

Team OpenFact at PAN 2024: Fine-Tuning BERT Models with Stylometric Enhancements

Notebook for the PAN Lab at CLEF 2024

Ewelina Książniak¹, Krzysztof Węcel¹ and Marcin Sawiński¹

¹Poznań University of Economics and Business, Al. Niepodległości 10, 61-875 Poznań, Poland

Abstract

This paper presents our solution for the Multi-Author Style Change Detection task at PAN 2024. The task involves detecting paragraph-level writing style changes in texts, with datasets classified into easy, medium, and hard difficulty levels. We incorporated stylometric tags directly into the text to enhance the sensitivity of BERT-family models to stylistic features. Our approach aimed to improve the model's detection of authorship changes by adding these tags to the training dataset and model sensitivity to stylometric features. The results showed F1 improvements when training on smaller datasets, indicating the method's potential for hard-to-obtain data types.

Keywords

Stylometric Analysis, Style Change Detection, BERT Models

1. Introduction

Multi-author style change detection has been a task organized by PAN since 2016 [1]. Prior to the advent of BERT models, style-change detection techniques predominantly relied on traditional stylometric features, including lexical elements (n-grams), word frequencies, and syntactic characteristics like parts of speech or syntactic trees. For instance, in 2018, the leading approach for cross-domain authorship attribution task involved text distortion and extraction of character n-grams to emphasize punctuation, numbers, and diacritic characters [2], as demonstrated by [3]. The top performance in the 2018 style detection task was achieved by [4], who utilized features such as repetition, contracted wordforms, frequent words, quotation marks, vocabulary richness, and readability to train an ensemble classifier. However, starting from 2020, most participants have submitted solutions by fine-tuning pre-trained models [5]. For example, in the 2023 edition, the highest accuracy on easy and medium datasets was achieved using BERT, RoBERTa, and ELECTRA combined with a binary classification layer [6].

This paper describes the solution submitted for the Multi-Author Style Change Detection task, part of the PAN 2024 workshop series. This task aims to identify paragraph-level writing style changes in a given text between consecutive paragraphs. It includes three levels of difficulty: easy, medium and hard [1]. For each subtask, distinct datasets were provided: 1) Easy – paragraphs cover various topics, allowing topic information to aid in detecting authorship changes; 2) Medium – there is limited topical variety, requiring a greater emphasis on stylistic differences; 3) Hard – all paragraphs cover the same topic, thus relying solely on stylistic cues to identify changes. The entire dataset was in English and sourced from comments on Reddit. It included metadata indicating the points of author change between paragraphs, as well as the total number of authors within each set [7].

Given the stylometric nature of the task and the importance of stylistics in detecting authorship changes, we decided to employ a method that directly adds stylometric tags to the texts in the training dataset used for training models from BERT-family. Our approach aims to enhance the model's sensitivity to stylistic features, acknowledging that in authorship change detection, semantic content alone could be insufficient. This paper presents our methodology and findings, offering insights into the effectiveness of

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

✉ ewelina.ksiezniak@ue.poznan.pl (E. Książniak); krzysztof.wecel@ue.poznan.pl (K. Węcel); marcin.sawinski@ue.poznan.pl (M. Sawiński)

ORCID 0000-0003-1953-8014 (E. Książniak); 0000-0001-5641-3160 (K. Węcel); 0000-0002-1226-4850 (M. Sawiński)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

using proposed stylometric enhancements in training language models for authorship change detection. Additionally, it presents background studies aimed at determining whether the proposed method enhances the sensitivity of BERT-family language models to stylometric features.

2. Background

2.1. Background studies

To explain the rationale behind the proposed method, we conducted an additional experiment to determine which stylometric features are important in the author style change detection task. We calculated stylometric features that describe text complexity, text formality, text fogginess, and patterns related to punctuation and grammar. Subsequently, we computed the absolute values of differences for specific features between text pairs created from consecutive paragraphs. We hypothesized that absolute differences between pairs of texts would be smaller when there is no change in authorship and larger in the case of authorship change.

