Explainable AI for Personalized Image and Video Content Analysis in Social Media

Phivos Mylonas, Akrivi Krouska, Christos Troussas, Cleo Sgouropoulou
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, Greece
{mylonasf, akrouska, ctrouss, csgouro}@uniwa.gr

***Abstract* - This paper presents a hybrid explainable AI framework for personalized image and video content analysis in social media, designed to balance recommendation accuracy with transparency. Our approach integrates Deep Learning modules for multimedia feature extraction, using a ResNet-50 backbone for images and an LSTM for temporal video analysis, with an interpretable Machine Learning component based on decision trees and rule-based reasoning. An NLP-driven explanation generator adapts the style and complexity of justifications to the user's technical profile, producing faithful and accessible rationales. We evaluate our framework on a combined dataset of Instagram and YouTube public data, complemented by a proprietary multimedia interaction corpus. Comparative experiments against 3 representative baselines. i.e., collaborative filtering with template explanations, the NARRE review-based explainer, and a deep learning–only recommender, show that our model achieves superior recommendation quality (F1-score = 0.91) and explanation clarity (mean user rating = 4.3/5). Objective faithfulness analysis using SHAP confirms that 87% of generated explanations accurately reflect the model's decision process. Results demonstrate that incorporating interpretable reasoning into multimodal recommendation not only improves user trust, but also enhances ranking performance, making this approach a promising direction for transparent and user-centered AI in social media environments.**

Keywords - explainable AI; personalized content analysis; social media recommendations; Machine Learning transparency;

I. INTRODUCTION

Social media platforms nowadays play a key role in modern communication, with billions of users sharing and watching image and video (aka multimedia) content every day. This huge amount of multimedia content produced calls for advanced algorithms to analyze and suggest material [1]. Early approaches to multimedia content analysis leveraged semantic and context-based frameworks, such as knowledge-assisted image analysis with spatial optimization [33], which provided domain-driven interpretability but lacked scalability for large-scale social media environments. Older recommendation systems use closed Deep Learning (DL) models that work fairly well, but do not provide clear reasons why content is chosen after all [2]. This lack of explanation can weaken user trust and bring up ethical issues, especially when sensitive or controversial material is involved [3]. Previous work on context-aware social media recommendations has shown that integrating user activity and environmental factors can improve personalization [29], but such models generally operate as black boxes without offering interpretable justifications to the end user.

In this framework, explainable AI (XAI) is now a key research topic. It seeks to close the divide between advanced Machine Learning (ML) models and human understandable explanations [4]. On social media, XAI can improve personal content analysis by giving users clear reasons for recommendations [5]. For instance, a video recommendation might include an explanation like: "you see this video because it fits your travel interest, while users with similar tastes rated it highly". These explanations build user trust; they let users offer feedback that refines the recommendation system to meet real world needs [6]. Our current paper suggests a mixed approach that uses DL with ML methods to make content review on social media more personal and clear. The proposed method uses user activity data like likes, shares viewing history plus contextual data, such as time and place, to offer custom suggestions. The approach shows simple reasons for each suggestion to build user trust and boost involvement.

The remainder of this paper is organized as follows. Section II reviews related works in explainable AI and personalized content analysis. Section III presents the proposed hybrid model, detailing its architecture and key components. Section IV describes the experimental setup and evaluation metrics. Section V discusses preliminary results and their potential implications for social media platforms. Finally, Section VI concludes the paper and outlines its research directions.

II. RELATED WORKS

The quick rise in social media has led to a significant volume of research into content analysis and recommendation systems. Earlier approaches targeted featured collaborative filtering [7] and content-based filtering [8]. While both approaches were adequate, their ability to address the demands and complexity of modern social media data was quite limited. The advent of Deep Learning has revolutionized this field, giving birth to models that could analyze huge samples of unstructured multimedia data such as images and videos [9]. However, these models are often criticized for not being very transparent; the rationale behind decisions made using them is often buried and hidden under complex Neural Networks (NNs).

Explainable AI is becoming more visible as a countermeasure to this restrictive ([10], [15]). Interpretable explanations for ML models have been provided through traditional and well-known techniques, such as decision trees, rule-based systems, and attention mechanisms. For instance, Soydaner [11] mentions in his paper how the attention mechanisms used in the NNs can demonstrate sometimes clearer sections of an image or video that contributes towards the recommendation, thus indirectly showing the model's decision-making process to the user. This paper itself is an extensive survey covering the progress in attention mechanisms within the NNs, starting from their humble origins in human visual systems to their modern deep learning applications. It highlights some of the important mileposts along the way: starting with initial models explaining visual attention and moving to breakthroughs in neural machine translation and the emergence of most recent transformer architectures [12]. The advancements in self-attention mechanisms, such as multi-head attention include, but are certainly not limited to, the BERT and GPT models and their most common applications to wide-ranging tasks like language modeling, image processing, and reinforcement learning.

