

Multimodal Transformer-based Fusion for Personality Prediction from Social Media: A Gender-aware Framework Integrating mBERT and DINOv2

Siham El Bahy ^{*}, Nouredine Aboutabit , and Imad Hafidi 

Department of Informatique, National School of Applied Sciences Sultan Moulay Slimane University, Khouribga, Morocco

Email: siham.elbahy@gmail.com (S.E.B.); n.aboutabit@usms.ma (N.A.); imad.hafidi@gmail.com (I.H.)

^{*}Corresponding author

Abstract—Accurate prediction of user personality from social media data represents a challenging problem in applied artificial intelligence due to the heterogeneity of multimodal data and multilingual user behavior. This paper proposes a scalable machine learning framework for multimodal data processing and analysis within real-world information technology environments, integrating textual and visual information for personality trait prediction in practical, IT-driven environments. The proposed system combines multilingual Bidirectional Encoder Representations from Transformers (mBERT) for deep semantic text representation and Distillation with No Labels v2 (DINOv2) for visual feature extraction, followed by a lightweight Transformer-based fusion architecture designed to capture cross-modal interactions while ensuring computational efficiency and deployment scalability. The framework is evaluated on an original dataset comprising 32,679 captions and 79,968 images collected from 1290 Instagram users, annotated using the NEO (Neuroticism, Extraversion, Openness to Experience)-FFI (Five Factor Inventory) personality model. To address variability in user behavior, a gender-aware modeling strategy is introduced, enabling more precise representation of heterogeneous user profiles. Experimental results, assessed using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), demonstrate that deep semantic representations significantly outperform surface-level features, while the proposed multimodal fusion model achieves state-of-the-art predictive performance, reaching an RMSE of 0.37 for Agreeableness and 0.52 for Conscientiousness. These results highlight the effectiveness of multimodal machine learning approaches for processing complex real-world data and demonstrate the potential of the proposed system for intelligent decision support, personalized services, and scalable Information Technology (IT)-based applications such as digital mental health and automated recruitment platforms within modern information technology infrastructures.

Keywords—multimodal data fusion, transformer-based models, machine learning, social media analytics, personality trait prediction, natural language processing, computer vision

I. INTRODUCTION

Digital traces on social media platforms provide a rich substrate for personality research, with Instagram standing out due to its dual-layered content structure [1–3]. Driven by advancements in Artificial Intelligence, researchers are increasingly leveraging these multimodal footprints to predict psychological profiles [4–6]. However, a significant portion of prior literature is restricted to single-modality analysis, which may fail to capture the complexity of human behavior.

While imagery typically reflects lifestyle patterns [7–9] and textual data encapsulates linguistic and affective markers [4, 10], true personality inference requires an integrated approach. Current studies that combine these signals often neglect the pivotal role of demographic factors. Our study addresses this limitation by exploring the synergy between captions and visual content in predicting Big Five traits, explicitly accounting for the variations introduced by gender.

The majority of extant literature relies on culturally homogeneous, Western student samples [1, 6, 7, 11–14], which limits the generalizability of predictive models. To overcome this, we utilize an original corpus comprising 1290 Moroccan Instagram profiles, characterized by various demographic characteristics including diverse gender representations and backgrounds. This work provides a more representative and broadened cultural framework for modeling individual personality differences in a multilingual environment.

This work significantly extends our previous studies [15–17]. While our earlier research demonstrated the feasibility of predicting personality traits from Instagram data using visual descriptors extracted through conventional computer vision techniques (e.g., InceptionV3, You Only Look Once (YOLO), cvlib, color and emotion analysis) together with basic caption-level

linguistic features on relatively small datasets, the present study represents a substantial methodological advancement. First, the dataset has been expanded to 1290 Moroccan Instagram users, providing a more representative basis for culturally contextualized personality analysis. Second, manually engineered visual and textual descriptors have been replaced by deep representations learned through Distillation with No Labels v2 (DINOv2) and multilingual Bidirectional Encoder Representations from Transformers (mBERT), enabling richer semantic modeling. Finally, instead of analyzing each modality independently or combining them using conventional machine learning strategies, we introduce a unified triple-input Transformer-based fusion architecture that explicitly models interactions among visual, semantic, and stylistic information through self-attention. These methodological advances substantially extend our previous work by enabling richer feature representations and explicit cross-modal interaction modeling, thereby providing a more powerful and culturally contextualized framework for personality prediction in multilingual social media environments.

Our methodology is organized into four distinct experimental axes to isolate the impact of each feature type. The first experiment focuses exclusively on deep visual features extracted via self-supervised learning (DINOv2). The second experiment analyzes classical stylistic linguistic markers (sentiment, emojis, and hashtags), while the third experiment leverages multilingual BERT (mBERT) to extract deep semantic characteristics, addressing the multilingual nature of our data. Finally, the fourth experiment implements a novel multimodal transformer-based fusion of both text and images. By treating these modalities as a sequence of global embeddings, the transformer model captures complex cross-modal interactions.

Each of these four experiments is conducted independently for each gender to account for gender-specific communication styles. Furthermore, we quantify the relative importance and contribution of these diverse features to the prediction of personality traits obtained through the validated NEO (Neuroticism, Extraversion, Openness to Experience) Five-Factor Inventory instrument. This approach enhances the interpretability of the multimodal decision-making process, showing how visual and textual signals interconnectedly contribute to psychological profiling.

The principal contributions of this work are therefore threefold: (i) the construction of a substantially expanded and culturally representative Moroccan Instagram dataset; (ii) the integration of deep visual, semantic, and stylistic representations; and (iii) the proposal of a Transformer-based multimodal fusion framework capable of modeling cross-modal interactions for personality prediction. The remainder of this paper is structured as follows. Section II presents a comprehensive review of related work. Section III details the proposed methodology, including dataset construction and feature extraction procedures. Section IV provides an extensive experimental evaluation along with a detailed discussion

of the results. Section V outlines the limitations of the study and presents directions for future work. Finally, the paper concludes with a summary of the main findings and contributions.

II. LITERATURE REVIEW

The field of computational personality profiling within online networks has evolved significantly, moving beyond unimodal analysis toward hybrid architectures. These systems are largely grounded in the Five-Factor Model (OCEAN) [18–22]. Contemporary literature has examined diverse facets of this field, ranging from authentic self-representation on Instagram [23] to advanced ML-driven predictive modeling [24] and cross-platform engagement dynamics [25]. Furthermore, these techniques have found critical applications in mental health, such as identifying depressive indicators through multimodal Instagram data [26]. We categorize existing literature into three distinct domains: textual, visual, and hybrid multimodal methodologies.

A. Text-based Methodologies

Initial efforts in this domain utilized psycholinguistic lexicons like Linguistic Inquiry and Word Count [10], though recent trends favor deep learning architectures. For instance, a dual-enhanced network was introduced in Ref. [10] that merges psychological graphs with BERT to refine prediction accuracy. Similarly, semantic structural analysis using LSTM was explored in Ref. [14]. Recent multilingual research [4] demonstrates that models like multilingual BERT (mBERT) and Cross-lingual Language Model-RoBERTa can effectively capture personality traits across different languages without necessitating translation, emphasizing the importance of linguistic variety in global datasets. These Transformer-based models outperform traditional word-level analyses by capturing the contextual and cultural nuances inherent in multilingual social media interactions.

B. Visual-based Methodologies

Visual attributes serve as potent predictors of psychological profiles. Research has established significant correlations between color palettes and personality, such as the link between blue tones and conscientiousness, or the preference for red-orange hues among individuals scoring high in openness and extraversion [1]. Furthermore, gender-specific aesthetic patterns [8] and the semantic categorization of imagery [5, 7] provide critical insights into user behavior. Modern computational models, particularly those leveraging Vision Transformers and self-supervised learning, have been shown to surpass human judgment in detecting traits like extraversion [27, 28]. Recent advancements in vision encoders, such as DINOv2, have further improved the extraction of rich, high-level features from uncurated social media images, moving beyond simple object detection to more complex scene understanding. Within this domain, our exploratory works laid the groundwork for regional personality computing. Specifically, preliminary analysis demonstrated the

feasibility of extracting psychological traits strictly from Instagram images [15], while subsequent research investigated the specific impact of user gender on image content density and prediction accuracy [16]. Expanding upon these visual baselines, our recent study [17] initiated a bimodal shift by exploring how combined images and captions act as complementary windows into personality while factoring in basic demographic influences.

C. Multimodal Integration

The fusion of linguistic and visual streams consistently yields superior performance compared to isolated analysis [13, 29, 30]. Ashawa [11] unified frameworks significantly enhance model reliability. More recent developments have expanded this scope to include tri-modal inputs, audio, text, and visuals [31, 32], utilizing transformer architectures to decode complex cross-modal interdependencies [33]. Our previous study [17] aligned with this multimodal shift by exploring how combined images and captions act as windows into personality while factoring in basic demographic influences. While transformer-based architectures have recently been explored for multimodal fusion [34], our approach uniquely introduces a triple-input framework tailored for text, self-supervised visual features, and explicit stylistic vectors, optimized for culturally contextualized data.

To further push the boundaries of cross-modal learning, recent literature has increasingly favored advanced structural fusion and foundation architectures. For instance, Wang [34] proposed an adaptive graph transformer network integrated with multi-task learning, mapping multi-modal interactions as topological graphs to capture non-linear behavioral correlations. To capture explicit communicative patterns, Fodor [35] enhanced apparent personality trait analysis by explicitly aligning cross-modal embeddings, proving that structural context improves the stability of downstream regressors. Furthermore, Bama [36] expanded the architectural diversity of the field by introducing a hierarchical transformer network with a label attention mechanism designed for Myers-Briggs Type Indicator (MBTI) classification, highlighting the versatility of attention frameworks across different psychological taxonomies.