To assess the statistical significance of the differences in data distribution for each specific feature, we employed the Mann-Whitney U test. This is a non-parametric test used to determine whether there is a difference between two groups (i.e., these with changing and non-changing authorship) by comparing the rank sums of the two samples rather than the means. The null hypothesis (H0) states that the distributions of the two groups are identical, while the alternative hypothesis (H1) suggests that the distributions differ [8]. We chose this test because our data did not follow a normal distribution which is required for the t-test for independent samples. Furthermore, we analyzed differences across two dimensions: real labels (0 and 1), predicted labels (0 and 1).

The background experiment was conducted on the validation datasets for the specific subtasks, using fine-tuned RoBERTa models, which served as our internal baselines and based on the following features: the number of sentences in the text, the average number of words per sentence, the count of punctuation marks, the count of personal words, the count of reported speech, the formality score, the Flesch Reading Ease score, the SMOG index, the Flesch-Kincaid Grade level, the Coleman-Liau index, the Automated Readability Index, the count of difficult words, and the frequencies of nouns, verbs, adjectives, adverbs, and prepositions.

Table 1 presents normalized mean absolute differences obtained for easy, medium, and hard validation datasets for features that showed statistically significant differences across the authorship change (label: 1) and no authorship change (label: 0) groups within actual and predicted labels. Most measures consistently exhibited higher absolute differences when there was a change in authorship, evident in both real and predicted label distributions. Surprisingly, for the medium and hard datasets, higher diversity was observed for some features without author changes. For the medium dataset, this was seen in sentence complexity, and the frequency of nouns and verbs. For the hard dataset, it was noted in the frequency of nouns and adverbs, as well as the Coleman-Liau readability index.

Table 1

Comparison of mean absolute difference for groups: no authorsip change (label: 0), authorship change (label: 1)

Dataset	Features	Real label		Prediction	
		0	1	0	1
Easy	sentence complexity	0.047	0.090	0.047	0.090
	punctuation	0.018	0.027	0.017	0.027
	nouns frequency	0.132	0.191	0.133	0.191
	difficult words frequency	0.079	0.140	0.074	0.140
	verbs frequency	0.175	0.242	0.177	0.242
	adverbs frequency	0.132	0.149	0.131	0.149
	coleman liau index	0.003	0.008	0.003	0.008
	automated readability index	0.006	0.019	0.006	0.019
	text length	0.028	0.065	0.027	0.065
	personal words	0.021	0.052	0.020	0.052
	adjectives frequency	0.096	0.114	0.098	0.114
Medium	sentence complexity	0.086	0.078	0.085	0.078
	punctuation	0.041	0.062	0.042	0.064
	nouns frequency	0.226	0.250	0.225	0.254
	verbs frequency	0.276	0.208	0.262	0.213
	adverbs frequency	0.136	0.182	0.146	0.178
	coleman liau index	0.134	0.186	0.140	0.187
	personal words	0.020	0.038	0.021	0.039
	formality	0.198	0.230	0.194	0.239
	prepositions frequency	0.205	0.228	0.209	0.227
Hard	sentence complexity	0.121	0.111	0.225	0.231
	punctuation	0.055	0.063	0.055	0.062
	nouns frequency	0.167	0.161	0.166	0.163
	verbs frequency	0.202	0.210	0.150	0.131
	adverbs frequency	0.182	0.180	0.180	0.182
	coleman liau index	0.176	0.157	0.175	0.161
	personal words	0.049	0.050	0.050	0.049
	formality	0.203	0.220	0.202	0.218
	prepositions frequency	0.175	0.181	0.175	0.180

3. System Overview

3.1. Methodology

Based on the results presented in [9], there are indications that models from BERT-family capture certain stylometric features. Building on these observations, we developed a method that enriches the text by directly incorporating stylometric tags. This approach aims to determine if this enhancement can improve model classification and make BERT family models more sensitive to stylometric characteristics.

The experiment was carried out in four phases:

1. Fine-tuning of models and selection of baseline models.
2. Feature engineering.
3. Training models with data augmented by stylometric tags on entire dataset and subsamples.
4. Conducting experiments by combining multiple tags within a single text.

3.1.1. Baseline models selection

For each dataset variant (easy, medium, hard), we fine-tuned the RoBERTa base and DeBERTa v3 base models in the initial phase and selected baseline models based on the results obtained. We experimented

with hyperparameters, including learning rates of 1e-5, 2e-5, and 3e-5, and seed values of 42, 100, and 1111. The default AdamW optimizer and a batch size of 4 were used. Additionally, we tested an approach implementing layer-wise decay in the RoBERTa base model, applying different learning rates to each layer to capture either general language information or task-specific details.

