Multimodal sentiment analysis with Grad-CAM for urban revitalization*

Serhii Dolhopolov[1,†] and Tetyana Honcharenko[1,*,†]

[1] *Kyiv National University of Construction and Architecture, 31, Air Force Avenue, Kyiv, 03037, Ukraine*

Abstract

The application of Artificial Intelligence (AI) in urban planning offers unprecedented opportunities for assessing public sentiment towards revitalization projects by analyzing the vast stream of multimodal data generated by citizens. However, the increasing complexity of these models, particularly those that fuse text, image, and geospatial data, often results in "black box" systems. This lack of transparency hinders their adoption by non-expert stakeholders like urban planners, who require trustworthy and interpretable insights to make informed decisions. To address this critical gap, this paper introduces a novel framework centered on an XAI Orchestrator. This system couples a high-performance multimodal sentiment analysis model with a suite of modality-specific explainability techniques, namely Gradient-weighted Class Activation Mapping (Grad-CAM) for visual saliency and SHapley Additive exPlanations (SHAP) for textual attribution. These disjointed technical outputs are then synthesized by a Large Language Model (LLM) composer into a cohesive, human-readable narrative, complete with a visual attribution map. Our quantitative evaluation demonstrates that the underlying predictive model achieves high classification accuracy, with a weighted F1-score of 0.9, validating the efficacy of the multimodal fusion approach. Qualitative case studies further reveal the framework's ability to generate clear and intuitive explanations, successfully deconstructing the model's predictions and grounding them in specific visual and textual evidence. By bridging the gap between prediction and interpretation, this work presents a viable methodology for deploying trustworthy AI systems in civic tech, fostering more transparent, accountable, and human-centric urban planning.

Keywords

Explainable AI (XAI), Multimodal Sentiment Analysis, Urban Planning, Grad-CAM, SHAP

1. Introduction

The revitalization of urban territories stands as a paramount challenge in contemporary global development, profoundly shaping the socio-economic vitality, environmental sustainability, and overall quality of life for an increasingly urbanized global population. For decades, the evaluation of such large-scale initiatives was predominantly governed by a techno-economic rationale, where success was quantified through tangible, lagging indicators: metrics of economic growth, infrastructure completion timelines, and return on investment. While valuable, this paradigm often neglected the most critical component of any urban ecosystem: its inhabitants. A modern, human-centric approach to city planning has since emerged, championing the principle that the true efficacy of urban transformation is not merely reflected in steel and concrete but is best measured through the public's perception, sentiment, and lived experience. In this evolving landscape of participatory governance and smart city development, the ability to accurately, dynamically, and ethically assess citizen sentiment is no longer a supplementary objective but a core requirement. Transforming the subjective, often unstructured feedback of a diverse populace into actionable, data-driven intelligence

*ITTAP'2025: 5th International Workshop on Information Technologies: Theoretical and Applied Problems, October 22-24, 2025, Ternopil, Ukraine, Opole, Poland

[1*] Corresponding author.

[†] These authors contributed equally.

dolhopolov@icloud.com (S. Dolhopolov); iust511@ukr.net (T. Honcharenko)

0000-0001-9418-0943 (S. Dolhopolov); 0000-0003-2577-6916 (T. Honcharenko)

is fundamental to fostering urban environments that are not only functional and efficient but also inclusive, equitable, and responsive to the genuine needs of their communities.

The traditional mechanisms for gauging public opinion, such as formally structured municipal surveys, town hall meetings, and public hearings, have long served as the bedrock of civic engagement. These methods, however, are increasingly misaligned with the pace and complexity of modern urban dynamics. They are inherently characterized by significant operational latency, often taking months to plan, execute, and analyze, by which time public sentiment may have already shifted. Furthermore, their high operational costs and reliance on limited sample sizes introduce significant potential for demographic, selection, and response biases, frequently failing to capture the nuanced, granular, and real-time pulse of community sentiment. Consequently, decisions based on this data, while well-intentioned, may not fully represent the multifaceted and often conflicting views present within a city.

The digital transformation of civic life has precipitated a paradigm shift, unlocking an alternative and profoundly richer source of public insight: a vast, unstructured, and continuous multimodal stream of data organically generated by citizens. This digital torrent comprises geo-tagged social media posts, intricate textual comments, and a formidable volume of digital imagery and video. This wealth of information presents an unprecedented opportunity to monitor public opinion with a level of granularity, scope, and immediacy previously unattainable. The imperative to transform this raw digital exhaust into structured, actionable intelligence is a challenge that extends far beyond urban planning. Parallel research efforts in critical domains such as emergency and disaster management have pioneered methods for enhancing situational awareness by contextually enriching dense mobility data from mobile phones with the sparse, yet semantically rich, content from geo-social media platforms [1]. Such work demonstrates a growing recognition that fusing disparate data sources is key to understanding complex human-environment interactions, whether in the context of a natural disaster [2–3] or the daily life of a city.

However, the effective utilization of this data demands the development of sophisticated analytical techniques capable of navigating its inherent complexities. The data is not only massive in volume but is also fundamentally multimodal, with each modality presenting unique challenges and opportunities. Textual data from social media is often short, informal, and rife with colloquialisms, sarcasm, and sentiment-laden emojis that defy traditional natural language processing. Visual data, in the form of images and video, provides powerful, direct evidence of the urban condition – a photograph of urban decay, a vibrant community event, or a dilapidated playground can convey a more potent and immediate sentiment than a textual description alone. The third critical modality, geospatial information, grounds this digital feedback in the physical world. Yet, as research by Honcharenko et al. highlights, the formal representation of spatial and topological relations through multidimensional data models remains a non-trivial technical challenge [4]. Successfully interpreting this data requires more than just technical proficiency; it necessitates an approach that can augment the inherent spatial awareness and cognitive decision-making capabilities of human stakeholders, a complex process explored by Riabchun et al. [5].

