Team karami-sh at PAN: Transformer-based Ensemble Learning for Multi-Author Writing Style Analysis

Notebook for the PAN Lab at CLEF 2024

Mohammad Karami Sheykhlan[1,*], Saleh Kheiri Abdoljabbar[2] and Mona Nouri Mahmoudabad[1]

[1]*University of Mohaghegh Ardabili, Daneshgah St., Ardabil, 5619911367, Iran*
[2]*University of Tabriz, Bahman Boulevard, Tabriz, 5166616471, Iran*

Abstract

Our study addresses the intricate task of detecting style changes within documents authored by multiple individuals. The primary aim is to pinpoint instances where authors transition within the text. This task holds significant importance in author identification, particularly in situations where no comparative texts are available. By discerning variations in writing style, we can unveil instances of plagiarism, identify instances of gift authorship, authenticate claimed authorships, and potentially develop novel technologies to support writing endeavors. Our methodology employs sophisticated ensemble learning techniques, incorporating fine-tuned Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa) and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) models for medium and hard subtask and fine-tuned RoBERTa for easy subtask, to effectively address this complex challenge across diverse datasets, such as those found in Reddit comments. Our team's approach achieved F1-scores of 97.2%, 66.4%, and 64.2% for the Easy, Medium, and Hard subtasks, respectively.

Keywords

Ensemble learning, Multi-author style change detection, Natural language processing, Transformers,

1. Introduction

Multi-Author Writing Style Analysis is a challenging field aiming to dissect documents authored collaboratively. It involves discerning authorship within a text and pinpointing stylistic shifts, which is essential for plagiarism detection and validating authorship. With the proliferation of digital content, ensuring robust mechanisms to prevent misappropriation becomes increasingly critical. Automated solutions, like Style Change Detection((SCD) algorithms, show promise but require ongoing refinement. Initiatives like PAN@CLEF 2024 [1, 2] underscore efforts to enhance style analysis techniques.

In recent years, PAN@CLEF has organized numerous international competitions focusing on SCD tasks, attracting a diverse array of participants who have contributed innovative methodologies. While some participants have relied on conventional natural language processing techniques to address SCD challenges, others have opted for the utilization of deep learning models for feature extraction. Additionally, there exists a cohort of researchers who have leveraged advanced and intricate algorithms, including BERT and its analogous models, alongside collaborative learning techniques.

Jacobo et al. [3] employed two feature extraction techniques, Prediction by Partial Matching (PPM) and Term-document matrix (TD), alongside Support Vector Machine (SVM) and Logistic Regression (LR) algorithms for model training. Their experiments across all three datasets revealed that PPM combined with LR achieved the highest F1-score for the easy and medium subtasks, while Term Frequency-Inverse Document Frequency (TF-IDF) coupled with SVM yielded the highest F1-score for the hard subtask. Zi et al. [4] proposed a method utilizing a combination of convolutional neural networks, BERT, and bidirectional long short-term memory networks for style change detection and author attribution tasks, employing binary classification. Hashemi and Shi [5] utilized fine-tuning on transformer-based models

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
mohammadkaramisheykhlan@gmail.com (M. K. Sheykhlan); salehkheiri@gmail.com (S. K. Abdoljabbar); monanouri.m@gmail.com (M. N. Mahmoudabad)
0000-0003-2316-545X (M. K. Sheykhlan); 0009-0009-3328-7486 (S. K. Abdoljabbar)

A RoBERTa and employed data augmentation and ensemble learning techniques. Their models ranked first in two sub-tasks and second in others, showcasing strong performance. Chen et al. [6] present a study employing comparative learning techniques for analyzing writing style. They enhance sentence segment embeddings produced by a pre-trained model's encoder to ensure closer proximity for similar stylistic sentences and greater separation for dissimilar ones. Utilizing this optimized encoder, they generate sentence embeddings by combining tag data with paragraph sample pairs and classifying them through a full connection layer. Experimental results demonstrate F1-scores of 0.9145, 0.8203, and 0.6755 on Task 1, Task 2, and Task 3 of the official test set, respectively. Kucukkaya et al. [7] delve into the task of multi-author writing style detection, which involves identifying shifts in writing style within text documents. They frame this challenge as a natural language inference task, pairing consecutive paragraphs. Their strategy emphasizes paragraph transitions and token truncation for input. Using various Transformer-based encoders with warmup training, they submit a model version that surpasses baselines and other proposed versions in experimentation. Specifically, they employ a transition-focused natural language inference approach based on Decoding-enhanced BERT with Disentangled Attention (DeBERTa) with warmup training for easy and medium setups, while opting for the same model without transitions for the hard setup.

While feature extraction methods like TF-IDF have yielded satisfactory results in author profiling and Authorship Attribution tasks [8, 9], their performance is not as effective, particularly on the Hard dataset in the SCD task. This is due to the significant similarity observed among sample texts, limiting their effectiveness. Recent progress in this field has shown that BERT and similar language models, known for their complexity with extensive parameters, perform exceptionally well in SCD tasks.

