UDC 004.8:005.8:316.644 **doi: 10.32620/reks.2025.4.07**

Serhii DOLHOPOLOV, Yuliia RIABCHUN,
Maksym DELEMBOVSKYI, Oleksandr MOLODID

Kyiv National University of Construction and Architecture, Kyiv, Ukraine

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR MULTIMODAL SENTIMENT ANALYSIS IN REVITALIZATION PROJECT MANAGEMENT

This study focuses on the development and evaluation of an explainable artificial intelligence (XAI) framework for multimodal sentiment analysis, specifically applied to territorial revitalization project management. The research addresses the critical problem of "black box" AI models, whose lack of transparency hinders their adoption by project managers who require trustworthy information for high-stakes decision-making in complex social environments. The goal of this study is to propose and rigorously validate a novel framework for multimodal sentiment analysis that is tailored to provide transparent, trustworthy, and actionable insights for decision-making in territorial revitalization project management. The tasks to be solved include developing a hybrid XAI technique that fuses insights from cross-modal attention and gradient-based attribution, designing a cohesive, user-centric explanation format combining highlighted text and image heatmaps, constructing a custom RevitalizeSent-MM dataset for this specific domain, and empirically evaluating the framework's predictive accuracy and, crucially, the fidelity of its explanations. The methods used involve a transformer-based Multimodal Sentiment Analysis (MSA) model using BERT and ViT with cross-modal attention for information fusion. The explainability component is a hybrid XAI technique that integrates cross-modal attention analysis with Integrated Gradients to assign importance scores to input features. Evaluation was performed using standard classification metrics for performance and the "Accuracy Drop on Perturbation" metric for explanation fidelity. The results confirmed the efficacy of the framework. The multimodal model demonstrated superior accuracy over unimodal baselines, and the proposed XAI method achieved significantly higher fidelity than naive explanation approaches, demonstrating its ability to accurately reflect the model's internal reasoning. The scientific novelty lies in three areas: the development of a fused, hybrid XAI technique specifically for transformer-based multimodal models, creation of a unique, domain-specific dataset for revitalization analysis, and validation of a methodology for adapting advanced XAI to solve critical trust and adoption barriers, thereby confirming its practical significance in project management.

Keywords: *explainable AI (XAI); multimodal sentiment analysis; project management; deep learning; revitalization.*

1. Introduction

1.1. Motivation

Territorial revitalization, particularly in post-conflict or post-disaster scenarios, represents a profoundly complex undertaking. As highlighted by S. Yao and L. Wang, such initiatives extend far beyond mere physical reconstruction, encompassing intricate social, economic, and political dimensions [1]. Therefore, successfully navigating these projects requires not only effective resource allocation and infrastructure development but also careful management of public perception, stakeholder alignment, and community engagement. These social factors are shown by B. Guo et al., are significantly influenced by social capital [2]. Gaining social acceptance, a concept explored by X. In urban revitalization contexts, Jin et al. [3] stated that ensuring that efforts meet the actual needs and address the concerns of affected populations are paramount for achieving long-term success and stability [3]. In this high-stakes environment, understanding the dynamic landscape of public sentiment becomes a critical project management function, enabling timely risk identification, proactive communication, and adaptive planning.

The digital age offers unprecedented access to vast streams of public discourse through social media, news outlets, official reports, and community forums. As X. In the context of healthcare, Chen et al. observed that this information is increasingly multimodal, frequently combining textual narratives with powerful visual elements, such as photographs and videos [4]. Images depicting destruction, reconstruction progress, community gatherings, or protests provide rich contextual layers, as demonstrated by U. In machine translation, Sulubacak et al. often supplement, contradict, or nuance the accompanying text [5]. This fused multimodal data stream was noted by K. For health monitoring, Singh et al. holds immense potential for project managers seeking a holistic and timely understanding of stakeholder sentiment, emerging issues, and the overall social climate surrounding revitalization projects [6].

However, effectively leveraging this rich data source presents significant challenges. Traditional project monitoring methods, which rely on surveys, focus groups, or manual media scanning, are often slow, costly, resource-intensive, and resource-intensive, as exemplified by S. De Vito et al. struggled to capture the scale, speed, and nuances of online multimodal discourse in air quality monitoring [7]. The integration of AI and machine learning (ML) into project management has become a rapidly growing field, offering potential enhancements across various project phases, to address these limitations, as discussed by D. Chernyshev et al. in the context of smart construction [8–9]. Applications range from improved project duration and cost forecasting [10–11] to automated progress monitoring and risk prediction. Sentiment analysis, in particular, has been applied, primarily utilizing text data from stakeholder communications or social media platforms, to gauge project reception or identify potential issues, with recent proposals such as that of R. M. Omas-as and Encarnacion focusing on unified feedback management using text analytics [12].

Despite their analytical power, these modern multimodal models frequently suffer from a critical limitation: the “black box” problem, a philosophical challenge discussed by von Eschenbach [13]. Their internal decision-making processes are opaque, making it difficult, if not impossible, for users to understand why a particular sentiment prediction was made. This lack of transparency poses a fundamental barrier to adoption, particularly in high-stakes domains such as territorial revitalization project management, where AI and BIM integration are being explored, as seen in the work of S. Dolhopolov et al. [14]. Project managers are understandably hesitant to base critical decisions on resource deployment, risk mitigation strategies, or public communication adjustments, areas also investigated by S. Dolhopolov et al. regarding site analysis [15] on AI outputs they cannot scrutinize or trust. An incorrect or misunderstood AI assessment could lead to misallocation of resources, ineffective interventions, or even worsen social tensions. Therefore, the inability to explain how multimodal sentiment is derived prevents these powerful AI tools from reaching their full potential as trusted decision-support aids for revitalization project managers.

1.2. State of the art

The challenges of analyzing public sentiment have driven significant advancements in Artificial Intelligence, particularly in the fields of Multimodal Sentiment Analysis (MSA) and Explainable Artificial Intelligence (XAI). An examination of the state of the art in these areas reveals both the potential of current technologies and the critical gaps that hinder their practical application in project management.

Multimodal Sentiment Analysis (MSA) has progressed significantly from its early focus on unimodal textual data, as noted in foundational work on modality-invariant representations by D. Hazarika, R. Zimmermann, and S. Poria [16]. The proliferation of multimedia content spurred intensive research into MSA, which aims to leverage information from text, visuals, and audio to achieve a more robust and context-aware understanding of sentiment – a challenge tackled early on by A. Zadeh et al. with their Tensor Fusion Network [17]. While Artificial Intelligence (AI), particularly sentiment analysis based on advanced models like the Bidirectional GRUs explored by W. Xu et al., offers powerful tools for processing large volumes of text data [18], unimodal approaches fall short. Text-only analysis, as reviewed by Z. Tang, misses the crucial context embedded in visuals, potentially misinterpreting sarcasm, overlooking visual evidence contradicting text, or failing to grasp the emotional impact conveyed by an image [19]. Conversely, analyzing images alone lacks the specific details, opinions, and arguments expressed in text. Consequently, advanced multimodal AI models, often employing sophisticated transformer architectures and fusion mechanisms, sometimes questioning if “captions are all you need”, have emerged to analyze text and images jointly. These models offer significantly more accurate and context-aware sentiment assessments, though their application requires careful consideration, as shown by M. Aldeen et al. regarding adversarial attacks on multimodal systems [20]. Contemporary approaches within MSA are frequently distinguished by their fusion strategy. Early fusion techniques combine features from different modalities at the input level before prediction. While straightforward, researchers like A. A. Beserra, R. M. Kishi, and R. Goularte have noted that this approach can struggle with heterogeneous feature spaces and dimensionality [21]. Conversely, late fusion involves training separate models for each modality and subsequently combining their outputs, often at the decision level. As demonstrated by J. Cheng et al. in RGB-Thermal crowd counting, this allows for modality-specific modeling but risks missing complex inter-modal interactions [22]. Seeking to capture these crucial interactions, hybrid fusion methods have gained prominence, operating at intermediate representational stages. Dominant contemporary approaches, surveyed by M. Shaikh et al. in the context of action recognition, heavily involve attention mechanisms and transformer-based architectures [23]. Seminal models such as ViLBERT, developed by J. Lu et al. [24], LXMERT by H. H. Tan and M. Bansal [25], and

UNITER proposed by Y. Chen et al. [26], employ sophisticated cross-modal attention layers. Recent research continues to refine these mechanisms, with novel approaches like the dual-attention model by Wang et al. [27], which simultaneously models intra- and inter-modality dynamics and introduces techniques like cross-correlation loss to improve feature fusion. These allow textual and visual features to dynamically influence each other during processing, leading to high performance on diverse vision-and-language tasks, including MSA.

Despite these advancements, significant hurdles remain in handling modality imbalance, noisy data, and the scarcity of domain-specific annotated datasets like MOSI or MOSEI, as discussed by W. Yu et al. [28] and Z. Liu et al. [29].