In response to this data-rich environment, the field of Multimodal Sentiment Analysis (MSA) has advanced significantly. Initial research efforts, which were predominantly focused on unimodal, text-centric analyses, have given way to more sophisticated architectures designed to integrate and interpret heterogeneous data streams. Foundational work in this area has explored powerful fusion strategies, moving from early-stage feature concatenation to more advanced techniques. These include the development of dedicated representation learning frameworks like MISA, which learns both modality-invariant and modality-specific features [6], and complex fusion mechanisms like the Tensor Fusion Network, which explicitly models the intricate inter-modality dynamics between language, visual, and acoustic signals [7]. Further research has advanced this frontier by exploring multi-task learning frameworks to disentangle these complex signal characteristics [8] and, more recently, by leveraging the immense capabilities of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) to handle text-centric multimodal tasks [9] and incorporate external, contextual world knowledge to resolve ambiguities in visual scenes [10].

Despite these remarkable technical achievements, a persistent and critical challenge shadows the field: the inherent opacity of the models themselves. Even as these systems achieve state-of-the-art performance, they often function as inscrutable "black boxes." This issue is further compounded by systemic vulnerabilities within the data itself; research by Yang et al. has shown that even the most robust models can be misled by spurious correlations stemming from latent label and context biases within datasets, a problem they propose to mitigate using a causal inference framework [11]. The "black box" nature of these models is not a mere academic curiosity; it represents a fundamental barrier to their real-world adoption. The urgency of this problem is magnified by the rapid and widespread proliferation of complex AI into disparate, high-stakes domains, with critical applications ranging from automated software modernization [12] and the behavioral modeling of autonomous robotic systems [13] to the multi-stage classification of construction site elements [14] and the modeling of complex environmental systems using ensemble machine learning [15]. In all such applications, predictions are delivered without transparent justification, severely limiting their utility and struggling to earn the trust of the domain experts and non-expert stakeholders who are ultimately responsible for making decisions.

This fundamental trade-off between predictive accuracy and model transparency has catalyzed the growth of a dedicated and vital research discipline: Explainable Artificial Intelligence (XAI), which seeks to develop theories and methods for interpreting the decisions of complex models, as comprehensively surveyed by Linardatos et al. [16]. The resulting "explainability gap" is arguably the single most significant barrier to the responsible and effective deployment of AI in civic planning and governance. This sentiment is echoed in technical research focused on specific explainability techniques; for instance, Sharma and Kumar emphasize that methods like Grad-CAM are essential for visualizing model activations and building a foundational understanding of model behavior in classification tasks [17]. Furthermore, the work of Dhore et al. demonstrates that individual XAI techniques often have their own limitations, such as noisy outputs from Grad-CAM, and that hybrid approaches combining the strengths of multiple explainers can yield clearer and more reliable insights [18]. Ultimately, if an urban planner, a policy-maker, or a community representative cannot understand why an AI model has flagged a specific neighborhood as exhibiting a spike in negative sentiment, the prediction, no matter how accurate, remains an unactionable and untrustworthy piece of data.

To address the critical challenges of model opacity and stakeholder trust, this paper introduces a novel, integrated system for multimodal sentiment analysis explicitly designed for the domain of urban revitalization. Our approach moves beyond the singular pursuit of predictive accuracy and instead prioritizes the generation of transparent, human-intelligible, and actionable explanations. The proposed framework, termed the "XAI Orchestrator," is architected to ingest, process, and interpret data from three distinct yet complementary modalities: the textual content of social media posts, the visual information contained in associated images, and the critical context provided by geospatial data. The inclusion of the geospatial modality is particularly crucial, as it grounds abstract digital feedback in the tangible reality of the physical urban environment. However, as noted, this integration is not without its own ethical complexities. The demonstrated ability of modern Large Multimodal Models to infer precise geographic locations from user-posted imagery with startling accuracy raises significant geo-privacy concerns that must be proactively addressed [19]. Our framework's emphasis on transparency is therefore not merely a technical feature but a core tenet for the ethical and trustworthy deployment of such technologies in a civic context.

The central innovation of our work is the concept of an "orchestrator" that intelligently synergizes the outputs of multiple, state-of-the-art explainability techniques rather than relying on a single method. By doing so, we aim to construct a more holistic and robust explanatory narrative that is greater than the sum of its parts. For the visual modality, our system employs Gradient-weighted Class Activation Mapping (Grad-CAM), a canonical technique for producing visual heatmaps that highlight the specific regions within an image that were most salient to the model's classification decision. For the textual modality, we utilize SHapley Additive exPlanations (SHAP), a game-

theoretic approach that attributes a prediction to individual words or tokens, quantifying the precise contribution of each to the final sentiment score.