Each model was trained for 5 epochs using the original dataset provided by the task organizers. The data preparation involved concatenating two consecutive paragraphs with a separator token. After completing this phase, we chose the fine-tuned RoBERTa-base model as the baseline for the second part of the experiment. The training was conducted on server equipped with four NVIDIA GeForce RTX 2080 Ti GPU cards, each with 11 GB of memory.

3.1.2. Features Engineering

In the second iteration, we enhanced the original datasets by integrating stylometric feature tags. The features were chosen based on a manual review of prediction errors from the baseline models.

We decided to augment the dataset by adding tags related to the following stylometric dimensions:

- text complexity
- text formality
- punctuation.

To quantify text complexity, we developed two metrics: **text length**, determined by counting the number of sentences, and **sentence complexity**, calculated as the average number of words per sentence.

Text formality was evaluated using the method proposed by [9]. This method introduced “seed words” with the same semantic meaning but different levels of formality. The seed pairs included among others: *my gosh* - *Jesus*, *breathing* - *respiratory*, *yeah* - *yes*, *ten years* - *decade*, *first of all* - *foremost*, *a whole bunch* - *full*, and *my dad* - *father*. The method involved calculating the mean difference between the word embeddings of each pair to create a “stylometric embedding”. The formality level of a text was then determined by measuring the cosine similarity between the input text embedding and the “stylometric embedding”. We also created features to measure text formality by analyzing **reported speech occurrences** and **personal style**. The degree of personal style was measured by the frequency of words used to express personal opinions or experiences (e.g., I, me, my).

We also analyzed **punctuation patterns**, focusing specifically on infrequently used punctuation marks: the ampersand, ellipsis, question mark, quotation mark, and semicolon.

To generate tags for the dataset based on text length, sentence complexity, and formality, we computed descriptive statistics: mean, standard deviation, and quantiles across the entire training dataset. These statistics were then used to establish thresholds for embedding stylometric tags into the original text.

For text length, we prefixed the original text with *The text is long.* if it contained at least three sentences and with *The text is short.* if it contained only one sentence. For sentence complexity, we added phrase *The text contains long sentences.* at the beginning if the average sentence length exceeded 21 words, and *The text contains short sentences.* if it was below 15 words. For formality, we used the phrase *The text is highly informal.* if the formality measure exceeded 0.2, and *The text is formal.* if it was below 0.05. Additionally, we added the tag *This text contains reported speech* if any reported speech patterns were detected.

To generate tags related to punctuation and personal style, we added specific tag when a designated word or punctuation mark occurred in the text. For personal style, we identified words such as “I”, “me” and “my”. For punctuation, we looked for marks including the ampersand, ellipsis, question mark, quotation mark, and semicolon.

3.1.3. Training models with data augmented by stylometric tags on entire dataset and subsamples

In the subsequent phase, we trained models using datasets augmented with specific tags. To evaluate the impact of these tags, we used the same hyperparameters as those employed for the baseline models.

To expedite the training process, initial experiments were conducted on a randomly reduced dataset comprising of 10,000 observations.

Following these preliminary experiments, we proceeded to train on the entire dataset, exploring several variants: datasets modified by the addition of tags: the number of sentences, average words per sentence, punctuation, personal style, reported speech, and formality level, each tested separately. Additionally, models were trained on data incorporating various combinations of these tags.

Initially, we combined all engineered tags; however, this approach introduced excessive noise into the data. Given the RoBERTa base model's token limit of 512, we then tested combinations of tags related to personal words and punctuation, which were relatively short. As a result, three additional models were trained for each task level (easy, medium, and hard) in this phase.

3.2. Text Augmentation

Using TIRA [10], we opted to make only the final submission for models that demonstrated the best performance during the internal testing phase. The software used for the final submission processes a pair of texts as input and adds different tags depending on the dataset level:

- Hard dataset: The input text pairs are modified by adding a punctuation tag.
- Medium dataset: The input text pairs are modified by adding a tag related to sentence complexity (average words per sentence).
- Easy dataset: The input text pairs are modified by adding both the punctuation tag and the tag related to personal words

Here are examples of original and tagged text using the proposed approach for easy, medium, and hard datasets:

Modification by adding punctuation and personal style tag

Original text:

Why did I think hemp production started earlier? I wonder if it was just government controlled back then and the 2014 farm bill opened it up more for private business...

