Language Processing 2 Exam Submission

Part 1

Abstract

In the domain of Natural Language Processing (NLP), the task of multi-author writing style analysis is important. It plays a role in plagiarism detection, forensics, history, and more. This paper introduces an experiment designed to detect changes in authorship within Reddit comment threads at the paragraph level. We use two types of neural network architectures: our baseline model, a Multi-Layer Perceptron (MLP), processes pairs of paragraphs, while our more advanced model utilizes a Gated Recurrent Unit (GRU) layer to handle sequences of paragraphs. Paragraphs are represented by either a semantic vector, a stylometric vector, or the concatenation of the two, depending on the setup. The semantic vectors are extracted from Sentence-BERT embeddings, and the stylometric vectors are constructed by feeding paragraphs to a function that determines various features, such as average sentence length, and the frequencies of particular kinds of words. Our findings, evaluated with F1 scores, suggest that incorporating sequential information enhances performance in multi-author writing style analysis. Additionally, as paragraphs become more similar in content, the contribution of stylometric information becomes increasingly important.

Introduction

Multi-author writing style analysis is a fascinating and challenging field within Natural Language Processing (NLP). It has a rich history. The concept of distinguishing authorship based on stylistic nuances has roots predating the digital text age, tracing back to fields such as cryptography, literary history, and forensics, where features like word choice, sentence structure, and rhythmic patterns of writing were analyzed for signs of individual authorship.

An interesting early instance of authorship detection occurred during World War II. Telegraph operators were frequently identified by their unique Morse code patterns. Despite Morse code's standardized system of dots and dashes, operators developed individualized patterns discernable by experts. This is mentioned to illustrate the idea

that personal idiosyncrasies could seep through even the most formalized communication modes.

Today, authorship attribution and style analysis have found wide applications, helping with uncovering ghostwriting, and plagiarism, and contributing to forensics. With the advent of digital communication and generative artificial intelligence vast amounts of text data are being produced, making authorship attribution an increasingly relevant field.

This paper describes an experiment in detecting authorship changes within sequences of Reddit comment paragraphs. We represent each paragraph as a semantic, or stylometric vector, or as the concatenation of the two. Semantic features are extracted from pre-trained Sentence-BERT embeddings, while the stylometric features are constructed by feeding paragraphs to a function written as part of the experiment. We initiate our analysis with a Siamese Multi-Layer Perceptron (MLP) model, which serves as our baseline. This model processes pairs of paragraphs, calculating a distance measure between the two, and determining if they share the same authorship. We then explore a more advanced model utilizing a Gated Recurrent Unit (GRU) layer to handle sequences of paragraphs. We do this to investigate how useful having the larger contexts is for the task, and if the additional complexity required of the model to process the context, yields an increase in performance. We define performance as an F1 score, though other metrics, like processing speed, are also interesting.

The paper is structured as follows: following a review of relevant literature, we introduce our data and task. This in turn is followed by a detailed overview of our methodology. Followed by that, we have a presentation and discussion of our results. We conclude with reflections on our findings and potential future work directions.

Literature Review

Multi-author writing style analysis combines several active areas within natural language processing and machine learning. It has advanced significantly in recent decades. Early high-profile use of statistical author style change detection can be traced back to Mosteller and Wallace (1963)'s analysis of the disputed Federalist Papers based on common word frequency alone.

With the advent of machine learning, new tools were made available for writing style analysis and authorship attribution. An extensive overview of these techniques is provided by Stamatatos (2009). Deep learning methods, particularly Siamese and GRU architectures, were later employed to further these tasks as described by Qian, He, and Zhang (2017). As with much of natural language processing, these tasks too have been influenced by the transformer architecture first introduced by Vaswani et al. (2017), though our approach is simpler, attempting to isolate the effect of including the full sequence of paragraphs in particular.