These specific techniques are prime examples of gradient-based feature attribution methods, a class of explainers that has become central to the scientific endeavor of interpreting neural networks, as detailed in the technical review by Wang et al. [20]. Our framework deliberately employs these methods within their established classification context. However, we acknowledge that the applicability and boundaries of these tools are an active and evolving area of research. For example, recent work by Bachhawat has explored the novel challenge of generalizing Grad-CAM to embedding networks, which do not produce the discrete class scores on which the method traditionally relies [21]. This ongoing innovation underscores the dynamism of the XAI field. Our approach is thus conceptually similar to integrated spatio-temporal topic-sentiment models proposed in related GeoAI domains, which seek to overcome the limitations of sequential analysis by creating a unified output [3]. The ultimate goal of our XAI Orchestrator is to translate the complex, high-dimensional outputs of both the predictive model and its explainers into clear, composite insights that augment the inherent spatial awareness and cognitive decision-making capabilities of human stakeholders [5].

The overarching goal of this research is to bridge the critical gap between the high predictive power of modern multimodal AI systems and the pressing need for their transparent and trustworthy application in urban planning. We contend that for AI to become a truly collaborative tool in civic governance, its outputs must be scrutable, its reasoning legible, and its potential biases identifiable by the human experts who bear the ultimate responsibility for urban outcomes. Our work is motivated by the hypothesis that a multi-explainer, multimodal framework can provide a more comprehensive, reliable, and actionable form of insight than any unimodal or single-explainer system alone.

To that end, the primary contributions of this paper are threefold and are designed to address specific gaps in the current body of research:

- We first propose and implement a robust deep learning architecture specifically designed to fuse textual, visual, and geographic features for superior sentiment classification in the urban domain. This model serves as the high-performance predictive engine for which our explainability framework provides interpretive oversight.
- Our core contribution is the introduction of the XAI Orchestrator, a system that synergizes multiple, modality-specific explainability techniques (Grad-CAM for visual saliency, SHAP for textual attribution) and integrates them with a Large Language Model (LLM)-based composer. This orchestrator is designed to generate cohesive, multimodal explanations that present a unified narrative explaining a given prediction.
- We demonstrate the efficacy of the complete system through a series of experiments on a simulated but realistic dataset reflecting the challenges of urban sentiment analysis. The validation moves beyond standard accuracy metrics to assess the quality and utility of the generated explanations themselves, providing tangible examples of visual heatmaps, textual attributions, and composed natural language summaries that justify the model's conclusions.

Through this work, we aim to present a viable and extensible methodology for building and deploying explainable AI systems in the complex, high-stakes environment of urban revitalization. By prioritizing interpretability, we seek to foster greater trust, facilitate more effective human-AI collaboration, and ultimately contribute to the development of more equitable and responsive cities.

2. Methodology

The efficacy of any data-driven system for urban analysis is fundamentally predicated on the quality, diversity, and integrity of its underlying data. To this end, our proposed framework is built upon a robust and scalable data acquisition and preprocessing pipeline, engineered to systematically ingest, cleanse, and structure heterogeneous data streams into an analysis-ready format. This foundational

layer, depicted in the preprocessing pipeline diagram (Figure 1), ensures that the subsequent sentiment analysis and explainability models are supplied with data that is not only rich in contextual information but also compliant with stringent ethical and privacy standards.

2.1. Data Acquisition and Preprocessing Pipeline

The pipeline is designed to address the inherent challenges of multimodal urban data, which originates from a variety of disparate and often unstructured sources, as depicted in Figure 1. The initial stage of our methodology focuses on the acquisition of this diverse data. We identify four primary sources of information. First, official municipal data provides a structured, authoritative baseline, including administrative boundaries, infrastructure logs, and demographic statistics. Second, public surveys offer a source of explicitly solicited citizen feedback, providing deep qualitative insights, albeit on a limited scale. The third and most voluminous source is geotagged social media posts (Geo-posts), an unsolicited stream of real-time public expression from platforms such as X, Instagram, Telegram, and Facebook. Finally, direct image and video uploads from citizens or dedicated inspection platforms offer high-fidelity visual evidence of the urban environment.

Given the disparate formats and velocities of these data streams, a centralized Ingestion Bus serves as the unified entry point into our system. This layer, implemented using a distributed event streaming platform such as Apache Kafka, is responsible for collecting and queuing incoming data from various APIs, scrapers, and direct uploads. This architecture decouples the data sources from the processing logic, ensuring fault tolerance and scalability. Upon ingestion, all raw, unmodified data is persisted in a Raw Data Lake. This repository acts as an immutable, chronological source of truth, allowing for data lineage tracking and reprocessing should the downstream analytical requirements evolve.

Once ingested, the raw data is channeled into a multi-stage parallel processing pipeline where each modality undergoes specialized preparation. For textual content, a Text NLP module performs a series of linguistic normalization procedures. This includes cleaning raw text by removing HTML tags and special characters, tokenization, stop-word removal, and lemmatization to reduce inflectional variations. The primary objective of this stage is to convert unstructured textual data into a structured format suitable for ingestion by advanced language models.

In parallel, the Image Analysis module handles the initial processing of visual data. This stage involves standard computer vision preprocessing steps, including image resizing to a uniform resolution required by our neural network encoders, pixel value normalization, and data augmentation where necessary to improve model robustness. This ensures that all visual inputs are standardized before feature extraction.

A critical component of the pipeline is the Geo Normalization module. Geospatial information arrives in various forms, including precise latitude-longitude coordinates, place names (e.g., “City Hall”), or the bounding boxes of administrative districts. This module is responsible for resolving these different formats into a unified, standardized coordinate system (e.g., WGS 84). It performs geocoding for place names and calculates centroids for polygons, ensuring that every data point has a consistent and machine-readable spatial representation.

