A., Ionescu, R. T., & Popescu, M. (2022). Rethinking the Authorship Verification Experimental Setups. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2022.emnlp-main.380

Corbara, S., Moreo, A., & Sebastiani, F. (2023, January 24). Same or Different? Diff-Vectors for Authorship Analysis. ArXiv.org. https://doi.org/10.48550/arXiv.2301.09862

Gungor, A. (2016). Victorian Era Authorship Attribution. UCI Machine Learning Repository. https://doi.org/10.24432/C5SW4H

Ibrahim, M., Akram, A., Radwan, M., Ayman, R., Abd-El-Hameed, M., El-Makky, N., & Torki, M. (2023). Enhancing Authorship Verification using Sentence-Transformers Notebook for PAN at CLEF 2023. https://downloads.webis.de/pan/publications/papers/ibrahim_2023.pdf

Juola, P., & Stamatatos, E. (2013). PAN13 Author Identification: Verification [Data set]. In CLEF 2013 Labs and Workshops, Notebook Papers. Conference title: PAN at Conference and Labs of the Evaluation Forum 2013 (PAN at CLEF 2013). Zenodo. https://doi.org/10.5281/zenodo.3715999

Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M., & Stein, B. (2020). Overview of the Cross-Domain Authorship Verification Task at PAN 2020. CLEF (Working Notes).

Liu, Z. (2016). Reuter_50_50. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/217/reuter+50+50

Moreo, A. (2022). arXiv abstracts and titles from 1,469 single-authored papers (100 unique authors) in computer science [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7404702

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2008). British Academic Written English Corpus. Ota.bodleian.ox.ac.uk. https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2539

Nguyen, T., Alperin, K., Dagli, C., Vandam, C., & Singer, E. (2023). Improving Long-Text Authorship Verification via Model Selection and Data Tuning (pp. 28–37). https://aclanthology.org/2023.latechclfl-1.4.pdf

PAN. (n.d.). Pan.webis.de. Retrieved November 3, 2022, from https://pan.webis.de/

Prasad, R. S., & Chakkaravarthy, M. (2022). State of the Art in Authorship Attribution With Impact Analysis of Stylometric Features on Style Breach Prediction. Journal of Cases on Information Technology, 24(4), 1–12. https://doi.org/10.4018/jcit.296716

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of Age and Gender on Blogging. Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. http://www.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship Attribution with Topic Models. Computational Linguistics, 40(2), 269–310. https://doi.org/10.1162/coli_a_00173

Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2015). PAN15 Author Identification: Verification [Data set]. In CLEF 2015 Labs and Workshops, Notebook Papers. Conference title: PAN at Conference and Labs of the Evaluation Forum 2015 (PAN at CLEF 2015). Zenodo. https://doi.org/10.5281/zenodo.3737563

Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., Sanchez-Perez, M. A., & Barrón-Cedeño, A. (2014). PAN14 Author Identification: Verification [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3716033

Weerasinghe, J., Singh, R., & Greenstadt, R. (2021). Feature Vector Difference based Authorship Verification for Open-World Settings Notebook for PAN at CLEF 2021. https://ceur-ws.org/Vol-2936/paper-197.pdf