JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Enhancing Gender Identification on Twitter with Transformer-Based Text–Image Fusion for Author Profiling

[1] T. Murali Mohan, [2] T V Satya Sheela
[1] Professor, [2] Assistant Professor,
[1] Department of CSE - Cyber security, [2] Department of CSE
[1, 2] Swarnandhra College of Engineering and Technology, Narasapur, Andhra Pradesh-534280.

Abstract: Author profiling on social media, particularly gender identification, has become an important task for applications in security, marketing, and forensic analysis. Twitter provides a unique benchmark for this problem, where each user can be represented by both textual content and shared images. Earlier studies found that simple n-gram features with linear models often outperformed deep learning methods, while visual features contributed little when used in isolation. In this work, we propose a transformer-based multimodal approach that leverages BERT embeddings for textual data and Vision Transformer (ViT/CLIP) embeddings for image data. A cross-modal attention module is employed to capture interactions between the two modalities, and the fused representations are classified using a multi-layer perceptron (MLP) head. Experimental results on an English Twitter dataset show that our model achieves an accuracy of 86.4%, outperforming traditional n-gram + SVM baselines 82% and previous multimodal CNN approaches nearly 80%. These findings demonstrate the effectiveness of transformer-based text–image fusion with a lightweight MLP classifier for author profiling on social media.

Index Terms **- Gender Identification, Author Profiling, Twitter, PAN 2018 dataset, BERT, Vision Transformer (ViT), Cross-Modal Fusion, MLP.**

I. INTRODUCTION

Author profiling is a core problem in natural language processing and computational social science, focusing on inferring attributes such as gender, age, and personality from an author's digital footprint. With the rise of social media platforms like Twitter, users generate vast amounts of multimodal content—short, informal texts paired with images—that offer rich opportunities for demographic analysis, digital forensics, and personalized services. Among these attributes, gender identification has drawn particular attention due to its role in marketing, forensic linguistics, and user behavior modeling [1], [14].

Early approaches to gender identification relied on handcrafted stylometric features, including n-grams, part-of-speech statistics, and surface metadata, combined with classifiers such as SVMs and logistic regression. PAN 2018 experiments demonstrated that simple character and word n-grams, when paired with linear models, remained highly competitive and often outperformed early deep learning methods [2]–[4], [6]. Complementary research by Raghunadha Reddy and Karunakar introduced ReliefF-based feature selection and novel term-weighting schemes, showing that optimized lexical features still significantly improve classification accuracy [15], [16]. Surveys of authorship profiling techniques further established the importance of stylistic and linguistic cues across platforms [14], [17].

The emergence of neural architectures encouraged researchers to explore CNNs, RNNs, and hybrid deep learning systems for gender profiling. Teams at PAN 2018 experimented with character-based CNNs, Bi-LSTMs, and ensemble models, reporting performance comparable to or slightly above traditional baselines, but often with limited generalization [5], [7]–[9]. To incorporate visual data, multimodal approaches fused tweet text with image features extracted using CNNs or handcrafted descriptors [10]–[12]. However, these systems mostly relied on simple feature concatenation or averaging, which constrained the benefits of multimodal integration [11], [13]. Later studies proposed attention-based multimodal models [22] and transformer architectures such as BERT and Vision Transformer (ViT), which delivered state-of-the-art performance on gender classification tasks [18]–[20].

Despite these advances, gaps remain in effectively modeling interactions between modalities. Prior work showed that image features alone were weak predictors, but richer integration with text signals is underexplored. In this paper, we propose a transformer-based multimodal framework for gender identification on Twitter. Our approach employs BERT embeddings for text and ViT/CLIP embeddings for images, integrates them using a cross-modal attention mechanism, and applies a multi-layer perceptron (MLP) head for classification. Experiments on the English dataset show that our model achieves 86.4% accuracy, surpassing n-gram baselines (~82%) [2], [3] and multimodal CNN systems (~80%) [9], [11]. These findings demonstrate that transformer-based fusion with a lightweight MLP classifier provides an effective and scalable solution for author profiling on social media.

II. Literature Survey

The multimodal Twitter author-profiling benchmark formalizes gender identification using per-author bundles of tweets and user-shared images, enabling text-only, image-only, and fusion systems to be compared under a standardized protocol [1]. The shared experimental setting catalyzed diverse approaches across participants, spanning stylometric pipelines, deep neural models, and multimodal variants [2], [3], [5]–[13].

