DISTILLED AND FUSED STYLE EMBEDDING FOR AUTHORSHIP ATTRIBUTION WITH WEIGHTED MULTI-KERNEL ENSEMBLE MODEL

B. Lavanya, R. Sowmiya*

Department of Computer Science, University of Madras, Guindy, Chennai
India

* Corresponding Author, e-mail: lavanyaunom@gmail.com

Abstract: In this age of Artificial Intelligence, the dispute over ownership of the web contents are increasing. In forensic studies and plagiarism cases, determining the author of a document undoubtedly plays a crucial role. The task of Authorship Attribution (AA) involves learning the style of the writers based on available documents to predict the ownership of unknown documents. Hence, capturing the style of writing is primary besides challenging. In this work, key features are systematically extracted, multiple deep-learning models are trained on various aspects of the text, and two novelties: Distilled and Fused Style Embedding (DFSE) and a multi-kernel ensemble model with kernels weighted based on Z-scores are proposed. The performance of the proposed model is superior to the baseline model and many contemporary models for in-genre and multi-topic datasets.

Key words: authorship classification, natural language processing, style embedding, neural network, ensemble model.

1. INTRODUCTION

Informing the audience about the creator of the content is not only their right but also fair recognition for the creators. In a forensic investigation, determining the authors of emails, anonymous letters, etc., might help understand the truth. Similarly, in the Artificial Intelligence (AI)-driven world, to prevent the spread of misinformation and the risk of plagiarism by an AI author, it is crucial to be informed about who wrote what on the web. Also, in curbing the spread of rumours or hate comments on social media, knowing the authors of the comments would prove efficient. Not limited to these applications, the Authorship Attribution (AA) task has a wider scope of usability. To classify the text based on authorship, a pattern outlining the style of the author must be traced out. That implicitly puts forth the question, “What are the stylistic features which distinguishably describe an author?”

In the literature, many studies have shown the impact of lexical and syntactic features [1-5]. Several studies [6-9] also hint at the efficiency of various aspects of the text in stylistic problems. For classification problems, broadly there are two important steps: one corresponds to the informativeness of the feature and the other pertains to the classifier used. Linear Support Vector Classifier (SVC) is a popular kernel-based machine-learning algorithm that has also been useful in AA studies [5]. Considering Linear SVC as the baseline model, this work attempts to unveil the usage of a mixture of distilled features in AA and the strength of the multi-kernel ensemble SVC model. Deep-learning neural models are exploited to distil the features and derive the fused embedding, termed Distilled and Fused Style Embedding (DFSE). Then, a novel ensemble classifier made of multiple kernels, weighted by Z-scores is developed (taking into account the significance of ensemble learning in recent studies [10]) to carry out the task of identifying the authorship of the input text.

The datasets utilised have writings in the same genre in both train and test sets (in-genre). A custom dataset of fan fiction is derived from a bigger dataset built for a different task [11, 12]. It contains uniformly clipped excerpts and is a multi-topic dataset (i.e., for a few authors, the topic in their train set differs from their test set). This work presents several notable contributions, including:

- Extracting information-rich key style features through a systematic process.
- Deriving a new dataset for AA (which will be made public in the repository: [13] upon publication)
- Extracting a novel style embedding achieved by learning from various aspects of the text: lexical, phonological, morphological, semantic and syntactic features.
- Building a multi-kernel ensemble model weighted by Z-Scores.

Moreover, the proposed classifier proves to be more effective than many contemporary models and through proper training, our model can also help in other classification tasks. The remaining section of the paper is divided into the following sections:

Section 2, discusses the related works. Section 3, explains the nature of the datasets used and the procedure employed to extract key features. Section 4, consists of the experimental procedures – preparation of DFSE and ensemble model. Section 5 discusses the results and analyses errors and finally, Section 6, summarises the work.