The increasing complexity and "black box" nature of these models have catalyzed the development of Explainable AI (XAI) [30–31]. As B. Pradhan et al. emphasize that XAI methods aim to render AI predictions transparent and understandable to humans, thereby fostering trust, enabling debugging, and promoting accountability [31]. The major paradigms within XAI include feature attribution methods that assign importance scores to input features. Prominent examples of text include LIME, SHAP, and attention visualization, as applied by H. Jang et al. [32], whose utility was also explored using R. Hasan et al. [33] and reviewed in the context of genomics by S. R. Choi and M. Lee [34]. For image data, gradient-based techniques such as Saliency Maps and Grad-CAM, along with its enhancements like Grad-CAM++ proposed by Y. Gao et al. [35] highlighted salient pixel regions. Recent comprehensive reviews, such as the one by Cheng et al. [36], categorize these computer vision XAI methods into attribution-based, perturbation-based, and transformer-based approaches, systematically evaluating their trade-offs in terms of key characteristics, such as faithfulness and computational efficiency. Other XAI approaches include example-based explanations, which reference influential training instances, a method compared to rule-based explanations by van der J. V. Waa et al. [37], and surrogate models, where simpler models mimic the complex model, as explored by A. Engel et al. [38]. Furthermore, emerging research is exploring novel paradigms for achieving interpretability, such as the self-supervised learning frameworks proposed by Sun et al. [39], which aim to automatically extract key sentiment cues from text to provide global, rather than instance-specific, explanations.

XAI techniques for unimodal natural language processing and computer vision are relatively mature, as surveyed by R. Gipiškis et al. [40], achieving explainability for multimodal models remains a significant challenge. Recent research has demonstrated the value of applying XAI techniques, such as LIME and SHAP, to transformer-based language models for sentiment analysis, particularly to enhance transparency and trust in contexts such as low-resourced languages, as shown by Mabokela et al. [41]. However, these studies primarily focus on unimodal text and do not address the unique challenges of explaining fused, multimodal predictions. This has been highlighted in recent studies by F. Cerasuolo et al. focused on network traffic classification [42]. Applying unimodal XAI techniques independently to each modality fails to explain the crucial fusion process where modalities interact. This creates a clear and compelling research gap: there is a notable lack of established, robust methods specifically designed to generate fused, intuitive explanations from complex multimodal models that are readily interpretable and actionable for domain experts, such as project managers. This paper addresses this gap by adapting and combining XAI techniques to produce explanations tailored specifically for decision support in revitalization project management.

1.3. Objectives and tasks

To bridge the critical gap identified between the capabilities of advanced multimodal AI and the practical needs of project managers for trustworthy, interpretable insights, this study aims to develop and rigorously evaluate a novel framework for explainable multimodal sentiment analysis. The primary objective of this study is to create a methodology tailored to provide transparent, trustworthy, and actionable insights for decision-making in territorial revitalization project management.

To achieve this overarching objective, this research undertakes the following specific tasks:

1. To develop and implement a novel hybrid XAI technique specifically adapted for modern transformer-based multimodal sentiment analysis models, fusing insights from cross-modal attention analysis and gradient-based attribution methods.

2. To design and propose a cohesive, user-centric explanation format that presents highlighted salient text words alongside image region heatmaps, making the model's reasoning process transparent and intuitive for project managers.

3. To construct and validate RevitalizeSent-MM, a custom, domain-specific dataset that accurately reflects the real-world challenges and linguistic nuances of territorial revitalization, providing a robust foundation for model training and evaluation.

4. To rigorously evaluate the proposed framework through comprehensive experiments, assess its sentiment prediction accuracy against unimodal baselines, and measure the fidelity and utility of its explanations via quantitative metrics.

By addressing these tasks, this study aims to fill the identified research gap and generate valuable insights into the design principles for creating effective, domain-specific explanations. The application is firmly grounded within the domain of territorial revitalization project management, tailoring the generated explanations to support relevant decision-making tasks, such as risk identification and stakeholder analysis. Crucially, through empirical evaluation, the impact of these explanations on interpretability is explicitly investigated from a project management perspective. Ultimately, this work seeks to provide a validated methodology and a tangible pathway toward more responsible, transparent, and effective AI-assisted decision support for managers navigating the complexities of territorial revitalization.

The remainder of this article is structured to logically present the research findings. Section 2 details the proposed methodology, outlining the overall architecture, data preprocessing steps, the multimodal sentiment analysis model, and the core XAI techniques. Section 3 describes the experimental setup, including the custom dataset construction, implementation details, evaluation metrics, and baselines for comparison. Section 4 presents the results of the quantitative experiments and includes a practical Case Study to demonstrate the application of the framework. Section 5 provides a comprehensive Discussion of the findings, their practical implications, and limitations of this study. Finally, Section 6 offers the Conclusion, summarizing the key contributions and outlining promising directions for future research.

2. Methodology

This section details the proposed framework for explainable multimodal sentiment analysis, specifically tailored to support decision-making in revitalization project management. This paper presents the overall architecture, outlines the data preprocessing steps, describes the chosen multimodal sentiment analysis (MSA) model, elaborate on the core XAI techniques adapted for explaining fused predictions, and defines the format of the generated explanations.

2.1. Overall Framework Architecture

The proposed framework is designed to process pairs of text and images related to revitalization projects, predict the associated sentiment polarity (Positive, Negative, or Neutral), and provide interpretable explanations that link this prediction back to salient features within both the textual and visual modalities. The workflow consists of several interconnected modules, as shown in Figure 1. The process begins with the Input Module accepting text-image pairs, such as social media posts or news snippets with accompanying images. The Preprocessing Module then independently prepares the text (through tokenization and cleaning) and images (through resizing and normalization) for their respective encoders. The multimodal sentiment analysis (MSA) module, which employs a deep learning model featuring separate text and image encoders followed by a fusion mechanism to integrate the information before predicting the sentiment, is central to the framework. Subsequently, the XAI Explanation Module receives the original input pair and the prediction of the MSA model. This module applies specifically adapted XAI techniques to identify and quantify the contribution of textual elements (words or sub-word tokens) and visual regions (pixels or image patches) to the final sentiment outcome. Then, the Explanation Output Module renders these generated explanations into a user-friendly, combined format optimized for project managers' interpretability. Finally, a conceptual Interpretation Layer for PM represents the interface or cognitive process through which project managers utilize the sentiment prediction and its accompanying explanation to inform critical decisions, such as risk assessment or communication strategy adjustments.

2.2. Data Preprocessing

Standard preprocessing steps are applied to ensure compatibility with deep learning models. The input text is cleaned to remove artifacts like URLs or excessive special characters. It is then tokenized using a subword tokenizer compatible with the chosen text encoder, such as WordPiece for BERT models. The token sequences are padded or truncated to a fixed maximum length, and special tokens (e.g., [CLS], [SEP]) are incorporated as required by the specific transformer architecture. The input images are decoded and resized to the precise input dimensions expected by the visual encoder (e.g., 224x224 for a standard ViT model).

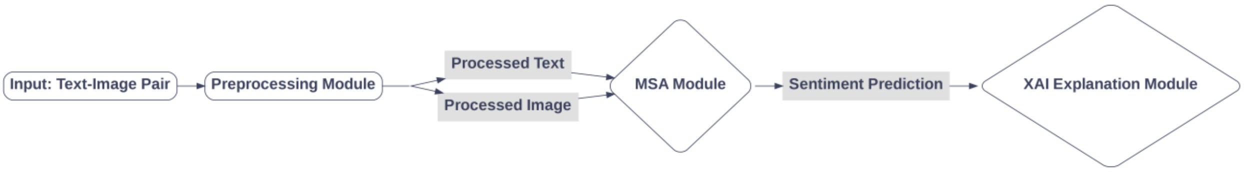


Fig. 1. Proposed Framework Architecture. Text and image inputs are processed by the MSA module for sentiment prediction. The XAI module, accessing the MSA model, generates explanations by analyzing attention and gradients, which are then formatted for project manager interpretation and decision support

2.3. Multimodal Sentiment Analysis (MSA) Model

A transformer-based architecture is employed to effectively capture the complex interplay between textual and visual information, which uses powerful pre-trained unimodal encoders coupled with a dedicated fusion mechanism. Specifically, the proposed model comprises a pre-trained BERT model (or a suitable multilingual or domain-adapted variant, e.g., bert-base-uncased) serving as the Text Encoder. This model processes the tokenized text to produce contextualized token embeddings. A pre-trained vision transformer (ViT) model (e.g., google/vit-base-patch16-224-in21k) [42] acts as the image encoder for the visual modality. It divides the input image into fixed-size patches, linearly embeds them, adds positional information, and processes these sequences through transformer layers to generate patch embeddings, typically including a [CLS] token embedding that summarizes the global image context. The core of the multimodal integration lies in the Fusion Mechanism, for which cross-modal attention layers are utilized, drawing inspiration from seminal works such as ViLBERT and LXMERT [24–25]. These layers enable representations from one modality to dynamically attend to representations from the other (e.g., text tokens attending to image patches and vice versa), facilitating the learning of joint representations that capture inter-modal dependencies crucial for accurate sentiment prediction. Multiple such fusion layers can be stacked to deepen the interaction. Finally, the fused representation, often derived from the [CLS] tokens of both modalities or through a pooling operation over the fused sequence, is passed to a simple Prediction Head, typically a Multi-Layer Perceptron (MLP) classifier terminating in a softmax layer, which outputs the probability distribution over the predefined sentiment classes (Positive, Negative, Neutral). The entire model is trained end-to-end using a standard cross-entropy loss function. This architectural choice leverages strong pre-trained unimodal representations while employing cross-modal attention to effectively capture the nuanced interplay between text and visuals, which is essential for high-performance MSA.