Tagged text:

Why did I (personal style) think hemp production started earlier? (question mark) I (personal style) wonder if it was just government controlled back then and the 2014 farm bill opened it up more for private business... (ellipse mark)

Modification by adding only a sentence complexity tag

Original text:

If the Russian soldiers (or their wives back in Russia) hear this, it could keep the already low morale of the Russian solders low...

Tagged text:

The text contains long sentences. If the Russian soldiers (or their wives back in Russia) hear this, it could keep the already low morale of the Russian solders low...

Modification by adding only a punctuation tag

Original text:

The tu quoque defense (Latin for 'you too') asserts that the authority trying a defendant has committed the same crimes of which they are accused. It is related to the legal principle of clean hands, reprisal, and "an eye for an eye". The tu quoque defense does not exist in international criminal law and has never been accepted by an international court.

Tagged text:

The tu quoque defense (Latin for 'you too') asserts that the authority trying a defendant has committed the same crimes of which they are accused. It is related to the legal principle of clean hands, reprisal, and

(quotation mark) "an eye for an eye" (quotation mark). The tu quoque defense does not exist in international criminal law and has never been accepted by an international court.

After preprocessing the text pairs according to a specific schema, the system makes predictions using models trained on tagged versions of the dataset. Each subtask utilizes a separate model, and the training methodology is detailed in Section 3.1.

4. Results

4.1. Internal testing phase

During our internal testing phase, we trained the models on subsamples and the entire dataset (separately for easy, medium, and hard tasks) on original and tagged data. Table 2 presents the macro F1 scores for hard, medium, and easy datasets obtained by training on a randomly selected subsample (10,000 observations). The results indicate that incorporating stylometric tags - in most cases - led to an improvement in the F1-macro score, with the best results achieved by adding reported speech tag and text length tag for hard, and easy datasets, respectively. Augmenting the dataset with stylometric tags improved the macro F1 score by 2% and 1% for the hard, and easy datasets, respectively. Adding tags for the medium dataset does not affect the results, except for the sentence complexity tag, which decreases the macro F1 score by approximately 1%.

Table 2

Comparison of macro F1 score achieved on the validation datasets by training on the subsamples of training datasets for easy, medium and hard subtasks

	hard	medium	easy
baseline	0,75	0,82	0,97
personal	0,76	0,82	0,97
sentence complexity	0,76	0,81	0,96
text length	0,74	0,82	0,98
formality	0,76	0,82	0,96
reported speech	0,77	0,82	0,97

Table 3 presents the classification results for the hard, medium, and easy validation datasets trained on the entire tagged dataset, alongside the baseline model outcomes. Notably, the impact of adding tags was significantly lower than the results achieved by training on dataset subsamples. The best results for the hard dataset were obtained by adding the punctuation tag, achieving an macro F1 score of 0.813 on validation dataset, while adding other tags surprisingly worsened the baseline results. For the medium dataset, the most significant improvement was observed compared to the baseline, with the highest difference from 0.819 to 0.836. Each tag or combination of tags improved over the baseline results. For the easy dataset, the best results were achieved by adding personal style, formality level, and punctuation tags, as well as a combination of punctuation and personal tags. However, the improvement over the baseline was only 0.001.

Building on previous studies, we aimed to determine whether incorporating stylometric tags directly into the text enhances model sensitivity to stylometric features. Our objective was to establish whether the observed improvements from adding stylometric features were genuinely due to increased model sensitivity to specific stylometric aspects (e.g., better understanding of punctuation marks and their significance in detecting author style changes) or if other factors, such as randomness, played a role.

To test this, we validated the distribution of mean absolute differences between the real labels for no authorship change group and the correct predictions for no authorship change group in both baseline models and those trained on the tagged dataset. The assumption was that if adding tags enhances the model's sensitivity to stylometric features, the difference between the trend observed for actual labels and model predictions would be smaller for the model trained with tags.