2. RELATED WORKS

Across different studies set out to enhance the success rates in AA tasks, the lexical features are ubiquitous. This pattern is widely seen since it has been promising in capturing certain valuable authorial writing styles. However, the combination of various aspects of the text has also proved to improve the accuracy of identifying the authors. In [4], the authors fused the lexical and syntactic information (like Parts-of-Speech (POS), etc.,) of texts and introduced the structural information with the help of a Hierarchical Attention Network (HAN), developing the Style HAN. The ablation study showed that individually Lexical HAN and Syntactic HAN performed inferiorly compared to their fusion Style HAN. Similar options of using lexical and syntactic features to accomplish AA are observed in the work of [3] where the Siamese network, Bi-directional Long-short Term Memory (BLSTM) and attention mechanisms are incorporated.

There are works which ventured to produce general embedding that could be used for style-based classification of text. One such notable work is from the authors of [14] in which a universal style encoder was built using the BLSTM encoder and soft voting ensemble classifier. Another important tool for extracting stylistic information is StyloMetrix [7, 15] which is an open-source tool to compute and collect lexical, syntactic, tenses-related, psycholinguistic and other information about the input text. The readings are normalised over the length of the text and it is made available for multiple languages as well including English. For English, 196 features are available. In our work, a subset of those features along with an assortment of other features are employed.

In the work of [1] a strong opinion is put forward about the usefulness of n-grams in stylistic classification as opposed to topical classification and the authors achieved state-of-the-art (SOTA) performance in AA with continuous n-gram representations. Our work finds relevance to the potential of n-grams by considering them in the feature group. Although lexical features and syntax information are prevalent in AA tasks, the efficiency of varied aspects of texts – like the inclusion of semantic features – are emphasised in the work of [6] in which the authors added multifarious features to the feature set formed basically with [1] to achieve SOTA performance in the AA task. In [2] the authors carried out the AA task by utilising syntax trees, syntax-level encoding, Convolutional Neural Networks (CNN) with Max pooling mechanism and content-level encoding. Authors of [16] analysed the relationship between different features and the characteristics of datasets in the AA problem. They concluded that for datasets with high topical variance, content-related features are more helpful and style-related features would help for datasets with less topical variance. The authors of [17] explored the efficiency of the multi-channel self-attention network incorporating multiple features to distinguish one author from another. Our work (illustrated in Figure 1) shares a commonality with [17] in terms of employing n-grams at both character and word levels.

3. DATASET PROCESSING

Four balanced datasets used in this study as shown in Table 1 pertain to in-genre and multi-topic setups. They are as follows:

- CCAT_50 and CCAT_10 are benchmark in-genre datasets (i.e.) all the articles are from news columns. (The underscore precedes the number of authors present in the dataset.)
- FANFIC_22 is a custom-prepared dataset derived from the works of [12] and [11] which was originally built for the author verification problem. This is a relatively smaller dataset with 22 authors each having 20 instances of fanfiction excerpts. 9% of the authors are considered multi-topic authors, indicating that they have written for multiple fanbases. If the models have sufficient generality to encompass the style, they can classify these writers accurately. Furthermore, it should be acknowledged that certain authors can write across different sub-genres, such as romance and horror. This would allow the model to assimilate and engage with various styles without bias towards specific sub-genres.
- IMDB_62 is also a well-known benchmark dataset for AA containing around 61987 movie reviews from 62 different users. Given the substantial size of the

original dataset and the computational resources needed for its processing, a random sample of this dataset is utilised for the study.

3.1. Feature Extraction

Stylistic features include lexical features regarding passive voice, active voice, type-token ratio, repetitive word, possessives, punctuation counts, first, second and third-person singular pronouns and third-person plural pronouns, a total of 13 features arrived with a subset of StyloMetrix embeddings [7], in addition to syntactic features (POS unigrams and bigrams), linguistic features (frequency of hapax legomena and dis legomena), phonological features (syllable counts), semantic features (sentence labels through semantic clustering), and frequency information at word levels and character levels (n-grams) [5]. As for the retrieval of sentence labels, the below steps are followed:

- a sentence transformer [18] is employed to extract the embedding
- the embedding (vector, say e_i) is normalised over the L2 norm as shown in the formula: $\frac{e_i}{\|e_i\|}$ $where\ \ \|e_i\| = \sqrt{\sum_{j=1}^{n} e_{ij}^2}$ and e_{ij} is jth component of e_i
- the embeddings are passed into the agglomerative clustering algorithm
- the labels produced for both the train and test sets are collected

As a starting point for semantic clustering, we initially utilised the default configuration of two clusters in the scikit-learn implementation of Agglomerative clustering [19]. This bi-cluster setup proved informative when validated through a high mutual information score. Moreover, increasing the number of clusters could introduce superfluous complexity without substantial enhancements in performance, as evaluated through further trials.