2.4 XAI Technique Adaptation for Multimodal Explanations (Core Novelty)

The central challenge addressed by this methodology is explaining how the MSA model's fused representation leads to a specific sentiment prediction, attributing importance back to the original input features, namely, text tokens (or words) and image regions (or patches). To achieve this, this study proposes adapting and combining two complementary XAI approaches, as detailed below.

2.4.1. Cross-Modal Attention Analysis

Transformer architectures inherently rely on attention mechanisms that calculate relevance scores between different elements in the input sequences. The proposed approach analyzes these weights, focusing on both self-attention (interactions within a single modality) and, more critically, cross-attention (interactions between text and image modalities) derived from the trained MSA model's fusion layers. The adaptation involves moving beyond visualizing raw attention weights, which can be noisy or misleading. Instead, techniques potentially inspired by methods such as Attention Rollout are employed, or simple attention weight aggregation strategies across relevant layers and heads are used. The flow of information can be traced more reliably by aggregating attention across multiple layers, smoothing out anomalies from single layers and capturing a more holistic view of the model's reasoning. This process estimates the effective contribution or focus directed from specific input text tokens to specific image patches, and vice versa, as relevant to the final prediction. The process involves identifying the attention heads and layers most influential in the final classification step and aggregating the self-attention weights that are cross-modal and potentially modulated. The output yields scores indicating the model's "attention focus" on specific words and image patches resulting from its internal interaction patterns.

2.4.2. Multimodal Gradient-based Attribution

Gradient-based methods form another cornerstone of the presented XAI approach. These techniques are used to calculate the gradient of the model's output score (e.g., the predicted probability for the target sentiment class) with respect to its input features. The magnitude of this gradient signifies how a small change in a particular feature would influence the prediction, indicating its importance. Established gradient-based methods are adapted to the multimodal architecture, with a preference for Integrated Gradients (IG) due to its desirable theoretical properties, such as obvious satisfaction. The adaptation process involves calculating the target class probability gradient with respect to the final fused representation generated before the classification head. This gradient is then backpropagated through the fusion layers and subsequently routed separately down through the text and image encoder pathways to their respective input embedding layers (i.e., token embeddings for text and patch embeddings for images) or even to the raw input level (i.e., input IDs and pixel values). For text, attribution scores derived at the sub-word token level are aggregated to produce word-level importance scores. For images, attributions calculated for patch embeddings are mapped back to the original pixel space and typically visualized as a heatmap, analogous to methods like Grad-CAM but derived differently based on the transformer architecture. This process yields attribution scores signifying the direct influence or importance of each word and image region in driving the model's specific sentiment prediction.

The format typically includes a Visual Component, where the input image is displayed overlaid with a heatmap. This heatmap, primarily derived from gradient-based attribution methods (or potentially attention maps), highlights the pixel regions or patches deemed most influential for the predicted sentiment, with intensity or color indicating the influence's strength. Complementing this is the Textual Component, where the original input text is presented with highlighted important words or phrases. Highlighting intensity or color reflects the aggregated importance scores derived from a combination of attention analysis and gradient-based methods for the corresponding sub-word tokens. Figure 2 shows an example of this target fused explanation format. This dual-method approach captures both the model's "focus" (from attention) and its "feature influence" (from gradients). Optionally, this core visual-textual explanation can be augmented with Summary Metrics or Text, such as a brief automatically generated textual summary (e.g., "Positive sentiment linked to 'reconstruction' in text and bridge structure in image") or feature importance bar charts enumerating the top contributing words and image regions. This combined format is designed to allow project managers to quickly grasp which parts of the text and image contributed most significantly to the AI's sentiment assessment, thereby facilitating a more grounded, transparent, and trustworthy interpretation compared to receiving only a simple prediction label. The design emphasizes direct visual evidence (heatmaps) and textual pointers (highlights) as potentially more intuitive for non-AI experts than abstract numerical scores alone.

Figures 3 and 4 illustrate the process of mapping abstract model outputs back to a user-interpretable visual format. The initial output, shown in Figure 3, is a raw, low-resolution attribution map, where each square represents a patch of the original image, and its color indicates the importance score. This raw map is then upscaled and smoothly interpolated to be overlaid on the original image (Figure 4). This final heatmap provides an intuitive visual guide to the regions considered most salient by the model for its decision. While these figures showcase the visualization of attention scores, the principle for mapping gradient-based attributions is analogous.

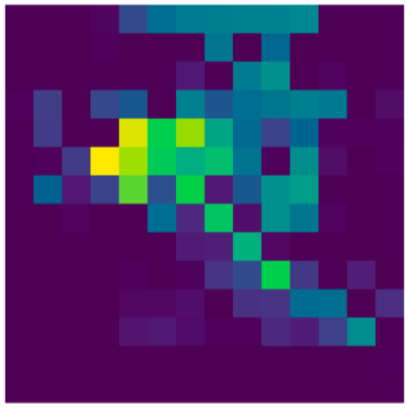
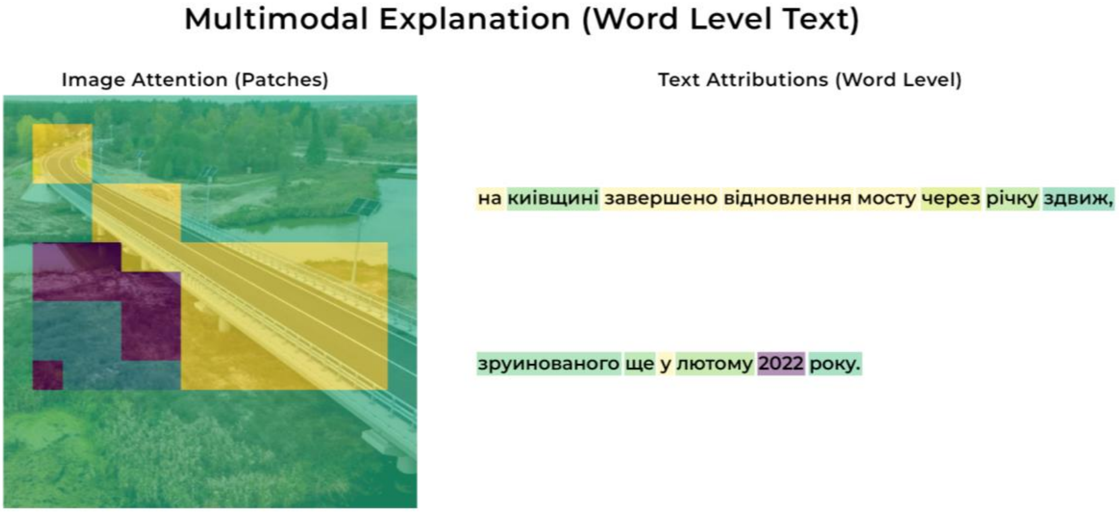


Fig. 3. A raw, patch-based attention map

2.5. Mathematical Formulation of the Hybrid XAI Method

This section provides the detailed mathematical and algorithmic formulations for the hybrid XAI approach to ensure methodological rigor and experimental reproducibility, covering attention aggregation, gradient routing, and the final score combination.



Multimodal Explanation (Word Level Text)

Image Attention (Patches)

Text Attributions (Word Level)

на київщині завершено відновлення мосту через річку здвиж,

зруинованого ще у лютому 2022 року.

Fig. 2. Example of the target fused explanation format, showing the etalon explanation for the Zdvizh bridge revitalization

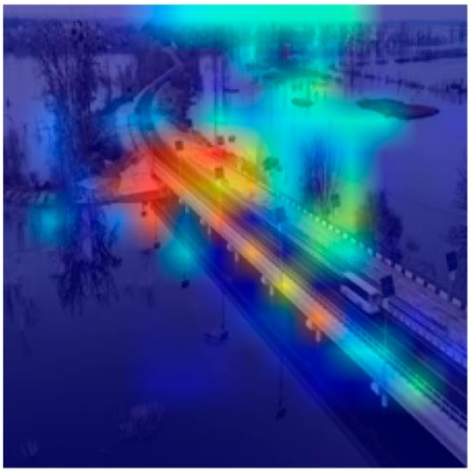


Fig. 4. The final attention heatmap overlaid on the original image

2.5.1. Cross-Modal Attention Score Aggregation

The attention scores are derived directly from the cross-modal attention mechanisms within the fusion layers of the model. For a model with L fusion layers and H attention heads per layer, the attention matrix for a given layer l and head h is denoted as $A^{(l,h)}$. These matrices are aggregated by averaging them to obtain a single, robust attention map:

$$A_{agg} = \frac{1}{|L'|\cdot H} \sum_{l\in L'} \sum_{h=1}^{H} A^{(l,h)}, \qquad (1)$$

where L′ is the subset of "relevant" layers selected for explanation. For this analysis, the attention matrices from the final two cross-modal fusion layers (L′ = {L-1, L}) were exclusively utilized because they have the most direct influence on the final fused representation that feeds into the prediction head. The resulting aggregated matrix A_{agg} provides the raw attention-based importance scores, S_{attn}.