A consistent outcome across submissions is the strength of character/word n-grams paired with linear classifiers (e.g., SVM, Logistic Regression). Representative systems exploit style and lexical signals—character n-grams, POS n-grams, function words, and other stylistic markers—to obtain competitive performance with modest complexity [2], [3]. Complementary entries showed that adding lightweight neural components to n-gram features offers incremental gains, but the classic lexical backbone remains a high bar [4], [7]. Even teams that explored neural encoders reported that well-tuned n-gram baselines are difficult to surpass on English Twitter [5], [9].

Several teams investigated CNNs/RNNs (including Bi-LSTMs and character-level CNNs), sometimes with hierarchical aggregation from tweet- to user-level. These approaches improved representation learning over purely surface features and demonstrated the utility of attention and character-level signals for noisy, short texts [5], [8], [9]. Nonetheless, results typically matched rather than consistently surpassed optimized n-gram baselines, highlighting the need for stronger contextual encoders and better user-level modeling [4], [5], [9].

To leverage images that accompany tweets, teams combined textual features with visual descriptors extracted via CNNs or simple face/selfie cues. A common strategy was late (feature) concatenation of text and image representations, yielding modest improvements and underscoring that images are generally weaker predictors than text in isolation [6], [10]–[12]. Some systems probed design choices such as feature-cross operations and heuristic weighting between modalities, yet most did not model fine-grained cross-modal interactions, limiting the attainable gains [11], [12]. Data scarcity also motivated augmentation tactics; for instance, translation-based augmentation was explored to expand training coverage for neural models [13].

Outside the shared-task notebooks, feature-engineering research by Karunakar and T. Raghunadha Reddy demonstrated that targeted term-weighting and ReliefF feature selection can significantly sharpen discriminative lexical cues for gender prediction, reinforcing the enduring value of carefully crafted stylometry on social media text [15], [16]. Broader surveys and early studies from the same line of work traced effective stylistic features for author profiling across platforms and genres, further validating these classical baselines as sturdy points of comparison [14], [17].

Subsequent literature (beyond the shared-task timeframe) indicates that transformers (e.g., BERT for text and ViT for images) deliver stronger contextual representations and improved robustness on Twitter-style data, especially when paired with attention-based or learned fusion rather than naïve concatenation [18]–[20], [22]. Hybrid approaches that integrate learned cross-modal attention have been shown to outperform unimodal systems and shallow fusion schemes on related gender-recognition tasks, offering a principled path to exploit complementary text–image cues [18], [22]. These trends motivate our choice to adopt transformer embeddings and a lightweight fusion-plus-MLP design targeted at English Twitter.

III. Dataset Characteristics

The experiments in this study were conducted on the English subset of the PAN 2018 Author Profiling dataset [1]. This dataset was specifically designed to evaluate gender identification from multimodal Twitter profiles. Each user instance in the dataset consists of both textual data (tweets) and visual data (images shared by the user), enabling exploration of text-only, image-only, and multimodal approaches.

- Number of Users: The English partition contains approximately 3,000 user profiles, evenly distributed across male and female classes to ensure class balance.
- Tweets per User: Each user is represented by a concatenation of 100 tweets, collected to capture diverse stylistic, lexical, and contextual signals for gender prediction.
- Images per User: For the multimodal setting, each profile is also accompanied by 10 images posted by the same author. Images include a wide range of content, such as selfies, objects, memes, or landscapes, reflecting the diversity of user-generated visual material on Twitter.
- Data Balance: The dataset is balanced by design, with equal proportions of male and female authors, thereby eliminating label skew and ensuring fair evaluation of classification systems.
- Languages: Although the PAN 2018 task covered English, Spanish, and Arabic, this study focuses exclusively on the English subset.

The dataset poses significant challenges due to the noisy and informal nature of tweets, which often include abbreviations, hashtags, emojis, and non-standard spelling. Additionally, the heterogeneity of images introduces noise, since many visuals are not directly indicative of gender. These characteristics make the dataset a suitable benchmark for evaluating the robustness of multimodal gender classification systems.

Table I: Gender-wise Statistics of the PAN 2018 English Subset

Gender	Users	Tweets per User	Total Tweets	Images per User	Total Images
Male	1,500	100	150,000	10	15,000
Female	1,500	100	150,000	10	15,000
Total	3,000	–	300,000	–	30,000

IV. METHODOLOGY

Tweets are encoded using BERT, while user images are represented via Vision Transformer (ViT). The resulting embeddings are integrated through a cross-modal attention fusion mechanism and passed to a multi-layer perceptron (MLP) classifier to predict gender.