For syntax representation, POS tags formed the basis. Unigrams and bigrams of 10 POS tags are recorded. Moreover, the linguistic features popular in AA research: hapax legomena and dis legomena, which record rare word occurrences in an author's writing are extracted. For selecting character and word level n-grams (bigrams, trigrams, quadrigrams and pentagrams), K-Best features are selected using the F-test and grid search of K (where K = 500, 600, 700, 800) based on accuracy.

Further, average sentence length, total uppercase letters, total short words and digit frequencies are also incorporated for preparing the stylistic feature group. Firstly, mutual information scores of the above information are computed during which, the features scoring 0.0 are eliminated and the remaining are retained (say, Set A). All the features in Set A, except the sentence labels, are normalised based on the text length for consistency. Secondly, those scored above or equal to 0.09 are taken (say, Set B). As a customary analysis process, both Sets are evaluated by the baseline model. Ultimately, Set A (explained in **Select_stylistic_features()** in pseudocode in Figure 2) is carried forward as the *finalised key stylistic feature group* based on its better performance.

Lower-cased and tokenised text is used for training Glove [20] and Fasttext [21] models (obtained from the functions **Get_glove_embed()** and **Get_fasttext_embed()** respectively as per pseudocode in Figure 2). The maximum sequence length is calculated from the mean ($\bar{x}_{dl}$) and standard deviation (σ_{dl}) of the document length (dl) by the formula: $round(\bar{x}_{dl} + \sigma_{dl})$. Details about the vocabulary size and maximum sequence length, alongside the extracted stylistic feature set size, are presented in Table 1.

Since accuracy provides a reliable representation of the model performance with balanced datasets, it is primarily used for evaluation in this study, besides F1 score, precision and recall for error analysis.

Table 1. Dataset details

Dataset	*No. of. authors*	*Genre*	*Instances per author*	*max sequence length*	*vocabulary size*	*No. of. Finalised stylistic features*
33	*50*	*News*	*100*	*636*	*17693*	*702*
CCAT_10	*10*	*News*	*100*	*643*	*17643*	*743*
Fanfic_22	*22*	*Fanfiction*	*20*	*1000*	*17355*	*569*
Imdb_62	*62*	*Reviews*	*100*	*561*	*49386*	*599*

4. DFSE AND ENSEMBLE CLASSIFIER

In addition to the extraction of finalised stylistic features, the proposed methodology comprises two phases, as depicted in Figure 1 and detailed in the pseudocode in Figure 2. They are as follows:

- Deriving DFSE through fusing features of several aspects and distilling them through deep learning algorithms (proposed feature) and
- Building a Z-score-weighted ensemble SVM model with multiple kernels for classifying the authors (proposed method/ feature vector model).

4.1. Deriving Distilled and Fused Style Embedding

In this phase, three deep learning models are trained with different aspects of text and their weights are extracted to form the proposed feature, DFSE.

Style NN (S) – This is a feed-forward neural network with five hidden layers activated with the ReLU activation function and three dropout layers trained with the group of finalised stylistic features prepared as mentioned in Section 3 (and dimensions mentioned in Table 1). The output layer is operated by the softmax activation function. Also, early stopping is used.

Glove-LSTM (L) – As the name implies, pre-trained Glove embeddings [20] of 100 dimensions (D) available on the Stanford website are used to train a Long-Short Term Memory (LSTM) model. This is supposed to capture the contextual aspect of the text or the semantics.

Fasttext-CNN (C) – The Fasttext [21] model trained in an unsupervised fashion over individual corpora is used to capture the morphological aspect of the text. The extracted embeddings of 200 D are used for training the CNN model with a max pooling layer.