Concurrently, the domain has experienced a profound paradigm shift driven by the maturation of vision-language architectures. As detailed in a comprehensive review by Singh [37], modern multimodal Vision-Language Models (VLMs) have redefined standard feature extraction pipelines, establishing critical foundation frameworks that allow models to capture rich, open-vocabulary concepts from uncurated social media assets. This shift is further illustrated by Bama [36] in their “Traits Run Deep” framework, which enhances personality assessment by combining psychology-guided Large Language Model (LLM) representations with multimodal apparent behaviors. Their work demonstrates that leveraging the contextual depth of LLMs alongside visual cues provides a more robust and granular psychological profiling mechanism than traditional surface-level linguistic features.

Despite these recent advancements, several challenges remain, including limited demographic representation and the inherent biases of pre-trained architectures. This study bridges these gaps by exploring multimodal fusion within a culturally specific Moroccan dataset, incorporating a gender-stratified analysis to offer a more nuanced understanding of real-world social media dynamics.

III. METHODOLOGY

The following sections present the systematic framework adopted for dataset construction, multimodal feature extraction, and the analytical techniques used to model personality traits and assess the contribution of different data modalities.

To conduct a rigorous appraisal of the predictive capacity inherent in various data modalities, the study is structured as a four-stage experimental sequence: (1) an examination of visual attributes via DINOv2, (2) the processing of stylistic linguistic markers, (3) deep semantic embedding via mBERT, and (4) an integrated multimodal fusion utilizing a Transformer-based architecture.

The remaining part of this section is organized as follows. First, the procedures for the collection and processing of the dataset are described. We then detail feature extraction for visual and textual data, categorized according to the four experimental stages previously mentioned. We also present the machine learning models and regression techniques used for personality estimation, including the implementation of the Multimodal Transformer. Finally, we outline the evaluation metrics and experimental setup used to ensure the robustness and reproducibility of our findings.

A. Database

To construct a robust dataset for our analysis, we gathered data from a diverse sample of Instagram users in Morocco. The final cohort consists of 1290 participants aged 18 to 40, reflecting a balanced gender distribution (631 women and 659 men). This study integrates 473 profiles from our earlier research [15–17] with 817 newly recruited individuals to ensure a comprehensive and statistically significant set of features.

The recruitment strategy targeted a broad demographic by engaging students from Moulay Ismail University and professionals within various research networks. To ensure the representativeness and quality of the sample, we applied specific inclusion criteria: participants were required to have at least one published post and provide explicit consent or maintain a public profile. Furthermore, a limited number of authorized private and inactive accounts were intentionally retained to capture a more diverse range of user behaviors.

The entirety of the data acquisition and processing was managed directly by the research team. For psychological profiling, we employed the Big Five framework [20] through a 60-item NEO (Neuroticism, Extraversion, Openness to Experience)-FFI (Five Factor Inventory) inventory [18], which was adapted into Moroccan Arabic (Darija) to suit the local context. This model defines

human personality through five broad, continuous dimensions: Openness, which reflects intellectual curiosity, creativity, and a preference for novelty; Conscientiousness, measured by organization, self-discipline, and goal-directed behavior; Extraversion, characterized by sociability, high energy, and assertiveness; Agreeableness, which indicates compassion, cooperativeness, and trust toward others; and Neuroticism, representing emotional instability and a predisposition to experiencing negative emotions such as anxiety or stress. This instrument was administered via

Google Forms, and the collected metadata, including usernames, age, and gender, was processed through the NEO-FFI platform. The resulting raw scores were subsequently normalized to a standard [1–5] scale. To strengthen the reliability of our findings, participants were given access to their personality profiles for validation and feedback. The final corpus, comprising 32,679 captions and 79,968 images, was systematically retrieved using Turbo Downloader and structured into individual user directories (see Table I).

TABLE I. SAMPLE CAPTIONS FROM THREE REPRESENTATIVE PARTICIPANTS IN THE MOROCCAN INSTAGRAM DATASET

Users	Caption	English Annotation of Non-English Text
User 1	الحمد لله على كل حال 🙏🙏	Praise be to Allah in all circumstances 🙏🙏
	Bonne fin de semaine à tous... RtL RtL RtL RtL #foryou #stylish	Have a great end of the week, everyone... RtL RtL RtL RtL #foryou #stylish
	صباح الخير من مدينة خريبكة RtL . RtL . RtL . RtL . \#khouribga	Good morning from the city of Khouribga RtL . RtL . RtL . RtL . \#khouribga
	#zen #nature #morocco	-
User 2	Stay focused on your goals, the rest is just noise. 😊😊	-
	Work hard in silence 🙏	-
	Let success be your noise	-
	Coffee first, schemes later ☕️🧐💖💖💖	-
User 3	#motivation #success	-
	عش حياتك كما تحب ❤️❤️❤️	Live your life the way you love ❤️❤️❤️
	Live your life your way! 🗣️ RtL	-
	☀️ #summer #vibes	-

The resulting dataset is organized into three primary modules: (i) a repository of visual assets structured by user, (ii) individual text files containing user-generated captions, and (iii) a consolidated feature matrix in spreadsheet format. This matrix integrates multimodal variables, including deep visual descriptors (DINOv2 embeddings), semantic textual embeddings (mBERT), linguistic attributes (sentiment polarity, emoji semantics, and script identification), and validated Big Five personality scores. To support the reproducibility of this research, the complete processing pipeline and source code have been made publicly accessible via GitHub link: https://github.com/siham-elbahi/Code_personality_analysis_prediction

However, due to privacy constraints and the sensitive nature of social media content, the raw media files are not publicly hosted. Access to the anonymized dataset may be granted for academic purposes upon a formal and justified request, contingent on ethical clearance and adherence to Instagram’s Terms of Service.

B. Feature Extraction

The following segment describes the diverse set of variables derived from our dataset. We have categorized these attributes into two primary domains: text-based metadata from captions and visual properties extracted from images. These features are analyzed incrementally across four experimental stages to evaluate their individual and collective predictive power.

1) Visual features

To capture the complex psychological cues embedded in visual content, we transitioned from heuristic-based

descriptors to a deep semantic representation using DINOv2 [38]. DINOv2 is a state-of-the-art foundation model based on the Vision Transformer (ViT) architecture, pre-trained via self-supervised learning on a massive dataset of 142 million images. This model is particularly effective for personality prediction as it learns robust, high-level visual concepts (textures, compositions, and contextual relationships) without being limited by predefined labels.

In our framework, we utilize the DINOv2-Small (ViT-S/14) backbone to extract a global feature vector for each image. Because each Instagram user may publish multiple images, DINOv2 embeddings are first extracted independently for every image associated with a given user. These image-level embeddings are subsequently aggregated using mean pooling to obtain a single 384-dimensional visual representation per user before dimensionality reduction. While the model natively produces a 384-dimensional embedding, we implemented an independent dimensionality reduction phase to align with our statistical constraints. Specifically, the initial embeddings were projected into a 128-dimensional latent space using Principal Component Analysis (PCA). To ensure strict evaluation isolation and eliminate any risk of data leakage, the PCA algorithm was fitted exclusively on the training folds within each cross-validation iteration, and subsequently applied to transform the remaining validation and test subsets.

This decoupled approach ensures two critical methodological advantages:

Statistical Robustness: By reducing the feature space to 128 dimensions for our 1290-user dataset, we maintain a

favorable sample-to-feature ratio (approximately 10:1), effectively mitigating the risk of overfitting.

Feature Independence: The extraction and compression process remains strictly unsupervised and independent of the personality labels, ensuring that the resulting visual “digital signature” represents the intrinsic characteristics of the user’s imagery.

These standardized 128-dimensional vectors serve as the primary input for the unimodal visual analysis (Experiment 1) and the subsequent multimodal Transformer-based fusion (Experiment 4).

2) *Stylistic text-based attributes*

To transform raw Instagram captions into structured data for Experiment 2, we extracted a multidimensional set of features categorized into the following five analytical domains:

- **Surface and Engagement Metrics:** This category captures the frequency and density of hashtags and emojis across all captions to measure digital interactivity.
- **Morphological and Statistical Analysis:** We computed structural descriptors, including total and average caption lengths, average word length, and the percentage of textless captions.
- **Linguistic and Script Diversity:** Given the Moroccan context, we integrated automatic language identification and script detection to distinguish between Latin, Arabic scripts, and code-switching.
- **Granular Emoji Taxonomy:** Emojis are classified into seven thematic categories (e.g., emotions, objects, symbols) to capture non-verbal semantic nuances.
- **Affective Profiling:** Each caption is processed for sentiment polarity (positive, negative, or neutral) to model the user’s emotional baseline.
- The resulting handcrafted feature vector contains 23 stylistic descriptors that summarize each user’s writing habits, language usage, and emotional expression.

3) *Deep semantic feature extraction*

While surface attributes such as hashtags and emojis provide explicit behavioral markers, they often fail to capture the complex nuances and underlying psychological cues of natural language. To address this, we leverage a pre-trained Multilingual BERT model. This architecture is specifically suited for datasets characterized by high linguistic diversity and code-switching, such as the frequent alternation between Arabic, French, and English observed in the Moroccan context.

The process begins with text pre-processing, where caption files are cleaned of structural noise and whitespace is normalized to ensure a continuous and coherent semantic flow for the Transformer encoder. Each caption is independently encoded using mBERT to obtain a contextualized sentence embedding. Subsequently, all caption embeddings belonging to the same user are aggregated through mean pooling, producing a single 768-dimensional semantic representation before PCA-based dimensionality reduction.