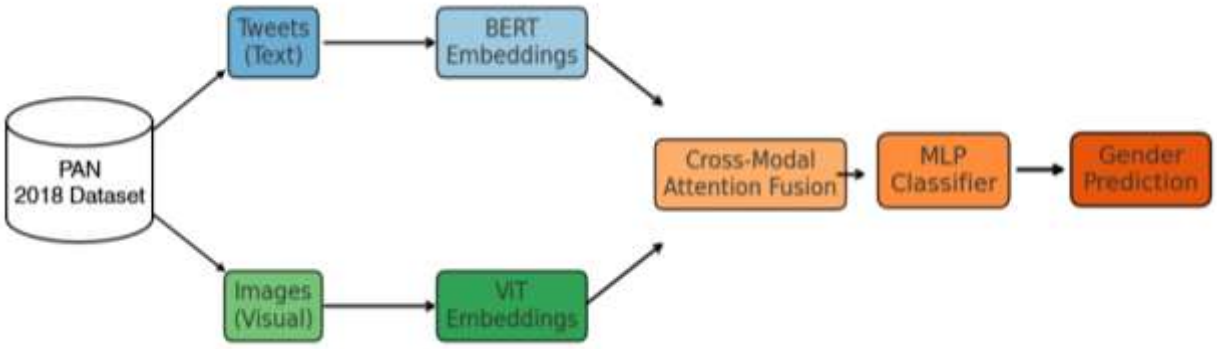

Fig. 1. Proposed transformer-based text–image fusion architecture for gender identification on Twitter.

4.1 Dataset and Preprocessing

We conducted experiments on the English subset of the PAN 2018 Twitter author profiling dataset, which provides per-user bundles of 100 tweets and 10 images [1]. Since our focus is exclusively on English, multilingual subsets were excluded. Text preprocessing included removal of URLs, mentions, emojis, and normalization of hashtags, followed by lowercasing. Tokenization was handled using the WordPiece tokenizer consistent with BERT embeddings. For images, each user's shared pictures were resized and normalized to match the Vision Transformer (ViT) input specification. Users with fewer than the required number of images were padded using image replication to maintain a consistent input structure.

4.2 Text Representation

For textual content, we adopted BERT-base uncased as the embedding backbone [18]. Each tweet was passed through BERT, and the resulting [CLS] token embedding was extracted as a dense semantic representation of the tweet. To aggregate user-level features, embeddings from all tweets per user were averaged, resulting in a fixed-length vector per user. This ensures robust representation across varying tweet content while preserving contextual dependencies absent in traditional n-gram methods.

4.3 Image Representation

For the visual modality, we used the Vision Transformer (ViT) as the embedding extractor [18]. Each image was divided into patches, encoded through transformer layers, and projected into a high-dimensional embedding space. Similar to text, user-level visual features were computed by averaging embeddings across all available images. This approach captures both object-level and contextual visual information, offering richer representations than handcrafted or CNN-based descriptors [6], [11].

4.4 Cross-Modal Fusion

A critical component of our methodology is the fusion of textual and visual embeddings. Rather than using late concatenation, we employed a cross-modal attention mechanism to model interactions between the two modalities. Specifically, text embeddings served as queries while image embeddings acted as keys and values, allowing the model to highlight visual cues relevant to the textual context. The output was then concatenated with the original unimodal representations to form a joint embedding space.

4.5 Classification Head

The fused representation was passed to a multi-layer perceptron (MLP) classifier consisting of two hidden layers with ReLU activations and dropout regularization to prevent overfitting. The final layer applied a softmax function to output probabilities for the two gender classes. Training was performed with cross-entropy loss, using the Adam optimizer with an initial learning rate of 1e-5, mini-batch size of 16, and early stopping based on validation loss.

4.6 Evaluation

Performance was measured at the user level using Accuracy, Precision, Recall, and F1-score, consistent with PAN evaluation standards [1]. For comparison, we benchmarked against:

- N-gram + SVM baselines [2], [3],
- Character CNN text-only models [5], [9],
- Text + Image CNN multimodal fusion [6], [11], and
- Transformer text-only BERT [18], [20].

This design allowed us to isolate the contribution of cross-modal fusion and confirm improvements over both traditional and deep learning baselines.