2.5.2. Multimodal Gradient-based Attribution

The gradient-based component is calculated using Integrated Gradients (IG), which require defining a baseline (an information less input).

For the textual modality, the baseline ($T_{baseline}$) was a sequence of embeddings corresponding to the [PAD] token.

For the visual modality, the baseline ($V_{baseline}$) was a black image with all pixel values set to zero.

The attribution for a given input feature embedding is calculated by integrating the gradients along a straight-line path from the baseline to the input. In practice, this integral is approximated using a summation over several steps, as shown in the following formula:

$$\mathrm{Attr}(E_i) \approx (E_i - (E_i') \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial y_c\left(E' + \frac{k}{m}(E - E')\right)}{\partial E_i}, \qquad (2)$$

where $\mathrm{Attr}(E_i)$ is the Integrated Gradients attribution score for the i-th feature embedding. E_i is the embedding of the input feature (a text token or an image patch). E_i' is the corresponding baseline embedding. m is the number of steps used to approximate the integral. $\partial y_c(\ldots) \,/\, \partial E_i$ is the output gradient of the model for predicted class c with respect to the feature embedding, evaluated at step k along the path from the baseline to the input.

2.5.3. Hybrid Explanation Generation

The raw scores from both attention and gradient methods are combined to generate the final, unified explanation. First, both sets of scores are independently normalized to a [0, 1] range using min-max scaling to ensure that they are comparable.

$$S_i' = \frac{S_i - \min(S)}{\max(S) - \min(S)}, \qquad (3)$$

where S_i' is the normalized importance score for the i-th feature. S_i is the raw importance score (either from attention, S_{attn}, or gradients, S_{grad}). $\min(S)$ and $\max(S)$ are the minimum and maximum scores across all features for the method.

Then, the normalized scores are linearly combined to produce the final hybrid score:

$$S_{hybrid} = \alpha \cdot S_{attn}' + (1 - \alpha) \cdot S_{grad}', \qquad (4)$$

where S_{hybrid} is the final hybrid importance score. S_{attn}' and S_{grad}' are the normalized scores obtained from the attention and gradient methods, respectively. α is a weighting coefficient between 0 and 1. In the experiments, α was set to 0.5 to give equal importance to both the model's focus (attention) and the features' influence (gradients).

Finally, since the text encoder operates on sub-word tokens, the resulting attribution scores S_{hybrid} for text must be aggregated to the word level for human interpretability. For any given word w, which comprises a set of sub-word tokens T_w, its final score is determined by the maximum attribution score among its constituent tokens:

$$\mathrm{Score}_w = \max\left(S_{hybrid}(t) \mid t \in T_w\right), \qquad (5)$$

where Score_w is the final importance score for the word w. $S_{hybrid}(t)$ is the hybrid score for an individual sub-word token t.

This aggregation method ensures that the word's most influential part determines its overall importance in the final explanation.

3. Experimental Setup

This section details the experimental design employed to rigorously evaluate the proposed framework for explainable multimodal sentiment analysis within the specific context of territorial revitalization project management. This section describes the construction and characteristics of the RevitalizeSent-MM domain-specific dataset, the implementation specifics for the MSA model and XAI techniques, the evaluation metrics used for both quantitative and qualitative assessment, and the baseline methods used for comparative analysis.

3.1. Dataset: RevitalizeSent-MM

Recognizing the absence of publicly available datasets tailored to multimodal sentiment capture during territorial revitalization, particularly in post-conflict settings like Ukraine post-2022, a new dataset named RevitalizeSent-MM was constructed. Data pairs consisting of text snippets and associated images were collected from publicly accessible sources pertinent to ongoing revitalization efforts. These sources primarily included publicly available news articles and reports from reputable national and international news outlets covering reconstruction and recovery activities, public posts from official government and municipal channels detailing project progress, and publicly shared content from social media platforms (such as X, Facebook, and relevant Telegram channels, adhering to platform Terms of Service and user privacy settings) identified using keywords such as #reconstruction, #recovery, and relevant project names. Only pairs in which an image was directly associated with the text were retained.

An initial large collection of data pairs was filtered to ensure relevance to the revitalization context. The selection criteria included keywords indicating revitalization activities (e.g., rebuilding, repair, recovery, and restored), associated geotags or explicit location mentions within the targeted revitalization zones, and the requirement for both meaningful textual content (exceeding simple captions) and a contextually relevant image. Pairs containing generic stock photos that were unrelated to the textual narrative were discarded where feasible.

A crucial phase involved the manual annotation of the filtered text-image pairs to assign the sentiment labels. A team of five annotators, fluent in both Ukrainian and English and specifically briefed on the revitalization context's nuances, performed the labeling task. Each pair was assigned one of three sentiment labels: Positive (expressing satisfaction, hope, progress, and successful completion), Negative (expressing dissatisfaction, criticism, despair, destruction, lack of progress, and danger), or Neutral (typically objective reporting or factual statements lacking strong sentiment). Annotators were explicitly instructed to consider the combined meaning conveyed by both the text and the image. To ensure annotation quality, a subset comprising 15% of the data was independently annotated by at least two annotators, allowing for the calculation of inter-annotator agreement (IAA). Using Cohen's Kappa, a score of 0.78 was achieved, indicating substantial agreement among the annotators. Instances with disagreements were subsequently resolved through moderated discussion to reach a consensus label. The final dataset was randomly partitioned into training (70%), validation (15%), and test (15%) sets, ensuring no overlap between the splits for unbiased evaluation. Data collection strictly adhered to ethical considerations, using only publicly available sources. For social media data, identifiable user information was anonymized during the analysis and was not present in any published examples, in accordance with privacy norms and platform guidelines. The analysis focuses on aggregated sentiment trends rather than individual user profiling. Table 1 summarizes the key statistics characterizing the resulting RevitalizeSent-MM dataset.

Table 1

Summary Statistics for the RevitalizeSent-MM Dataset

Statistic	Training Set	Validation Set	Test Set	Total
Total Samples (Text+Img)	5.810	1.245	1.245	8.300
Positive Samples	1.980	415	425	2.820
Negative Samples	2.250	490	485	3.225
Neutral Samples	1.580	340	335	2.255
Avg. Text Length (Tokens)	65	68	66	66
Unique Tokens (Vocabulary)	~25.000	-	-	~25.000
Image Dimensions (Avg.)	~600x450	~600x450	~600x450	~600x450
IAA (Kappa Score)	-	-	-	0.78

3.2. Implementation Details

The models and experimental workflow were implemented using Python 3.11. The primary deep learning framework employed was PyTorch (version 1.13+). The Hugging Face transformers library (version 4.25+) was utilized to leverage pre-trained transformer models, specifically for accessing BERT (bert-base-multilingual-cased chosen for broader language support) and ViT (google/vit-base-patch16-224-in21k) architectures. The Captum library (version 0.6+) was used to implement explainability functionalities, particularly gradient-based methods like Integrated Gradients, supplemented by custom scripts for extracting and analyzing attention weights where applicable. Standard image processing tasks were performed using Pillow and OpenCV, while data manipulation and visualization relied on Pandas, NumPy, Matplotlib, and Seaborn.

The Multimodal Sentiment Analysis (MSA) model, featuring the BERT and ViT backbones combined via cross-modal attention fusion layers, was specifically fine-tuned on the RevitalizeSent-MM training dataset. The text and image encoders were initialized with their respective pre-trained weights to leverage transfer learning. Model optimization was performed using the AdamW optimizer with a learning rate of 2e-5, incorporating a linear warmup phase over the first 10% of training steps followed by linear decay. Training was conducted with a batch size of 16, which was constrained by the available GPU memory (NVIDIA GeForce RTX 4080 SUPER). The model was trained for a maximum of 10 epochs, employing early stopping based on the validation set's F1-score performance, with a patience of 2 epochs to prevent overfitting.

For the XAI implementation, cross-modal attention analysis involved extracting aggregated attention weights from the trained MSA model's final fusion layer. Multimodal Integrated Gradients were implemented using Captum's LayerIntegratedGradients targeting word embeddings for text and standard IntegratedGradients targeting pixel values for images, relative to appropriate baselines (padding token embeddings for text, zero-pixel image for visuals). Fifty steps were used for the Integrated Gradients approximation path integral.