To maintain consistency with the visual modality and mitigate the risk of overfitting given our sample size of

1290 users, we apply Principal Component Analysis to project these initial embeddings into a refined 128-dimensional latent space. As with the visual pipeline, the PCA reduction was strictly fitted within each cross-validation loop exclusively on the training folds to prevent data leakage and ensure rigorous evaluation isolation. This reduction ensures that we preserve the core semantic relationships that reflect stable personality traits while providing the statistical robustness necessary for the subsequent multimodal Transformer-based fusion phases (Experiment 4).

C. *Analytical Methods*

To evaluate the relationships between personality traits and extracted features, predictive modeling was combined with feature importance analysis to capture both predictive performance and the relative contribution of different modalities.

1) *Prediction models*

To capture complex patterns in the data, we implemented a variety of linear and non-linear regression techniques. Among these are Support Vector Regression (SVR) kernels with Radial Basis Function (RBF) [39], Bagging Regressors and Quantile Regression [40]. They were selected for flexibility and robustness against outliers. To ensure the generalizability of our results, all models were evaluated using a 10-fold cross-validation method.

2) *Multimodal fusion strategy*

Beyond regressions we implemented a Transformer-based Fusion model [41]. It uses a self-attention mechanism to integrate visual basic linguistic and semantic features into a unified predictive space capturing cross-modal dependencies that single-modality models may miss. Specifically, to retain the distinction between the inputs within the unified space, each 128-dimensional modality vector is augmented with a dedicated, learnable modality-type embedding before being fed into the Transformer encoder. This approach explicitly allows the self-attention mechanism to identify the underlying source (DINOv2 visual features, mBERT semantic features, or explicit stylistic vectors) while modeling cross-modal interactions. Because the input tokens strictly follow a fixed and predefined sequence (Visual, Textual, and then Stylistic), these modality-type embeddings are sufficient for the network to uniquely identify and map each input representation.

While Transformer-based architectures are typically associated with large-scale datasets, the proposed design remains suitable for our experimental setting. The use of compact 128-dimensional embeddings combined with a shallow Transformer configuration (single layer, limited attention heads) ensures a controlled model capacity.

3) *Evaluation metrics*

The model performance was measured on the scale of Refs. [1, 5] using root mean squared error and mean absolute error.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

where:

n : represents the total number of users in the evaluation test set ($n = 1290$).

y_i : denotes the actual personality trait score of the i -th user, obtained from the NEO-FFI inventory.

$\hat{y}_i$: represents the predicted personality trait score generated by the model for the i -th user.

To assess the statistical reliability and stability of the reported performance metrics, 95% confidence intervals $CI_{95\%}$ are computed for both Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) scores. Assuming an asymptotic normal distribution of the prediction errors across the test sample, the interval is formulated as follows:

$$CI_{95\%} = Metric \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

where:

Metric: represents the calculated performance score (either RMSE or MAE).

σ : denotes the standard deviation of the individual prediction errors across the test instances.

n : represents the total number of users in the evaluation test set ($n = 1290$).

D. Experimental Protocol

To systematically evaluate the predictive contribution of each modality, four complementary experiments were conducted following an incremental evaluation strategy. The first three experiments investigate the predictive capability of each modality independently, allowing the respective contributions of visual features, handcrafted stylistic features, and deep semantic representations to be quantified. The final experiment combines these complementary sources of information through the proposed Transformer-based multimodal fusion architecture to model cross-modal interactions and maximize personality prediction performance. To ensure a fair comparison across all experimental settings, the same prediction models and evaluation protocol were employed throughout the study.

1) Experiment 1: Visual features

The first experiment establishes a unimodal baseline to evaluate the predictive capability of visual information alone. Each Instagram user is represented by the standardized 128-dimensional visual embedding obtained from the DINOv2-Small backbone following user-level mean pooling and PCA-based dimensionality reduction, as described in Section III.B.1. These visual representations are used independently to predict the five Big Five personality traits, allowing us to assess the contribution of

visual content without incorporating any textual information. Personality prediction is performed using the three regression models described in Section III.C.1, namely Bagging Regressor (BR), Support Vector Regression with a Radial Basis Function (RBF) kernel, and Quantile Regression (QR). All models are evaluated using the 10-fold cross-validation protocol described in Section III.C.3, and their predictive performance is assessed using RMSE, MAE, standard deviation, and 95% confidence intervals.

2) Experiment 2: Stylistic features

The second experiment transitions to evaluating the predictive capacity of explicit textual features extracted from Instagram captions. This analysis focuses on linguistic markers such as hashtag usage, emoji statistics, caption length, language and script distribution, and sentiment proportions. We examine how these surface-level linguistic features contribute to personality estimation for female and male users by conducting separate analyses for each gender. This experiment aims to identify direct behavioral cues embedded in written content before moving toward deeper semantic representations in the subsequent stage.

The resulting 23-dimensional stylistic feature vectors are subsequently used as inputs to the same regression models (Bagging Regressor, Support Vector Regression with an RBF kernel, and Quantile Regression) and evaluated using the identical 10-fold cross-validation protocol.

3) Experiment 3: Semantic features

The third experiment shifts the focus to deep linguistic representations by evaluating the predictive capability of contextual embeddings extracted with mBERT. Using the standardized 128-dimensional semantic vectors obtained after user-level embedding aggregation and PCA-based dimensionality reduction, we investigate how high-level semantic information captures the psychological characteristics expressed in the multilingual Moroccan social media context. Unlike handcrafted stylistic features, these contextual embeddings encode the semantic relationships and latent meaning embedded in users' captions. The resulting standardized 128-dimensional semantic vectors are subsequently used as inputs to the same regression models (Bagging Regressor, Support Vector Regression with an RBF kernel, and Quantile Regression) and evaluated using the identical 10-fold cross-validation protocol.

4) Experiment 4: Proposed multimodal transformer fusion

The final experiment introduces a triple-input Transformer-based fusion model designed to maximize predictive accuracy by exploiting the synergies between visual and textual modalities. Moving beyond classical late fusion techniques, this architecture simultaneously processes deep semantic embeddings from mBERT, structural textual features, and deep visual representations from DINOv2 through a Transformer encoder.

The model operates at the user level by integrating three distinct feature vectors. To ensure statistical robustness

and mitigate the risk of overfitting, each modality is standardized before fusion. Both the visual descriptors from DINOv2 and the deep textual embeddings from mBERT are represented by 128-dimensional latent vectors. These are complemented by a third vector containing 23 structural features, which encompass

general caption metrics, language and script detection, emoji categories, and sentiment proportions.

The overall architecture of the proposed Transformer-based multimodal fusion framework is illustrated in Fig. 1.

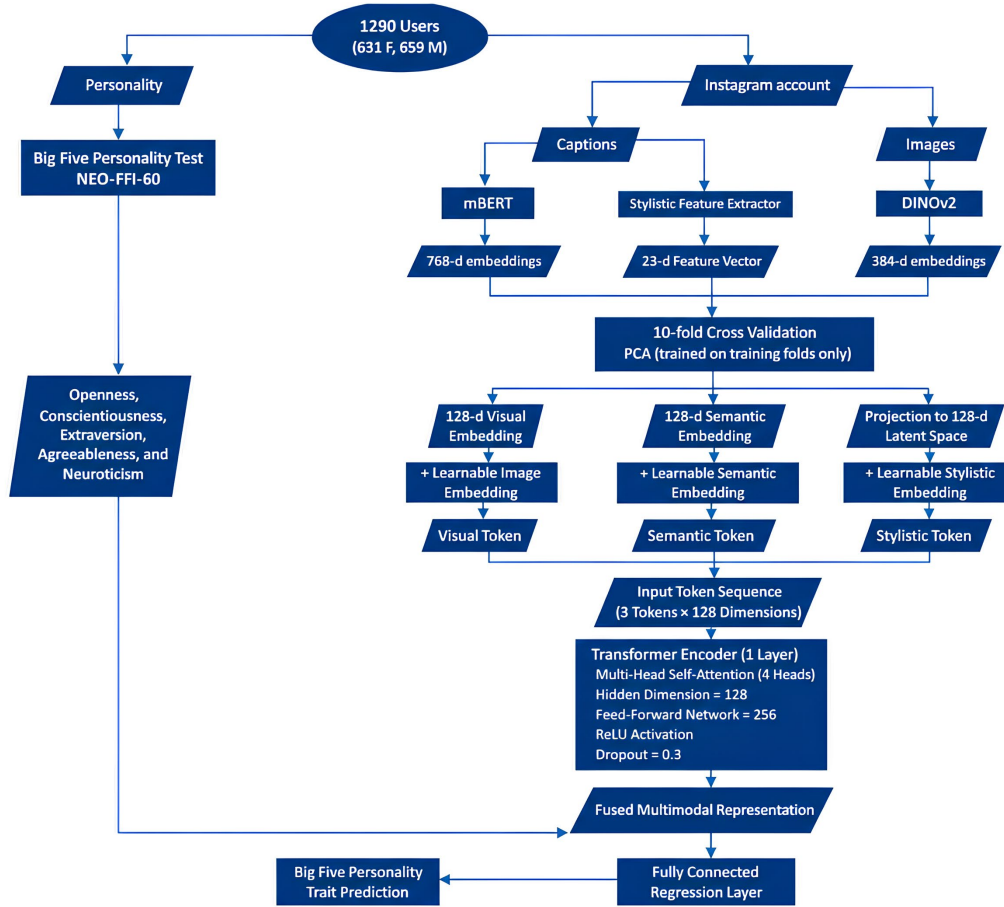


Fig. 1. Proposed Transformer-based multimodal fusion architecture integrating DINOv2 visual embeddings, mBERT semantic embeddings, and stylistic textual features for Big Five personality prediction.