The application of machine learning to natural language processing tasks entails transforming raw text into processable, numerical forms, typically achieved through text embeddings. The influential work by Devlin et al. (2019) on BERT, a context-aware language model, and its extension for sentence embeddings—Sentence-BERT (SBERT)—by Reimers and Gurevych (2019) forms the basis for semantic vectors in our study.

The architecture of our neural networks stems from the work of Bromley et al. (1993) on Siamese networks. A Siamese network takes two samples, and feeds them through the same layer, before computing their difference. Originally developed for signature verification, this architecture's ability to learn distance functions has found considerable application in comparison-based tasks. To handle sequence data effectively, we augment our model with Gated Recurrent Units (GRUs), introduced by Chung et al. (2014), due to their ability to handle long-term dependencies.

The relationship between authorship and stylistic variation is an integral part of our research, guided by Bergsma and Van Durme (2013)'s (CHANGE CITATION) findings of consistent stylistic choices by authors across varying topics.

Thus, it is with a robust foundation in multi-author writing style analysis that we attempt this experiment to test the effect of processing entire sequences of paragraphs, instead of pairs. To further elucidate this effect we explore how it varies for semantic, stylometric, and combined text representations.

Data & Task

The foundation for our experiment is the 2023 PAN challenge on Multi Author Writing Style Analysis. The challenge is based on data derived from Reddit, a social media platform with many "subreddits" for different topics, which inturn consists of threads to which many different users contribute. Threads are thus types of multi-author documents where each comment represents a paragraph in a sequentially arranged text. Threads makes Reddit is a good environment for studying authorship transitions.

The PAN challenge provides three datasets, each presenting varying levels of difficulty—"Easy," "Medium," and "Hard". Each dataset varies in its degree of topical diversity between paragraphs in documents, ranging from a broad spectrum in the "Easy" dataset to a single topic in the "Hard" dataset. We primarily focus on the 'Medium' dataset, which has limited topical variation, encouraging the detection of authorship changes based on stylistic and syntactic cues over content based semantic cues.

The datasets are each made up of a training set (4200 samples) and a validation set (800 samples). The validation set is further partitioned into a development set and a test set, each containing 400 samples. These sets come with ground truth data indicating authorship change points. The true test data for the challenge is unavailable, which is why the validation set was further split in two.

Our key task is detecting style changes at the paragraph level within documents,

indicating authorship changes. These changes can only occur *between* paragraphs. A single paragraph belongs to one author. Each task problem, is identified by an ID, and includes two files—one with the text and the other with ground truth data. The latter specifies whether a style change occurs (1) or not (0) between every pair of consecutive paragraphs. Our objective is to train models that can accurately predict these ground truth values, enabling effective identification of authorship transitions. We evaluate this effectiveness using the F1 score metric.

Methodology

Our approach to authorship change detection involves two neural models, both three layers deep. The baseline model, referred to as the Siamese model, is a Siamese Multi-Layer Perceptron (MLP) designed to determine whether two given paragraphs are written by the same author. The advanced model, called the Recurrent model, replace the Siamese layer with gated recurrent unit (GRU) layer. This change enables the model to process sequences paragraph, offering additional information about each author's writing style. Three versions—semantic, stylometric, and concatenated—of the two models were trained for each of the three datasets, yielding 18 distinct models (see Appendix B for details).

Vector representations

The stylometric vector is a, manually-constructed feature vector with various lingustic and stylistic elements of the text. It consists of 61 features, with the first six addressing fundamental textual metrics such as the number of sentences, words, unique words, and the ratios of unique words to total words and unique part-of-speech (POS) tags to total words in a paragraph. The rest of the features delve into specific POS tag usage, featuring counts, word ratios, and sentence ratios for various forms of nouns, verbs, adjectives, adverbs, and pronouns. This includes tags like "NN" for common nouns, "VB" for base form verbs, or "JJ" for adjectives. The semantic vectors are the corresppnding paragraph's Sentence-BERT (SBERT) embedding. The SBERT embeddings are 384-dimensional, making it about six times larger than the stylometric vector. Concatenating the two vectors yilded our largest paragraph represetion of size 445.