3.3. Evaluation Metrics

The evaluation strategy employed a combination of quantitative metrics to assess both the performance of the underlying MSA model and the characteristics of the generated explanations, alongside qualitative user-based assessments focused on interpretability and trust.

3.4. MSA Model Performance

The predictive capability of the fine-tuned MSA model was evaluated on the held-out test set using the following standard multi-class classification metrics: overall Accuracy, Precision, Recall, and F1-Score. To account for potential class imbalance, precision, recall, and F1-Score were calculated both per class (positive, negative, neutral) and as macro and weighted averages.

3.5. Quantitative XAI Evaluation

The "quality" of explanations is inherently challenging to quantify. Two metrics commonly used in XAI literature were adopted to assess the properties of the explanation: Fidelity and Sparsity. Fidelity was measured using the Accuracy Drop on Perturbation. This involved identifying the top-k% (with k=15%) most important features (i.e., aggregated words for text based on token attributions, image patches based on aggregated pixel attributions, or attention scores) according to the XAI method. These features were then removed or masked from the input (e.g., replacing text tokens with [MASK], blacking out image patches), and the resulting drop in the prediction accuracy of the MSA model was measured. A higher accuracy drop suggests higher fidelity, indicating that the explanation successfully identified features that were genuinely influential to the decision of the model. This was compared against the accuracy drop caused by removing random features of the same proportion. Sparsity, often linked to comprehensibility, was measured by the average percentage of features (words for text, patches for image) highlighted as belonging to the top-k% importance level. Simpler explanations (lower sparsity) are generally preferred, provided that fidelity is maintained.

The chosen metrics are aligned with a formal model of explainability to ground the evaluation in established quality frameworks for AI systems, such as the hierarchical quality model proposed by Kharchenko et al. [43] In this model, the high-level characteristic "Explainability" (EXP) is decomposed into several measurable subcharacteristics, including Comprehensibility (CMH), Interpretability (INP), and Verifiability (VFB). The selection of Fidelity and Sparsity was deliberately driven by the need to quantify these key aspects within the application context.

Fidelity, which measures the faithfulness of an explanation to the model's internal logic, serves as a quantitative measure of Verifiability. Establishing that an explanation accurately reflects the model's reasoning is a prerequisite for user trust in a high-stakes domain such as territorial revitalization. Sparsity was chosen as a critical proxy for Comprehensibility and Interpretability. For end-users who are not AI experts, concise explanations that isolate the most critical sentiment drivers are more effective and less prone to information overload than dense, complex explanations. While other XAI evaluation metrics, such as stability or consistency, exist, Fidelity and Sparsity were considered the most relevant for assessing the practical utility of explanations in this specific decision-support context, as they directly measure the core components of a formal XAI quality model and align with the primary goals of the research: providing trustworthy and actionable insights.

3.6. Baselines for Comparison

Several baseline methods were established for comparison to effectively evaluate the contribution of the proposed explainable multimodal framework:

1. MSA Model (No XAI). Performance and user

perception (trust, understanding) of the base multimodal sentiment analysis model when predictions are presented without any accompanying explanation. This serves as the fundamental baseline.

2. Unimodal XAI (Text-Only) A BERT model was trained exclusively on the textual data from RevitalizeSent-MM. Its predictions were explained using a standard text XAI method (Integrated Gradients attributing to word embeddings, visualized via text highlighting).

3. Unimodal XAI (Image-Only). A ViT model trained solely on the image data. Explanations were generated using a standard vision XAI method (Integrated Gradients attributing to input pixels, visualized as a heatmap).

4. Naive Multimodal XAI. Applying the chosen unimodal XAI techniques (IG for text embeddings, IG for image pixels) independently to the respective streams within the trained multimodal model, but visualizing them separately without the fused explanation format or cross-modal consideration during the XAI calculation itself.

5. Attention Rollout. It was included as a direct baseline for attention-based explanation. This technique, which is readily applicable to transformer architectures, generates explanations by aggregating attention scores across all model layers. The attention mechanism is treated as a flow network and the contribution of input tokens to the final representation is computed by recursively multiplying attention matrices. This method was applied to the trained MSA model's cross-modal attention layers to generate separate importance scores for text and image features. Including this baseline is crucial as it allows for a direct comparison of the hybrid gradient-attention approach against a well-established method that relies solely on attention, highlighting the specific benefits of incorporating gradient-based information.

These baselines enable a multi-faceted evaluation: assessing the benefit of multimodality (comparing MSA vs. Unimodal models), the value of adding any explanation (comparing No XAI vs. XAI baselines), and the specific advantage of the adapted multimodal XAI approach and fused explanation format over simpler, non-integrated, and purely attention-based explanation strategies.

4. Case Study

This section presents the experimental evaluation findings, comparing the performance of the proposed multimodal sentiment analysis model and its associated explainability framework against established baselines. This section reports on the predictive accuracy of the core model and then delve into quantitative and qualitative assessments of the generated explanations.

4.1. Multimodal Sentiment Analysis Model Performance

The proposed MSA model, which integrates BERT and ViT encoders with cross-attention fusion, was first evaluated against unimodal baselines (Text-Only BERT, Image-Only ViT) trained on the RevitalizeSent-MM dataset. Table 2 summarizes the performance, measured on the held-out test set using weighted average metrics across the three sentiment classes (positive, negative, and neutral).

As shown in **Ошибка! Источник ссылки не найден.**, the proposed multimodal model significantly outperforms both the text-only and image-only baselines across all standard evaluation metrics. The substantial improvement in F1-score from 0.716 (text-only) and 0.570 (image-only) to 0.803 for the fused model confirms the synergistic benefit of integrating both textual narratives and visual cues for accurately assessing sentiment within the dataset's complex revitalization context. This highlights the need for a multimodal approach, as visual information often provides crucial context absent in text alone and vice versa.

4.2. Quantitative XAI Evaluation

The explanations produced by the adapted XAI framework, which combines insights from cross-modal attention and Integrated Gradients (IG), were then quantitatively assessed, and its characteristics were compared against baseline explanation methods using Fidelity and Sparsity metrics.

Table 2

Sentiment Classification Performance on the RevitalizeSent-MM Test Set (Weighted Averages)

Model	Accuracy	Precision	Recall	F1-Score
Text-Only (BERT Baseline)	0.718	0.715	0.718	0.716
Image-Only (ViT Baseline)	0.572	0.569	0.572	0.57
Proposed MSA (BERT+ViT+Fusion)	0.805	0.802	0.805	0.803

4.2.1. Fidelity

Fidelity measures how well an explanation reflects the internal reasoning of the model by evaluating the impact of removing the most important features identified using the XAI method. The drop in the accuracy of the MSA model on the test set was calculated after masking the top 15% most important features (tokens aggregated to words for text, image patches derived from pixel/embedding attributions for image) identified by each XAI approach. A larger accuracy drop indicates higher fidelity. The results, compared against removing random features, are shown in Figure 5.

The proposed fused XAI method demonstrated significantly higher fidelity, inducing an average accuracy drop of 17.6%, compared to only 1.9% when removing random features ($p < 0.001$, via paired t-test). It also significantly outperformed the naive application of multimodal IG (which yielded a 11.8% drop, $p < 0.01$) and the unimodal baselines applied within the multimodal framework (Text-Only IG: 8.5% drop; Image-Only IG: 6.9% drop). This strongly suggests that the adapted method, designed to account for cross-modal interactions during explanation generation, is more effective at identifying the input features that are truly critical to the complex multimodal model's final prediction.

The hybrid approach also showed a substantial improvement over the Attention Rollout baseline, which yielded a 14.2% accuracy drop. While Attention Rollout proved more faithful than naive methods by effectively propagating attention scores through the model's layers, its lower fidelity compared to the proposed method highlights the limitations of relying solely on attention mechanisms. This result strongly suggests that the gradient-based information integrated into the proposed approach captures signals of critical feature importance that are missed by purely attention-based explanations. This confirms that the adapted method, which is designed to account for both attention patterns and feature influence via gradients, is more effective at identifying the input features that are critical to the final prediction of the complex multimodal model.

4.2.2. Sparsity

Sparsity is related to the conciseness of an explanation, which is often correlated with comprehensibility. The average percentage of input features (words for text, patches for image) was measured and highlighted as belonging to the top 15% importance level by each method. The results are presented in Table 3.

As expected, the sparsity levels were broadly comparable across the different attribution-based methods when selecting a fixed percentage (top 15%), as shown in Table 3. The proposed method provides focused explanations by highlighting approximately 14.2% of text words and 14.5% of image patches, achieving high fidelity without excessive visual clutter, thus balancing faithfulness to the model with potential user comprehensibility.