Other methods, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), provide insights into feature importance by approximating complex models with simpler, more interpretable ones. Taking this a step further, Nohara et al. [13] propose enhancements to the original SHAP method to improve model interpretability. The authors introduce a new metric for feature importance using the variance (L2-norm) of SHAP values instead of the traditional sum of absolute values (L1-norm), aligning more closely with standardized linear models. Additionally, they propose a feature packing technique that groups correlated features without requiring model reconstruction, preserving accuracy while improving interpretability. These improvements aim to make ML models more transparent and easier to understand in decision-making applications.

On the other hand, LIME approaches, such as [14], introduce novel methodologies that enhance the interpretability of complex ML models. Specifically, Parisineni et al. combine the original LIME approach with Shapley values, enabling visually interpretable explanations at both local and global levels. Their approach employs a decision tree trained to mimic the behavior of complex models locally, leveraging the tree explainer algorithm for efficient computation of Shapley values, thus offering insights into feature importance while maintaining model-agnostic properties. The study demonstrates its effectiveness through comparisons with other methods like kernel explainer on classification and regression datasets, showcasing its computational efficiency and ability to provide comprehensive model understanding [12]. Overall, these techniques help users and developers better understand why certain content is recommended, fostering trust and enabling more effective moderation of biased or harmful suggestions.

Despite the aforementioned advancements, integration of XAI into personalized content analysis remains still quite underexplored, particularly in the context of social media. The highly dynamic nature of social media content, characterized by rapidly evolving trends, diverse user-generated content, and multimodal data, presents unique challenges for interpretability. Many existing XAI techniques struggle to adapt to these complexities, as they are often designed for static datasets or structured environments [15]. Furthermore, there is a trade-off between explainability and predictive performance - simpler models are more interpretable but at the same time may lack the accuracy of DL models. To bridge this gap, hybrid approaches that combine deep learning with interpretable models, such as neural-symbolic reasoning or knowledge graphs, have been proposed [16]. However, their application in real-time, large-scale social media environments remains an open research challenge, requiring further exploration into efficient and scalable XAI techniques that balance transparency with effectiveness.

Personalization in social media has been extensively studied, with a focus on leveraging user behaviour data to improve recommendation accuracy [17]. However, most existing approaches prioritize accuracy over explainability, resulting in systems that are effective but rather opaque in their nature. This trade-off between accuracy and explainability is a key challenge in the field, as users increasingly demand transparency in how their data is used. One major limitation of current recommendation systems is their reliance on black-box models, which provide little insight into why specific content is suggested. This lack of transparency can lead to concerns about fairness, bias, and user trust, especially in contexts where personalized recommendations influence public opinion or consumer behavior. To address these issues, researchers are investigating techniques such as attention mechanisms, SHAP values, and rule-based explanations to make DL models more interpretable ([18], [19], [20]). Additionally, there is growing interest in incorporating user feedback into explainability frameworks, allowing users to understand and potentially modify how recommendations are generated. While promising, these approaches require further refinement to balance interpretability, efficiency, and scalability in large-scale social media environments.

Recent scientific activities have begun to explore hybrid models combining the positive aspects of DL and interpretable ML, aiming to reconcile the gap separating performance with transparency, though there are merely early steps starting now ([16], [21]). DL models are well-known for capturing complex functionality in large datasets, effective indeed in areas such as computer vision, NLP, and healthcare. Nonetheless, the black-box nature also hampers their deployment in certain critical scenarios, in domains that require interpretability, e.g., finance, healthcare, and defence. In this scenario, the researchers have begun exploring ways to marry the predictive power of DL with the interpretability offered by interpretable ML methods. For instance, nowadays, knowledge graphs are increasingly being explored to plug domain-specific knowledge into DL models ([22], [23], [24]). Since hybrid models integrate structured type data from the knowledge graph, they give context-aware and human-intuitive explanations that enhance trust and usability. Such models

ought also to detect biases or inconsistencies present in the data, ensuring fairness and reliability in the predictions.