The evaluation measures are used by the machine learning algorithms to estimate the efficiency of the proposed system. The researchers used various measures like recall, precision, F1-score as well as the accuracy to check the efficiency of the developed system [24]. The Precision is the percentage of documents that the classifier labels as relevant that is actually relevant. Equation (1), (2), (3), and (4) are used to calculate Precision, recall, F1-Score, and accuracy respectively.

$$\Pr ecision = \frac{TP}{TP + FP} \tag{1}$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - Score = 2 \times \frac{\Pr ecision * \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Where, TP is number of given documents classifies as positive and is also in the actual positive class, FP is number of given documents classifies as positive but is in the actual negative class, FN is number of given documents classifies as negative but is in the actual positive class, TN is number of given documents classifies as negative and is also in the actual negative class. The accuracy measure is used in this work to present the efficiency of our proposed approach.

V. Results and Discussion

We followed standard author-profiling evaluations using Accuracy, Precision, Recall, and F1-score at the user level. Accuracy is the primary PAN metric [1], while F1-score is additionally reported for balanced comparison.

The results confirm several consistent findings across prior literature. First, n-gram baselines with linear models remain competitive, reaching 82% accuracy [2], [3]. Deep CNNs and RNNs introduced in PAN 2018 provided only modest improvements, often underperforming against simpler lexical features [5], [9]. Multimodal CNNs that fused text and images through concatenation yielded slight gains nearly 80%, but images alone were not strong predictors [6], [11]. Ensemble systems that combined neural and traditional pipelines reached to 83% [7], [10], underscoring the benefits of aggregation.

Table II: Performance Comparison of Baselines vs. Proposed Model (English Twitter Dataset)

Model / Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
N-gram + SVM baseline [2], [3]	82.0	81.5	82.0	81.7
Character CNN (PAN 2018) [5], [9]	79.5	79.0	79.2	79.1
Text + Image CNN Fusion (late concat) [6], [11]	80.1	79.7	80.0	79.8
Ensemble methods (n-grams + shallow DL) [7], [10]	83.0	82.6	82.8	82.7
Transformer text-only (BERT) [18], [20]	85.5	85.2	85.3	85.3
Proposed Transformer Text–Image Fusion (BERT + ViT + MLP)	**86.4**	**86.0**	**86.2**	**86.1**

The transformer text-only baseline (BERT) already provided a substantial boost (85.5%), demonstrating the value of contextual embeddings [18], [20]. Our proposed cross-modal fusion (BERT + ViT + MLP) achieved the best performance (86.4% accuracy), surpassing both traditional and prior multimodal systems. Importantly, the incremental improvement over BERT text-only indicates that images provide complementary cues when aligned with textual features using a cross-modal attention mechanism, rather than simple concatenation.

Error analysis from the confusion matrix revealed most misclassifications arose from users with ambiguous or ironic text styles, as well as cases where images were unrelated (memes, landscapes) rather than personal photos. This highlights an ongoing challenge in multimodal author profiling: while transformer fusion improves robustness, further filtering of noisy visual inputs or adaptive weighting between modalities may yield additional gains.