Table 3

Sparsity of Explanations (Average % of Features Highlighted in Top 15% Attribution)

XAI Model	Modality	Avg. % Features Highlighted
Proposed Fused XAI	Text	14.2%
	Image	14.5%
Naive Multimodal IG	Text	14.3%
	Image	14.6%
Attention Rollout	Text	14.3%
	Image	14.6%
Unimodal Text-Only IG	Text	14.1%
Unimodal Image-Only IG	Image	14.4%

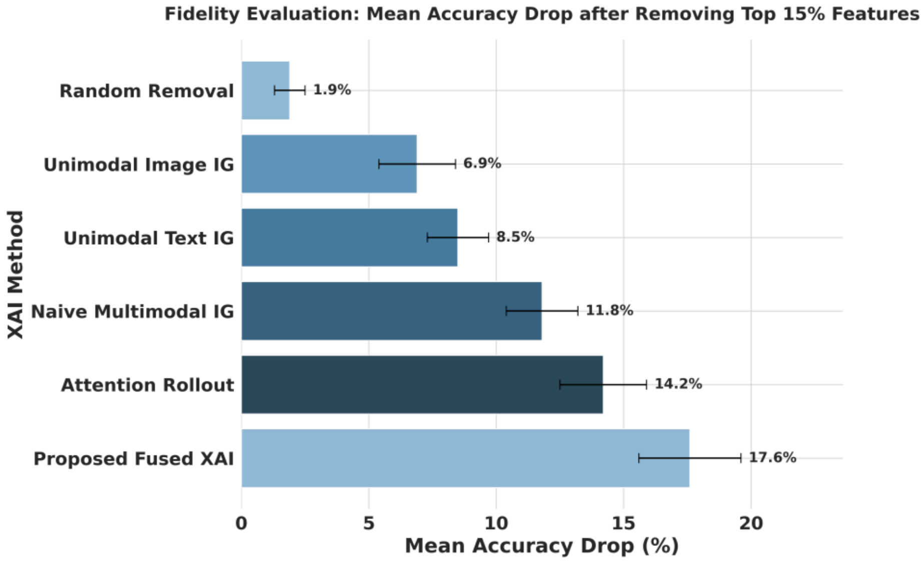


Fig. 5. Fidelity evaluation: Mean accuracy drop (%) on the test set after removing the top 15% of features identified using different XAI methods. Higher bars indicate higher fidelity. Error bars represent 95% Confidence Intervals

5. Discussion

The results of this study offer compelling evidence for the efficacy of the proposed framework for explainable multimodal sentiment analysis in the context of territorial revitalization. This section dissects these findings, interprets their broader implications for both project management practice and the academic field of Explainable AI (XAI), and acknowledges the inherent limitations of this study.

5.1. Interpretation of Key Findings

The experimental evaluation yielded two central findings. The first is the superior predictive performance of the multimodal sentiment analysis (MSA) model compared to its unimodal counterparts. The second is the significantly higher fidelity of the proposed hybrid XAI method compared with naive or unimodal explanation techniques and, notably, a strong attention-based baseline.

The synergy of multimodality is clearly demonstrated by the marked improvement in accuracy and F1-score of the fused BERT-ViT model. This underscores a fundamental truth about human communication, which is particularly salient in revitalization contexts, that is, meaning is co-constructed from multiple channels. Text alone can be ambiguous. A statement such as “Work is progressing on the bridge” could be neutral, but when paired with an image of a fully reconstructed bridge, the sentiment becomes unequivocally positive. Conversely, an image of a desolate site could render the same text sarcastic. The ability of the model to capture these synergistic and sometimes contradictory signals through its cross-modal attention mechanism is the primary driver of its enhanced performance. It moves beyond simple feature concatenation toward a more profound, contextual understanding that is essential for the nuanced domain of public sentiment.

The critical importance of fused explanations represents the most significant contribution of this study. The quantitative evaluation of explanation fidelity revealed that the proposed hybrid XAI method, which back-propagates gradients through fusion layers, dramatically outperforms naive approaches. More importantly, it surpassed the performance of Attention Rollout, a well-established method that relies solely on attention scores. This specific comparison is particularly insightful because it demonstrates that while attention mechanisms effectively show where the model is looking during the fusion process, gradient-based methods reveal the influence of those features on the final decision. The hybrid approach’s superior fidelity proves that combining both sources of information – the model’s focus (attention) and the features’ influence (gradients) provides a more complete and faithful explanation of its reasoning than either method alone. This confirms that gradients contribute critical information that is not captured by attention analysis alone. The core of true multimodal explainability is this ability to trace sentiment back to the interplay between modalities. Achieving this high fidelity without sacrificing conciseness also ensures that the explanations are both faithful and comprehensible.

5.2. Implications for Territorial Revitalization Project Management

The practical implications of this research for project managers are substantial. The proposed framework can transform how decision-makers interact with public feedback by moving beyond opaque sentiment scores toward transparent, evidence-based insights. Facilitates enhanced situational awareness and proactive risk mitigation. Project managers can move from reactive problem-solving to proactive management. If the framework consistently highlights negative sentiment linked to keywords like “delay” paired with images of stalled progress, managers receive a clear, early warning to investigate and adjust strategies before public discontent escalates.

The framework also provides a data-driven justification for decisions. Every decision in high-stakes environments requires justification. A manager can defend allocating more resources to a sub-project by presenting data showing it is a major source of positive public sentiment, with explanations pointing directly to specific words and images. This strengthens accountability and builds trust with stakeholders. Most importantly, it fosters trust in AI-assisted tools. Lack of trust is the primary barrier to AI adoption in critical domains. By demystifying the “black box,” the XAI framework presented here serves as a crucial trust-building mechanism. When managers can see why the AI concluded and verify its reasoning, they are far more likely to integrate the tool into their workflows.

5.3. Contributions to the Field of Explainable AI

This work contributes a validated methodology for the challenging subfield of multimodal XAI from a scientific standpoint. This demonstrates that simply extending unimodal XAI techniques is insufficient for models in which cross-modal fusion is the centerpiece. The hybrid approach of combining gradient-based attribution with an analysis of cross-modal attention provides a template for developing explanations for other sophisticated fusion models. This underscores the principle that an explanation method must be architecturally aware of the model it seeks to explain, particularly of the mechanisms that integrate disparate data streams.

5.4. Limitations of the Study

Despite the promising results, the limitations of this study should be acknowledged. The RevitalizeSent-MM dataset, while a crucial asset for this domain-specific task, is inherently limited in size and scope. This study is primarily focused on the Ukrainian post-2022 context, which may introduce cultural and linguistic biases that could affect the model's generalizability to other revitalization scenarios. Furthermore, the evaluation of explanation quality relied on quantitative proxy metrics, such as fidelity. While these metrics are valuable for assessing the model's internal logic's faithfulness, they do not directly measure the "goodness" or utility of an explanation from a human perspective. The ultimate test of an explanation's value lies in its comprehension and use by the target audience, which was not formally assessed in this study. Finally, the proposed approach was validated on a specific transformer-based architecture, and its applicability to other model families with different fusion mechanisms is yet to be explored. These limitations naturally lead to several avenues for future work, which will be discussed in the concluding section.

6. Conclusion

This study tackled the critical challenge of enhancing the trustworthiness and practical utility of advanced AI within the high-stakes domain of territorial revitalization project management. While powerful at processing diverse data streams, standard multimodal sentiment analysis models often function as opaque "black boxes," creating a significant barrier to adoption for project managers who depend on transparent and reliable insights for effective decision-making in complex social environments.

To overcome this limitation, a novel framework specifically designed for explainable multimodal sentiment analysis tailored to the revitalization context was introduced and rigorously evaluated. In doing so, the research successfully achieved its primary objective of creating a methodology that provides project managers with transparent and actionable insights. The proposed approach synergizes a high-performing transformer-based multimodal sentiment model with a custom-adapted XAI methodology, integrating BERT and ViT via cross-modal attention. This methodology fuses insights from both cross-modal attention analysis and gradient-based attributions (Integrated Gradients) to generate cohesive, human-understandable explanations that pinpoint key sentiment drivers across both textual narratives and visual evidence.

Comprehensive experiments, performed on the specially constructed RevitalizeSent-MM dataset reflecting real-world revitalization scenarios, empirically validated the proposed system's efficacy. The underlying multimodal model demonstrated superior sentiment prediction accuracy compared with unimodal baselines, confirming the value of the fused data. More importantly, the adapted XAI method exhibited significantly higher fidelity in identifying critical predictive features than naive or unimodal explanation techniques, demonstrating its ability to accurately represent the model's complex internal reasoning. These results confirm the successful completion of the core tasks set out in this study: developing a hybrid XAI technique, designing a user-centric explanation format, creating a domain-specific dataset, and rigorously evaluating the framework's performance and fidelity.

In conclusion, this study contributes a validated methodology for developing explainable multimodal AI systems specifically designed for deployment in sensitive, high-stakes domains. By effectively bridging the gap between sophisticated AI capabilities and project managers' practical requirements for transparent, actionable intelligence, the proposed framework offers a tangible pathway toward more responsible, trustworthy, and ultimately more effective AI-assisted decision-making in the vital efforts of territorial revitalization.