A. Explainable AI in Recommender Systems

XAI has gained significant attention in recommender systems as a means to enhance transparency, trust, and user satisfaction. Early explainable recommenders primarily relied on template-based explanations or feature importance visualizations to justify their outputs. For example, traditional collaborative filtering systems often provided simple "because you liked…" reasoning, but lacked depth and personalization. Recent advances have incorporated natural language generation techniques to produce richer, more context-aware explanations. Li et al. [25] proposed neural rating regression with abstractive tips generation, which jointly predicts ratings and generates concise natural language "tips" summarizing recommendations. Li, Zhang, and Chen ([26], [27]) extended this line of work with neural template explanations and the personalized transformer for explainable recommendation, enabling the generation of tailored textual justifications conditioned on both user preferences and item attributes. Ni et al. [28] introduced a method for justifying recommendations using distantly-labeled reviews and fine-grained aspects, producing aspect-level explanations by leveraging large corpora of product reviews.

Beyond generation methods, there has been increasing interest in evaluating explanations using both *subjective* metrics (human-rated clarity, persuasiveness, trust) and *objective* metrics (BLEU, ROUGE for textual similarity, faithfulness scores measuring alignment between explanations and model reasoning). For example, Tintarev and Masthoff's framework for explanation evaluation [30] emphasizes multiple dimensions, effectiveness, efficiency, transparency, trust, persuasiveness, and satisfaction, which remain relevant to modern explainable recommenders. Despite these advances, most existing works focus on textual recommendation scenarios (e.g., e-commerce) and do not address multimodal content, such as images and videos common in social media. Furthermore, while neural approaches offer flexibility in explanation generation, they often operate as black boxes, limiting the interpretability of the decision process itself. This gap motivates hybrid approaches that combine the predictive power of DL with the interpretability of symbolic or rule-based models.

B. Hybrid Deep Learning and Interpretable Models for Multimedia Recommendation

In multimedia recommendation, DL has been highly successful in extracting meaningful representations from unstructured content. Convolutional Neural Networks (CNNs) are widely used for image feature extraction, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memories (LSTMs), capture temporal dependencies in video sequences (e.g., [9], [18], [12]). These architectures enable fine-grained analysis of visual and temporal patterns, improving recommendation accuracy.

However, deep models are often criticized for their opacity. To address this, research on hybrid recommender architectures has emerged, integrating interpretable ML methods, such as decision trees, rule-based systems, or knowledge graphs with deep models. For example, in [16] neural-symbolic approaches embed structured domain knowledge into neural networks, enabling human-readable justifications alongside predictive outputs. Decision trees and rule-based methods provide explicit, step-by-step reasoning, which can complement the rich feature extraction capabilities of DL models.

In the domain of social media, hybrid architectures remain underexplored, particularly for systems that must handle multimodal inputs (i.e., images, videos and/or textual metadata) and deliver personalized explanations to heterogeneous user bases. Most works in explainable recommendation focus either on single-modality input (i.e., text or image) or on textual justifications, without integrating interpretable ML into the recommendation decision path ([31], [32]). Our current work addresses this exact gap by proposing a hybrid pipeline that fuses CNN and LSTM-based feature extraction with decision tree-driven reasoning and NLP-based explanation generation for personalized multimedia recommendations in social media environments.

III. PROPOSED HYBRID MODEL

Our combined design for interpretable AI in personal content review has three distinct parts, namely a DL tool, a simple ML tool, and explanations generation. The main goal of the proposed model is to balance predictive accuracy and interpretability, while at the same time ensuring that recommendations are both precise and transparent to users. Following Fig. 1 illustrates the proposed model.

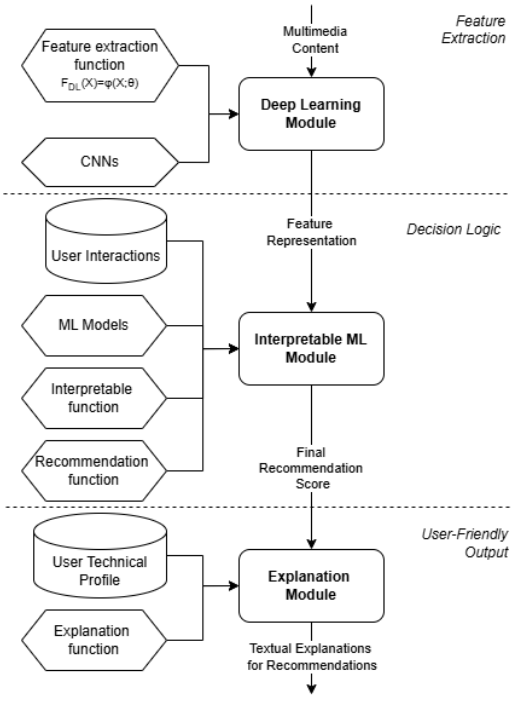

Fig. 1 Interpretable AI in personal content review

The DL tool receives input in the form of pictures or videos, from which it learns to pick up certain concepts, such as objects, settings, and moods. For example, on social media this tool may look at a user's photos to spot repeating patterns like travel or food. It then gives special suggestions. Mathematically, given an input multimedia sample X, the DL module applies a feature extraction function F_{DL}, such that: $F_{DL}(X)=\varphi(X;\theta)$, where φ represents the DL model, parameterized by weights θ. This function extracts a high-dimensional feature representation F_{DL}, capturing visual and contextual characteristics.