In this study, we investigated the efficacy of transformer-based models for detecting changes in authorial style at the paragraph level. Initially, we fine-tuned three transformer-based models on the training dataset. Furthermore, we employed Ensemble learning techniques to augment the performance of our approach.

2. Task and Datasets

The datasets provided by PAN@CLEF for writing SCD are categorized into three difficulty levels: Easy, Medium, and Hard. Each level presents unique challenges for detecting shifts in authorship within documents. On the Easy level, documents contain paragraphs covering diverse topics, allowing approaches to utilize topic information effectively in detecting authorship changes. Conversely, the Medium level entails documents with limited topical variety, requiring a stronger emphasis on stylistic analysis to address the detection task. Finally, the Hard level poses the most challenging scenario, where all paragraphs within a document focus on the same topic, demanding approaches to rely solely on stylistic cues for accurate detection.

Moreover, for each subtask, PAN@CLEF provides distinct datasets comprising multiple documents, each containing paragraphs. Accompanying these datasets are ground truth files, which furnish essential information: the number of authors associated with each document and the identification of consecutive paragraphs where style changes occur, signifying transitions in authorship. These datasets are partitioned into training and validation sets to facilitate experimental setup. In our approach, we treat every pair of consecutive paragraphs as a sample, concatenating them and assigning a label indicating whether a style change occurred between the two paragraphs (labeled as 1) or if they were written by the same author (labeled as 0).

3. System Overview

3.1. Data preparation

For each subtask, our methodology involves concatenating two consecutive paragraphs and assigning them a binary label, thereby framing each subtask as a binary classification problem. In our study, we harness the power of three state-of-the-art transformer-based models: BERT, RoBERTa, and ELECTRA complemented by their corresponding tokenizers. Given the constraint of 512 tokens imposed by the medium-sized models and the rarity of samples exceeding this limit, we adopt a uniform approach by adhering to the 512-token threshold and truncating longer samples as necessary. Additionally, to ensure consistent attention distribution across paragraphs within each sample, we implement a truncation strategy that involves removing tokens from the end of the sequence. This meticulous approach enables us to effectively leverage the capabilities of these advanced models while addressing the practical constraints of our dataset.

3.2. Transformer-based Models

Transformer models have revolutionized natural language processing, capturing complex relationships in text data through self-attention mechanisms. Unlike traditional models, they process all words in a sentence simultaneously, enabling efficient handling of longer sequences. Pre-trained on large corpora, models like BERT and Generative Pre-trained Transformer (GPT) learn rich language representations, adaptable to various tasks with minimal labeled data. Widely used in machine translation, summarization [10], and text classification like hate speech detection [11], transformers are fundamental in modern NLP systems. In our study, we employed three popular pre-trained transformer models, namely BERT, RoBERTa, and ELECTRA.

BERT, RoBERTa, and ELECTRA are prominent transformer-based models in NLP. BERT, introduced by Google, utilizes bidirectional training and transformer architecture to capture contextual information effectively. RoBERTa, an improvement upon BERT, optimizes training strategies and scales model size, achieving superior performance on various NLP tasks. ELECTRA, employing a novel pre-training approach, replaces tokens with plausible alternatives and trains a discriminator to distinguish between real and replaced tokens, enhancing efficiency and learning effectiveness.

Ultimately, we augmented each language model with a binary classification layer to detect changes in writing style. We fine-tuned the models for each subtask using their respective datasets. Additionally, we aggregated the datasets from all three subtasks and fine-tuned the RoBERTa and ELECTRA models on this combined dataset.

3.3. Ensemble learning

Hard Ensemble Learning refers to a sophisticated approach in machine learning where multiple models, known as base learners, are combined to form a stronger, more robust predictive model. Unlike traditional ensemble methods, such as bagging and boosting, which primarily focus on combining diverse but weak models, Hard Ensemble Learning integrates several complex and high-performing models to tackle challenging tasks. This technique leverages the collective intelligence of diverse models, each trained on different aspects of the data, to improve overall predictive accuracy and generalization.

In this study, we developed five single models for each subtask, all based on three fine-tuned models: BERT, RoBERTa, and ELECTRA. Initially, we fine-tuned all three models with the dataset corresponding to each subtask. Then, we combined the datasets of all three subtasks and fine-tuned the RoBERTa and ELECTRA models on them. Our goal with ensemble learning is to leverage the strengths of each approach to enhance system performance. Detailed specifics will be provided in the next section.

4. Experiments

4.1. Hyperparameter tuning and Evaluation

In this study, we utilized Google Colaboratory's GPU to fine-tune the BERT, RoBERTa, and ELECTRA models. Due to token constraints, we limited our consideration to 512 tokens. The hyperparameter learning rate for all three transformer models was set to 2e-5. For the dataset associated with the "easy" subtask, we employed an epoch value of 10, while for the other two subtasks, we used 7 epochs.

Table 1
Macro F1 Score on the validation dataset. The best result for each dataset is given in bold. Combined1 and Combined2 refer to fine-tuned RoBERTa and ELECTRA on the entire tasks dataset, respectively.