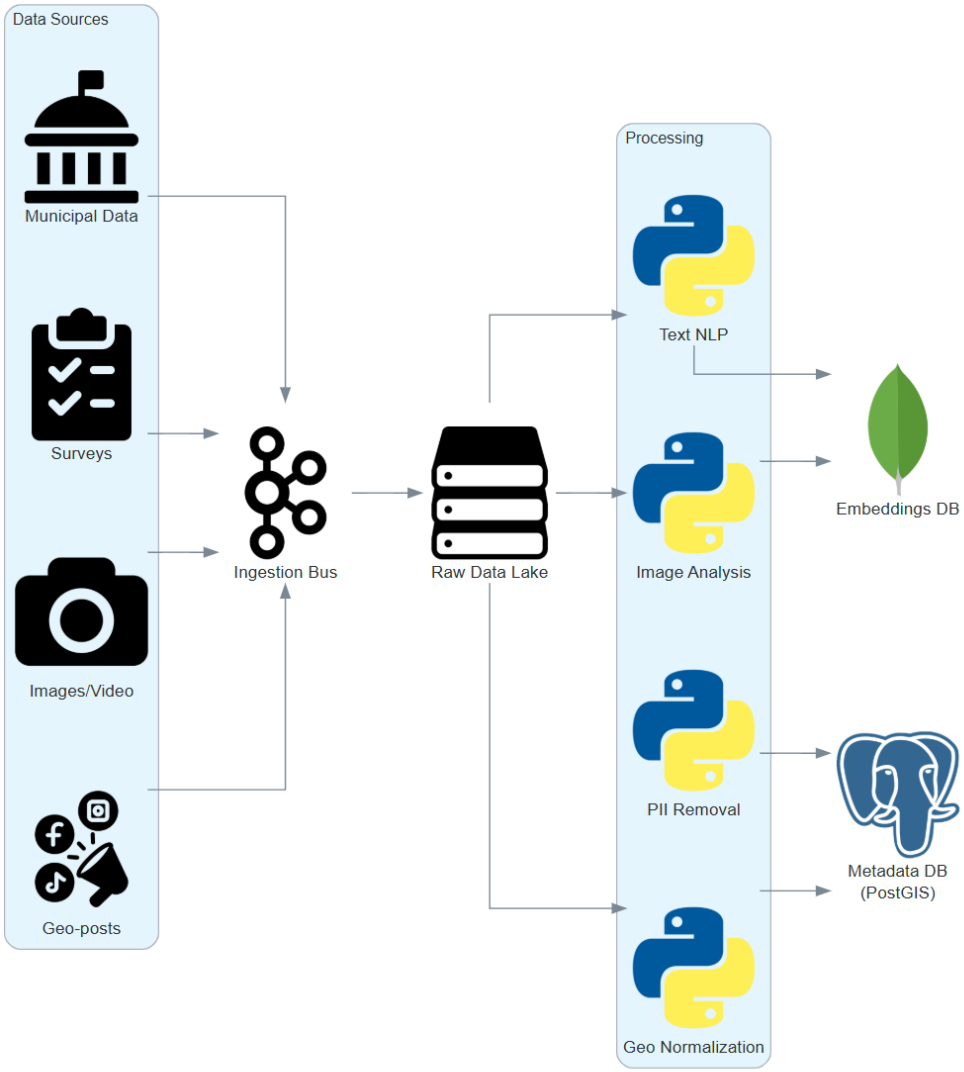

Figure 1: An architectural overview of the data acquisition and preprocessing pipeline (Diagram by the authors).

Crucially, an ethical and privacy-preserving layer for PII Removal is integrated into the pipeline. This dedicated anonymization module automatically identifies and redacts Personally Identifiable Information from all data modalities. It employs a combination of named entity recognition (NER) for text and object detection or blurring techniques for images to remove sensitive information such as names, faces, addresses, and license plates, ensuring compliance with data protection regulations like GDPR.

The culmination of this pipeline is the structured storage of the processed and anonymized data into two distinct, optimized databases. All structured metadata – including the cleaned text, normalized geospatial coordinates, timestamps, and pointers to the processed image files – is stored in a Metadata DB. This database is implemented using PostgreSQL with the PostGIS extension, enabling efficient and powerful geospatial querying. Concurrently, the high-dimensional vector representations (embeddings) generated by the downstream deep learning models are stored in a dedicated Embeddings DB, optimized for high-speed similarity search. This dual-database architecture separates structured metadata from high-dimensional feature vectors, ensuring optimal performance for both relational queries and machine learning operations.

2.2. The Multimodal Sentiment Analysis Model

Following the initial data preprocessing, the core of our framework's predictive capability resides in the Multimodal Sentiment Analysis Model. This sophisticated engine is architected to transform the prepared, unimodal data streams into high-dimensional, semantically rich representations, fuse them into a unified vector, and ultimately render a sentiment classification. The complete architecture of this model, from its modality-specific encoders to the final classification head, is illustrated in Figure 2. The model is logically partitioned into two primary stages: a set of parallel Multimodal Encoders responsible for feature extraction and a sequential Feature Fusion Architecture that integrates these features for prediction.