Siamese MLP – Baseline Model

Our Siamese MLP, functioning as the baseline, accepts a pair of consecutive paragraphs (represented by their extended SBERT embeddings) and determines whether a style change, implying an authorship change, occurs between them. The choice of MLPs is due to their proven success in complex function modeling and classification tasks. The model first feeds the two vectors through the same lienar layer, before mering them into one vector by subtraction. This single vector is then fed through the reamining two layers of the network.

GRU Model

To outperform the baseline, we implement a GRU model. Unlike the Siamese MLP, the GRU model processes the entire sequence of paragraph embeddings for a document, thus capturing broader context and making more informed predictions about authorship changes. The GRU model does not use a Siamese architecture, as the relationship between a paragraph and it's neightbours is handled by the recurrence layer.

Hyperparameter Optimization

To enhance our models' performance, we conduct a random hyperparameter search across variables such as layer sizes, learning rates, batch sizes, and dropout rates. Good hyperparameters were determined after training a total of 378 models, with 21 hyperparameter samples drawn and trained with for each model. The hyperparameter sample spaces are in the table below.

Table 1: Hyperparameter sample spaces.

Hyperparameter	Sample Space
Learning rate	{0.001, 0.0001, 0.00001}
Dropout	{0.1, 0.2, 0.3}
Hidden dimension size	{32, 64, 128, 256}
Batch size	{16, 32, 64}
Number of steps	{2000, 4000, 6000, 8000}

Model Training & Evaluation

Models are trained on the training set and fine-tuned and evaluated on the validation set. The chosen evaluation metric, in accordance with the competition's requirements, is the F1 score—an aggregate measure of precision and recall. Our aim is to maximize this score for accurate authorship change prediction. Binary Cross Entropy (BCE) was chosen as the loss function over Soft F1 loss, as early experiments indicated that it yilded a better performance, in spite of the latters mathematical similarity to the F1 score.

Results & Analysis

The final hyperparameters for each of the 18 models can be seen in Appendix A. It is noteworthy that, inspite of the similarity between tasks, models and input data, almost the entire sample space of each hyperparamter is present when looking at all 18 models. This might be an indication of model robustness, though more exploration would need to be done to confirm this.

The performance of the Siamese baseline on Dataset 2 are reported in **table 2**. When including both semantic and stylometric features the test F1 score is 0.6155. Excluding the semantic features drops the F1 score to 0.6538, while excluidng the stylometric feature drops the F1 score to 0.5123. This is an indication that the stylometric feature is more indicative of author change than the SBERT vector. There drop in performance from the training set to the validation set, which is not unexpected and suggests a degree of overfitting to the training data. The Recurrent model does outperform its Siamese baseline in on Dataset 2 when training on both modalities, and the purely symantic modality, reaching a test F1 score of 0.6807 when training on the former—significantly better than the Siamese's 0.6155. However, on the purely syntactic modality the Siamese model slightly outperforms the Recurrent model.

Table 2: Dataset 2 — Test and train results.

Model	Modality	Train F1	Test F1	Train Loss	Test Loss	Params
Recurrent	both	0.7300	0.6807	0.4995	0.5475	20,641
Recurrent	semantic	0.6460	0.5945	0.5850	0.6043	18,689
Recurrent	syntactic	0.6720	0.6462	0.5652	0.5780	107,137
Siamese	both	0.6660	0.6155	0.5617	0.5943	73,729
Siamese	semantic	0.7016	0.5123	0.4949	0.6655	28,865
Siamese	syntactic	0.6761	0.6538	0.5780	0.6112	24,577

While both models were successful at identifying changes in authorship based on stylistic differences, the superior performance of the GRU model suggests that considering the full context of the document, rather than only pairwise comparisons of paragraphs, can lead to more accurate predictions.