Recognizing the potential for further advancement, future work will prioritize several key directions. The first crucial step of human-centric evaluation is Although the quantitative metrics confirmed the fidelity of the explanations, future research must involve real-world deployment studies in collaboration with project teams. Such studies are necessary to assess the tangible impact of the proposed framework on decision-making processes and refine the practical utility of the user feedback-based explanations. Second, future work will focus on developing interactive explanation interfaces. Managers can probe model reasoning more deeply by adjusting input data to observe changes in predictions or by directly querying the relationships between highlighted text and image regions. Third, the framework should be extended to incorporate additional modalities, such as video and structured data (e.g., project timelines and budgets), to provide a more holistic analytical view. Finally, exploring the application of these explainability principles to other complex project management domains and different fusion architectures represents another promising avenue for continued research, ensuring that the benefits of transparent AI can be realized more broadly.

Contributions of authors: conceptualization, methodology, software, investigation, and writing – original draft preparation – **Serhii Dolhopolov**; data curation, formal analysis, and validation – **Yuliia Riabchun**, **Maksym Delembovskyi**; supervision, project administration, and conceptualization – **Oleksandr Molodid**. All authors contributed to writing – review and editing.

Conflict of Interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, author ship or otherwise, that could affect the research and its results presented in this paper.

Financing

The authors gratefully acknowledge the support provided by the Ministry of Education and Science of Ukraine. This research was conducted within the framework of the applied research project titled "Methodology for determining tonality and classification of multimodal content in territorial revitalization projects based on neural network methods" (State Registration Number: 0125U001683).

Data Availability

The manuscript contains no associated data.

Use of Artificial Intelligence

During the preparation of this work, the author(s) used Google Gemini Pro (version "Gemini 2.5 Pro Preview 03-25") in order to: Editing (improve grammar, clarity, and style). Further, the author(s) used Python code (assisted by Gemini Pro) to make the data visualization more informative. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

All the authors have read and agreed to the published version of this manuscript.

References

1. Yao, S., & Wang L. Difficulties and Challenges Faced by the Implementation of China's Rural Revitalization Strategy. *Frontiers in Sustainable Development,* 2023 vol. 3, no. 2, article no. 3768. DOI: 10.54691/fsd.v3i2.3768.

2. Guo, B., Yuan L., & Lu M. Analysis of Influencing Factors of Farmers' Homestead Revitalization Intention from the Perspective of Social Capital. *Land*, 2023, vol. 12, no. 4, article no. 812. DOI: 10.3390/land12040812.

3. Jin, X., Chin, T., Yu, J., Zhang, Y., & Shi, Y. How Government's Policy Implementation Methods Influence Urban Villager. Acceptance of Urban Revitalization Programs: Evidence from China. *Land*, 2020, vol. 9, no. 3, article no. 77. DOI: 10.3390/land9030077.

4. Chen, X., Xie, H., Tao, X., Wang, F. L., Leng, M., & Lei, B. Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics. *Artificial Intelligence Review*, 2024, vol. 57, article no. 91. DOI: 10.1007/s10462-024-10712-7.

5. Sulubacak, U., Caglayan, O., Gronroos, S., Rouhe, A., Elliott, D., Specia, L., & Tiedemann, J. Multimodal machine translation through visuals and speech. *Machine Translation*, 2020, vol. 34, pp. 97-147. DOI: 10.1007/s10590-020-09250-0.

6. Singh, K., Piyush, P., Kumar, R., Chhabra, S., Goomer, N., & Kashyap, A. Multimodal Data Extraction & Fusion for Health Monitoring System and Early Diagnosis. *2024 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, 2024, vol. 1, pp. 216-220. DOI: 10.1109/ICCICA60014.2024.10585027.

7. De Vito, S., Del Giudice, A., D'Elia, G., Esposito, E., Fattoruso, G., Ferlito, S., Formisano, F., Loffredo, G., Massera, E., D'Auria, P., & Di Francia, G. Future Low-Cost Urban Air Quality Monitoring Networks: Insights from the EU's AirHeritage Project. *Atmosphere*, 2024, vol. 15, no. 11, article no. 1351. DOI: 10.3390/atmos15111351.

8. Chernyshev, D., Ryzhakova, G., Honcharenko, T., Petrenko, H., Chupryna, I., Reznik, N. Digital Administration of the Project Based on the Concept of Smart Construction. *Explore Business, Technology Opportunities and Challenges After the Covid- 19 Pandemic .ICBT 2022. Lecture Notes in Networks and Systems*, 2022, vol. 495, pp. 1316–1331. DOI: 10.1007/978-3-031-08954-1_114.

9. Honcharenko, T., Mihaylenko, V., Borodavka, Y., Dolya, E., & Savenko, V. Information Tools for Project Management of the Building Territory at the Stage of Urban Planning. *CEUR Workshop Proceedings*, 2021, vol. 2851, pp. 22–33.

10. Matsiievskyi, O., Achkasov, I., Borodavka, Y., & Mazurenko, R. Behavioral model of autonomous robotic systems using reinforcement learning methods. *International Workshop on Information Technologies: Theoretical and Applied Problems*, 2024, vol. 3896, pp. 1–9. Available at: https://ceur-ws.org/Vol-3896/short14.pdf (accessed 12.05.2025).

11. Riabchun, Y., Honcharenko, T., Honta, V., Chupryna, K., & Fedusenko, O. Methods and means of evaluation and development for prospective students' spatial awareness. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2019, vol. 8, no. 11, pp. 4050-4058. DOI: 10.35940/ijitee.k1532.0981119.

12. Omas-as, R. M., & Encarnacion, R. E. Stakeholders' Satisfaction on Institutional Assessment: A Proposal for Unified Feedback Management System with Text Analytics and Sentiment Analysis. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, 2024, vol. 4, no. 1, pp. 50-56. DOI: 10.48175/ijarsct-18716.

13. von Eschenbach, W. J. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 2021, vol. 34, no. 4, pp. 1607–1622. DOI: 10.1007/s13347-021-00477-0.

14. Dolhopolov, S., Honcharenko, T., Savenko, V., Balina, O., Bezklubenko, I.S., & Liashchenko, T. Construction Site Modeling Objects Using Artificial Intelligence and BIM Technology: A Multi-Stage Approach. *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, 2023, pp. 174-179. DOI: 10.1109/SIST58284.2023.10223543.

15. Dolhopolov, S., Honcharenko, T., Terentyev, O., Predun, K., & Rosynskyi, A. Information system of multi-stage analysis of the building of object models on a construction site. *IOP Conference Series: Earth and Environmental Science*, 2023, vol. 1254, no. 1, article no. 012075. DOI: 10.1088/1755-1315/1254/1/012075.

16. Hazarika, D., Zimmermann, R., & Poria, S. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 2020, pp. 1363–1371. DOI: 10.1145/3394171.3413678.

17. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. Tensor Fusion Network for Multimodal Sentiment Analysis. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 1103–1114. DOI: 10.18653/v1/D17-1115.

18. Xu, W., Chen, J., Ding, Z., & Wang, J. Text sentiment analysis and classification based on bidirectional Gated Recurrent Units (GRUs) model. *Applied and Computational Engineering*, 2024, vol. 77, pp. 132–137. DOI: 10.54254/2755-2721/77/20240670.

19. Tang, Z. Review of Multimodal Sentiment Analysis Techniques. *Applied and Computational Engineering*, 2024, vol. 120, pp. 88–97. DOI: 10.54254/2755-2721/2025.18747.

20. Aldeen, M., MohajerAnsari, P., Ma, J., Chowdhury, M., Cheng, L., & Pesé, M. D. WIP: A First Look At Employing Large Multimodal Models Against Autonomous Vehicle Attacks. *Proceedings of the 2024 Symposium on Vehicle Security and Privacy (VehicleSec)*, 2024, pp. 1–7. DOI: 10.14722/vehiclesec.2024.23044.

21. Beserra, A. A., Kishi, R. M., & Goularte, R. Evaluating Early Fusion Operators at Mid-Level Feature Space. *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia '20)*, 2020, pp. 113–120. DOI: 10.1145/3428658.3431079.

22. Cheng, J., Feng, C., Xiao, Y., & Cao, Z. Late better than early: A decision-level information fusion approach for RGB-Thermal crowd counting with illumination awareness. *Neurocomputing*, 2024, vol. 594, article no. 127888. DOI: 10.1016/j.neucom.2024.127888.

23. Shaikh, M., Chai, D., Islam, S. M., & Akhtar, N. From CNNs to Transformers in Multimodal Human Action Recognition: A Survey. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2024, vol. 20, no. 8, article no. 260, pp. 1-24. DOI: 10.1145/3664815.

24. Lu, J., Batra, D., Parikh, D., & Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, article no. 2, pp. 13-23.

25. Tan, H. H., & Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5100-5111. DOI: 10.18653/v1/D19-1514.

26. Chen, Y., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. UNITER: UNiversal Image-TExt Representation Learning. *Computer Vision – ECCV 2020. Lecture Notes in Computer Science*, 2020, vol. 12375, pp. 104-120. DOI: 10.1007/978-3-030-58577-8_7.

27. Wang, P., Liu, S., & Chen, J. CCDA: A Novel Method to Explore the Cross-Correlation in Dual-Attention for Multimodal Sentiment Analysis. *Appl. Sci.*, 2024, vol. 14, 1934. DOI: 10.3390/app14051934.