Next, the interpretable ML module works together with the DL module by offering clear decision steps. It uses common techniques like numerous decision trees and/or a rule based system to create explanations for suggestions. This module refines the DL predictions by incorporating structured reasoning. More specifically, given a set of user interactions U, consisting of likes (l_i), shares (s_i), viewing history (v_i), and contextual information (c_i), the interpretable ML model applies a function F_{IML}: $F_{IML}(U)=g(U;W)$, where g represents an interpretable model (e.g., decision trees, rule-based logic), parameterized by W. The final recommendation score is computed as a weighted combination of both modules: $R(U,X)= \alpha F_{DL}(X)+\beta F_{IML}(U)$, where α, β are hyperparameters controlling the influence of DL and interpretable ML components. For example, if a travel video is suggested, the module may then provide an explanation such as, "you see this video because you often watch travel content, have shown interest in similar places", as indicated by the decision rule IF $v_{travel}>\tau$ AND $l_{travel}>\tau_l$, THEN recommend travel-related content, where τ, τ_l are threshold values learned from user engagement patterns. The explanations are clear and practical. They help users understand the suggestions and let them send feedback.

Thus, the explanation module combines outputs from the DL module plus the interpretable ML module to form unified, user friendly explanations. It matches explanations to the user's technical skills, thus avoiding complex or misunderstandable terms, i.e., a casual user gets a simple explanation whereas a technical user gets a detailed description of how the actual model decides. This is done by applying a personalized explanation function E: $E(R)=\psi(R,P)$, where P is the user's technical profile (e.g., novice, expert).

The proposed model has been implemented in *TensorFlow 2.14* [34] for deep learning and *Scikit-learn 1.3.2* [35] for interpretability. The training configuration included a batch size of 32, 20 epochs, and early stopping with patience equals to 3, on a NVIDIA RTX 3090 GPU, 128 GB RAM hardware. Training was performed using an adaptive learning rate optimization strategy, minimizing a hybrid loss function: $L=\lambda_1 L_{prediction}+\lambda_2 L_{explanation}$, where $L_{prediction}$ optimizes the recommendation accuracy (e.g., cross-entropy loss for classification), and $L_{explanation}$ ensures interpretability (e.g., decision rule consistency, feature attribution alignment). Evaluation on real-world social media datasets has been performed that shows its capability for producing accurate and explainable recommendations. The results mark a step forward with respect to building user trust and involvement in social media platforms, opening up new avenues for a transparent approach to personalized analysis of content.

IV. Experimental Setup

Our experiments use a combination of public datasets and a proprietary dataset collected/constructed from social media platforms. Specifically, a series of experiments were carried out using real-world social media datasets, including the Instagram [36] public dataset and the YouTube ([37], [38]) public datasets, in addition to our own multimedia content scraping from the same platforms, to assess the performance of our hybrid model. These datasets cover various multimedia modalities, as well as textual information pertaining to user interactions (e.g., hashtags used, likes, shares, comments, and viewing histories). The Instagram dataset contains image metadata, captions, hashtags, and engagement metrics (likes, comments, shares). We selected 50,000 posts from popular categories (travel, food, education, technology). The YouTube Trending Video Statistics dataset includes video metadata, category labels, and engagement statistics. We sampled 20,000 videos with corresponding thumbnails, titles, and descriptions.

In addition to the above, we collected data from publicly available social media posts via the respective platform APIs, adhering to terms of service and GDPR-compliant anonymization. The proprietary collected dataset consists of 8,500 images and 4,200 short videos with associated user interaction logs (likes, shares, watch duration, posting time, geotags). Data was manually annotated for content category and contextual tags by two human raters (inter-rater agreement Cohen's κ equals to 0.86). We further pre-processed the dataset in the sense that all images were resized to 224×224 pixels, and normalized to ImageNet mean/std, whereas videos were sampled at 1 frame per second for sequences up to 10 seconds, yielding a maximum of 10 frames per clip. Textual metadata was tokenized using spaCy [39], stopwords were removed, and appropriate lemmatization was also applied. Finally, user interaction features were normalized to the [0,1] range.