To ensure mathematical compatibility between these diverse inputs, the 23 structural features are projected into the same 128-dimensional shared latent space. Prior to entering the Transformer encoder, each of these three 128-dimensional vectors is combined with its corresponding learnable modality embedding. As illustrated in Fig. 1 each 128-dimensional modality vector is first combined with its corresponding learnable modality embedding to form a modality token. The resulting sequence of three modality tokens is then processed through a lightweight Transformer encoder to model cross-modal interactions. The network relies entirely on these fixed-order modality tokens coupled with their type embeddings to compute sequence dependencies. The encoder configuration consists of a single layer with four attention heads, utilizing a hidden size of 128 and a feed-forward network dimension of 256. To further enhance the model's generalization capabilities, a dropout rate of 0.3 is applied. This architecture allows the model to learn complex cross-modal interactions, effectively weighting the importance of each feature set for

personality trait prediction. The complete hyperparameter configuration and training settings are summarized in Table II.

TABLE II. HYPERPARAMETERS OF THE TRANSFORMER-BASED FUSION MODEL

Component	Hyperparameter	Value
Transformer Encoder	Number of layers	1
	Number of attention heads	4
	Hidden dimension	128
	Feed-forward network size	256
	Activation function	ReLU
	Dropout rate	0.3
Optimization	Optimizer	Adam
	Learning rate	1×10^{-4}
	Batch size	16
	Early stopping	Yes

By exploiting multi-head self-attention, the model effectively weighs the relative importance of textual and visual cues while modeling complex inter-modal dependencies. This mechanism allows the architecture to

remain expressive while adapting to the relatively small size of the dataset. During training, to further control overfitting, early stopping based on validation loss was employed along with a 10-fold cross-validation protocol. Finally, a fully connected layer is utilized to map the aggregated Transformer representations to the final personality trait scores.

E. Ethical Considerations

This study involves the collection and analysis of Instagram user-generated content combined with a psychological evaluation using the Big Five personality model. Given the sensitive nature of personality inferences, particular attention was paid to ethical considerations throughout the research process.

All participants were informed about the purpose of the study and consented to the use of their data for research purposes. The data collection process complied with academic research standards and institutional guidelines. All data were anonymized before analysis to protect user privacy. To ensure that individuals cannot be traced from the dataset, usernames and personally identifiable information were removed and replaced with unique identifiers.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental results and provides a detailed discussion. A comprehensive

evaluation of our four-stage predictive framework, ranging from unimodal feature analysis to multimodal Transformer-based fusion, is conducted. In addition, a feature importance analysis is performed to investigate the relative contribution of different digital markers to personality trait prediction, enhancing the interpretability of the proposed model. Each subsection provides further details on the methodology, performance metrics and linguistic patterns relevant to the Moroccan context.

A. Predicting Personality Based on Extracted Features

This sub-section presents the experimental results for predicting Big Five personality traits from Instagram data across four progressive stages: visual features, basic linguistic markers, deep semantic embeddings (mBERT), and the proposed multimodal transformer-based fusion model. Results are reported separately for female and male users.

1) Experiment 1: Visual feature analysis and image based personality inference

The predictive performance obtained using the visual modality is summarized in Table III, which reports the RMSE and MAE for each of the Big Five traits. To ensure a robust interpretation of these findings, the analysis includes standard deviations and 95% confidence intervals, providing a comprehensive view of the model's stability and its ability to generalize across the user dataset.

TABLE III. THE PERFORMANCE OF PERSONALITY PREDICTION FOR MALE AND FEMALE USERS USING VISUAL FEATURES, EVALUATED USING RMSE AND MAE ACROSS PERSONALITY TRAITS (MEAN $\pm$ STD, 95% CI)

Trait	Metric	F			M		
		BR	RBF	QR	BR	RBF	QR
O	RMSE	0.63 $\pm$ 0.03	0.61 $\pm$ 0.02	0.63 $\pm$ 0.03	0.62 $\pm$ 0.03	0.63 $\pm$ 0.03	0.61 $\pm$ 0.02
	CI	[0.61 – 0.65]	[0.60 – 0.62]	[0.61 – 0.65]	[0.60 – 0.64]	[0.61 – 0.65]	[0.60 – 0.62]
	MAE	0.52 $\pm$ 0.02	0.50 $\pm$ 0.02	0.52 $\pm$ 0.02	0.53 $\pm$ 0.02	0.50 $\pm$ 0.02	0.51 $\pm$ 0.02
	CI	[0.51 – 0.53]	[0.49 – 0.51]	[0.51 – 0.53]	[0.52 – 0.54]	[0.49 – 0.51]	[0.50 – 0.52]
C	RMSE	0.76 $\pm$ 0.04	0.67 $\pm$ 0.03	0.78 $\pm$ 0.04	0.66 $\pm$ 0.03	0.61 $\pm$ 0.02	0.68 $\pm$ 0.03
	CI	[0.73 – 0.79]	[0.65 – 0.69]	[0.75 – 0.81]	[0.64 – 0.68]	[0.60 – 0.62]	[0.66 – 0.70]
	MAE	0.66 $\pm$ 0.03	0.57 $\pm$ 0.02	0.67 $\pm$ 0.03	0.56 $\pm$ 0.02	0.51 $\pm$ 0.02	0.58 $\pm$ 0.02
	CI	[0.64 – 0.68]	[0.56 – 0.58]	[0.65 – 0.69]	[0.55 – 0.57]	[0.50 – 0.52]	[0.57 – 0.59]
E	RMSE	0.74 $\pm$ 0.04	0.73 $\pm$ 0.04	0.66 $\pm$ 0.03	0.79 $\pm$ 0.05	0.95 $\pm$ 0.06	0.79 $\pm$ 0.05
	CI	[0.71 – 0.77]	[0.70 – 0.76]	[0.64 – 0.68]	[0.76 – 0.82]	[0.91 – 0.99]	[0.76 – 0.82]
	MAE	0.65 $\pm$ 0.03	0.64 $\pm$ 0.03	0.56 $\pm$ 0.02	0.68 $\pm$ 0.03	0.82 $\pm$ 0.04	0.68 $\pm$ 0.03
	CI	[0.63 – 0.67]	[0.62 – 0.66]	[0.55 – 0.57]	[0.66 – 0.70]	[0.80 – 0.84]	[0.66 – 0.70]
A	RMSE	0.58 $\pm$ 0.03	0.60 $\pm$ 0.03	0.49 $\pm$ 0.02	0.59 $\pm$ 0.03	0.67 $\pm$ 0.03	0.60 $\pm$ 0.03
	CI	[0.56 – 0.60]	[0.58 – 0.62]	[0.48 – 0.50]	[0.57 – 0.61]	[0.65 – 0.69]	[0.58 – 0.62]
	MAE	0.47 $\pm$ 0.02	0.50 $\pm$ 0.02	0.47 $\pm$ 0.02	0.48 $\pm$ 0.02	0.56 $\pm$ 0.02	0.46 $\pm$ 0.02
	CI	[0.46 – 0.48]	[0.49 – 0.51]	[0.46 – 0.48]	[0.47 – 0.49]	[0.55 – 0.57]	[0.45 – 0.47]
N	RMSE	0.79 $\pm$ 0.05	0.95 $\pm$ 0.06	0.82 $\pm$ 0.04	0.79 $\pm$ 0.04	0.84 $\pm$ 0.05	0.83 $\pm$ 0.04
	CI	[0.76 – 0.82]	[0.91 – 0.99]	[0.79 – 0.85]	[0.76 – 0.82]	[0.81 – 0.87]	[0.80 – 0.86]
	MAE	0.68 $\pm$ 0.03	0.84 $\pm$ 0.04	0.75 $\pm$ 0.03	0.68 $\pm$ 0.03	0.74 $\pm$ 0.03	0.75 $\pm$ 0.03
	CI	[0.66 – 0.70]	[0.82 – 0.86]	[0.73 – 0.77]	[0.66 – 0.70]	[0.72 – 0.76]	[0.73 – 0.77]

In Table III, the results demonstrate that aesthetic and semantic image attributes carry a significant predictive signal. However, the predictive power varies notably between personality traits and genders. Interestingly, the superior performance of the RBF and QR models over the Bagging Regressor in several categories suggests that the relationship between deep visual embeddings and personality is often non-linear, requiring more flexible kernels to capture subtle aesthetic patterns.

Moreover, the variability indicators reported in Table III further support the consistency of the observations. Standard deviations remain in a moderate range generally between ± 0.02 and ± 0.06 depending on the trait and model reflecting a controlled level of variation across the cross-validation folds. In parallel, 95% confidence intervals are relatively narrow typically covering intervals of approximately 0.02 to 0.04 for most configurations. This suggests the estimated performance of the system is

stable and not driven by isolated fluctuations. For certain traits such as Neuroticism (N) and Extraversion (E) slightly wider intervals indicate a higher level of prediction difficulty which is consistent with their higher RMSE values. Overall these statistical indicators confirm that visual features' results exhibit a reasonable degree of robustness while still reflecting the inherent variability of personality prediction from imagery alone.

For Female users, the trait most accurately captured by visual signals is Agreeableness (A) at a minimum RMSE of 0.57. This suggests that visual markers in women's posts are strong indicators of this trait such as specific color palettes or scene types. Neuroticism (N) proved to be the most difficult to predict with a RMSE of 0.79. This difficulty aligns with the "invisible trait" hypothesis in psychology, where emotional instability is often masked by the positive self-presentation bias prevalent in Instagram imagery.

Conscientiousness is the most accurately predicted trait for male users with an RMSE of 0.61 using the RBF

model. This indicates that organizational or lifestyle indicators are particularly telling in men's imagery. Similar to the female cohort Extraversion (E) and Neuroticism (N) have higher error rates with RMSE values reaching up to 0.79 and 0.84 respectively.

