TOPIC AND STYLE: AUTHORSHIP ATTRIBUTION WITH MULTI-VECTOR META-FEATURES AND STACKED CLASSIFIER

B. Lavanya, R. Sowmiya*

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

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

Abstract: Authorial rights are meaningful only when a document is promptly attributed to its legitimate creator. While an author's unique topical preferences may pinpoint them within a group of thematically-diverse writers, attribution becomes considerably challenging in cases such as judicial opinions – e.g., attributing a criminal case judgement to one among hundreds of judges heading criminal courts – due to overlapping contextual jargon. In such situations, integrating both topical and stylistic cues can enhance the reliability and generalisability of authorship attribution. This work proposes a custom stacking ensemble model trained on diverse aspects of text, later giving rise to novel features: Multi-Vector Meta-Features. The proposed method demonstrates statistically significant boosted accuracy scores across three out of four datasets, outperforming baseline model and recent stacking ensemble models.

Key words: authorship attribution, bertopic, stacked classifier, multi-vector, natural language processing.

1. INTRODUCTION

As a commonly used method of detecting texts' provenance, the stylometric methods are advancing in the direction of Machine Learning (ML) and Deep Learning (DL) in recent trends [1]. A wide range of features based on textual aspects and statistics utilized in traditional and DL methods contribute to stylometric studies. Attributing a text to its author based on enough historical evidence of the author's works – Authorship Attribution (AA) – has an indispensable role in unveiling the authors of posthumous/debated literary works [2], digital forensic pieces of evidence, offensive social media messages, fake news and ghostwritten academic works, etc. Research shows that stylometric methods can be employed with Named Entity Recognition models [3] to serve the purpose and have wide applications as unique as in mediumistic writings [4].

Although the interests of finding the best distinguishing style markers are the driving factor of AA studies, critical variables of datasets were emphasised in recent studies, such as the topical similarity among the authors [5] and the dataset size [6]. This effort focuses on the former parameter and strives to find the best generalised solution.

The role played by the authors' topical preferences in AA datasets has raised concern about analysing the topical nature of the dataset before carrying out feature extraction. AA may be skewed towards topic categorisation in some datasets due to authors' apparent topical interests [5]. Consequently, topic modelling can be used to evaluate the authors' topical dissimilarity, which further could hint at the effective features to be utilised to achieve better results. i.e. If the dataset has greater topical dissimilarity, then content/topical features may yield superior performance. However, it is less feasible in real-time scenarios due to this ongoing concern about the data characteristics in order to carry out the attribution task. The goal of this work is to develop a better set of features that are sufficiently generalisable to function effectively in datasets that have a topical and/or stylistic inclination.

To achieve this, using a Pretrained Language Model (PLM), Bi-directional Encoder Representations from Transformers (BERT), which has proved to perform better in capturing the contextual and topical aspects of the text, and BERTopic, a topic modelling technique and set of quantitative stylistic features provided by an open-source tool, StyloMetrix [7], a novel multi-vector-based feature is developed. It significantly outperforms the baseline [5]. The main contributions of this paper are as follows:

- An effective pre-processing technique to facilitate more coherent topical clustering of text using BERTopic is recommended.
- A custom stacked classifier is developed, with each optimised meta-classifier trained on different types of textual features, respectively.
- A powerful combination of generalised meta-features, namely Multi-Vector Meta-Features (MVMF), is proposed for authorial classification of text that can handle datasets biased towards authors with topical similarity and/or stylistic similarity. Supporting files will be available at [8] upon publication.
- The findings highlight that although a dataset is topically diverse (author-wise), best results can be achieved by employing topical, content and stylistic features appropriately, thereby eliminating the overhead of considering the topical divergence of the dataset before working on it.
- Ablation study and error analysis have been carried out to emphasise the significance of MVMF and the established stacking method.

The paper proceeds with the following subsequent sections: Section 2 throws light on existing works, Section 3 describes the dataset employed in the study, the components and preparation of MVMF are explained in Section 4, the results and ablation study come under Section 5 and Section 6 respectively, and the conclusion is provided in Section 7.