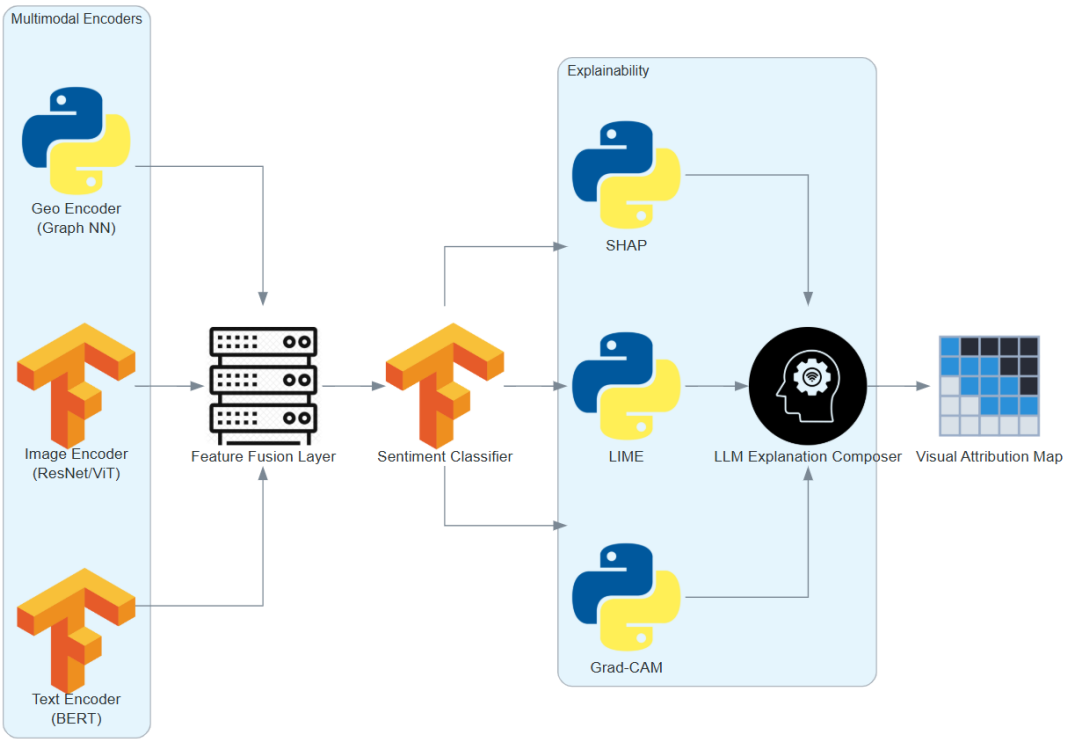

Figure 2: The architecture of the Multimodal Sentiment Analysis Model and the XAI Orchestrator (Diagram by the authors).

2.2.1. Multimodal Encoders

The initial task of the model is to independently process each of the three data modalities – text, image, and geospatial data – by projecting them from their raw, heterogeneous formats into a shared, high-dimensional latent space. This process, known as embedding, is performed by three specialized neural network encoders, each tailored to the unique characteristics of its corresponding data type.

The Text Encoder is responsible for converting the preprocessed textual data into dense vector representations. For this task, we employ a Transformer-based architecture, specifically a pre-trained model from the BERT (Bidirectional Encoder Representations from Transformers) family. BERT is exceptionally well-suited for this purpose due to its ability to capture deep, bidirectional contextual relationships between words in a sentence. By leveraging its pre-training on vast linguistic corpora, the model can generate a fixed-length embedding for each input text that encapsulates not only its lexical content but also its nuanced semantic and syntactic structure.

Concurrently, the Image Encoder processes the visual data. This component is built upon a state-of-the-art computer vision model, such as a deep Convolutional Neural Network (CNN) like ResNet or a Vision Transformer (ViT). These architectures are designed to extract a rich hierarchy of visual features, ranging from low-level patterns like edges and textures to high-level semantic concepts

such as objects, scenes, and their spatial arrangements. The output of the Image Encoder is a dense feature vector that serves as a quantitative summary of the visual content pertinent to urban sentiment.

The most novel of the encoders is the Geo Encoder, which is tasked with interpreting the normalized geospatial information. To move beyond the limitations of simple coordinate-based features, we model the urban environment as a complex graph, where locations of interest (e.g., specific intersections, public squares, or building sites) are represented as nodes and their spatial relationships (e.g., adjacency, street network connectivity, administrative containment) are represented as edges. We employ a Graph Neural Network (GNN) to process this structured data. A GNN is uniquely capable of learning from the relational topology of the graph, allowing it to generate a context-aware embedding for each location that reflects its position and significance within the broader urban fabric. This approach enables the model to capture spatial context that would be lost in a simple coordinate representation.

2.2.2. Feature Fusion and Classification

Following the generation of these unimodal representations, the next critical step is to integrate them into a single, holistic feature vector that captures the combined sentiment signal. This is the primary function of the Feature Fusion Layer. This layer receives the fixed-length embedding vectors from the Text, Image, and Geo Encoders and synergizes them. While simple fusion techniques such as vector concatenation provide a baseline, our framework employs a more sophisticated cross-attention mechanism. This allows the representations of different modalities to interact and "attend" to one another, enabling the model to dynamically weight the importance of each modality in the context of the others. For instance, the model can learn to amplify the importance of a negative textual sentiment when the accompanying image visually confirms the presence of urban decay.

The conceptual goal of this fusion process is to project the combined multimodal data into a well-structured latent space where semantically similar data points cluster together. As illustrated in Figure 3, this process maps individual multimodal inputs – each comprising an image, text, and geo-location – to a specific point in an abstract vector space. The model is trained such that inputs corresponding to the same sentiment class (e.g., "Public Complaint," "Infrastructure Failure," or "Community Appreciation") form dense, separable clusters within this space. This conceptualization of projecting data into modality-invariant or shared subspaces is a powerful paradigm in multimodal learning.