While visual data provides a useful baseline the relatively high error margins for traits like Extraversion suggest that imagery alone lacks the granularity required to fully map complex psychological profiles, justifying the transition to linguistic analysis in the next stage.

2) Experiment 2: Personality trait inference based on textual features

Table IV presents the prediction performance obtained using the handcrafted stylistic features extracted from Instagram captions. The results are reported separately for female and male users in terms of RMSE and MAE, allowing the predictive capability of explicit linguistic markers to be compared with the visual representations evaluated in Experiment 1.

TABLE IV. EVALUATION OF LINGUISTIC-DRIVEN PREDICTIVE OUTCOMES FOR GENDER-SPECIFIC GROUPS (MEAN $\pm$ STD, 95% CI)

Trait	Metric	F			M		
		BR	RBF	QR	BR	RBF	QR
O	RMSE	0.59 $\pm$ 0.03	0.56 $\pm$ 0.02	0.62 $\pm$ 0.03	0.61 $\pm$ 0.03	0.60 $\pm$ 0.03	0.64 $\pm$ 0.03
	CI	[0.57 – 0.61]	[0.55 – 0.57]	[0.60 – 0.64]	[0.59 – 0.63]	[0.58 – 0.62]	[0.62 – 0.66]
	MAE	0.49 $\pm$ 0.02	0.46 $\pm$ 0.02	0.48 $\pm$ 0.02	0.52 $\pm$ 0.02	0.53 $\pm$ 0.03	0.53 $\pm$ 0.02
	CI	[0.48 – 0.50]	[0.45 – 0.47]	[0.47 – 0.49]	[0.51 – 0.53]	[0.51 – 0.55]	[0.52 – 0.54]
C	RMSE	0.67 $\pm$ 0.03	0.65 $\pm$ 0.03	0.68 $\pm$ 0.03	0.57 $\pm$ 0.02	0.56 $\pm$ 0.02	0.58 $\pm$ 0.02
	CI	[0.65 – 0.69]	[0.63 – 0.67]	[0.66 – 0.70]	[0.56 – 0.58]	[0.55 – 0.57]	[0.57 – 0.59]
	MAE	0.57 $\pm$ 0.02	0.56 $\pm$ 0.02	0.59 $\pm$ 0.02	0.46 $\pm$ 0.02	0.46 $\pm$ 0.02	0.48 $\pm$ 0.02
	CI	[0.56 – 0.58]	[0.55 – 0.57]	[0.58 – 0.60]	[0.45 – 0.47]	[0.45 – 0.47]	[0.47 – 0.49]
E	RMSE	0.64 $\pm$ 0.03	0.65 $\pm$ 0.03	0.67 $\pm$ 0.03	0.57 $\pm$ 0.02	0.63 $\pm$ 0.03	0.58 $\pm$ 0.02
	CI	[0.62 – 0.66]	[0.63 – 0.67]	[0.65 – 0.69]	[0.56 – 0.58]	[0.61 – 0.65]	[0.57 – 0.59]
	MAE	0.54 $\pm$ 0.02	0.56 $\pm$ 0.02	0.59 $\pm$ 0.03	0.48 $\pm$ 0.02	0.51 $\pm$ 0.02	0.48 $\pm$ 0.02
	CI	[0.53 – 0.55]	[0.55 – 0.57]	[0.57 – 0.61]	[0.47 – 0.49]	[0.50 – 0.52]	[0.47 – 0.49]
A	RMSE	0.46 $\pm$ 0.02	0.49 $\pm$ 0.02	0.48 $\pm$ 0.02	0.51 $\pm$ 0.02	0.49 $\pm$ 0.02	0.56 $\pm$ 0.03
	CI	[0.45 – 0.47]	[0.48 – 0.50]	[0.47 – 0.49]	[0.50 – 0.52]	[0.48 – 0.50]	[0.54 – 0.58]
	MAE	0.35 $\pm$ 0.02	0.33 $\pm$ 0.02	0.37 $\pm$ 0.02	0.42 $\pm$ 0.02	0.40 $\pm$ 0.02	0.46 $\pm$ 0.02
	CI	[0.34 – 0.36]	[0.32 – 0.34]	[0.36 – 0.38]	[0.41 – 0.43]	[0.39 – 0.41]	[0.45 – 0.47]
N	RMSE	0.67 $\pm$ 0.03	0.68 $\pm$ 0.03	0.70 $\pm$ 0.03	0.70 $\pm$ 0.03	0.69 $\pm$ 0.03	0.71 $\pm$ 0.03
	CI	[0.65 – 0.69]	[0.66 – 0.70]	[0.68 – 0.72]	[0.68 – 0.72]	[0.67 – 0.71]	[0.69 – 0.73]
	MAE	0.55 $\pm$ 0.02	0.55 $\pm$ 0.02	0.56 $\pm$ 0.02	0.60 $\pm$ 0.02	0.57 $\pm$ 0.02	0.62 $\pm$ 0.03
	CI	[0.54 – 0.56]	[0.54 – 0.56]	[0.55 – 0.57]	[0.59 – 0.61]	[0.56 – 0.58]	[0.60 – 0.64]

The transition to Surface-Level Linguistic features, such as emoji density and hashtag usage, led to a significant improvement in prediction stability. In Table IV the results reveal distinct linguistic patterns for each gender group. The notable decrease in RMSE across most traits indicates that these surface-level markers act as direct behavioral cues, providing more explicit signals of personality than the latent aesthetic attributes analyzed in the previous stage.

The most accurately predicted trait for female users Agreeableness (A) remains the most with an RMSE of 0.46 (a consistent result that underscores how linguistic politeness and positive emojis serve as stable predictors) but a significant gain is observed for Openness (O), which achieved an RMSE of 0.56 using the RBF model a clear improvement over visual-based. These findings suggest that the choice of specific emojis and hashtags is a strong indicator of intellectual curiosity and openness in women.

In the case of male user textual features proved particularly effective for predicting Conscientiousness (C) and Extraversion (E) both reaching an RMSE of 0.57. The structural style of communication (caption length and social tagging) is a more precise marker of social engagement for men than visual aesthetics.

Overall, these surface level markers reduced the error margin for extraversion compared to the first experiment. This shift suggests that while images capture a curated "atmosphere", the explicit way users structure their captions provides a more direct reflection of their social tendencies and behavioral discipline.

3) Experiment 3: Deep semantic feature extraction using multilingual BERT

The results of this deep semantic evaluation are presented in Table V. By reporting the RMSE and MAE across all five personality traits.

TABLE V. PERFORMANCE PREDICTION BASED ON mBERT SEMANTIC FEATURES (MEAN $\pm$ STD, 95% CI)

Trait	Metric	F			M		
		BR	RBF	QR	BR	RBF	QR
O	RMSE	0.56 $\pm$ 0.03	0.59 $\pm$ 0.03	0.62 $\pm$ 0.03	0.61 $\pm$ 0.03	0.63 $\pm$ 0.03	0.62 $\pm$ 0.03
	CI	[0.54 – 0.58]	[0.57 – 0.61]	[0.60 – 0.64]	[0.59 – 0.63]	[0.61 – 0.65]	[0.60 – 0.64]
	MAE	0.48 $\pm$ 0.02	0.51 $\pm$ 0.02	0.51 $\pm$ 0.02	0.50 $\pm$ 0.02	0.53 $\pm$ 0.02	0.52 $\pm$ 0.02
	CI	[0.47 – 0.49]	[0.50 – 0.52]	[0.50 – 0.52]	[0.49 – 0.51]	[0.52 – 0.54]	[0.51 – 0.53]
C	RMSE	0.63 $\pm$ 0.03	0.66 $\pm$ 0.03	0.69 $\pm$ 0.03	0.62 $\pm$ 0.02	0.62 $\pm$ 0.02	0.63 $\pm$ 0.02
	CI	[0.61 – 0.65]	[0.64 – 0.68]	[0.67 – 0.71]	[0.61 – 0.63]	[0.61 – 0.63]	[0.62 – 0.64]
	MAE	0.53 $\pm$ 0.02	0.56 $\pm$ 0.02	0.60 $\pm$ 0.03	0.50 $\pm$ 0.02	0.49 $\pm$ 0.02	0.50 $\pm$ 0.02
	CI	[0.52 – 0.54]	[0.55 – 0.57]	[0.58 – 0.62]	[0.49 – 0.51]	[0.48 – 0.50]	[0.49 – 0.51]
E	RMSE	0.65 $\pm$ 0.03	0.68 $\pm$ 0.03	0.69 $\pm$ 0.03	0.65 $\pm$ 0.03	0.65 $\pm$ 0.03	0.67 $\pm$ 0.03
	CI	[0.63 – 0.67]	[0.66 – 0.70]	[0.67 – 0.71]	[0.63 – 0.67]	[0.63 – 0.67]	[0.65 – 0.69]
	MAE	0.55 $\pm$ 0.02	0.58 $\pm$ 0.02	0.60 $\pm$ 0.03	0.52 $\pm$ 0.02	0.52 $\pm$ 0.02	0.54 $\pm$ 0.02
	CI	[0.54 – 0.56]	[0.57 – 0.59]	[0.58 – 0.62]	[0.51 – 0.53]	[0.51 – 0.53]	[0.53 – 0.55]
A	RMSE	0.40 $\pm$ 0.02	0.40 $\pm$ 0.02	0.40 $\pm$ 0.02	0.53 $\pm$ 0.03	0.47 $\pm$ 0.02	0.53 $\pm$ 0.03
	CI	[0.39 – 0.41]	[0.39 – 0.41]	[0.39 – 0.41]	[0.51 – 0.55]	[0.46 – 0.48]	[0.51 – 0.55]
	MAE	0.29 $\pm$ 0.01	0.29 $\pm$ 0.01	0.29 $\pm$ 0.01	0.39 $\pm$ 0.02	0.36 $\pm$ 0.02	0.38 $\pm$ 0.02
	CI	[0.28 – 0.30]	[0.28 – 0.30]	[0.28 – 0.30]	[0.38 – 0.40]	[0.35 – 0.37]	[0.37 – 0.39]
N	RMSE	0.75 $\pm$ 0.04	0.71 $\pm$ 0.03	0.71 $\pm$ 0.03	0.78 $\pm$ 0.04	0.69 $\pm$ 0.03	0.74 $\pm$ 0.04
	CI	[0.72 – 0.78]	[0.69 – 0.73]	[0.69 – 0.73]	[0.75 – 0.81]	[0.67 – 0.71]	[0.71 – 0.77]
	MAE	0.62 $\pm$ 0.03	0.58 $\pm$ 0.02	0.58 $\pm$ 0.02	0.66 $\pm$ 0.03	0.58 $\pm$ 0.02	0.63 $\pm$ 0.03
	CI	[0.60 – 0.64]	[0.57 – 0.59]	[0.57 – 0.59]	[0.64 – 0.68]	[0.57 – 0.59]	[0.61 – 0.65]