2. RELATED WORKS

As evident in many existing AA studies, multiple combinations of features or multi-vector representations are found to produce better accuracy: using the lexical, syntactic and semantic features (output of a knowledge parser, semantic roles, etc.) combined [9] showed that higher accuracy could be achieved compared to standalone continuous n-grams [10]. A few of the multiple aspects of text collectively contributing to AA efficiency, as found in literature, include Parts-of-Speech (POS), mean measure of lexical attributes (like average word count and average sentence count, etc.), lexical features such as n-grams [11], etc. Although only a few works employ topic modelling as a tool to enhance AA, the Latent Dirichlet Allocation (LDA) [5] model is ubiquitous among them (while this presented work uses the state-of-the-art (SOTA) BERTopic modelling technique). There are several DL models, such as Hierarchical Attention Networks (HAN) [12], Convolutional Neural Networks (CNN) [13], and Siamese networks, as well as innovative ML models using ensemble architectures [14-16] that show better results in AA tasks.

Employing BERT in AA studies is a very recent practice. In BERTAA [14], two individual Logistic Regression (LR) models are trained on stylistic and n-gram features, respectively, and their output probabilities are concatenated with those of the BERT classifier (i.e. output from a linear network appended to BERT). It implements the stacking ensemble technique employing BERT as similar to another study [15] where BERT is a baseline and a dynamic selection of base models (LRs) is incorporated to build the ensemble. Table 1 reports the different techniques employed in various existing works compared to the proposed work, for a better overview.

Table 1. Techniques employed in existing works and the proposed work

Reference	***Multi-vector representations***	***BERT***	***Stacking ensemble***	***Topic modelling***
[14]	✓	✓	✓	
[15]	✓	✓	✓	
[5]	✓			✓
[9]	✓			
[16]	✓			
[17]	✓			
MVMF	✓	✓	✓	✓

3. DATASETS

Three of the benchmark AA datasets, namely, CCAT_10, CCAT_50 and Judgement_3, are used in this study, along with Fanfic_22, the fan fiction AA dataset derived[16] from original work[18] , whose details are provided in Table 2. Judgement_3 dataset[19] is sampled appropriately to attain balance in each class. In the names of the datasets, following the '_' or a space is the number of authors in each dataset.

Table 2. Dataset details

Dataset(dt)	***Genre***	***No. of. authors***	***Nature of dt***	***Average words/doc***	***Instances per author***
CCAT_50	*News*	*50*	*Topical*	*584*	*100*
CCAT_10	*News*	*10*	*Topical*	*580*	*100*
Fanfic_22	*Fiction*	*22*	*Stylistic*	*1000*	*20*
Judgement_3	*Legal*	*3*	*Stylistic*	*2367*	*187*

3.1. Technologies utilised

As topical, semantic and stylistic aspects of the text are holistically considered to handle the skewness present in either topical and/or stylistic direction, three latest efficient technologies are utilised in this study. They are as follows:

BERT: A well-known PLM called ‘bert-base-uncased’ was created by Google as part of the [20] study. It was developed with twelve transformer encoder layers, or blocks, and attention heads. The dimensions of the embeddings produced by this model are 768. The term ‘uncased’ in its name describes its inability to differentiate between cases, such as ‘Apple’ and ‘apple’. It splits the words into smaller sub-word chunks before encoding using the ‘WordPiece’ tokenisation technique.

In this regard, BERTAA [14], as discussed in Section 2, comprises three parts (base models) from which class probabilities are collected: a pre-trained BERT classification model, an LR trained with stylistic features such as text length, average word length, word count, short word count, character frequencies, punctuation frequencies, hapax-legomena and percentage of capital letters and digits, and another LR trained with character bi-grams and tri-grams (a total of 100 each). To train the final (meta-) LR model, the obtained probabilities are subsequently concatenated and fed as input.

The present work, building a stacking ensemble technique involving BERT, BERTAA, is included in the comparison of results. The training of the first base model in the present study differs from BERTAA by the utilisation of an LR (base) model instead of BERT’s (sequence model’s) classification head. Having this difference, it was intriguing to examine whether switching BERTAA’s first base model to LR could produce a notable change in the outcomes. Hence, a revised model named ‘mod_BERTAA’ is also developed and compared.