For both the Glove-LSTM and Fasttext-CNN models, two embedding layers – one trainable and the other non-trainable – are set and merged to connect to the rest of the layers. As a common setup for all the models, a modified exponential scheduler ($LR(t)$) as in Equation 1, with epoch (t) and a customisable parameter (num) is implemented with categorical cross-entropy loss, and the Adam optimiser with the initial learning rate (η) set to 0.001.

$$LR(t) = \begin{cases} \eta & if\ t < num \\ \eta \cdot e^{-0 \cdot 1} & if\ t >= num \end{cases} \tag{1}$$

After the training, the outputs of the train and test set from the last hidden layer ($h_{(n-1)y}$) are extracted from each model and fused (concatenated $\oplus$) to obtain 448 dimensions of distilled style embedding, termed DFSE. Here, in Equation 2, n-1 denotes the last but one of 'n' dense layers and y is the output of the hidden layer h. Before and after fusing, the weights are tested against the baseline Linear SVC model.

$$DFSE : S\left(h_{(n-1)y}\right) \oplus \mathrm{L}\left(h_{(n-1)y}\right) \oplus \mathrm{C}\left(h_{(n-1)y}\right) \tag{2}$$

The stratified 5-fold cross-validated performance of DFSE and StyloMetrix [7] embedding, variedly treated with the Baseline Linear SVC model are tabulated in Table 2. As for the feature treatments of StyloMetrix features, adding features mentioned in Section 3.1, deleting the features with no information (or with zeroes in all the rows) and retaining features with standard deviation (σ) and mutual information score (mutual info) greater than or equal to threshold 0.09 are considered.

Table 2. Comparison of DFSE with variants of StyloMetrix embeddings based on accuracy

Features with Linear SVC	***CCAT_10***	***CCAT_50***	***Fanfic_22***	***Imdb_62***
Proposed DFSE	***82***	***76.54***	***54.77***	***49.66***
StyloMetrix (all 196 features)	*38.1*	*19.12*	*36.36*	*43.21*
StyloMetrix with σ!=0	*38.1*	*19.10*	*36.36*	*43.21*
StyloMetrix with σ>=0.09	*38.1*	*18.76*	*36.13*	*42.95*
subset of StyloMetrix with ngrams and mutual info !=0	*46.7*	*19*	*42.26*	*27.27*
subset of StyloMetrix with ngrams and mutual info >=0.09	*46.6*	*18*	*42.26*	*27.27*

4.2. Weighted Ensemble Model

In the second phase, 10 different kernels such as additive chi-square, chi-square, linear, poly kernel, polynomial kernel, Radial Basis Function kernel (RBF), Laplacian kernel, sigmoid, cosine and Neural Tangent Kernel (NTK) are set up to train SVC models. The weights of these kernels are decided by the Z-score of their performance scores as detailed in the pseudocodes in Figure 2.

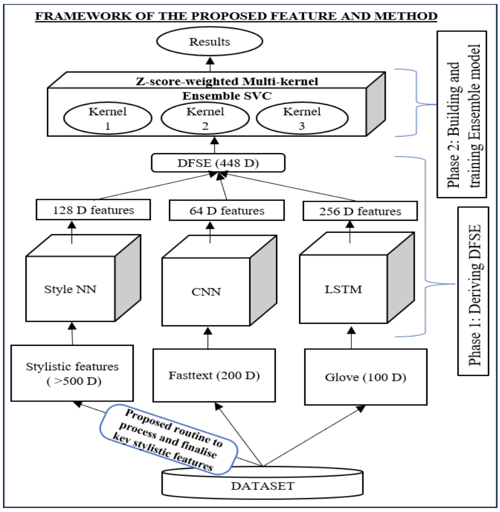

Figure 1. The architecture of the proposed methodology.