Table 3

Comparison of macro F1 score achieved on the validation datasets by training on the entire training datasets for easy, medium and hard subtasks

	hard	medium	easy
Baseline	0.810	0.820	0.982
Number of sentences	0.792	0.825	0.978
Average words in sentences	0.778	0.836	0.982
Punctuation	0.813	0.827	0.980
Personal style	0.807	0.825	0.983
Reported speech	0.798	0.829	0.982
Formality level	0.781	0.824	0.983
Punctuation and personal tags combined	0.795	0.834	0.983

Table 4 presents the results of the mean absolute differences across two dimensions: real labels (0) and correct predictions (0) from baseline models and models trained on tagged data. Surprisingly, models trained on tagged datasets showed higher differences between predictions and real labels compared to baseline models. This suggests that the observed improvement in macro F1 scores may not be due to increased sensitivity to stylometric features, but rather other factors.

Table 4

Comparison of Mean Absolute Differences between Correct Predictions and Real Labels for Tagged and Standard Datasets

Dataset/Feature	0 (actual label)	0 (correct prediction)	Difference
Hard/punctuation tagged	0.663	0.666	0.003
Hard/punctuation standard	0.663	0.661	0.002
Medium/sentence complexity tagged	11.050	11.463	0.414
Medium/sentence complexity standard	11.050	11.078	0.028
Easy/personal style tagged	0.413	0.375	0.038
Easy/personal style standard	0.413	0.388	0.024

4.2. Final submission

Table 5 presents the results from the wary-pita system, which was our final submission via Tira, evaluated on training, validation, and test datasets. The results on the unseen test dataset are slightly lower than on the validation set. However, our internal testing indicates that the method has potential for further improvements.

Table 5

Performance metrics across different datasets and difficulty levels

Dataset	Easy	Medium	Hard
Train dataset	0.999	0.896	0.922
Validation dataset	0.983	0.836	0.813
Test dataset	0.981	0.821	0.805

5. Conclusions

This study introduced a method for directly integrating text with stylometric information.

Main Findings:

- The method yielded the most significant F1 improvements when training on smaller datasets, suggesting its potential use for data types that are difficult to obtain (e.g., authorship for insurance claims). While there were tags that improved F1 when training on the entire dataset, the improvements over the baseline were minimal, especially for the hard and easy datasets.
- An attempt was made to determine if adding tags with stylometric information genuinely increased the model’s sensitivity to specific stylometric features. However, the analysis did not confirm this hypothesis, indicating the need for further exploration in this area.

Future Work:

- Given the observed improvements on the medium dataset and the training on subsamples, we see potential in the proposed method. Future research should focus on a detailed analysis of which stylometric features are significant. We hypothesize that adding tags may be particularly beneficial for stylometric features that BERT-family models cannot “learn” independently by itself.
- Additionally, BERT-family models typically have a limited number of tokens they can process. Therefore, future work should focus on constructing tags in a concise or implicit manner to accommodate this limitation.

Acknowledgments

The research is supported by the project “OpenFact – artificial intelligence tools for verification of veracity of information sources and fake news detection” (INFOSTRATEG-I/0035/2021-00), granted within the INFOSTRATEG I program of the National Center for Research and Development, under the topic: Verifying information sources and detecting fake news.

References

- [1] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024, p. 3.
- [2] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection, in: *Working Notes Papers of the CLEF 2018 Evaluation Labs*. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al., 2018, pp. 1–25.
- [3] J. E. Custódio, I. Paraboni, Each-usp ensemble cross-domain authorship attribution, *Working Notes Papers of the CLEF (2018)*.
- [4] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An ensemble-rich multi-aspect approach for robust style change detection, in: *CLEF 2018 Evaluation Labs and Workshop–Working Notes Papers*, CEUR-WS. org, 2018, p. 3.
- [5] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at pan 2023, *Working Notes of CLEF (2023)*.
- [6] A. Hashemi, W. Shi, Enhancing writing style change detection using transformer-based models and data augmentation, *Working Notes of CLEF (2023)*.

- [7] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024, p. 2.
- [8] P. E. McKnight, J. Najab, Mann-whitney u test, The Corsini encyclopedia of psychology (2010) 1–1.
- [9] Q. Lyu, M. Apidianaki, C. Callison-Burch, Representation of lexical stylistic features in language models' embedding space, arXiv preprint arXiv:2305.18657 (2023).
- [10] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.