BERTopic modelling: As introduced in [21] BERTopic creates topic representations through the following phases: employs a PLM to convert each document to its embedding representation, applies a dimensionality reduction technique on the embeddings and implements the Class-based variation of Term Frequency and Inverse Document Frequency technique (C-TF-IDF) to derive topic representations from the cluster of documents. Topic distributions place documents in a multi-vector space based on their semantics; i.e. semantically comparable documents are presumed to have the same topics. Document embeddings are dense representations of sentences in the document produced with the help of PLM and Sentence-BERT. Since the PLM used in the BERTopic is customisable, this technique adapts to the advancement in the field.

Meanwhile, in [5], the Latent Dirichlet Allocation (LDA) topic modelling method was employed for assessing the characteristics of the datasets, where each word in every document is assigned a randomly chosen topic. This random assignment provides a subject representation for each document and word distributions for each topic. LDA iteratively refines these assignments by reallocating every word to a topic after examining it. The probability of a word being associated with a topic is influenced by the frequency of both the topic and the word within the document. Following multiple iterations, the algorithm progressively attains a stable state in which the assignments become more logical. However, unlike BERTopic used in the present work, LDA requires the user to mention the number of topics.

StyloMetrix: StyloMetrix is an open-source tool for generating text representations in the form of a set of 196 quantitative metrics (for text in English) [7]. Each metric in the vector quantifies a linguistic attribute in the text, such as syntactic (parts-of-speech frequencies), lexical (frequencies of plural nouns, punctuations, superlatives, etc.) and more. Consequently, comprehensive information regarding the text's style can be converted into numerical values and utilised for stylistic tasks.

4. PROPOSED METHODOLOGY

In the preliminary analysis, on comparing with popular contemporary ML models such as Support Vector Machines with linear and radial basis function kernels, and Random Forest classifiers, LR is found to perform better for the given set of features.

Recommended preprocessing steps for optimised BERTopic-based text clustering: As one of the foremost contributions of this study, preprocessing of text input to BERTopic is recommended as shown in Figure 1 (in which 'normalize_punct_space' function makes certain punctuations (comma, semicolon, etc.) left-affined to letters and removes extra space between letters and punctuations, and 'remove()', 'replace()' and 'any()' behave like respective pythonic functions with modifications in syntax.

```
Pseudocode 1  clean_text_for_bertopic()
Require:  text x, list of stop words sw, digits d
for word in x : do
    if word in sw : then
        x :← remove(word)
    end if
    if any(d) in word : then
        word :← replace('#')
    end if
end for
x :← remove(' ')
x :← normalise_punct_space(x) /*Return updated x*/
Return:  x
```

Figure 1. Recommended pre-processing for text input to BERTopic

Removing stop words (or common function words) is found to have a profound impact on the distribution of topics and also on the Jensen-Shannon Divergence (JSD) score; the difference between two authors' topic probability distributions is calculated using JSD. The average of JSD for all author pairs in each of the datasets is tabulated with the help of the topic distributions. High JSD scores indicate more topical dissimilarity between authors in the dataset. For illustrative purposes, entire datasets (including train, validation and test sets put together) are carried forward in two forms: 'uncleaned' (UC) or the raw form of text (i.e. widely utilised form of text input to BERTopic) and the recommended 'cleaned' (CL) text, are compared in Table 3. It highlights the effectiveness of the proposed pre-processing of input to BERTopic. Consequently, in agreement with [5], CCAT_50 and CCAT_10 have greater JSD, followed by Fanfic_22 and then Judgement_3, brilliantly highlighting the topical divergence in the former two and the possibilities of high stylistic differences in the latter two. A diagrammatic comparison of BERTopic distributions (model prior fit on train sets) in CCAT_10 and Judgement_3 is given in Figure 2.

Table 3. Effects of recommended pre-processing on the topical divergence of datasets

Dataset	***JSD***		***# Outliers***		***# Topics***		***Accuracy with LR***	
	UC	***CL***	***UC***	***CL***	***UC***	***CL***	***UC***	***CL***
CCAT_50	*0.0325*	*0.4331*	*2477*	*847*	*12*	*149*	*20.94*	*62*
CCAT_10	*0.0636*	*0.3885*	*289*	*156*	*9*	*33*	*56.3*	*77.7*
FANFIC_22	*0.0033*	*0.2305*	*258*	*61*	*3*	*9*	*13.18*	*43.18*
JUDGEMENT_3	*2.48E-6*	*0.0086*	*518*	*213*	*2*	*11*	*35.6*	*44.2*

In Figure 2, topical dissimilarity between authors is more evident in the former, while nearly a uniform distribution across topics can be observed in the latter.