28. Yu, W., Xu, H., Yuan, Z., & Wu, J. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 12, pp. 10798–10806. DOI: 10.1609/aaai.v35i12.17289.

29. Liu, Z., Braytee, A., Anaissi, A., Zhang, G., Qin, L., & Akram, J. Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion. *WWW '24: Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 1841–1848. DOI: 10.1145/3589335.3651971.

30. Keane, M. T., & Smyth, B. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). *Case-Based Reasoning Research and Development. ICCBR 2020. Lecture Notes in Computer Science*, 2020, vol. 12311, pp. 163–178. DOI: 10.1007/978-3-030-58342-2_11.

31. Pradhan, B., Dikshit, A., Lee, S., & Kim, H. An explainable AI (XAI) model for landslide susceptibility modeling. *Applied Soft Computing*, 2023, vol. 142, article no. 110324. DOI: 10.1016/j.asoc.2023.110324.

32. Jang, H., Kim, S., & Yoon, B. An Explainable AI (XAI) model for text-based patent novelty analysis. *Expert Systems with Applications*, 2023, vol. 231, article no. 120839. DOI: 10.1016/j.eswa.2023.120839.

33. Hasan, R., Dattana, V., Mahmood, S., & Hussain, S. Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights. *Information*, 2025, vol. 16, no. 1, article no. 7. DOI: 10.3390/info16010007.

34. Choi, S. R., & Lee, M. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review. *Biology*, 2023, vol. 12, no. 7, article no. 1033. DOI: 10.3390/biology12071033.

35. Gao, Y., Liu, J., Li, W., Hou, M., Li, Y., & Zhao, H. Augmented Grad-CAM++: Super-Resolution Saliency Maps for Visual Interpretation of Deep Neural Network. *Electronics*, 2023, vol. 12, no. 23, article no. 4846. DOI: 10.3390/electronics12234846.

36. Cheng, Z., Wu, Y., Li, Y., Cai, L., & Ihnaini, B. A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision. *Sensors*, 2025, vol. 25, article no. 4166. DOI: 10.3390/s25134166.

37. Waa, J. V., Nieuwburg, E., Cremers, A. H., & Neerincx, M. A. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence,* 2021, vol. 291, article no. 103404. DOI: 10.1016/j.artint.2020.103404.

38. Engel, A., Wang, Z., Frank, N.S., Dumitriu, I., Choudhury, S., Sarwate, A.D., & Chiang, T. Faithful and Efficient Explanations for Neural Networks via Neural Tangent Kernel Surrogate Models. *arXiv preprint,* 2023, pp. 1–54. DOI: 10.48550/arXiv.2305.14585.

39. Sun, Y., He, S., Han, X., & Luo, Y. Interpretability in Sentiment Analysis: A Self-Supervised Approach to Sentiment Cue Extraction. *Appl. Sci.*, 2024, vol. 14, 2737. DOI: 10.3390/app14072737.

40. Gipiškis, R., Tsai, C., & Kurasova, O. Explainable AI (XAI) in Image Segmentation in Medicine, Industry, and Beyond: A Survey. *arXiv preprint,* 2024, pp. 1-35. DOI: 10.48550/arXiv.2405.01636.

41. Mabokela, K. R., Primus, M., & Celik, T. Explainable Pre-Trained Language Models for Sentiment Analysis in Low-Resourced Languages. *Big Data Cogn. Comput.*, 2024, vol. 8, article no. 160. DOI: 10.3390/bdcc8110160.

42. Cerasuolo, F., Guarino, I., Spadari, V., Aceto, G., & Pescapé, A. XAI for Interpretable Multimodal Architectures with Contextual Input in Mobile Network Traffic Classification. *2024 IFIP Networking Conference (IFIP Networking),* 2024, pp. 757–762. DOI: 10.23919/IFIPNetworking62109.2024.10619769.

43. Kharchenko, V., Fesenko, H., & Illiashenko, O. Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. *Sensors,* 2022, vol. 22, 4865. DOI: 10.3390/s22134865.

Received 27.06.2025, Received in revised form 10.11.2025
Accepted date 17.11.2025, Published date 08.12.2025

ПОЯСНЕННИЙ ШТУЧНИЙ ІНТЕЛЕКТ ДЛЯ МУЛЬТИМОДАЛЬНОГО СЕНТИМЕНТ-АНАЛІЗУ В УПРАВЛІННІ ПРОЄКТАМИ РЕВІТАЛІЗАЦІЇ

С. Ю. Долгополов, Ю. В. Рябчун,
М. М. Делембовський, О. С. Молодід

Предметом статті є розробка та оцінка фреймворку поясненного штучного інтелекту (XAI) для мультимодального сентимент-аналізу, що застосовується спеціально для управління проєктами територіальної ревіталізації. Дослідження розв'язує критичну проблему моделей штучного інтелекту, що функціонують за принципом «чорної скриньки», чия непрозорість перешкоджає їхньому впровадженню менеджерами проєктів, які потребують достовірної інформації для ухвалення важливих рішень у складних соціальних умовах. Метою є запропонувати та ретельно перевірити новий фреймворк для поясненного мультимодального сентимент-аналізу, адаптований для надання прозорих, достовірних та дієвих інсайтів для ухвалення рішень в управлінні проєктами територіальної ревіталізації. Завдання, які необхідно вирішити, включають розробку гібридної техніки XAI, що поєднує дані з крос-модальної уваги та градієнтної атрибуції; проєктування цілісного, орієнтованого на користувача формату пояснень, що комбінує підсвічений текст та теплові карти зображень; створення спеціалізованого набору даних RevitalizeSent-MM для цієї конкретної галузі; а також емпіричну оцінку прогностичної точності фреймворку та, що найважливіше, точності його пояснень. Використані методи включають трансформерну модель мультимодального сентимент-аналізу (MSA) на базі BERT і ViT з крос-модальною увагою для злиття інформації. Компонент поясненності – це гібридна техніка XAI, що інтегрує аналіз крос-модальної уваги з методом інтегрованих градієнтів для присвоєння ваг важливості вхідним ознакам.

Оцінка проводилася з використанням стандартних метрик класифікації для визначення продуктивності та метрики «Падіння точності при збуренні» для оцінки точності пояснень. Результати підтвердили ефективність запропонованого фреймворку. Мультимодальна модель продемонструвала вищу точність порівняно з унімодальними базовими моделями, а запропонований метод XAI досяг значно вищої точності, ніж наївні підходи до пояснення, що доводить його здатність точно відображати внутрішню логіку моделі. Наукова новизна полягає у трьох аспектах: розробка об’єднаної гібридної техніки XAI спеціально для трансформерних мультимодальних моделей; створення унікального, предметно-орієнтованого набору даних для аналізу ревіталізації; та валідація методології адаптації передового XAI для розв’язання критичних бар’єрів довіри та впровадження, що підтверджує його практичну значущість в управлінні проєктами.

Ключові слова: поясненний штучний інтелект; мультимодальний аналіз настроїв; управління проєктами; глибоке навчання; ревіталізація.

Долгополов Сергій Юрійович – асп. каф. Інформаційних технологій, Київський національний університет будівництва і архітектури, Київ, Україна.

Рябчун Юлія Володимирівна – д-р філос., доц. каф. Інформаційних технологій, Київський національний університет будівництва і архітектури, Київ, Україна.

Делембовський Максим Михайлович – канд. техн. наук, доц., доц. каф. Кібербезпеки та комп’ютерної інженерії, Київський національний університет будівництва і архітектури, Київ, Україна.

Молодід Олександр Станіславович – д-р техн. наук, проф., проф. каф. будівельних технологій, Київський національний університет будівництва і архітектури, Київ, Україна.

Serhii Dolhopolov – PhD Student of the Department of Information Technologies, Kyiv National University of Construction and Architecture, Kyiv, Ukraine,
e-mail: dolhopolov_sy@knuba.edu.ua, ORCID: 0000-0001-9418-0943, Scopus Author ID: 57994271400.

Yuliia Riabchun – PhD, Associate Professor at the Department of Information Technologies, Kyiv National University of Construction and Architecture, Kyiv, Ukraine,
e-mail: super.etsy@ukr.net, ORCID: 0000-0002-8320-4038, Scopus Author ID: 57211403226

Maksym Delembovskyi – Candidate of Technical Sciences, Associate Professor, Associate Professor at the Department of Cybersecurity and Computer Engineering, Kyiv National University of Construction and Architecture, Kyiv, Ukraine,
e-mail: delembovskyi.mm@knuba.edu.ua, ORCID: 0000-0002-6543-0701, Scopus Author ID: 57222122288.

Oleksandr Molodid – Doctor of Technical Sciences, Professor, Professor at the Department of Construction Technologies, Kyiv National University of Civil Engineering and Architecture, Kyiv, Ukraine,
e-mail: molodid.os@knuba.edu.ua, ORCID: 0000-0001-8781-6579, Scopus Author ID: 57219054089.