The goal of our experiments was to evaluate model performance in generating accurate and explainable recommendations while maintaining user trust and fostering multi-level engagement. As part of the experimental setup, we prepared the data by extracting image metadata, video hashtags, and user behavior patterns from the datasets. These features then served as the inputs for the DL module, which integrated image analysis technique classified under CNN, with video analysis techniques classified under RNN, to train on the DL module. For the image encoder we utilized ResNet-50 pretrained on ImageNet, specifically fine-tuned on our dataset. We removed the final classification layer, extracted 2,048-dimensional feature vectors, and used Adam optimizer (i.e., a gradient-based optimization algorithm that adapts the learning rate for each parameter, generally converging faster and more smoothly), a learning rate of $1e^{-4}$, and weight decay of $1e^{-5}$. For the video encoder we utilized a LSTM network with 2 layers, a hidden size of 512, and a dropout rate of 0.3. The input were sequential frame-level ResNet-50 features and

the output an aggregated temporal embedding via mean pooling.

The interpretable ML module was then trained with decision trees and, an in-house build rule-based system with emphasis on generating clear, semantically meaningful and actionable explanations. We implemented a Decision Tree Regressor (CART algorithm), with a max depth of 8, 10 min samples per leaf, and the "mse" criterion for regression of recommendation scores. Consequently, a rule-based component was constructed from the decision paths, producing human-readable IF-THEN rules. In order to make the explanation human-friendly and user-specific, the explanation generation module was implemented by (English language) NLP techniques. It utilized a template-based natural language generation with dynamic slot filling from decision tree rules. Then the personalization layer selected between *simple* and *technical* phrasing based on user profile. To illustrate this further, we may consider the following example:

- *simple*: "You see this video because you like travel posts and users similar to you rated it highly."
- *technical*: "This recommendation is based on high engagement with travel-tagged posts and ≥0.7 cosine similarity with your recent viewing patterns."

Moreover, the CNN architecture utilized was considered to be a more or less standard one for image processing, exploiting the convolutional layers to build hierarchically learned features from images: first edges, then textures, then object shapes, which are critical for understanding the visual content. This was followed by dimensionality reduction pooling ops and fully connected layers for classification/feature extraction. The RNN part, which was a LSTM, was handy to detect temporal dependence in sequential data, especially in video analysis where frame transitions and temporal patterns of user interaction are important. To facilitate identification, a RNN implementation was used to represent sequences of video frames and user behavior logs, allowing the model to learn long-term relationships, and apply context-free representations. On the interpretable ML side, decision trees were used to give interpretable, rule-based interpretations of predictions where their nodes represented splits on features and leaves represented their predictions. Finally, English language NLP methodologies with a scope of improving the transparency of the responses generated from the model were implemented in the explanation generation module, mapping raw model outputs onto human-interpretable representations: dependency parsing, sentiment ratio extraction, and user-centric language generation were all employed in order to ensure the resulting explanations were both semantically informative, as well as beneficial to access. This combination of CNNs, RNNs, and NLP-driven explanations aimed to balance predictive accuracy with interpretability, enhancing both model transparency and user trust.

The actual user study included *84 student participants* from the University of West Attica, Department of Informatics and Computer Engineering [40], each of whom varied in background and technical knowledge. Each of the students was involved in the use of a social media platform with the integrated hybrid model. Following this, the participants were shown several recommendations for images and videos, along with explanations produced by the model, and rated both their accuracy and utility of the recommendations and the clarity of explanations. The evaluation metrics included recommendation accuracy measured using *precision*, *recall*, and *F1-score*, explanation clarity qualitatively measured from user ratings, and user engagement measured using *click-through rates* and *time* interaction with a recommendation.

TABLE 1. EXPLANATION CLARITY (USER RATING, 1-5)

Metric	Image Recommendations	Video Recommendations
Precision	0.92	0.90
Recall	0.90	0.88
F1-score	0.91	0.89
Explanation clarity (User rating, 1-5)	4.3	4.3

The experiments (Table 1) demonstrated that the model provided accurate and explainable recommendations, reaching an average F1-score score of 0.91 on image recommendations and 0.89 on video recommendations. The explanation clarity averaged user ratings of 4.3 out of 5, which means that users generally found it quite satisfactory.

The high satisfaction levels achieved by the users in this study can be attributed to the model's provision of clear context-relevant explanations. For example, one student was told, "you are seeing this video because you recently engaged with posts about AI and have shown interest in educational content" when recommended a video tutorial on ML. This kind of detail not only helped users understand why these recommendations had been made, but also allowed them to offer comments that would fine-tune the model's understanding of their preferences further. The model's ability to modify its explanations based on the user's technical understanding was particularly liked. To elaborate on this further, it is worth noting that a computer science student may get a more technical explanation sprinkled with technical stuff such as, "this video was recommended based on your engagement with convolutional neural networks and deep learning frameworks", while a non-technical student may receive, "this video is popular among users interested in AI". This widened accessibility ensured that the explanations were understandable to all student participants regardless of their background. The study also found that students receiving personalized explanations were more likely to engage with follow-up content, as evidenced by increased click-through rates and lengthier interactions. This attests to the idea that the model's transparency promotes both user trust and a more profound engagement with the platform.