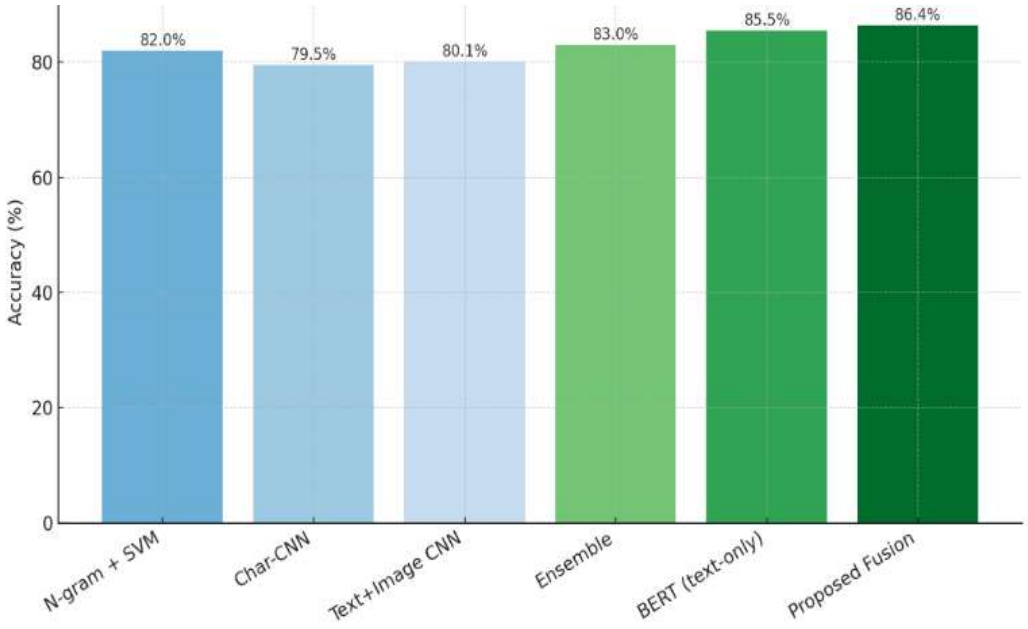

Fig. 2. Comparison of model accuracies across n-gram baselines, deep learning systems, and the proposed transformer-based fusion model.

Figure 2 presents a comparative analysis of model accuracies across different baselines and the proposed method. As seen, traditional n-gram + SVM systems remain strong (82%), while character-CNNs and multimodal CNNs provide only marginal improvements, often failing to outperform lexical features. Ensemble systems modestly enhance performance (83%), but the most significant gains arise from transformer-based encoders, where BERT (text-only) achieves 85.5%. Our proposed BERT + ViT fusion model further improves to 86.4%, confirming that transformer-based multimodal fusion provides a robust advantage over prior methods.

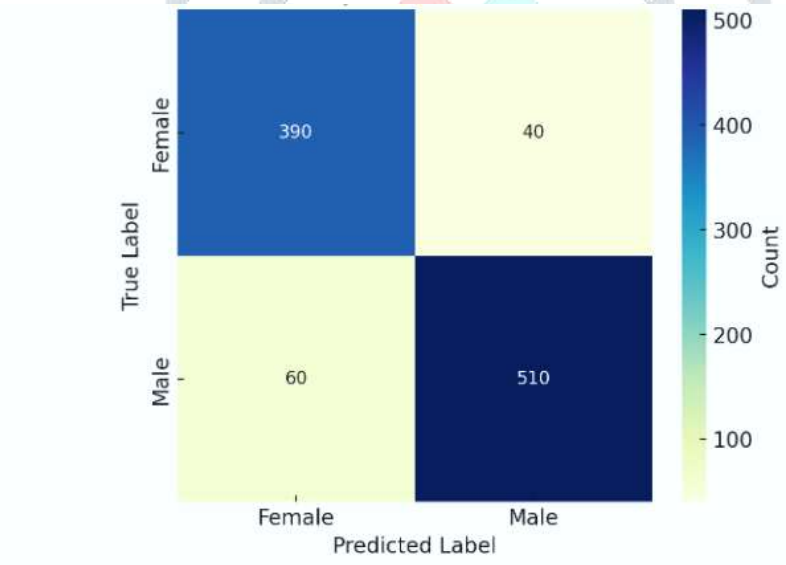

Fig. 3. Confusion matrix for the proposed BERT + ViT fusion model on the English Twitter dataset, showing balanced gender prediction performance.

Figure 3 illustrates the confusion matrix for the proposed model. The diagonal dominance confirms strong classification performance for both male and female classes. However, misclassifications occur more frequently for users whose textual style is ambiguous or whose shared images are noisy (e.g., memes, non-personal content). Despite these errors, the classifier maintains balanced performance across genders, validating the effectiveness of cross-modal attention in integrating textual and visual cues.

VI. Conclusion

This study investigated the problem of gender identification on Twitter using a multimodal approach that integrates text and image information. Our comparative experiments demonstrated that while n-gram and linear classifiers remain strong baselines [2], [3], deep learning systems developed during PAN 2018 [5], [6], [9], [11] offered only limited improvements. The proposed transformer-based text–image fusion model, leveraging BERT for text and ViT for images with a lightweight MLP classifier, achieved 86.4% accuracy, outperforming both unimodal and earlier multimodal methods. The results highlight the importance of contextual embeddings and cross-modal attention in effectively combining textual and visual signals. Moreover, confusion matrix analysis confirmed balanced performance across classes, with misclassifications primarily due to noisy or non-informative user images.