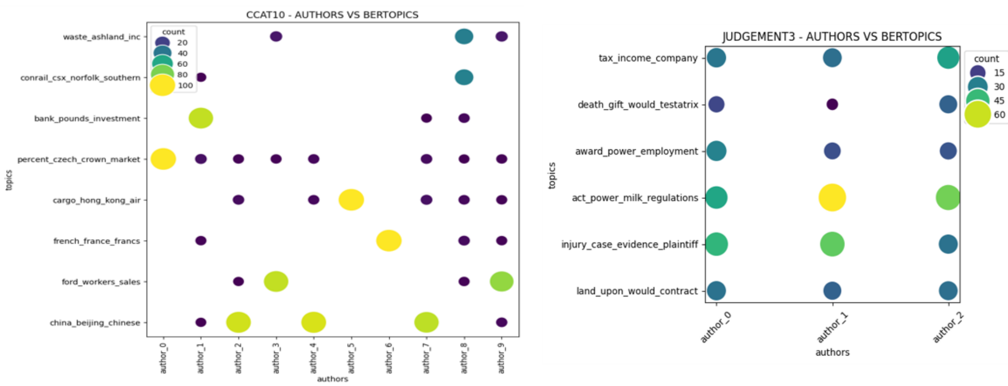

Figure 2. BERTopic distributions in CCAT_10 and Judgement_3 datasets

4.1. Stacking Ensemble

In popular ML libraries like scikit-learn, the stacking classifiers can be built in such a way that all the classifiers are trained on the same features and later proceed with a meta-learning task. As illustrated in the proposed framework, Figure 3 (where P(LR) in the diagram denotes probabilities collected from the respective LR), a custom stacking ensemble model is built. Optimal 'solver' parameter for base classifiers is found to be 'sag' solver. Datasets are split in a 60:20:20 ratio for training, validation and testing.

The necessity of employing a stacking ensemble model is crucial and warranted due to its advantages over adhering to conventional training methods. Eliminating meta-learning and encouraging direct training by treating base model input features as final features may pose the following risks:

- Topic distributions of certain datasets with more topically dissimilar authors might increase the dimensions of BERTopic output, which further need to undergo proper feature reduction treatment.
- On the whole, if not feature-engineered properly, the dimensions of BERT, StyloMetrix embeddings also contribute to the curse of dimensionality, whereas the proposed MVMF has a fixed dimension of 3*k, where k is the number of authors in the dataset, and the stacking model distributes the training overheads to the base models.

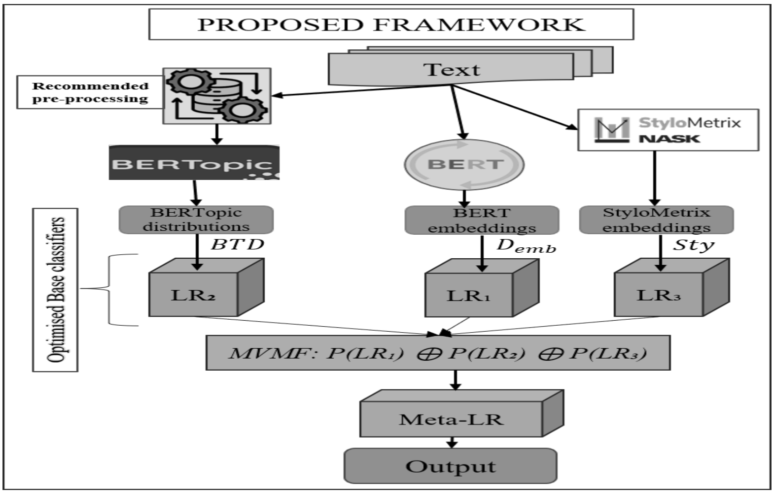

Figure 3. Proposed framework

4.2. Extraction of Multi-Vector Meta-Features:

Multi-Vector in MVMF denotes the multiple aspects of the text (mainly, topical, semantic and stylistic aspects) involved in creating the Meta-Features. Let $S_j \in S$ where S is representative of sentences in Document D with $D = \{S_1, S_2, \cdots S_M\}$. Deriving the document embeddings for each sentence yields $E_s = \{E_{s_1}, E_{s_2}, \dots E_{s_M}\}$. The sentence embedding is $E_{s_j} = h_{CLS}^{(L)}(S_j)$ where $h_{CLS}^{(L)}(S_j)$ is the last layer (namely, [CLS]) embedding when the sentence S_j is the input to the BERT model. To arrive at the entire document representation D_{emb} , mean pooling over the sentence embeddings is done as follows, starting with Equation 1:

$$D_{emb} = \frac{1}{M}\sum_{j=1}^{M} E_{s_j} \tag{1}$$

Substituting the sentence definition in Equation (1) yields,

$D_{emb} = \frac{1}{M}\sum_{j=1}^{M} h_{CLS}^{(L)}(S_j)$. Having extracted the embeddings, an optimised LR model (LR_1) is trained, and the output probabilities as shown in Equation 2 are collected:

For each author class $C_i \in C\ where\ C = \{C_1, C_2, \cdots C_N\}$ in an N-class classification AA problem, with input embedding of the j[th] document is,

$$P(C_i | D_{j\ emb}; LR_1) \tag{2}$$

such that $\sum_{i=1}^{N} P(C_i | D_{j\ emb}; LR_1) = 1$. (For BERTopic, firstly, the model is fit on the train set, then the distributions are derived from that model on the train, validation and test sets.) Similarly, BERTopic Distributions (BTD) are collected and trained with optimised LR_2 as shown in the notations below:

$$P(C_i | BTD_j; LR_2) \text{ such that } \sum_{i=1}^{N} P(C_i | BTD_j; LR_2) = 1$$

And the standardised StyloMetrix embeddings (Sty) are trained with optimised LR_3 model to obtain:

$$P(C_i|Sty_j; LR_3) \text{ such that } \sum_{i=1}^{N} P(C_i|Sty_j; LR_3) = 1$$

Finally, the output probabilities are concatenated to train the meta-model (Meta-LR) as shown in Equation 3 to form MVMF:

$$P(C_i|D_{j\,emb}; LR_1) \oplus P(C_i|BTD_j; LR_2) \oplus P(C_i|Sty_j; LR_3) \quad (3)$$

4.3. Strength of Base Features

In the analysis stage, with BERT base (BB) document embeddings $D_{emb}, BTD\ and\ Sty$, performance evaluation (based on accuracy) on their respective LRs (LR_1, LR_2 and LR_3) is recorded as shown in Table 4, where the top values in each column are given in **bold**.

Table 4. Accuracy scores of base features from respective base models

Base feature/accuracy of LR	*CCAT_50*	*CCAT_10*	*Fanfic_22*	*Judgement_3*
BB	***79.7***	***87***	*68.63*	*70.05*
BTD	*57.62*	*67.4*	*31.13*	*42.25*
Sty	*55.14*	*74.4*	***71.13***	***80.03***

The results in Table 3 are in congruence with the idea of baseline[5] that the topical features (BB, BTD) score more in CCAT_50 and CCAT_10 - datasets with topically dissimilar authors, as compared to Fanfic_22 and Judgement_3, where the style-related feature (Sty) tops. The development of MVMF is to make these features supplement one another and contribute to generalised and reliable embeddings irrespective of topic skewness.

As an attempt to create a better pattern in the feature, the normalised predicted class labels were also concatenated with the MVMF. It produced results comparable to or lower than the scores of MVMF and thus was removed. Although the results of the attempt proved it ineffective, it revealed that predicted probabilities (as proposed in MVMF) are themselves sufficiently reliable and far superior.

5. RESULTS AND DISCUSSIONS

The accuracy scores tabulated in Table 5 demonstrate the effectiveness of MVMF in comparison with many contemporary works, including BERTAA and mod_BERTAA, which are quite similar in terms of adopting a stacking ensemble utilising BERT. In Table 5, the bold middle section highlights the comparison of MVMF versus Baseline and bold values represent top scores in the respective sections. The '*' represents statistical significance - p-value <0.05 achieved in the right-tailed t-test with baseline and MVMF scores. The statistical significance in three out of four datasets justifies that the application of MVMF is promising. It is to be noted that these results are obtained by computing the mean of five cross-validation (CV) runs.