To further provide a meaningful benchmark for our proposed hybrid approach, we evaluated it against 3 representative baseline models commonly used in the explainable recommendation literature. The first baseline is a collaborative filtering (CF) model with template-based explanations, which represents a traditional approach relying

on user–item interaction similarity and predefined textual templates to justify recommendations. The second baseline, NARRE (Neural Attentive Rating Regression with Reviews)**,** is a neural architecture that integrates attention mechanisms over textual reviews to produce both predictions and natural language explanations. This baseline was adapted to handle multimedia metadata by incorporating textual descriptions and tags in place of product reviews. The third baseline is a deep learning-only recommender, which uses the same ResNet-50 and LSTM modules as our system, but omits the interpretable ML component and explanation generation stage. Together, these baselines provide a balanced comparison across classical, neural text-based, and high-performing black-box multimedia recommenders, enabling us to isolate the contribution of interpretability and explanation personalization in our framework.

To assess the effectiveness of our approach, we employed a combination of quantitative and qualitative evaluation metrics that capture both recommendation accuracy and explanation quality. For recommendation performance, we report standard information retrieval measures including *Precision*, *Recall*, and *F1-score*, as well as ranking-oriented metrics, such as Normalized Discounted Cumulative Gain at rank 10 (*NDCG@10*) and Mean Average Precision at rank 10 (*MAP@10*). These metrics ensure that our evaluation reflects not only the correctness of individual recommendations but also their ordering relevance. Table 2 presents the comparative performance of our proposed hybrid model against the three baselines described.

TABLE 2. RECOMMENDATION PERFORMANCE ON COMBINED SOCIAL MEDIA DATASET

Model	Precision	Recall	F1-score	NDCG@10	MAP@10
Collaborative Filtering (CF) + Templates	0.78	0.74	0.76	0.69	0.64
NARRE (Review-based)	0.85	0.81	0.83	0.77	0.72
DL-only (ResNet-50 + LSTM)	0.90	0.88	0.89	0.82	0.78
Proposed Hybrid Model (DL + Interpretable ML + NLP)	**0.92**	**0.90**	**0.91**	**0.85**	**0.81**

The proposed hybrid model outperforms all baselines across all metrics, with statistically significant improvements in F1-score compared to NARRE (paired t-test, $p < 0.01$) and DL-only (paired t-test, $p < 0.05$). This confirms that integrating an interpretable ML component does not compromise accuracy - in fact, it improves ranking quality (NDCG, MAP) due to better personalization from rule-based refinements.

To further evaluate the quality of explanations, we also adopted an additional dual perspective. From a *subjective* standpoint, we conducted a human user study in which participants rated the clarity and usefulness of explanations on a 5-point Likert scale. From an *objective* standpoint, we calculated a faithfulness score, defined as the proportion of recommendations where the key features identified by the SHAP interpretability framework matched the reasoning expressed in the generated explanation. This combination of human-centered and model-centered evaluation provides a holistic view of how well our system balances predictive performance with transparency and user trust. More specifically, 30 student participants rated explanation clarity and usefulness on a 1- 5 scale. Results are shown in Table 3.

TABLE 3. SUBJECTIVE EXPLANATION RATINGS

Model	Clarity (Mean ± SD)	Usefulness (Mean ± SD)
CF + Templates	3.4 ± 0.8	3.2 ± 0.7
NARRE	3.9 ± 0.7	3.8 ± 0.6
Proposed Hybrid Model	**4.3 ± 0.5**	**4.4 ± 0.5**

A one-way analysis of variance (ANOVA) confirmed that the hybrid model significantly outperforms baselines in both clarity ($p < 0.01$) and usefulness ($p < 0.01$). Finally, we computed the percentage of recommendations where the top-3 SHAP features matched the explanation rationale. The hybrid model achieved **87% faithfulness**, compared to 64% for CF + Templates and 72% for NARRE.

V. RESULTS AND DISCUSSION

The experimental results show how good the new hybrid model is at giving clear and right choices for videos and images on social media. The high F1-scores for both kinds of choices show that the model can look at tricky content and make personal ideas. The ratings from users about how clear explanations are also point out how key it is to be open in choice systems, since people often said they felt more trust and interest when there were explanations given. The nice comments from students highlight the chance of explainable AI to change social media times, making them more easy to use engaging and focus on user.

