Frequency Analysis of Embeddings for Author Style Change Detection

Thom Lazor, Maksym Kravchuk, Serkan Karatastan

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{thomas.lazor,maksym.kravchuk,serkan.karatastan}@fer.hr

Abstract

In this study, we aim to explore the application of standard signal processing techniques to NLP, specifically on the task of author change detection. We use Fourier analysis and filtering as layers in a fully connected neural network. Our findings demonstrate that including frequency domain features and filtering certain frequencies significantly improves model generalization and accuracy in detecting author changes. This approach not only advances the performance in author change detection but also suggests the potential of signal processing methods in various NLP tasks.

1. Introduction

Style change detection involves analyzing the writing style of a document to determine whether it was written by more than one author. Three problems formulated and defined in the Style Change Detection task, organized by PAN 2021. First one requires us to decide whether a text has one or more authors, second one is to decide if the author changed between consecutive paragraphs, and finally the last problem is deciding which paragraph is written by which author.

Our research aims to take previous efforts one step further by using a new method: the application of the Fourier Transform to paragraphs of text, thus allowing the separation and analysis of high and low frequency components of writing style. Our idea is that this frequency based analysis will work for texts as it works for signals, and will further improve the detection of different authors and increase the accuracy in identifying style changes. In this way, we will test whether a significant improvement in the results given in PAN 2021 tasks can be determined by this method.

The objective of the research is to test how effectively the Fourier Transform distinguishes authorship changes and to answer the intriguing question: *Can frequency domain analysis of text provide new and robust means for style change detection?* If the research work is successful, we hope that the study of texts as signals in general can improve methods not only for the task of style change detection, but also for different tasks.

2. Background

The Style Change Detection task of the PAN21 competition focuses on the detection of style change in multi-author documents and aims to identify text positions where author changes occur. This task is crucial for a variety of purposes, such as detecting plagiarism or verifying authorship claims. The competition was held differently in previous years, ranging from distinguishing single-authored and multi-authored documents, to determining the number of authors, and then determining whether the author of consecutive paragraphs was the same. Success in these tasks allowed this task to focus on the main goal in 2021, namely identifying the exact locations of authorship changes. Participants are tasked with three questions: determining if a document is single or multi-authored (Task 1), identifying the positions of style changes between paragraphs (Task 2), and assigning paragraphs to authors in multi-author documents (Task 3) (Zangerle et al., 2021a).

In this competition *Zhang et al. 2021* succeeded in getting the top result, so we aimed to improve their work. Their proposed method for style change detection uses a binary classification approach that exploits writing style similarity between paragraphs. If there were any 1s in Task 2 (if a change of author was detected in one of the consecutive paragraphs), the document was considered multi-authored and Task 1 was marked as multi-authored (i.e. 1). Task 3 was converted into a binary label, allowing uniform framework application. Figure 1 explains how this is done. They estimated writing style similarity by using BERT pretraining and Fully Connected Neural Network Classifier. They examined the paragraphs in terms of style changes and made inferences about the identification of the author.

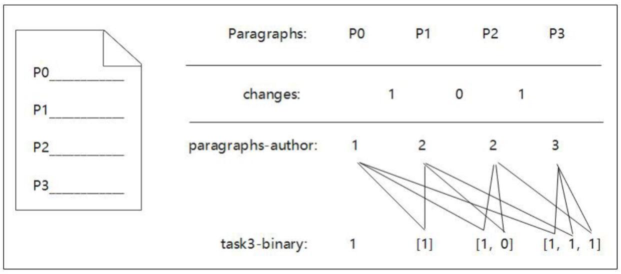
P0
P1
P2
P3
Paragraphs: P0 P1 P2 P3
changes: 1 0 1
paragraphs-author: 1 2 2 3
task3-binary: 1 [1] [1, 0] [1, 1, 1]

Figure 1: Task 3 binary labels are generated by comparing combinations of paragraphs. Task 3 solutions can then be created from the binary labels (Zhang et al., 2021).

Their method primarily uses the BERT pre-training model, taking into account resource constraints due to too much data. They compared maximum paragraph length of 256 and 512, and since results were similar they decided to set maximum paragraph length to 256, to balance classification effectiveness and computational efficiency. For fine-tuning, the training set and validation set used the Task 3 binary label and Task 2 label separately. With this, the models were fine tuned deeply because of the sufficient training data. The model was fed with three training stages and they aimed to achieve better results and integrated performance in all tasks (Zhang et al., 2021).