Model	Easy	Medium	Hard
BERT	90.95	78.3	66.93
RoBERTa	**97.76**	82.94	75.84
ELECTRA	93.86	81.21	75.84
Combined1	85.45	81.77	64.19
Combined2	82.66	80.37	65.54
RoBERTa+BERT+ELECTRA	95.45	82.84	77.08
RoBERTa+ELECTRA+Combined1	98.18	83.17	**77.54**
RoBERTa+ELECTRA+Combined2	95.9	82.89	77.38
Combined1+Combined2+RoBERTa	90.9	83.12	70.46
Combined1+Combined2+ELECTRA	89.44	82.28	70.62
Combined1+Combined2+BERT	89.21	82.68	68.85
Combined1+Combined2+RoBERTa+ELECTRA+BERT	95.49	**83.57**	75.50

Table 2
Evaluation of Final Performance on Unseen Test Sets

Task	Reported F1-score
Task 1	97.2
Task 2	66.4
Task 3	64.2

Moreover, for the fine-tuning models chosen for the entire subtasks dataset (RoBERTa, ELECTRA), the epoch value was set to 3.

The F1-score is a balanced measure of a model's accuracy, combining precision and recall. Precision measures the accuracy of positive predictions, while recall measures the model's ability to find all positive instances. The macro F1-score averages F1-scores across all classes, providing a balanced evaluation, especially in datasets with class imbalances. Following our experimentation and analysis of results from evaluation sets, we identify the most effective model for each subtask. Subsequently, we execute the selected model on an unseen test set through the TIRA platform [12].

4.2. Results

In this section, we present the results of our experiments on the SCD task. Table 1 summarizes the performance of the tested models. We conducted 12 different experiments for each subtask, both with standalone models and ensemble learning. Our findings show that for the "easy" subtask, the pre-trained RoBERTa model achieved the highest macro F1-score on the provided validation dataset. However, for the remaining subtasks, ensemble learning models outperformed others. Specifically, fine-tuning RoBERTa with the entire subtasks dataset (Combined1) and ELECTRA (Combined2), as well as training ELECTRA, RoBERTa, and BERT models on the medium subtask dataset, yielded the highest macro F1-score accuracy. Moreover, for the "Hard" subtask, models trained on RoBERTa and ELECTRA, along with Combined1, delivered the best performance across all experiments. Therefore, for each subtask's test dataset, we selected the approach with the best performance. The results of our work on the unseen dataset are presented in Table 2.

5. Conclusion

This paper provides an overview of our team's performance in the PAN Shared Task for the SCD task. We embarked on a series of diverse experiments to pinpoint the most effective strategy for identifying shifts in writing style within consecutive paragraphs. Initially, we trained five distinct language models, meticulously fine-tuning each to grasp the intricacies of our task. Following this, we delved into an extensive exploration of various combinations of these models, tailoring our approaches to suit the unique demands of each subtask. For the "easy" subtask, we opted to leverage the capabilities of the RoBERTa model exclusively, given its robust performance in preliminary assessments. However, for the more complex subtasks, we adopted an ensemble learning approach, harnessing the collective power of multiple models to tackle the nuanced challenges presented.

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: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[2] 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.

[3] G. Jacobo, V. Dehesa, D. Rojas, H. Gómez-Adorno, Authorship verification machine learning methods for Style Change Detection in texts, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023, pp. 2652–2658. URL: https://ceur-ws.org/Vol-3497/paper-217.pdf.

[4] L. Z. J. Zia, Z. Liua, Style Change Detection Based On Bi-LSTM And Bert, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022. URL: http://ceur-ws.org/Vol-3180/paper-234.pdf.

[5] A. Hashemi, W. Shi, Enhancing Writing Style Change Detection using Transformer-based Models and Data Augmentation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023, pp. 2613–2621. URL: https://ceur-ws.org/Vol-3497/paper-212.pdf.

[6] H. Chen, Z. Han, Z. Li, Y. Han, A Writing Style Embedding Based on Contrastive Learning for Multi-Author Writing Style Analysis, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023, pp. 2562–2567. URL: https://ceur-ws.org/Vol-3497/paper-206.pdf.

[7] I. E. Kucukkaya, U. Sahin, C. Toraman, ARC-NLP at PAN 23: Transition-Focused Natural Language Inference for Writing Style Detection, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023, pp. 2659–2668. URL: https://ceur-ws.org/Vol-3497/paper-218.pdf.

[8] M. Rahgouy, H. Giglou, T. Rahgooy, M. Sheykhlan, E. Mohammadzadeh, Cross-domain Authorship Attribution: Author Identification using a Multi-Aspect Ensemble Approach, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[9] H. B. Giglou, M. Rahgouy, T. Rahgooy, M. K. Sheykhlan, E. Mohammadzadeh, Author profiling: Bot and gender prediction using a multi-aspect ensemble approach, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_231.pdf.

[10] Y. Liu, M. Lapata, Text summarization with pretrained encoders, arXiv preprint arXiv:1908.08345 (2019).

[11] M. K. Sheykhlan, J. Shafi, S. Kosari, Pars-hao: Hate speech and offensive language detection on persian social media using ensemble learning, Authorea Preprints (2023).

[12] 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. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.