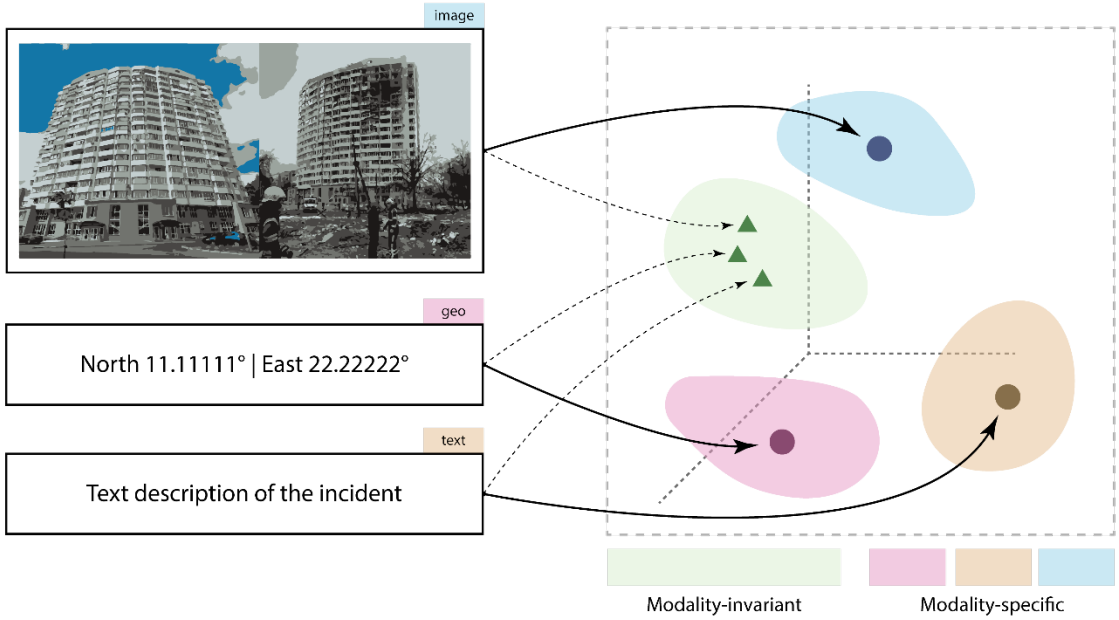

Figure 3: Conceptual illustration of the multimodal latent space (Diagram by the authors, adapted from the MISA framework by Hazarika, Zimmermann, and Poria [6], https://doi.org/10.1145/3394171.3413678. Photo of the building – author's adaptation of a photo by the National Police of Ukraine).

The culmination of the predictive model is the Sentiment Classifier. This final component, typically a multi-layer perceptron (MLP) with one or more hidden layers, takes the unified, fused feature vector as its input. Its function is to perform the final classification task by mapping this rich, multimodal representation to a probability distribution across the predefined sentiment categories (e.g., Positive, Neutral, Negative). The output of this classifier – the predicted sentiment label and its associated confidence score – serves as the primary input for the XAI Orchestrator, which is detailed in the subsequent section.

2.3. The XAI Orchestrator

While the Multimodal Sentiment Analysis Model provides the core predictive capability of our framework, its function as an opaque "black box" presents a significant barrier to its adoption in the trust-dependent domain of urban planning. To surmount this obstacle, we introduce the most novel component of our methodology: the XAI Orchestrator. This module, whose architecture is illustrated in the "Explainability" section of Figure 2, is engineered to deconstruct the model's complex decision-making process into a set of transparent, human-intelligible components. Its primary function is not to make predictions, but to explain them, thereby transforming the model from a mere analytical tool into a collaborative partner for human experts. The orchestrator operates in two distinct stages: first, it employs a suite of modality-specific explainers to generate attribution scores, and second, it uses a Large Language Model to synthesize these scores into a cohesive, multimodal narrative.

2.3.1. Modality-Specific Explainers

The initial step in generating an explanation is to attribute the model's final prediction back to the input features of each modality. To accomplish this, the XAI Orchestrator leverages three distinct, state-of-the-art explainability techniques, each selected for its suitability to a specific data type.

For generating visual explanations, we employ Gradient-weighted Class Activation Mapping (Grad-CAM). This technique is applied directly to the final convolutional layer of the Image Encoder (e.g., ResNet). Grad-CAM utilizes the gradients of the predicted sentiment class as they flow back into this layer, effectively weighting the importance of each activation map. The result is a class-discriminative localization map, or "heatmap," which highlights the specific pixels and regions within the input image that were most influential in the model's decision. This provides a direct and intuitive visual answer to the question: "Where in the image was the model looking when it made this prediction?"

For textual and geospatial explanations, we utilize methods that can provide feature-level attribution scores for structured and sequential data. Our framework incorporates both SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP, a method grounded in cooperative game theory, calculates the marginal contribution of each feature (e.g., a specific word in the text or a particular attribute from the Geo Encoder, such as proximity to a landmark) to the final prediction. It provides a theoretically robust and consistent measure of feature importance. LIME, a model-agnostic technique, operates by creating a local, interpretable surrogate model (such as a linear model) in the vicinity of the prediction being explained. It identifies the key features that are most influential within that local decision boundary. By applying these methods, we can precisely quantify the impact of individual words (e.g., "dilapidated," "vibrant") or geospatial characteristics on the sentiment score, answering the question: "Which specific textual or spatial features drove this prediction?".