Data set	Task1.F1	Task2.F1	Task3.F1
Validation set	0.85542	0.75193	0.39669

Table 1: Zhang et al. F1 scores for different tasks on validation Set

Evaluation on the validation set showed very good performance with F1 scores as we can see on Table 1. For the test set, they obtained the highest F1 scores on Task 2 and Task 3. Interestingly their performance kept increasing with the increasing number of authors (Zangerle et al., 2021a).

3. Methodology

3.1. Data

We used the dataset prepared for the Style Change Detection task of PAN 2021 (Zangerle et al., 2021b). This dataset is based on user posts on StackExchange sites. Each document in the dataset contains English conversations, mostly of on the topic of technology, separated into paragraphs. All paragraphs with fewer than 100 characters are dropped. There are at most four authors of each document, but the document can contain an arbitrary number of authorship changes. Authorship changes occur only between paragraphs, so each paragraph has a sole author. Ground truth data is provided for training and validation sets, while it is not provided for test sets. Table 2 shows some simple statistics of the dataset. Our source code is available on GitHub.[1]

3.2. Hardware

We trained and evaluated the models on a machine with an Intel Xeon Platinum 8358 and A10 GPU. Each model took roughly four hours to train.

3.3. Architecture

To enable comparison with Zhang et al. (2021), we implement their architecture as a pretrained BERT layer and FCNN. We use a three layer FCNN and train with an RMSProp optimizer and binary cross entropy loss.

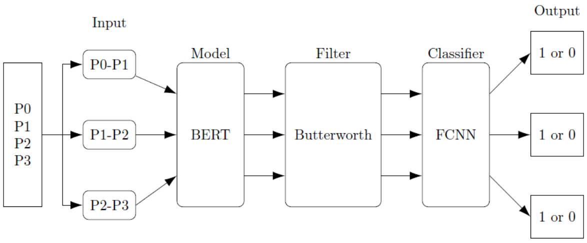

Figure 2: Model architecture with Butterworth filter

We truncate and zero pad the BERT embedding for each paragraph to a length of 256, since most of the stylistic information is contained in the beginning of the paragraph and there is little performance to be gained by using more (Zhang et al., 2021). This gives us 256 time steps of a signal with 768 dimensions per paragraph or 512 × 768 features per paragraph pair classification problem.

Dataset	#Docs	Documents / #Authors				Length / #Authors			
		1	2	3	4	1	2	3	4
Train	11,200	2,800	2,800	2,800	2,800	1,519	1,592	1,795	2,059
Valid.	2,400	600	600	600	600	1,549	1,599	1,785	2,039
Test	2,400	600	600	600	600	1,512	1,564	1,793	2,081

Table 2: Summary of datasets (Zangerle et al., 2021a)

We first investigate whether including the frequency spectrum as features improves performance. We keep the number of features per paragraph (256) the same, but replace some of the embedding features with 0, 64, 128, and 256 spectrum features. We perform the Fourier transform on all 256 embeddings, but decimate the spectrum to fit the number of spectrum features required. Additionally, in order to use the same model architecture as Zhang et al. (2021), we keep only the real portion of the spectrum. Since the real portion of the frequency spectrum contains most of the information, we think the trade off is acceptable (Almeida, 1994).

Next, to investigate which parts of the frequency spectrum contain the most useful information to determining author style, we sliced the spectrum into four equal parts and then use a Butterworth filter to bandstop filter each part in turn. The four equal parts are from 0 to the Nyquist frequency (for all our models, the Nyquist frequency is $\frac{256}{2} = 128Hz$). Butterworth filters have a maximally flat frequency response (no ripple in gain near the cutoff frequencies) and can efficiently approximate a linear phase response (all frequency components are time delayed by approximately the same amount) (Pal, 2017). Both properties are important to maintaining the shape of the frequency response and thus the unfiltered characteristics of author style. Our model architecture is shown in Figure 2.

4. Results

For brevity, we report only binary accuracy of the task3-binary labels and F1 scores for Task 1, 2, and 3 on the validation set. We were not provided with the test dataset for the Style Change Detection task of PAN 2021, but we directly compare our models with our implementation of Zhang et al. (2021) (referred to as ”base_model” in the figures).