Pseudocode 1 DFSE feature vector model

```
Require: Models: C,L,S, Kernels: K_1, K_2, ..., K_m and Input x
Train Model C with the output from Get_fasttext_embed(x)
Extract C(h_(n-1)y) /* vectors from last but one dense layer of C */
Train Model L with the output from Get_glove_embed(x)
Extract L(h_(n-1)y)
Train Model S with the output from Select_stylistic_features(x)
Extract S(h_(n-1)y)
DFSE := S(h_(n-1)y ⊕ L(h_(n-1)y ⊕ C(h_(n-1)y /* ⊕ denotes concatenation */
([Bk_1, Bk_2, Bk_3], [Ak_1, Ak_2, Ak_3]) ← Selective_kernel_optimisation([K_1, K_2, ... K_m], DFSE, evaluation_metric) /* Get top 3 kernels and their accuracy scores */
([Wk[0], Wk[1], Wk[2]] ← Set_weights_by_z_scores ([ Bk_1, Bk_2, Bk_3 ], [ Ak_1, Ak_2, Ak_3 ]) /* Get weights of the kernels */
Build weighted multi-kernel SVC ensemble classifier M: M([(Bk_1, Wk[0]), (Bk_2, Wk[1]), (Bk_3, Wk[2])])
Return: scores : M (Cross-validated accuracy, DFSE)
```

Pseudocode 2 Select_stylistic_features()

```
Require: Different aspects of the text x
Lex :← lexical features (i.e. subset of StyloMetrix)
Syn :← syntactic features
Ling :← linguistic features
Phon :← phonological features
Sem :← sentence cluster labels
Ngrams :← processed ngrams /* k-best-ngrams are selected through f-test and grid search on k parameter*/
Normalise Lex, Syn, Ling, Phon and Ngrams based on text length
Concatenate all the above features as Fs
Calculate mutual information score {Mu(fs)|fs ∈ Fs}
SetA :← {Mu(fs) <> 0|fs ∈ Fs}
SetB :← {Mu(fs) ≥ 0.09|fs ∈ Fs}
Compare the performance of SetA and SetB with the Baseline model
Return: Best performing Set as the finalised stylistic features
```

Pseudocode 4 Set_weights_by_z_scores()

```
Require: ([ Bk_1, Bk_2, Bk_3 ], [ Ak_1, Ak_2, Ak_3 ])
Z :← Z scores of [ Ak_1, Ak_2, Ak_3 ]
ƒ :← frequencies of unique values in Z
Wk :← [ ] /* For ease of understanding, indices searching and updating are not depicted below*/
if |ƒ| == 2 then
    if ƒ[0] == 1 then
        Wk[0] = 1 ; Wk[1] = Wk[2] = 3
    end if
    if ƒ[0] == 2 then
        Wk[0] = Wk[1] = 1 ; Wk[2] = 3
    end if
end if
if |ƒ| == 1 then
    Wk[0] = Wk[1] = Wk[2] = 1
end if
else
    if |{z ∈ Z|z <= 0}| == 2 then
        Wk[0] = 1 ; Wk[1] = 2 ; Wk[2] = 3
    end if
    if |{z ∈ Z|z <= 0}| == 1 then
        Wk[0] = 1 ; Wk[1] = Wk[2] = 2
    end if
Return: [Wk[0], Wk[1], Wk[2]]
```

Pseudocode 3 Selective_kernel_optimization()

```
Require: [K_1, K_2, ... K_m], DFSE, evaluation_metric
Build m SVC models with K_1, K_2, ... K_m
Train and test it on DFSE
Evaluate on the evaluation_metric /* default metric is accuracy */
Return: Top 3 kernels and their corresponding scores: ([Bk_1, Bk_2, Bk_3], [Ak_1, Ak_2, Ak_3])
```

Figure 2. Pseudo code of the proposed work

5. RESULTS AND DISCUSSION

The novel ensemble model is validated through 5-fold stratified cross-validation and the mean scores of the test sets are compared against contemporary works and tabulated in Table 3 where the readings in **bold** represent the highest scores in other works and proposed works. It is evident that the proposed feature and model outperform the baseline model and many SOTA models while they perform comparatively to some other SOTA models. This proves the applicability of DFSE and the proposed ensemble classifier in the AA tasks. Also, the performance of DFSE compared with the variants of StyloMetrix embeddings classified by the baseline Linear SVC model as recorded in Table 2 shows the relative superior performance of DFSE. To understand the reason for the shortcomings of the model, an error analysis is carried out.