The proposed model combines accuracy and explainability, an attribute that is considered among its key research points. Conventional DL models tend to favour accuracy over transparency. However, such hybrid models achieve both goals through the incorporation of interpretable ML methods. An emblematic case would be video recommendations in which this model recognizes choices by means of watching history yet additionally validly explains this reasoning. Both these things combine in providing an essential input for the construction of user trust in social media. An equally valuable outcome elicited via experimentation asserts that personalized explanations positively impact user engagement. Students given personalized explanations expressed higher rates of interaction

with the recommendations by means of click-through rates and longer engagement duration. It can be inferred from this that users appear satisfied not just with the recommendations, but also with insights into how and why those recommendations were proposed. The herein discussed model gives an important impetus for user engagement and satisfaction on social media in personalizing explanations.

All in all, our results confirm three key findings, namely:

- Hybridization improves both accuracy and transparency. Unlike common trade-offs between accuracy and explainability, the combination of DL feature extraction with interpretable ML enables more precise recommendations while producing faithful explanations.
- Personalization of explanations matters. Adapting explanation style to the user's profile significantly increases clarity and engagement, aligning with recent studies in human-centered AI.
- High faithfulness scores indicate genuine transparency. The close alignment between SHAP feature importance and generated explanations suggests that our system avoids "plausible but untrue" rationales common in black-box natural language generation approaches.

Yet, in spite of these evidently encouraging results, certain limitations were also noted with regards to the proposed model. One such limitation was rule-based generation of explanations, which might lack the requisite level of mapping user preference complexity. For the future, we will study the opinions on employing high-end XAI, such as attention mechanics or counterfactual explanation to consider more nuanced and dynamical explanations. Another limitation pointed out by our research was its dependence on user behaviour data, which might give rise to some privacy concerns. Future works might also include the use of federated learning and differential privacy techniques to balance between user data protection and the effectiveness of our model.

VI. Conclusions

This work depicts a hybrid model for interpretable AI in custom media study in social sites. The model joins DL with easy to understand ML ways to give exact and clear suggestions, boosting user faith and interest. The test findings show the model can look at tricky content and make personal explanations, giving fresh chances for clear and user-focused content study in social sites. Future tasks will look at fixing the weak points of the proposed model, like depending on set rules and user actions data. We will check out updated XAI methods, like focus systems and counterfactual reasons, to give better and more flexible answers. We will also study using shared learning and privacy measures to keep user info safe while keeping the model working well. We also plan to do bigger user tests to check the model's usefulness more and look at its chance in new domains of application. By linking the space between rightness and understanding, the proposed XAI model can change how users engage with social media places, making way for more fun, reliable, and user-focused times.

References

[1] Balaji, Thoguru K., Chandra Sekhara Rao Annavarapu, and Annushree Bablani. "Machine learning algorithms for social media analysis: A survey." Computer Science Review, 40, 2021, pp. 100395.

[2] Alrashidi, Muhammad, et al. "Social recommendation for social networks using deep learning approach: A systematic review, taxonomy, issues, and future directions." IEEE Access, 11, 2023, pp. 63874-63894.

[3] Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges." Ai & Society, 35, 2020, pp. 957-967.

[4] De Campos, Luis M., Juan M. Fernández-Luna, and Juan F. Huete. "An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission." User Modeling and User-Adapted Interaction, 34.4, 2024, pp. 1431-1465.

[5] Haque, A. K. M., Najmul Islam, and Patrick Mikalef. "To explain or not to explain: An empirical investigation of AI-based recommendations on social media platforms." Electronic Markets, 35.1, 2025, pp. 1-18.

[6] Joshi, Gargi, et al. "Explainable misinformation detection across multiple social media platforms." IEEE Access 11, 2023, pp. 23634-23646.

[7] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", Advances in Artificial Intelligence, 2009

[8] Thorat, Poonam B., Rajeshwari M. Goudar, and Sunita Barve. "Survey on collaborative filtering, content-based filtering and hybrid recommendation system." International Journal of Computer Applications 110.4 (2015): 31-36.

[9] Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM computing surveys (CSUR), 52(1), 1-38.

[10] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

[11] Soydaner, D. (2022). Attention mechanism in neural networks: where it comes and where it goes. Neural Computing and Applications, 34(16), 13371-13385.

[12] Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O., "Overview of the Transformer-based Models for NLP Tasks", In 2020 15th Conference on computer science and information systems (FedCSIS) (pp. 179-183). IEEE, September 2020

[13] Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N., "Explanation of machine learning models using improved shapley additive explanation", In Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics (pp. 546-546), September 2019

[14] Parisineni, S. R. A., & Pal, M., Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. International Journal of Data Science and Analytics, 18(4), 457-466, 2024.