The introduction of the mBERT model in our study represents a significant improvement. The underlying psychological context of Instagram captions have been successfully captured by moving beyond surface-level statistics to deep multilingual semantic embeddings. Table V shows that semantic depth is a far more powerful predictor than visual or stylistic attributes alone. By encoding contextual dependencies, mBERT effectively decodes latent psychological intent and emotional nuances that simple keyword frequencies often fail to capture.

Agreeableness (A) results are particularly notable for female users with a record low RMSE of 0.41 across all models. The results indicate that mBERT is highly effective in identifying the prosocial and cooperative language patterns typical of this trait. Moreover, Openness (O) also showed a strong performance (RMSE 0.56) suggesting that the semantic nuances of intellectual expression are well-preserved in the mBERT vector space.

For male users we observe a significant improvement in N prediction with the RBF model achieving an RMSE of 0.69. This result is noteworthy, as this trait was among the most difficult to predict in previous experiments. According to this study emotional instability in men is more readily reflected in the complex semantic structure of their captions rather than in simple hashtag usage or image choice.

Nevertheless, it is worth noting that for certain dimensions like Extraversion (E) the gains remain stable compared to Experiment 2. This suggests a “plateau of performance” where social traits are already highly visible through surface-level communicative styles (e.g., emoji density), and increased semantic depth offers marginal added value for these specific dimensions. For Conscientiousness (C) in men the structural discipline of their writing continues to be a primary indicator.

Overall, these observations indicate that while mBERT provides a robust linguistic representation, a multimodal approach remains necessary to capture the full complexity of personality traits, motivating the final fusion stage.

4) Experiment 4: Predicting personality traits from multimodal features

This experiment evaluates the proposed multimodal Transformer fusion framework described in Section III.C.4. The objective is to assess whether jointly modeling visual, semantic, and stylistic information improves personality prediction compared with the unimodal baselines.

Table VI reports the predictive performance of the proposed multimodal Transformer model.

TABLE VI. PREDICTIVE PERFORMANCE (RMSE AND MAE) OF THE TRANSFORMER-BASED MODEL (MEAN $\pm$ STD, 95% CI)

Trait	Metric	Female (F)	Male (M)
O	RMSE	0.52 $\pm$ 0.03	0.55 $\pm$ 0.03
	CI	[0.50 – 0.54]	[0.53 – 0.57]
	MAE	0.41 $\pm$ 0.02	0.45 $\pm$ 0.02
	CI	[0.40 – 0.42]	[0.44 – 0.46]
C	RMSE	0.58 $\pm$ 0.03	0.52 $\pm$ 0.03
	CI	[0.56 – 0.60]	[0.50 – 0.54]
	MAE	0.47 $\pm$ 0.02	0.41 $\pm$ 0.02
	CI	[0.46 – 0.48]	[0.40 – 0.42]
E	RMSE	0.55 $\pm$ 0.03	0.54 $\pm$ 0.03
	CI	[0.53 – 0.57]	[0.52 – 0.56]
	MAE	0.46 $\pm$ 0.02	0.43 $\pm$ 0.02
	CI	[0.45 – 0.47]	[0.42 – 0.44]
A	RMSE	0.37 $\pm$ 0.02	0.44 $\pm$ 0.02
	CI	[0.36 – 0.38]	[0.43 – 0.45]
	MAE	0.28 $\pm$ 0.01	0.35 $\pm$ 0.02
	CI	[0.27 – 0.29]	[0.34 – 0.36]
N	RMSE	0.53 $\pm$ 0.04	0.65 $\pm$ 0.04
	CI	[0.50 – 0.56]	[0.62 – 0.68]
	MAE	0.44 $\pm$ 0.03	0.56 $\pm$ 0.03
	CI	[0.42 – 0.46]	[0.54 – 0.58]

In the final phase of this research, the implementation of a multimodal transformer architecture combining the linguistic representations of mBERT with visual descriptors analyzed in earlier stages provides a unified framework for prediction. By utilizing a multi-head attention mechanism, the model captures cross-modal interactions that are not accessible through unimodal

analysis. The results show a consistent reduction in error rates across personality dimensions, highlighting the relevance of a holistic approach for modeling digital identity. The self-attention mechanism specifically enables the model to resolve ambiguities by dynamically weighting the most reliable modality for each trait, effectively acting as a cross-modal correction system.

In this configuration, the Transformer acts as a fusion mechanism, dynamically weighting the contribution of visual and textual features according to their predictive relevance. For female users, the combination of modalities leads to improved performance, particularly for Agreeableness (A), which achieves an RMSE of 0.37. This suggests that prosocial linguistic patterns are effectively aligned with corresponding visual cues. The performance observed for Openness (O) (RMSE 0.52) further indicates that semantic richness in captions is supported by visual diversity in images. Furthermore, the most significant gain is observed for Neuroticism (N) in females, dropping from 0.79 in the visual experiment to 0.53. This underscores that emotional instability is best captured through the dissonance or alignment between textual sentiment and visual aesthetics. These findings point to a degree of coherence in female digital self-presentation, where textual and visual signals jointly reflect underlying personality traits.

For male users, multimodal fusion helps address several limitations observed in previous experiments, particularly for Conscientiousness (C), where the RMSE reaches 0.52. This suggests that combining structural writing patterns with visual organization improves prediction compared to semantic features alone. Similarly, the prediction of

Extraversion (E) (RMSE 0.54) benefits from integrating textual indicators of social behavior with visual cues related to group presence. These results highlight the contribution of multimodal representations in capturing personality traits expressed across different modalities.

Finally, this experiment proves the robustness of the Transformer model in the complex Moroccan bilingual context. The architecture compensates for the inherent ambiguities of short or code-switched captions in Darija and French by using the visual modality as an “anchor of truth”. RMSE declines across both genders validate our central hypothesis that digital personality is not a fragmented phenomenon but a unified expression that manifests in the interaction between what is said and what is shown. This final model provides an effective framework for automated psychological profiling setting a new benchmark for cross-modal personality recognition in social media research.

5) Comparison with extant literature

To contextualize our methodology within the landscape of social media-based personality assessment, Table VII contrasts our RMSE results with those achieved in prior research. This comparison specifically highlights variations in data processing strategies. To maintain a uniform comparison, the RMSE metrics are sourced from the primary literature, all of which utilize Likert assessment format [1–5].

To maintain a uniform comparison, the RMSE metrics are sourced from the primary literature, all of which utilize the conventional five-point Likert measurement system.

TABLE VII. MINIMUM RMSE FOR EACH TRAIT ACROSS PREVIOUS STUDIES AND OUR RESULTS, REGARDLESS OF GENDER

Reference	[13]	[14]	[8]	[1]	[13]	Our study			
Trait	Text	Text	Text	Text	Image	Image	Text (Style)	Text (mBERT)	Transformer
O	0.62	0.54	0.64	0.62	0.80	0.68	0.64	0.56	0.52
C	0.66	0.84	0.61	0.67	0.70	0.72	0.68	0.62	0.52
E	0.92	0.57	0.65	0.72	0.98	0.60	0.58	0.54	0.54
A	0.69	0.75	0.56	0.55	0.74	0.65	0.61	0.40	0.37
N	0.92	0.63	0.74	0.79	0.90	0.85	0.82	0.69	0.63

Compare the results presented in Table VII indicates an improvement in personality trait prediction using the proposed incremental multimodal approach. Each of the three experimental phases is compared with established studies in the field. Despite providing competitive results for unimodal models (Experiments 1, 2 and 3) the multimodal transformer (Experiment 4) shows consistently improved performance across traits.

In this study the progressive reduction in the RMSE across the four stages suggests that integrating diverse feature sets ranging from visual descriptors to deep semantic embeddings captures provides a more comprehensive representation of personality than any single modality alone.

An important observation concerns agreeableness (A) and conscientiousness (C), where the model achieves RMSE values of 0.37 and 0.52 respectively. They improve previously reported ranges in the literature (0.55 to 0.84 for these traits). This suggests that personality-related

signals may be better captured through the interaction between linguistic patterns and visual self-presentation especially in the Moroccan digital context. While text-based studies (e.g., [8, 14]) provide strong baselines they do not incorporate visual information that can help disambiguate semantic content.

The progression from stylistic characteristics to the transformer-based fusion also highlights the limitations of unimodal approaches. mBERT’s transition from text to style shows that deeper semantic representations improve performance particularly in a bilingual context. The final integration through a Transformer further refines predictions, especially for more complex traits such as Neuroticism (N).