Table 3. Comparison of proposed methods and features with existing works based on accuracy

Methods	***CCAT_10***	***CCAT_50***	***Fanfic_22***	***Imdb_62***
POS-HAN [4]	*63.14*	*41.3*		
Structural-HAN [3]	*65.4*	*45.2*		
[1]	*74.8*	*72.6*		
content features [16]	*76.2*	*72.88*		
[17]	*77*	*65.64*		
with all their proposed features [6]	*80.07*	*77.75*		
Baseline Linear SVC with DFSE	*82*	*76.54*	*54.77*	*49.66*
The proposed model and DFSE	***83.3***	***79.04***	***55.91***	***51.46***
[2]	*88.2*	*81*		
***Style-HAN* [4]**	***90.58***	***82.35***		

5.1. Error Analysis and Ablation Study of DFSE

Ablation studies as reported in Table 4, manual analysis of the text of certain authors with classification reports and loss curves of the models are the tools exploited to analyse the errors.

Table 3 hints at the wide scope for improvement especially in Fanfic_22 and sampled IMDB_62. On investigating we can observe that both the former and the latter have limited data in terms of fewer instances per author and short texts per author respectively. On increasing the instances, IMDB_62 may yield better results; this could be explored in the future. Figure 3 (i) depicts that contextual features (from L) and morphological features (from C) help differentiate some authors correctly whereas some authors have more stylistic variations which are spotted by the style embeddings of DFSE. For instance, author ‘0’ has achieved a precision of 100% in Figure 3 (i) owing to the style embeddings (from S) whereas Figure 3 (ii) depicts scores as low as 0%. Also, the variant in Figure 3 (iii) could differentiate authors ‘0’, ‘1’, ‘7’, ‘9’ and ‘19’ as well, which CNN+LSTM in Figure 3 (ii) failed to achieve indicating strong stylistic variations in these authorial writings. Figure 3 (i, ii, iv) shows features that are capable of attributing author ‘20’ correctly unlike that in Figure (iii); this highlights the contextual and lexical-rich writing of that author. DFSE benefits from the mixture of embeddings, although the strength of stylistic features has not been utilised by it completely (eg. the scores against author ‘1’ across subplots in Figure 3). This could be due to the tuning and optimization required in the architecture of model S. This assumption is also validated through the Figure 4 (i) loss curves. However, while considering the multi-topic authors (authors ‘5’ and ‘6’) in Fanfic_22 (in Figure 3), reasonable generalisability is found to be achieved. Also, the performance of DFSE is observed to be very effective in CCAT_10 (Figure 4 (ii)).

Figure 3. Classification reports from the proposed ensemble model (from one of the cross-validation runs) with (i) DFSE, (ii)its constituent (CNN+LSTM), (iii) StyloMetrix variant and (iv) the input to model S of the Fanfic_22 dataset.

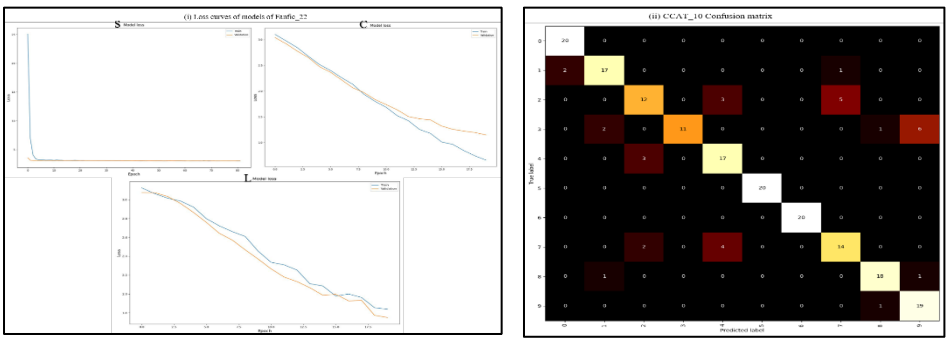

Figure 4. (i)Loss curves of Fanfic_22. The blue curves denote the train losses and the orange curves denote the validation losses (ii) CCAT_10 confusion matrix

Observing the loss curves of the models in all the datasets, chances of slight overfitting due to increased epochs, sufficient learning and underfitting due to early stopping are traceable in C, L and S models respectively. As put forth by the authors of [16], CCAT_10 and CCAT_50 are rich in author-representative context rather than style, thus model C and model L uplifted the performance alongside the minimal contribution of model S's feature vectors. The writers whose styles cannot be captured only through, for instance, n-grams or lexical features, still have possibilities to be represented well through DFSE. In other words, better generalisability of DFSE can be attributed to the various aspects of text captured to train the models which also adds to its robustness.