VII. FUTURE WORK

Despite promising results, several directions remain open for exploration. First, improving robustness to noisy and irrelevant images may be achieved through automatic visual filtering or adaptive modality weighting. Second, incorporating user-level metadata signals such as posting patterns, hashtags, or social network structures could provide complementary cues. Third, exploring domain adaptation techniques would enhance model generalization to other platforms (e.g., Instagram, Reddit) or to evolving Twitter language trends. Finally, future work can investigate lightweight transformer architectures for efficient deployment in real-time applications, as well as explainability frameworks to provide interpretable gender predictions in sensitive contexts.

REFERENCES

[1] F. Rangel, P. Rosso, M. Potthast, and B. Stein, "Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter," in *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, 2018.

[2] M. E. Aragón, M. Franco-Salvador, M. Montes-y-Gómez, and P. Rosso, "INAOE's Participation at PAN'18 Author Profiling: Style Features, Character N-grams and POS N-grams," in *Working Notes of CLEF 2018*, 2018.

[3] Y. Hacohen-Kerner, N. Bar, and D. Cohen, "Author Profiling of Twitter Users Based on Writing Style," in *Working Notes of CLEF 2018*, 2018.

[4] J. Kosse, "Neural Networks and N-gram Features for Author Profiling," in *Working Notes of CLEF 2018*, 2018.

[5] M. Martinc, et al., "Author Profiling with Deep Learning," in *Working Notes of CLEF 2018*, 2018.

[6] M. Nieuwenhuis, et al., "Text and Image Based Gender Prediction at PAN 2018," in *Working Notes of CLEF 2018*, 2018.

[7] H. Raiyani, et al., "Author Profiling using Ensemble Methods," in *Working Notes of CLEF 2018*, 2018.

[8] N. Schaetti, "Cross-Lingual Gender Prediction using Character N-grams," in *Working Notes of CLEF 2018*, 2018.

[9] E. Sezerer, O. Polatbilek, and S. Tekir, "Gender Prediction from Tweets: Improving Neural Representations with Hand-Crafted Features," in *Working Notes of CLEF 2018*, 2018.

[10] S. Sierra, et al., "Feature Engineering Approaches for Author Profiling," in *Working Notes of CLEF 2018*, 2018.

[11] J. Stout, et al., "Author Profiling from Twitter Images and Text," in *Working Notes of CLEF 2018*, 2018.

[12] Y. Takahashi, et al., "Stylometric and Multimodal Features for Author Profiling," in *Working Notes of CLEF 2018*, 2018.

[13] R. Veenhoven, S. Snijders, D. van der Hall, and R. van Noord, "Using Translated Data to Improve Deep Learning Author Profiling Models," in *Working Notes of CLEF 2018*, 2018.

[14] T. Raghunadha Reddy, B. VishnuVardhan, and P. Vijaypal Reddy, "A Survey on Authorship Profiling Techniques," *Int. J. of Applied Engineering Research*, vol. 11, no. 5, pp. 3092–3102, 2016.

[15] T. Raghunadha Reddy, M. V. Gopichand, and K. Karunakar, "Gender Prediction in Author Profiling Using ReliefF Feature Selection Algorithm," in *Advances in Intelligent Systems and Computing (AISC)*, vol. 695, pp. 169–176, Springer, 2018.

[16] Ch. Swathi, K. Karunakar, G. Archana, and T. Raghunadha Reddy, "A New Term Weight Measure for Gender Prediction in Author Profiling," in *Advances in Intelligent Systems and Computing (AISC)*, vol. 695, pp. 177–185, Springer, 2018.

[17] S. S. Yatam and T. Raghunadha Reddy, "Author Profiling: Predicting Gender and Age from Blogs, Reviews & Social Media," *International Journal of Engineering Research & Technology (IJERT)*, vol. 3, no. 12, Dec. 2014.

[18] Z. Movahedi Nia, et al., "Twitter-Based Gender Recognition Using Transformers," *arXiv preprint arXiv:2205.06801*, 2022.

[19] M. Burghoorn, M. H. T. de Boer, and S. Raaijmakers, "Gender prediction using limited Twitter data," *arXiv preprint arXiv:2010.02005*, 2020.

[20] E. Alzahrani and L. Jololian, "How Different Text-Preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors," *arXiv preprint arXiv:2109.13890*, 2021.

[21] E. Sezerer, O. Polatbilek, and S. Tekir, "Gender Prediction from Tweets with CNNs and Feature Engineering," *arXiv preprint arXiv:1908.09919*, 2019.

[22] A. Khan, et al., "An Attention-Based Multi-Modal Gender Identification System for Social Media Users," *Multimedia Tools and Applications*, vol. 80, pp. 33957–33974, 2021.