To further support the reliability of the reported improvements, an analysis of the 95% Confidence Intervals (CI) was conducted. As shown in Tables III–VI, the 95% CIs of the proposed Multimodal Transformer (Experiment 4) consistently demonstrate a separation with

those of the best-performing unimodal baselines (mBERT) across the majority of personality traits. While these indicators are intended as a measure of empirical robustness rather than a formal hypothesis test, the lack of significant overlap provides evidence that the performance gains, such as the reduction from 0.40 to 0.37 for Agreeableness, are unlikely to result from random variation. This stability is further evidenced by the narrow standard deviations (± 0.02 to ± 0.03), indicating that the model's superiority is maintained across the different cross-validation folds. These findings suggest that the Transformer-based fusion effectively captures complementary cross-modal signals that are not fully exploited by unimodal architectures.

This study supports the idea that personality prediction benefits from combining visual and textual modes reflecting the complementary nature of these information sources in social media environments.

However, it is important to note that while all compared studies utilize a standard 1–5 Likert scale, a direct cross-dataset RMSE comparison remains indicative rather than definitive. Because these benchmarks rely on different underlying datasets, the baseline variance and population distribution of personality scores within our Moroccan cohort may differ significantly from those of other international cohorts. Consequently, these variations should be considered when evaluating the relative predictive performance across different cultural and demographic contexts.

B. Feature Importance Analysis

Beyond the evaluation of predictive accuracy through RMSE metrics, it is essential to understand the underlying mechanisms that drive the model's decisions. Traditional statistical approaches often rely on linear correlation coefficients, which may fail to capture the complex, non-linear interactions inherent in deep learning architectures and multimodal data fusion. To address this, we conduct a comprehensive Feature Importance analysis

to quantify the relative contribution of each modality to the prediction of the Big Five personality traits.

This analysis aims to decipher the “black-box” nature of our Transformer-based fusion framework by identifying which information channels, textual or visual, are most influential for each specific trait. By normalizing the importance scores into percentages, we provide a clear interpretative map of how the model prioritizes deep semantic embeddings from mBERT, robust visual representations from DINOv2, and surface-level behavioral markers. This transition from traditional statistical approaches to feature contribution analysis not only enhances the interpretability of our findings but also highlights the synergy between linguistic and visual cues in the Moroccan social media context.

To operationalize this analysis, we extracted the attention weights from the Transformer's multi-head mechanism and combined them with permutation importance scores. This approach allows us to observe how the model dynamically allocates “importance” across the three input streams. Fig. 2 summarizes these contributions, normalized to 100% for each trait.

The relative contribution of each modality, visual, structural, and deep semantic, is visualized in Fig. 2, revealing a compelling structural symmetry between the male and female cohorts while highlighting subtle gender-specific nuances in digital self-expression. Across nearly all personality dimensions, the dominance of deep semantic features derived from mBERT is evident, particularly for traits like Agreeableness and Openness. In the female cohort, the semantic importance for Agreeableness reaches its zenith at 52%, suggesting that prosocial tendencies and cooperative behaviors are fundamentally encoded within nuanced linguistic choices. Similarly, the high reliance on semantic depth for Neuroticism across both genders confirms that emotional stability is best decoded through the latent intent and sentiment of the discourse, rather than through surface-level behavioral markers alone.

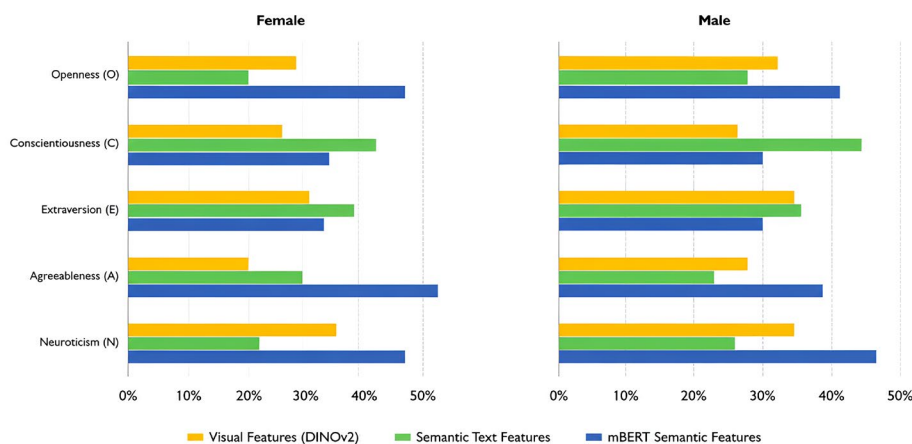


Fig. 2. Relative feature importance (%) per modality for each Big Five personality trait, stratified by gender (Female vs. Male).

In contrast to this semantic dominance, Conscientiousness and Extraversion exhibit a significantly higher reliance on structural text features. For male users,

these markers contribute up to 44% of the predictive weight for Conscientiousness, indicating that organizational discipline is more transparently reflected in

the “mechanics” of communication, such as caption length, tagging frequency, and hashtag usage, than in the semantic content itself. This structural stability across genders validates surface-level linguistic patterns as universal behavioral proxies for social energy and disciplinary traits, effectively capturing the user’s interactional style within the Instagram ecosystem.

The visual modality, represented by DINOv2 embeddings, maintains a crucial and stable role as a complementary anchor, particularly for traits that are traditionally difficult to capture. For Neuroticism, the visual contribution provides an essential aesthetic context that balances potential positivity biases in textual captions, acting as a “ground truth” for the model. In the male cohort, the visual weight for Extraversion is nearly equal to the semantic and structural components, implying that for men, the visual representation of social presence is as informative as their written communication. Ultimately, while the female model appears more “semantic-centric”, the male model displays a more balanced distribution across modalities, demonstrating that the multimodal transformer successfully adapts its weighting mechanism to the specific ways personality is externalized by different genders in the Moroccan context.

Although the proposed framework achieved promising predictive performance, the obtained results should be interpreted considering several limitations and potential sources of bias. The dataset is restricted to Moroccan Instagram users, which could introduce cultural and demographic biases; therefore, findings may not fully apply to other populations or social media contexts. Secondly, the use of pre-trained models such as mBERT and DINOv2 may carry inherent biases originating from the large-scale corpora on which these models were trained. These biases can subtly affect the representation of textual and visual features, and consequently, the prediction of personality traits.

Further, the interpretation of personality traits from digital traces remains inherently probabilistic and context-dependent. Social media behavior does not necessarily reflect stable psychological characteristics, and the model predictions should be interpreted cautiously.

In addition, the potential misuse of automated personality-inference systems raises important ethical concerns. Such models could be used without user consent for intrusive profiling, behavioral manipulation, or automated decision-making processes. This framework is intended for research purposes, with a focus on understanding behavioral patterns rather than labeling individuals, to mitigate these risks.

V. CONCLUSION

Using Instagram data this study presented a comprehensive incremental framework for predicting the Big Five personality traits in the specific cultural and linguistic context of Morocco. The integration of diverse data modalities is essential for accurate digital profiling by evolving from basic stylistic and visual features to advanced deep learning architectures.

The research core contribution lies in the successful implementation of a multimodal Transformer-based fusion mechanism. In contrast to traditional regression models, this architecture effectively captured the complex cross-modal interactions between visual self-presentation and linguistic expression. The adoption of mBERT proved to be a decisive factor in handling the complexity of the multilingual Moroccan environment. By effectively processing the diverse language landscape found in our dataset which includes various languages and dialects used by Moroccan users this model provided a level of semantic understanding that exceeds conventional stylistic.

We show that a multimodal approach consistently outperforms a unimodal model, achieving state-of-the-art accuracy with record minimal RMSE values such as 0.37 for Agreeableness and 0.52 for Conscientiousness. By incorporating a gender-stratified analysis, they show captions and images not just parallel streams of information but complementary windows into the human psyche.

The significant reduction observed in the RMSE observed in Experiment 4 highlights the superiority of feature-level fusion via Transformers. The final integration with deep visual representations from DINOv2 allowed the model to break the performance ceiling of unimodal analysis while mBERT already provided a strong baseline. Hence it appears that personality on Instagram is not just found in what is said or shown but in the specific synergy between the two which our Transformer model successfully decoded.

The findings of our predictive analysis further clarify the relationship between specific linguistic behaviors and personality traits. We identified significant patterns by examining static caption features like hashtags emojis and sentiment markers that explain how users reflect their personalities through digital vocabulary

These results provide a foundational understanding of the dataset suggesting that stylistic textual cues offer reliable indicators for traits such as Extraversion and Agreeableness even before advanced deep learning fusion. Using multilingual semantic depth and self-supervised visual intelligence the methodologies developed in this study offer a robust foundation for future applications in personalized systems digital mental health and cross-cultural social computing. Future work is expected to explore the temporal evolution of these traits to further refine the precision of automated personality inference.

Despite these promising findings, several limitations should be acknowledged, which also provide directions for future research. Despite our robust performance, our multimodal framework has some limitations. First, while the sample size of 1290 Moroccan users is significant for a culturally specific study, a larger and more geographically diverse dataset across different regions would further enhance the generalizability of our findings. In addition, our analysis is limited to static images and text; however, the growing popularity of short-form video content, such as Instagram Reels, presents a richer source of behavioral data for future exploration.