*Table 4. Ablation study report of components of DFSE with the proposed model (Here, * and ** denote statistical significance - p value <0.05 and <0.01 respectively when paired right-tailed t-test is performed between the scores of DFSE and its constituents)*

Dataset	Features	Dimension	Accuracy	F1 score
CCAT_50	**DFSE**	448	**79.04**	**78.72**
	CNN + LSTM	320	78.98	78.65
	CNN + STYLE NN	384	77.86**	77.43**
	STYLE NN + LSTM	192	61.84**	61.89**
	CNN	256	77.98**	77.58**
	LSTM	64	62.58**	62.59**
	STYLE NN	128	2.00**	0.08**
CCAT_10	**DFSE**	448	**83.3**	**83.06**
	CNN + LSTM	320	83	82.84
	CNN + STYLE NN	384	81.4*	80.82*
	STYLE NN + LSTM	192	72.6**	71.34**
	CNN	256	82.8	82.47
	LSTM	64	72.5**	72.07**
	STYLE NN	128	10.1**	8.25**
FANFIC_22	**DFSE**	448	**55.91**	**53.54**
	CNN + LSTM	320	55.45	52.73
	CNN + STYLE NN	384	51.82**	49.77*
	STYLE NN + LSTM	192	50.22**	49.96*
	CNN	256	51.13**	49.09**
	LSTM	64	49.77**	49.23**
	STYLE NN	128	7.95**	2.65**
IMDB_62	**DFSE**	448	**51.47**	**52.54**
	CNN + LSTM	320	50.04*	51.07*
	CNN + STYLE NN	384	43.67**	43.71**
	STYLE NN + LSTM	192	28.37**	30.07**
	CNN	256	44.35**	44.42**
	LSTM	64	29.11**	30.90**
	STYLE NN	128	1.5**	0.68**

6. CONCLUSION

This work aimed to enhance the accuracy of the authorship attribution task by using various aspects of the text and validated its findings using three benchmark datasets: CCAT_50, CCAT_10, IMDB_62 and a custom-derived Fanfic_22. Multiple aspects of the text are extracted systematically to capture the style print of the authors, later selectively distilled by deep neural networks to produce novel DFSE embeddings. This proposed feature produces results superior to SOTA AA embeddings. Secondly, a novel ensemble model with unique kernels weighted leveraging Z-scores is developed. The proposed ensemble model outperformed the baseline and several other SOTA models. Yet, optimisation in terms of the architecture of neural networks and hyperparameter tuning could be considered for enhancement in the future. Also, the applicability of DFSE in other style-related problems could be explored further.

ACKNOWLEDGEMENT

This work is supported by the University Grants Commission (UGC, India) under the Junior Research Fellowship scheme (JRF).

REFERENCES

[1] Sari, Y., Vlachos, A., Stevenson, M. Continuous N-gram Representations for Authorship Attribution. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, April 2017, pp. 267–273.

[2] Zhang, R., Hu, Z., Guo, H., Mao, Y. Syntax Encoding with Application in Authorship Attribution. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2018, pp. 2742–2753. DOI: 10.18653/v1/D18-1294.

[3] Jafariakinabad, F., Hua, K. A. A Self-Supervised Representation Learning of Sentence Structure for Authorship Attribution. *ACM Transactions on Knowledge Discovery from Data*, vol. 16, no. 4, 2022, pp. 1–16, DOI: 10.1145/3491203.

[4] Jafariakinabad, F., Hua, K. A. Style-aware Neural Model with Application in Authorship Attribution, *CoRR*, vol. abs/1909.06194, 2019, pp. 1-8.