Table 5. Comparison of accuracy scores of MVMF versus contemporary methods

Methods	***CCAT_10***	***CCAT_50***	***Fanfic_22***	***Judgement_3***
Syntactic-HAN [13]	*63.14*	*41.3*		
Structural-HAN [12]	*65.4*	*45.2*		
Continuous n-gram [10]	*74.8*	*72.6*		
Multi-level features [9]	*80.07*	*77.75*		
Distilled and Fused Style [16]	*83.3*	*79.04*	*55.91*	*72.5*
BertAA [14]	*85.5*	*76.3*	*69.31*	*71.17*
Modified BertAA	*86*	*79.4*	*73.86*	*81.08*
Baseline [5]	*85*	*73.1*	*70.45*	*72.9*
Proposed MVMF	***87.75***	***80.94****	***76.7****	***81.78****
Style-HAN [13]	*90.58*	*82.35*		
Structural-Lexical HAN [12]	***92.4***	***83.2***		

The mod_BERTAA has a neck-to-neck competition with MVMF, hinting at the possibility that BERT embeddings processed through LR might be a more powerful meta-feature compared to the one arrived through the BERT sequence classification model with a classification head (neural network) in BERTAA.

6. ABLATION STUDY

It is crucial to understand the contribution of each component of MVMF and its combined strength, so the ablation study is carried out as reported in Table 6. M_BB denotes the performance of the probabilities of BB carried forward as standalone next-level features (meta-features, thus prefixed with 'M'); similarly, M_BTD and M_Sty indicate the probabilities from base learners of BERTopic distribution and StyloMetrix, respectively. The combination of probabilities is represented with a '+' sign.

Table 6. Accuracy and F1 scores of ablated features against MVMF

Dataset	***Features***	***JSD***	***Accuracy***	***F1 score***
CCAT_50	***MVMF***	*0.4331*	***80.94***	***80.73***
	M_BB		*77.94*	*77.63*
	M_BTD		*52.22*	*47.08*
	M_Sty		*54.83*	*54.54*
	M_(BB+BTD)		*78.22*	*77.68*
	M_(BB+Sty)		*80.56*	*80.29*
	M_(BTD+Sty)		*61.5*	*60.98*
CCAT_10	***MVMF***	*0.3885*	***87.75***	***87.65***
	M_BB		*87*	*86.7*
	M_BTD		*66.75*	*61.96*
	M_Sty		*68.25*	*68.19*
	M_(BB+BTD)		*87*	*86.81*
	M_(BB+Sty)		*87.5*	*87.37*
	M_(BTD+Sty)		*80.25*	*80.1*

FANFIC_22	***MVMF***	*0.2305*	***76.7***	***75.26***
	M_BB		*69.69*	*66.68*
	M_BTD		*14.65*	*7.3*
	M_Sty		*72.67*	*69.03*
	M_(BB+BTD)		*72.2*	*69.35*
	M_(BB+Sty)		*74.69*	*72.81*
	M_(BTD+Sty)		*73.68*	*70.44*
JUDGEMENT_3	***MVMF***	*0.0086*	***81.78***	***81.83***
	M_BB		*64*	*63.52*
	M_BTD		*42.67*	*41.64*
	M_Sty		*79.56*	*79.57*
	M_(BB+BTD)		*63.56*	*63.1*
	M_(BB+Sty)		*81.33*	*81.37*
	M_(BTD+Sty)		*80.44*	*80.47*

It can be observed from Table 6 that M_(BB+Sty) is next best to MVMF, denoting the great strength of stylistic and topical/semantic features, regardless of the degree of topical skewness (as inferred by JSD) in datasets. The confusion matrix of Judgement_3 (from one of the CV runs) is presented in Figure 4, comparing M_BB, M_BTD, M_Sty with MVMF, showing how MVMF is enhanced by all three put together.