Table 3: Dataset 1 — Test and train results.

Model	Modality	Train F1	Test F1	Train Loss	Test Loss	Params
Recurrent	both	0.9720	0.9460	0.1281	0.2335	509,185
Recurrent	semantic	0.9403	0.9238	0.2823	0.3615	148,481
Recurrent	syntactic	0.9501	0.9462	0.2197	0.2441	8,353
Siamese	both	0.9551	0.9569	0.1855	0.2275	180,225
Siamese	semantic	0.9800	0.9449	0.0874	0.2226	28,865
Siamese	syntactic	0.9508	0.9398	0.2265	0.2917	24,577

Inspecting **table 3**, focusing on Dataset 1, we see that performance becomes similar for all models, while the the layer sizes fluctuate widely, is indicated by the paramter count column. In **table 4**, focused on Dataset 3, in which the topics are most similar, we see that, surprisingly, the semantic modality outperforms both the concatenated and the

stylometric modality. This is highly supring, since the semantic distance between the paragraph representations should decrease as we move from Dataset 1 through Dataset 2, to Dataset 3.

Table 4: Dataset 3 — Test and train results.

Model	Modality	Train F1	Test F1	Train Loss	Test Loss	Params
Recurrent	both	0.5575	0.5330	0.6458	0.6616	156,289
Recurrent	semantic	0.5282	0.4990	0.6562	0.6699	18,689
Recurrent	syntactic	0.5306	0.5117	0.6647	0.6757	410,881
Siamese	both	0.5343	0.5041	0.6500	0.6795	32,769
Siamese	semantic	0.6845	0.5250	0.5527	0.7260	65,921
Siamese	syntactic	0.4932	0.4671	0.6919	0.6799	81,921

Discussion

Our investigation into the problem of multi-author writing style analysis has yielded insightful results. We found that our approach of using Siamese networks and GRU models, coupled with SBERT embeddings, effectively detects the style changes that signal a shift in authorship.

The Siamese MLP served as a robust baseline model, but its pairwise comparison of paragraphs could not leverage the contextual information provided by the full sequence of paragraphs in a document. This limitation was addressed by the GRU model, which processes the entire sequence and was thereby better able to identify writing style changes. The improved performance of the GRU model substantiates the relevance of using sequence models for tasks that involve dependencies between data points, like text and time-series data.

Despite these promising results, several avenues remain open for further exploration. For example, additional gains may be possible by experimenting with other types of sentence or paragraph embeddings, such as Doc2Vec or Universal Sentence Encoder. Furthermore, exploring more advanced architectures, including attention-based models like the Transformer, could potentially lead to improvements. We could also attempt to create ensemble models that combine the strengths of different architectures. Previos work attemptiung similat tasks have also used diffent loss function, like soft F1 loss and focal loss.

Conclusion

In conclusion, the experiment indicates that processing the suroubnding paragraphs imprtoves performance, at least in the case of the Reddit data at hand. The increased complexity of the GRU is thus justified in this case.

Part 2

Q1.Explain the difference between topic change and style change. Based on this difference, discuss whether the documents in the datasets consistently display both changes, or whether they tend to show one single aspect. Please provide examples from one or more of the datasets where each of these specific changes happen.

Topic change refers to a shift in the *ideas* of a text. Style change, on the other hand, is a shift in the *way* these ideas are expressed. It may involve changes in grammar structure, word choice, and sentence lengths, and more. As mentioned in the literature review, style tends to vary between people even highly formal settings. As mentioned in the Data & Task section, author changes in Dataset 2 are also slightly changes in subject as the paragraphs are collected from quite similar reddit threads. Focusing on problem 42 from Dataset 2's validation set, we see the two following paragraphs:

You have the absolute right to breach a contract, but I don't see that you have any defence to doing so, so be prepared to be sued over it unless you work something out with the buyers, and potentially to be forced to close anyways and to pay damages due to the delay if the sellers sue for specific performance.

and

Sellers usually have few "outs" in real estate contracts, if any, but you'd have to read yours and see if you have any. I'm going to guess you don't. So if you cancel you may open yourself up to being sued for your buyer's expenses including temporary housing and re-arranging their move, as well as perhaps their deposit, inspection and/or appraisal costs. They could also sue you for specific performance which means you'd be ordered to sign the docs to conclude the sale. If the latter happens you will end up spending a lot of money on an attorney.