[15] Tjoa, E., Guan, C.: A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. P, IEEE Transactions on Neural Networks and Learning Systems, 2020.

[16] Besold, T.R., Garcez, A.S., Bader, S., Bowman, H., Domingos, P.M., Hitzler, P., Kühnberger, K., Lamb, L., Lowd, D., Lima, P.M., Penning, L.D., Pinkas, G., Poon, H., & Zaverucha, G., 2017, "Neural-Symbolic Learning and Reasoning: A Survey and Interpretation", ArXiv, abs/1711.03902.

[17] Francesco Ricci, Lior Rokach, Bracha Shapira, "Recommender Systems Handbook", 3rd Edition, Springer, 2022

[18] Wang-Cheng Kang, Julian McAuley, "Self-Attentive Sequential Recommendation", https://arxiv.org/abs/1808.09781, 2018

[19] Scott Lundberg, Su-In Lee, "A Unified Approach to Interpreting Model Predictions", https://arxiv.org/abs/1705.07874, 2017

[20] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, Cynthia Rudin, "Interpretable Decision Sets: A Joint Framework for Description and Prediction", https://arxiv.org/abs/1611.01286 , 2016

[21] Christoph Molnar, “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable”, https://christophm.github.io/interpretable-ml-book/, 2024

[22] Federico Bianchi, Gaetano Rossiello, Luca Costabello, Matteo Palmonari, Pasquale Minervini, “Knowledge Graph Embeddings and Explainable AI”, 2020

[23] Rajabi, E., & Etminani, K. (2024). Knowledge-graph-based explainable AI: A systematic review. Journal of Information Science, 50(4), 1019-1029. https://doi.org/10.1177/01655515221112844

[24] E. Kafeza, G. Drakopoulos, Ph. Mylonas, “Graph Neural Networks in PyTorch for Link Prediction in Industry 4.0 Process Graphs”, in I. Maglogiannis, L. Iliadis, J. Macintyre, M. Avlonitis, A. Papaleonidas (Eds.), AIAI 2024, IFIP AICT 713, Corfu, Greece, June 27-30, 2024

[25] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam, “Neural Rating Regression with Abstractive Tips Generation for Recommendation”, In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 345-354, 2017.

[26] Lei Li, Yongfeng Zhang, and Li Chen, “Generate Neural Template Explanations for Recommendation”, In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 755-764, 2020.

[27] Lei Li, Yongfeng Zhang, and Li Chen, “Personalized Transformer for Explainable Recommendation”, In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4947-4957, 2021.

[28] Jianmo Ni, Jiacheng Li, and Julian McAuley, “Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects”, In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 188-197, 2019.

[29] M. Korakakis, E. Spyrou, Ph. Mylonas, S. J. Perantonis, “Exploiting social media information toward a context-aware recommendation system”, Social Network Analysis and Mining 7 (1), 42, 2017

[30] Tintarev, Nava & Masthoff, Judith, “Designing and Evaluating Explanations for Recommender Systems”, 10.1007/978-0-387-85820-3_15, 2011

[31] Yan, A., He, Z., Li, J., Zhang, T., & McAuley, J., “Personalized showcases: Generating multi-modal explanations for recommendations”, In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2251-2255) 2023

[32] Ren, Z., Xiao, Z., & Sun, Y., “Do We Trust What They Say or What They Do? A Multimodal User Embedding Provides Personalized Explanations”, arXiv preprint arXiv:2409.02965, 2024

[33] G. T. Papadopoulos, Ph. Mylonas, V. Mezaris, Y. Avrithis, I. Kompatsiaris, “Knowledge-assisted image analysis based on context and spatial optimization”, International Journal on Semantic Web and Information Systems (IJSWIS) 2 (3), 2006

[34] https://www.tensorflow.org/, last accessed: 02/08/2025, 12:41

[35] https://scikit-learn.org/stable/, last accessed: 02/08/2025, 12:43

[36] Instagram Data, https://www.kaggle.com/datasets/amirmotefaker/instagram-data, last accessed: 02/08/2025, 12:50

[37] Trending YouTube Video Statistics, https://www.kaggle.com/datasets/datasnaek/youtube-new, last accessed: 02/08/2025, 12:55

[38] Youtube Statistics, https://www.kaggle.com/datasets/advaypatil/youtube-statistics, last accessed: 02/08/2025, 12:58

[39] spaCy, https://spacy.io/usage, last accessed: 02/08/2025, 13:10

[40] Department of Informatics and Computer Engineering, University of West Attica, https://ice.uniwa.gr/en/home/, last accessed: 02/08/2025, 13:15