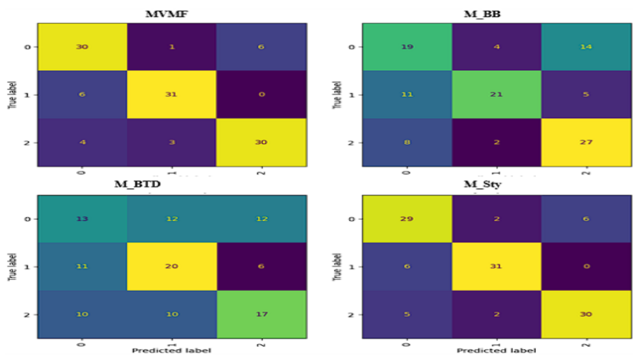

Figure 4. Confusion matrix of Judgement_3 comparing MVMF and its constituents

7. CONCLUSION

Using a pretrained BERT model, Bertopic and StyloMetrix library, we have successfully established a custom stacking method to derive novel Multi-Vector Meta-Features which effectively uses the topical distribution information of the data and produces superior results in datasets, irrespective of their topical or stylistic dominance. Thus, this method discourages the notion that the features are to be selected majorly based on the nature of topical divergence in the authorship attribution datasets, since this work proved that for the benchmark datasets with varied skewness, both the contextual and stylistic features successfully contribute in the proposed model, which outperforms the baseline and SOTA models with statistical significance. In high-stakes domains, as in legal settings, even marginal gains in reliability and confidence can be meaningful. However, there is room for improvement when compared with very recent methods. In the future, other aspects of text could be included in meta-model training. This research established a generalised embedding to employ for any AA dataset in terms of topic and style of authors. Other aspects of data, such as the text length, could also play a significant role; thus, an inclusive generalised method can be strived for.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Habib, N., Adewumi, T., Liwicki, M., Barney, E. Trends and Challenges in Authorship Analysis: A Review of ML, DL, and LLM Approaches, May 2025, URL: https://arxiv.org/pdf/2505.15422 (Visited on: 25.05.2025)

[2] Neocleous. A., Loizides. A. Machine Learning and Feature Selection for Authorship Attribution: The Case of Mill, Taylor Mill and Taylor, in the Nineteenth Century. *IEEE Access*, vol. 9, 2021, pp. 7143–7151, DOI: 10.1109/ACCESS.2020.3047583.

[3] Tsai, C. M., Stylometric Fake News Detection Based on Natural Language Processing Using Named Entity Recognition: In-Domain and Cross-Domain Analysis. *Electronics (Switzerland)*, vol. 12, no. 17, 2023, DOI: 10.3390/electronics12173676.

[4] Weibel, A. Mediumship and stylometry: Exploring a new way of attributing authorship to mediumistic writings. *Journal of Anomalistics / Zeitschrift für Anomalistik*, vol. 24, no. 1, 2024, pp. 55–79, DOI: 10.23793/zfa.2024.055.

[5] 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.

[6] Eder, M. Does size matter? Authorship attribution, small samples, big problem, *Digital Scholarship in the Humanities*, vol. 30, no. 2, 2015, pp. 167–182, DOI: 10.1093/llc/fqt066.

[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: 24.05.2025)

[8] *Github Repository to access supporting files*, Github, 31 March 2025 URL: https://github.com/kaapivalli/Research/tree/mvmf (Visited on 31.05.2025)

[9] 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: 25.05.2025)

[10] 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.

[11] 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.

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

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

[14] Fabien, M., Villatoro-Tello, E., Motlicek, P., Parida, S. BertAA: BERT fine-tuning for Authorship Attribution. *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, Indian Institute of Technology, Patna, India, December 2020, pp. 127–137.

[15] Custódio, J.E., Paraboni, I. Stacked authorship attribution of digital texts. *Expert Systems with Applications*, vol. 176, 2021, p. 1-13, DOI: 10.1016/j.eswa.2021.114866.

[16] Lavanya, B., Sowmiya, R. Distilled and fused style embedding for authorship attribution with weighted multi-kernel ensemble model. *International Journal on Information Technologies and Security*, vol. 16, no. 4, 2024, pp. 25–36, DOI: 10.59035/XDXI5594.

[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] 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.

[19] Seroussi, Y., Smyth, R., Zukerman, I. Ghosts from the High Court's past: Evidence from Computational Linguistics for Dixon ghosting for McTiernan and Rich. *University of New South Wales Law Journal*, vol. 34, no. 3, 2011, pp. 984–1005.

[20] Devlin. J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Oct 2018, URL: http://arxiv.org/abs/1810.04805 (Visited on: 24.05.2025)

[21] Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, Mar 2022, URL: http://arxiv.org/abs/2203.05794 (Visited on: 24.05.2025)

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 08 July 2025