These are about the same subjects, while their tones are quite different. The first is slightly provacative, presumably ironically indicating being fine with the person making a huge mistake, while the second paragraph has a kinder tone. Their conclusion is however the same. The most apparent difference between the two is length.

Q2.Explain what the task of authorship attribution is, and discuss how your approach for style change detection could be used and possibly modified for authorship attribution.

Style change is used in this experiment to detect *if* there has been a change in author. Authorship attribution focuses on detecting *who* a given text is authored by. The models developed in the experiment outlined above could be used in authorship attribution in at least two different ways: 1) Detecting points author change for later author classification, and 2) Chaning the output layer to, instead of predicting wheather a change has occured, who is the current writer of the peragraph. Pan 2020 focused on this challenge.

References

Bergsma, Shane, and Benjamin Van Durme. 2013. "Using Conceptual Class Attributes to Characterize Social Media Users." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 710–20. Sofia, Bulgaria: Association for Computational Linguistics. https://aclanthology.org/P13-1070.

Bromley, Jane, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. "Signature Verification Using a "Siamese" Time Delay Neural Network." In *Advances in Neural Information Processing Systems*. Vol. 6. Morgan-Kaufmann. https://proceedings.neurips.cc/paper/1993/hash/288cc0ff022877bd3df94bc9360b9c5d-Abstract.html.

Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." December 11, 2014. https://doi.org/10.48550/arXiv.1412.3555.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58 (302): 275–309. https://doi.org/10.2307/2283270.

Qian, Chen, Ting He, and R. Zhang. 2017. "Deep Learning Based Authorship Identification." In. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." August 27, 2019. https://doi.org/10.48550/arXiv.1908.10084.

Stamatatos, Efstathios. 2009. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60 (3): 538–56. https://doi.org/10.1002/asi.21001.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." December 5, 2017. https://doi.org/10.48550/arXiv.1706.03762.

Appendix

Combination	Batch Size	Dropout	Hidden Dim	Learning Rate	Number of Steps
Dataset 1 - Recurrent - Both	64	0.1	256	0.0001	6000
Dataset 1 - Recurrent - Semantic	64	0.3	128	0.0001	4000
Dataset 1 - Recurrent - Syntactic	32	0.3	32	0.001	8000
Dataset 1 - Siamese - Both	64	0.1	256	0.0001	6000

Combination	Batch Size	Dropout	Hidden Dim	Learning Rate	Number of Steps
Dataset 1 - Siamese - Semantic	16	0.1	64	0.001	4000
Dataset 1 - Siamese - Syntactic	32	0.2	128	0.001	8000
Dataset 2 - Recurrent - Both	64	0.3	32	0.001	4000
Dataset 2 - Recurrent - Semantic	64	0.3	32	0.0001	6000
Dataset 2 - Recurrent - Syntactic	32	0.2	128	0.0001	8000
Dataset 2 - Siamese - Both	32	0.2	128	0.001	6000
Dataset 2 - Siamese - Semantic	16	0.2	64	0.001	8000
Dataset 2 - Siamese - Syntactic	32	0.1	128	0.001	8000
Dataset 3 - Recurrent - Both	32	0.1	128	0.0001	4000
Dataset 3 - Recurrent - Semantic	16	0.2	32	0.001	2000
Dataset 3 - Recurrent - Syntactic	16	0.2	256	0.0001	8000
Dataset 3 - Siamese - Both	64	0.1	64	0.001	4000
Dataset 3 - Siamese - Semantic	64	0.3	128	0.001	4000
Dataset 3 - Siamese - Syntactic	32	0.2	256	0.0001	6000