As we can see in Figure 3, binary accuracy of models with different numbers of frequency features per-paragraph gives better results than the base model. But their binary accuracy is almost the same when epoch is 2. Figure 4 shows binary accuracy of the models with a Butterworth filter of different range of frequencies. We can see filtering of frequencies from 0% to 25% gives the best result. And almost all ranges give better results than the base model.

Table 3 illustrates the F1 scores for three different models on a validation set across three tasks. The models compared include the 0-25% filtered frequency model, the 256 frequency feature model, and the base model as implemented from Zhang et al. (2021). The table clearly shows that both the 0-25% filtered frequency model and the Fourier 256 model outperform the base model across all tasks.

[1] `https://github.com/tlazor/gerb_project`

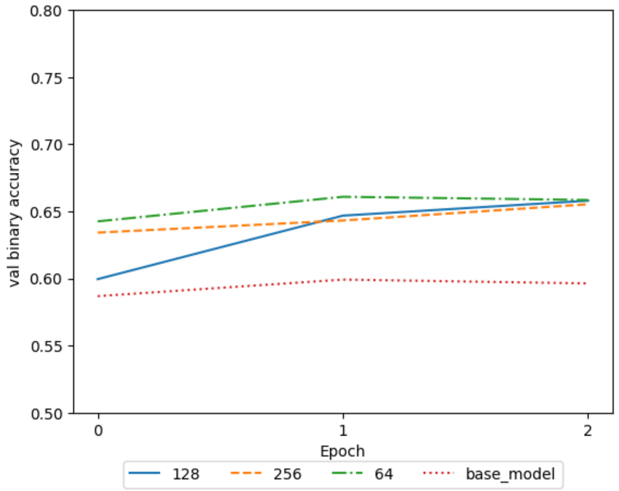

Figure 3: Binary accuracy of models with varying amount of per-paragraph frequency features. ”base_model” is our implementation of Zhang et al. (2021) with no frequency features.

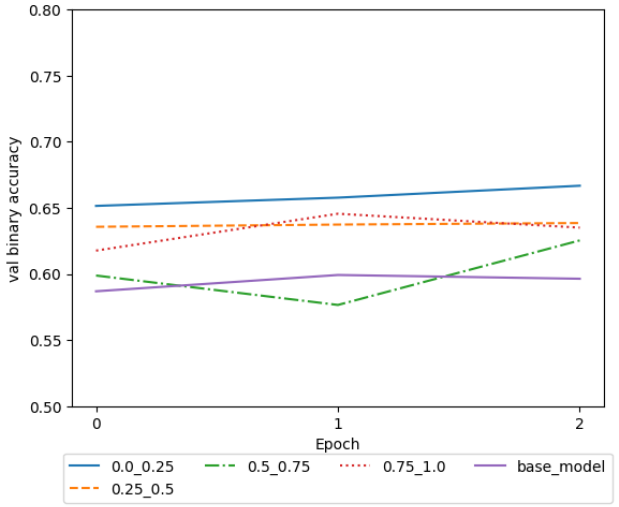

Figure 4: Binary accuracy of the models with a Butterworth filter. 0.0_0.25 refers to the model with filtering of the frequencies from 0% to 25% of the Nyquist frequency. ”base_model” is our implementation of Zhang et al. (2021) with no frequency filtering.

Specifically, the 0-25% filtered model achieves the highest F1 scores in Task 1 and Task 3, while the Fourier 256 model achieves a comparable performance with slightly lower scores in Task 1 and Task 3. Both advanced models demonstrate significant improvements over the base model, particularly in Task 1 and Task 2, with the base model scoring .451 and .469, respectively. This performance comparison highlights the effectiveness of incorporating frequency features and filtering techniques in enhancing model performance on diverse tasks.

Model	Task1.F1	Task2.F1	Task3.F1
0.0_0.25	.715	0.651	0.379
Fourier_256	.711	0.656	0.367
base_model	.451	0.469	0.287

Table 3: F1 scores for the 0-25% filtered model, the 256 frequency feature model, and our implementation of the Zhang et al. (2021) model on validation Set

5. Discussion

5.1. Score Discrepancy

Our implementation of Zhang et al. (2021) did not perform as well on the validation set as the version they created. We believe the differences can easily be explained by the particulars of the FCNN implementation, as well as our algorithm for converting the task3-binary labels into full Task 3 solutions.