2.3.2. LLM-based Explanation Composer

The raw outputs from these modality-specific explainers – a visual heatmap from Grad-CAM and numerical importance scores from SHAP/LIME – are powerful but disjointed. To be truly actionable for a non-technical stakeholder, they must be synthesized into a single, coherent narrative. This is the role of the LLM-based Explanation Composer.

This central component of the orchestrator first programmatically converts the structured outputs of the explainers into a detailed natural language prompt. For instance, the prompt might encode information such as: "The model predicted 'Negative' sentiment with 92% confidence. Grad-CAM analysis indicates a strong focus on the lower-left quadrant of the image, corresponding to a pile of rubble. SHAP analysis of the text identified the words 'neglected' and 'garbage' as having the highest negative contributions."

This rich, context-laden prompt is then fed into a pre-trained Large Language Model (e.g., Llama 3 or a model from the GPT series). The LLM is tasked with synthesizing this structured information into a fluid, easy-to-understand narrative. It is specifically instructed to connect the findings from different modalities, creating a story that explains the prediction. For example, it might generate a summary like: "The system has identified a strongly negative sentiment for this report. The visual evidence, which focuses on an area of accumulated waste and debris, corroborates the textual description, where the words 'neglected' and 'garbage' were the primary drivers of the negative classification."

The final output of the XAI Orchestrator is a two-part, composite explanation. The first part is a Visual Attribution Map, which is the original input image overlaid with the Grad-CAM heatmap to visually ground the explanation. The second part is the textual summary generated by the LLM, providing the narrative context. Together, these components deliver a holistic, multimodal, and interpretable justification for the AI's decision, designed to be directly usable by urban planners and citizens.

2.4. System Architecture and Implementation

To translate our methodological framework into a functional and scalable application, we have designed a robust, multi-tiered system architecture. This architecture, illustrated in Figure 4, delineates the end-to-end flow of information from the end-user to the core analytical services and back. It is engineered to support interactive, real-time analysis for stakeholders such as urban planners and citizens, ensuring that the insights generated by our models are both accessible and actionable. The system is comprised of four primary layers: the User Interface (UI), the API Layer, the Backend Service, and the Results Database.

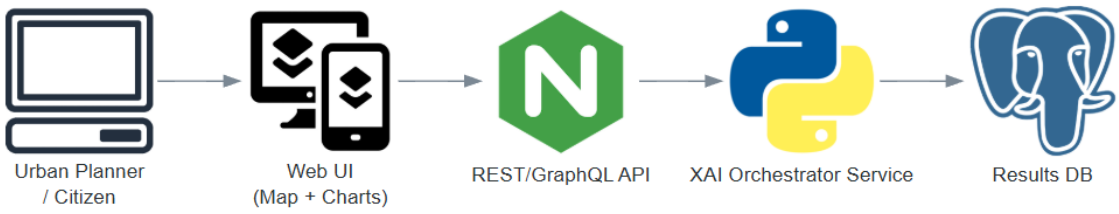

Figure 4: The end-to-end system architecture, illustrating the UI & API Layer (Diagram by the authors).

The primary point of interaction for end-users is the User Interface (UI). This layer is implemented as a responsive web application, accessible on both desktop and mobile devices, to cater to the distinct needs of its target audiences. For Urban Planners, the interface provides a comprehensive dashboard featuring interactive geospatial visualizations, such as heatmaps of sentiment distribution

overlaid on a city map, alongside temporal charts and detailed drill-down capabilities for individual data points. For Citizens, a simplified interface allows for the submission of reports and the viewing of aggregated, anonymized sentiment trends within their local communities. This dual-purpose design ensures that the system serves as both a professional analytical tool and a platform for civic engagement.

Mediating the communication between the user-facing application and the backend processing engine is the API Layer. This layer is implemented using a high-performance web server, such as Nginx, and exposes a well-defined REST/GraphQL API. The API serves as a formal contract, handling all incoming requests from the UI – for example, requests to submit new data, query for historical sentiment in a specific area, or retrieve the explanation for a particular prediction. Utilizing a standardized API decouples the frontend from the backend, allowing each to be developed, scaled, and maintained independently. This architectural choice is critical for ensuring system modularity and long-term maintainability.

The core computational logic of the system resides within the Backend Service, which we have implemented as the XAI Orchestrator Service. This service, developed in Python, encapsulates the entire analytical workflow detailed in the preceding sections. It houses the pre-trained multimodal sentiment model and the full suite of explainability modules (Grad-CAM, SHAP, LIME, and the LLM Composer). Upon receiving a request from the API layer, this service orchestrates the execution of the necessary models, generates both the sentiment prediction and its corresponding multimodal explanation, and packages the results for transmission back to the user. This service-based architecture allows for the complex machine learning workloads to be isolated on dedicated hardware, including GPUs for model inference, ensuring that the system can handle a high volume of requests efficiently.

Finally, the persistence and retrieval of all generated data are managed by the Results Database (DB). This layer is implemented using a robust relational database management system, such as PostgreSQL, which is well-suited for storing the structured outputs of our system. The database schema is designed to store the final sentiment predictions, the textual explanations generated by the LLM, pointers to the saved Visual Attribution Maps, and all associated metadata. By persisting these results, the system can serve historical queries rapidly without needing to re-run the computationally expensive models, ensuring a responsive and efficient user experience. This comprehensive, multi-layered architecture provides the foundation for a scalable, reliable, and user-centric platform for explainable multimodal urban analysis.