Appendix B: Stylometric Vector Features

Feature #	Description
0	Number of sentences in the paragraph
1	Number of words in the paragraph
2	Number of unique words in the paragraph
3	Ratio of unique words to total words in the paragraph
4	Number of unique part-of-speech tags in the paragraph
5	Ratio of unique part-of-speech tags to total words in the paragraph
6	Ratio of unique part-of-speech tags to total sentences in the paragraph
7-9	Counts, word ratio, and sentence ratio for ‘NN’
10-12	Counts, word ratio, and sentence ratio for ‘NNS’
13-15	Counts, word ratio, and sentence ratio for ‘NNP’
16-18	Counts, word ratio, and sentence ratio for ‘NNPS’
19-21	Counts, word ratio, and sentence ratio for ‘VB’
22-24	Counts, word ratio, and sentence ratio for ‘VBD’
25-27	Counts, word ratio, and sentence ratio for ‘VBG’
28-30	Counts, word ratio, and sentence ratio for ‘VBN’
31-33	Counts, word ratio, and sentence ratio for ‘VBP’
34-36	Counts, word ratio, and sentence ratio for ‘VBZ’
37-39	Counts, word ratio, and sentence ratio for ‘JJ’
40-42	Counts, word ratio, and sentence ratio for ‘JJR’
43-45	Counts, word ratio, and sentence ratio for ‘JJS’
46-48	Counts, word ratio, and sentence ratio for ‘RB’

Feature #	Description
49-51	Counts, word ratio, and sentence ratio for 'RBR'
52-54	Counts, word ratio, and sentence ratio for 'RBS'
55-57	Counts, word ratio, and sentence ratio for 'PRP'
58-60	Counts, word ratio, and sentence ratio for 'PRP$'

Appendix C: Problem 9 From Dataset 2's Validation set.

Change	Paragraph
-	As the child's grandmother, your mother has no particular legal obligation to care for her. (Likewise, if you are an adult, she doesn't have any legal obligation to provide for you either.) If you believe the child is not safe with her grandmother, you should not leave her in her grandmother's care, but you have not described anything that sounds like child abuse.
No	If you file a CPS report, they will investigate the child's living conditions holistically, and determine whether she is receiving adequate care in your home. What outcome are you hoping for?
Yes	Im not saying it is her responsibility. But she wants to claim the responsibility, she gives my son everything while my daughter gets nothing. Makes it so I can't work, but then keeps tabs like I'm the one abusing my kids. When she's the one who doesn't buy food and "provide" like she claims to others, and even tried to claim to the IRS.
No	I'm not even sure what outcome would come about it(that is why i asked the question) and I am fully aware she doesn't have legal obligations to anything have to do with my kids or me. But she's claiming it, tried to claim me and my kids on her taxes, yet REALLY doesn't do anything. So not sure if you can see what I'm saying still...
Yes	She's welcome to tell people she's responsible for you and your children all she wants, whether or not she actually is. Telling people she takes care of your children does not create a legal obligation to do so.
No	If she is committing tax fraud, you are welcome to report that, or file your own taxes on paper and let the IRS sort out the discrepancy.
No	She can't stop you from getting a job, and you're welcome to stop associating with her or letting her see your children.
No	It is your responsibility to ensure your children receive adequate care. If they are being neglected at home, you will be the one held responsible. CPS will not order your mother to provide more care for her grandchild, if that was what you were hoping for.
Yes	Why is it your mom's responsibility to provide childcare, feed, change diapers and comb your daughter's hair? What behaviors have led you to consider reporting your mom for child abuse?

Change	Paragraph
No	How is your mom making it so that you can't work? Since you think her treatment of you and your children is unfair, stop relying on her for support. Was your mom successful in claiming you and your kids as dependents?