We use a simple FCNN model of only three layers with 64, 32, and 1 unit each and trained for only 3 epochs. A larger FCNN, trained for longer, could very conceivably perform better, but the focus of this paper was to explore frequency analysis via comparison to the leading model architecture. We see no reason that our models would not also perform better given more resources. Indeed, the accuracy of our models had not yet dipped, indicating further training might be of benefit.

Another potential source of difference could be the translation of Task 3 binary labels into full Task 3 solutions. Zhang et al. (2021) do not extensively detail their algorithm for this translation and, since we do not achieve perfect accuracy on the Task 3 binary labels, there are many ways to resolve possibly conflicting binary labels. All of the models we implemented used the same binary label translation. For each paragraph, we look at the binary labels comparing that paragraph and all the preceding paragraphs in order. The first label that predicts the pair of paragraphs was written by the same author is taken to be true regardless of any conflicting labels in the following comparisons and the author of both paragraphs is set the same. If none of the preceding labels predict the paragraph was written by the same author, we set the paragraphs author to one higher than the current highest author encountered.

5.2. Frequency Domain Features

Our results show that a mix of Fourier Transform features and BERT embedding features performed better than both no transform features and no embedding features. This suggests that both time and frequency domain are important to consider. This may be because authors change their style in response to other authors and therefor the embedding signal is not time invariant.

In this paper we used the Fourier Transform because it is in widespread use for many domains and has very efficient algorithms for its computation, but it completely disregards the time domain. There are other transforms that can be used to analyse both time and frequency domains, like the Short-Time Fourier Transform and Discrete Wavelet Transform (Portnoff, 1980; Chun-Lin, 2010). Time-frequency analysis of the whole document may be a more fruitful avenue to capture the variation over the document. Additionally, looking at the document as a whole might improve performance, since structuring the problem as a chain of binary comparisons can lead to cascading misclassifications the longer the chain of paragraph pair comparisons.

One important caveat is that the task3-binary label accuracy score for the model with 256 frequency features presents a very different picture to the F1 score it achieves on the Task 3 labels. All the other models show a direct relation between the task3-binary labels and full Task 3 labels. We are not sure why.

5.3. Filtering

We were inspired by analogy with image processing techniques to analyse the frequency domain and, for natural images, the power spectra tends to be concentrated in the lower frequency components (van der Schaaf and van Hateren, 1996). This led us to hypothesize that filtering the high frequencies would lead to improved generalization, but our results show quite the opposite. Filtering the lowest 25% of the spectrum turned out to have the best results, which suggests the low frequencies represent noisy features or features orthogonal to author style.

We speculate that these frequencies could encode information about the conversation topic (topic specific words may occur sentence to sentence within a paragraph, but would also be present in responses to that paragraph written by a different author) or some aspects of conversational convergence (the tendency for conversational partners to adapt to each other's communication patterns) like syntactic complexity (Xu and Reitter, 2016). Further work to better understand what parts of the frequency domain encode specific features could be useful to selectively manipulate the frequencies for specific NLP tasks.

6. Conclusion

We found that including standard signal processing techniques such as Fourier analysis and filtering improves F1 score on a number of author change detection tasks by as much as .264 and even .092 on the hardest task, author attribution.

Signal and image processing are mature fields with many varied tools for analysis and manipulation. Our results show that the application of these well understood tools to NLP may be a very fruitful endeavor.

References

Luis B Almeida. 1994. The fractional fourier transform and time-frequency representations. *IEEE Transactions on signal processing*, 42(11):3084–3091.

Liu Chun-Lin. 2010. A tutorial of the wavelet transform. *NTUEE, Taiwan*, 21(22):2.

Ranjushree Pal. 2017. Comparison of the design of fir and iir filters for a given specification and removal of phase distortion from iir filters. In *2017 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–3. IEEE.

Michael Portnoff. 1980. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):55–69.

A. van der Schaaf and J.H. van Hateren. 1996. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770.

Yang Xu and David Reitter. 2016. Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–448.

Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2021a. Overview of the style change detection task at pan 2021. *CLEF (Working Notes)*, 2936.

Eva Zangerle, Maximilian Mayerl, Michael Tschuggnall, Martin Potthast, and Benno Stein. 2021b. Pan21 authorship analysis: Style change detection, March.

Zhijie Zhang, Zhongyuan Han, L Kong, X Miao, Z Peng, J Zeng, H Cao, J Zhang, Z Xiao, and X Peng. 2021. Style change detection based on writing style similarity—notebook for pan at clef 2021. In *CLEF*.