[5] Misini, A., Kadriu, A., Canhasi, E. Authorship classification techniques: Bridging textual domains and languages, *International Journal on Information Technologies and Security*, vol. 16, no. 1, 2024, pp. 27–38, DOI: 10.59035/UKBE1226.

[6] Verma, G., Srinivasan, B. V. A Lexical, Syntactic, and Semantic Perspective for Understanding Style in Text, Sep 2019, URL: http://arxiv.org/abs/1909.08349 (Visited on 08.12.2023)

[7] Okulska, I., Stetsenko, D., Kołos, A., Karlińska, A., Głąbińska, K., Nowakowski, A. StyloMetrix: An Open-Source Multilingual Tool for Representing Stylometric Vectors, Sep 2023, URL: http://arxiv.org/abs/2309.12810 (Visited on 10.12.2023)

[8] Nini, A., Grant, T. Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis. *International Journal of Speech, Language and the Law*, vol 20, no. 2, 2013, pp. 173-202. DOI: https://doi.org/10.1558/ijsll.v20i2.173

[9] Fourkioti, O., Symeonidis, S., Arampatzis, A. Language models and fusion for authorship attribution. *Information Processing & Management*, vol. 56, no. 6, 2019, pp. 102061. DOI: 10.1016/j.ipm.2019.102061

[10] Abbasi, A., Javed, A. R., Iqbal, F., Jalil, Z., Gadekallu, T. R., Kryvinska, N. Authorship identification using ensemble learning, *Scientific Reports*, vol. 12, no. 1, 2022, DOI: 10.1038/S41598-022-13690-4.

[11] Saedi, C., Dras, M. Siamese networks for large-scale author identification, *Computer Speech & Language*, vol. 70, pp. 1-15, 2021, DOI: 10.1016/j.csl.2021.101241.

[12] Fernandes, N., Dras, M., McIver, A. Generalised Differential Privacy for Text Document Processing. *Principles of Security and Trust (POST)*, April 2019, pp. 123–148. DOI: 10.1007/978-3-030-17138-4_6.

[13] *Github Repository to access fanfic_22 dataset*, Github, 1 August 2024 URL: https://github.com/kaapivalli/Research (Visited on 01.08.2024)

[14] Parikh, K. S., Venkataram, V., Kalita, J. Towards a Universal Document Encoder for Authorship Attribution, Aug 2018, URL: https://faculty.uccs.edu/jkalita/wp-content/uploads/sites/45/2024/02/00Proceedings2018Reduced.pdf#page=20 (Visited on 09.12.2023)

[15] Okulska, I., Zawadzka, A. Styles with Benefits. The StyloMetrix Vectors for Stylistic and Semantic Text Classification of Small-Scale Datasets and Different Sample Length. *Proceedings of the 3rd Polish Conference on Artificial Intelligence*, Gdynia, Poland, April. 2022, pp. 176-180.

[16] Sari, Y., Stevenson, M., Vlachos, A. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August 2018, pp. 343–353.

[17] Wu, H., Zhang, Z., Wu, Q. Exploring syntactic and semantic features for authorship attribution, *Applied Soft Computing*, vol. 111, 2021, pp. 1-11, DOI: 10.1016/j.asoc.2021.107815.

[18] Reimers, N., Gurevych, I. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, Aug 2019, URL: http://arxiv.org/abs/1908.10084 (Visited on 10.12.2023)

[19] *Agglomerative clustering: Scikit Learn API reference*, Scikit learn, URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html (Visited on 11.12.2023)

[20] Pennington, J., Socher, R., Manning, C. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA, October 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[21] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, vol. 5, 2017 pp. 135–146, DOI: 10.1162/tacl_a_00051.

Information about the authors:

Dr B. Lavanya is an Associate Professor at the Department of Computer Science at the University of Madras located in Guindy, Chennai, India with more than two decades of experience in teaching and research. Her areas of research interest include data mining, bioinformatics, big data analytics, and natural language processing.

R. Sowmiya is pursuing her PhD with a UGC JRF fellowship at the Department of Computer Science at the University of Madras, Chennai, India. She is an enthusiastic researcher with research interests in Natural language processing, Data mining, Stylometry and Kernel functions.

Manuscript received on 02 October 2024