Future research will focus on integrating temporal analysis to examine how personality expression evolves over time. Building on our gender-stratified findings, we aim to further investigate the intersectionality of demographic factors in digital self-presentation. We also plan to expand our architecture to include audio-visual features from video content, providing a more holistic view of the user. Finally, applying this cross-modal framework to other North African and Middle Eastern contexts would allow for a comparative analysis of how cultural nuances influence personality manifestation on social media platforms. Finally, although the proposed progressive evaluation protocol provides insight into the contribution of each modality, future work will include dedicated ablation studies and formal statistical significance tests to further quantify the individual impact of each modality and statistically validate the observed performance gains achieved by the multimodal Transformer-based fusion framework.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Siham El Bahy conceptualized the study, designed the methodology, conducted data collection and preprocessing, implemented the experimental framework, and drafted the original manuscript. Nouredine Aboutabit contributed to the supervision of the research, contributed to data analysis and interpretation, provided critical revisions of the manuscript, and guided the theoretical and methodological aspects of the study. Imad Hafidi supported the validation of the experimental results, and participated in reviewing and editing the manuscript. All authors have read and approved the final version of the manuscript.

REFERENCES

- [1] L. Branz, P. Brockmann, and A. Hinze, "Red is open-minded, blue is conscientious: Predicting user traits from instagram image data," in *Proc. the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, 2020, pp. 23–28.
- [2] S. Osterholz, E. I. Mosel, and B. Egloff, "#Insta personality: Personality expression in Instagram accounts, impression formation, and accuracy of personality judgments at zero acquaintance," *Journal of Personality*, vol. 91, no. 3, pp. 566–582, 2023. doi: 10.1111/jopy.12756
- [3] B. Subramani, "Predicting and analyzing personality traits from Instagram features," *Social Media and Business Analytics*, pp. 1–9, 2018.
- [4] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J. Big Data*, vol. 8, no. 1, 68, 2021. doi: 10.1186/s40537-021-00459-1
- [5] B. Ferwerda and M. Tkalcic, "You are what you post: What the content of instagram pictures tells about users' personality," in *Proc. the 23rd International on Intelligent User Interfaces, March 7–11, Tokyo, Japan*, CEUR-WS, vol. 2068, 2018.
- [6] Y. Kim and J. H. Kim, "Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: An exploratory study," *Information Processing & Management*, vol. 54, no. 6, pp. 1101–1114, 2018.
- [7] B. Ferwerda and M. Tkalcic, "Predicting Users' Personality from Instagram Pictures: Using Visual and/or Content Features?" in *Proc. the 26th Conference on User Modeling, Adaptation and Personalization*, Singapore: ACM, Jul. 2018, pp. 157–161. doi: 10.1145/3209219.3209248
- [8] J. H. Kim and Y. Kim, "Instagram user characteristics and the color of their photos: Colorfulness, color diversity, and color harmony," *Information Processing & Management*, vol. 56, no. 4, pp. 1494–1505, 2019.
- [9] C. Segalin, A. Perina, M. Cristani, and A. Vinciarelli, "The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 268–285, 2016.
- [10] H. Zhua *et al.*, "Enhancing textual personality detection toward social media: Integrating long-term and short-term perspectives," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 4563–4573, 2025.
- [11] M. Ashawa, J. D. Bryan, and N. Owoh, "Personality prediction model: An enhanced machine learning approach," *Electronics*, vol. 14, no. 13, 2558, 2025.
- [12] B. Ferwerda, M. Schedl, and M. Tkalcic, "Using instagram picture features to predict users' personality," in *Proc. International Conference on Multimedia Modeling*, Springer, 2016, pp. 850–861. doi: 10.1007/978-3-319-27671-7_71
- [13] M. Skowron, M. Tkalcic, B. Ferwerda, and M. Schedl, "Fusing Social Media Cues: Personality Prediction from Twitter and Instagram," in *Proc. the 25th International Conference Companion on World Wide Web*, Canada: ACM Press, 2016, pp. 107–108. doi: 10.1145/2872518.2889368
- [14] M. A. Kosan, H. Karacan, and B. A. Urgen, "Predicting personality traits with semantic structures and LSTM-based neural networks," *Alexandria Engineering Journal*, vol. 61, no. 10, pp. 8007–8025, 2022.
- [15] S. El Bahy, N. Aboutabit, and I. Hafidi, "Analyzing Instagram Images to Predict Personality Traits," in *Advances in Machine Intelligence and Computer Science Applications*, vol. 656, N. Aboutabit, M. Lazaar, and I. Hafidi, Eds. vol. 656, Cham: Springer Nature Switzerland, 2023, pp. 349–360. doi: 10.1007/978-3-031-29313-9_31
- [16] S. El Bahy, N. Aboutabit, and I. Hafidi, "Analysis and prediction of personality traits using a self-generated database of Moroccan instagram users: Impact of gender on image content and quantity on prediction accuracy," *Multimed. Tools Appl.*, vol. 83, no. 29, pp. 73583–73606, 2024. doi: 10.1007/s11042-024-19101-2
- [17] S. El Bahy, N. Aboutabit, and I. Hafidi, "Images and Captions as Windows into Personality: Exploring the Impact of Demographic Factors," in *Advances in Machine Intelligence and Computer Science Applications*, N. Aboutabit and I. Hafidi, Eds. Cham: Springer Nature Switzerland, vol. 1882, 2026, pp. 360–371. doi: 10.1007/978-3-032-19963-8_30
- [18] P. T. Costa Jr and R. R. McCrae. (2000). Neo Personality Inventory. *American Psychological Association*. [Online]. Available: <https://psycnet.apa.org/record/2004-12703-172>
- [19] O. John. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. *Published as*. [Online]. Available: <https://www.academia.edu/download/45432778/bigfive.pdf>
- [20] R. R. McCrae and P. T. Costa Jr, "Personality trait structure as a human universal," *American Psychologist*, vol. 52, no. 5, 509, 1997.
- [21] R. R. McCrae and O. P. John, "An Introduction to the Five-Factor Model and Its Applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, 1992. doi: 10.1111/j.1467-6494.1992.tb00970.x
- [22] L. A. Pervin and O. P. John. (1999). Handbook of personality. *Theory and Research*. [Online]. Available: http://jenni.uchicago.edu/econ-psych-traits/John_Srivastava_1995_big5.pdf
- [23] C. Geary, E. March, and R. Grieve, "Insta-identity: Dark personality traits as predictors of authentic self-presentation on Instagram," *Telematics and Informatics*, vol. 63, 101669, 2021.
- [24] M. D. Kamalesh and B. Bharathi, "Personality prediction model for social media using machine learning Technique," *Computers and Electrical Engineering*, vol. 100, 107852, 2022.
- [25] N. Mishnick and D. Wise, "Social Media Engagement: An Analysis of the Impact of Social Media Campaigns on Facebook, Instagram, and LinkedIn," *International Journal of Technology in Education*, vol. 7, no. 3, pp. 535–549, 2024.

- [26] C. Y. Chiu, H. Y. Lane, J. L. Koh, and A. L. P. Chen, "Multimodal depression detection on instagram considering time interval of posts," *J. Intell. Inf. Syst.*, vol. 56, no. 1, pp. 25–47, 2021. doi: 10.1007/s10844-020-00599-5
- [27] A. Naik and A. Kulkarni, "Personality prediction using profile pictures," *Int. J. Eng. Res. Comput. Sci. Eng.*, 2018.
- [28] C. Segalin *et al.*, "What your Facebook Profile Picture Reveals about your Personality," in *Proc. the 25th ACM International Conference on Multimedia*, Mountain View California USA: ACM, Oct. 2017, pp. 460–468. doi: 10.1145/3123266.3123331
- [29] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "Analysis and Prediction of Instagram Users Popularity using Regression Techniques based on Metadata, Media and Hashtags Analysis," *Engineering Letters*, vol. 28, no. 3, 2020.
- [30] Q. Yang, A. Farseev, and A. Filchenkov, "Two-Faced Humans on Twitter and Facebook: Harvesting Social Multimedia for Human Personality Profiling," in *Proc. the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, Taipei Taiwan: ACM, Aug. 2021, pp. 39–47. doi: 10.1145/3463944.3469270
- [31] C. O. Mawalim, S. Okada, Y. I. Nakano, and M. Unoki, "Personality trait estimation in group discussions using multimodal analysis and speaker embedding," *J. Multimodal User Interfaces*, vol. 17, no. 2, pp. 47–63, 2023. doi: 10.1007/s12193-023-00401-0
- [32] H. Salam, V. Manoranjan, J. Jiang, and O. Celiktutan, "Learning personalised models for automatic self-reported personality recognition," in *Proc. Understanding Social Behavior in Dyadic and Small Group Interactions*, PMLR, 2022, pp. 53–73.
- [33] T. Agrawal, D. Agarwal, M. Balazia, N. Sinha, and F. Bremond, "Multimodal Personality Recognition using Cross-Attention Transformer and Behaviour Encoding," in *Proc. the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2022, pp. 501–508. doi: 10.5220/0010841400003124
- [34] R. Wang, X. Zhao, X. Xu, and Y. Hao, "A multimodal personality prediction framework based on adaptive graph transformer network and multi-task learning," *Computer Graphics Forum*, vol. 44, no. 2, e70030, 2025. doi: 10.1111/cgf.70030
- [35] Á. Fodor, R. R. Saboundji, and A. Lőrincz, "Enhancing Apparent Personality Trait Analysis with Cross-Modal Embeddings," arXiv preprint, arXiv:2405.03846, 2024.
- [36] S. Bama *et al.*, "A hierarchical transformer network with label attention for personality prediction by MBTI classification," *Applied Soft Computing*, vol. 178, 113267, 2025. doi: 10.1016/j.asoc.2025.113267
- [37] G. Singh. (2025). A Review of Multimodal Vision-Language Models: Foundations, Applications, and Future Directions. [Online]. Available: <https://www.preprints.org/manuscript/202510.2511>
- [38] M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," arXiv preprint, arXiv:2304.07193, 2024.
- [39] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
- [40] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: Journal of the Econometric Society*, pp. 33–50, 1978.
- [41] A. Vaswani *et al.*, Attention is all you need. (2017). [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](#)).