3. Results

To empirically validate the performance of our proposed framework, we conducted a series of rigorous experiments on a comprehensive test dataset. The evaluation is designed to be twofold: first, a quantitative assessment of the multimodal model's predictive accuracy and the contribution of its components; and second, a qualitative analysis of the explanations generated by the XAI Orchestrator. This section is dedicated to the quantitative findings, which establish the model's efficacy and justify its architectural design.

3.1. Quantitative Performance Evaluation

The primary objective of this evaluation is to measure the predictive power of the fully integrated Multimodal Sentiment Analysis Model. We assess its ability to accurately classify citizen feedback into three distinct sentiment categories: Negative, Neutral, and Positive.

3.1.1. Classification Accuracy

To obtain a granular understanding of the model's classification performance, we first generated a confusion matrix, which provides a disaggregated view of its predictions against the ground-truth

labels. The resulting matrix, displayed in Figure 5, visualizes the distribution of correct and incorrect classifications for each sentiment class. The strong diagonal entries – 728, 395, and 218 for Negative, Neutral, and Positive classes, respectively – indicate a high rate of correct predictions across the board. The model demonstrates exceptional proficiency in identifying negative sentiment, which is often the most prevalent and critical category in urban monitoring applications.

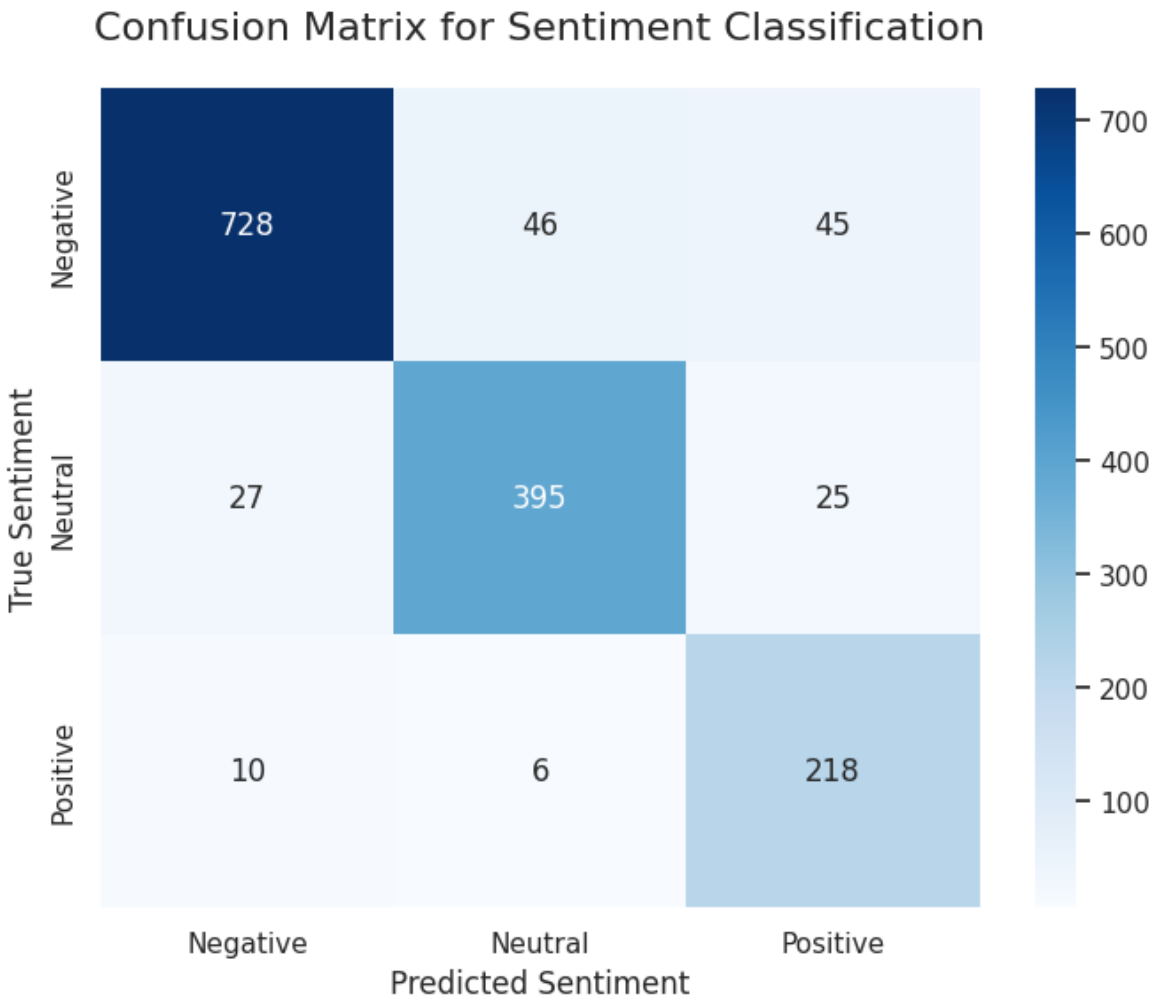

Figure 5: Confusion matrix illustrating the performance of the full multimodal sentiment classification model on the evaluation dataset (Figure by the authors).

Analysis of the off-diagonal elements reveals the model's specific confusion patterns. The most notable misclassification occurs between the Negative and Neutral classes, where 46 Negative instances were incorrectly predicted as Neutral, and 27 Neutral instances were misclassified as Negative. This suggests a degree of semantic overlap in the data for these categories, where expressions of mild dissatisfaction may be difficult to distinguish from neutral statements. The model exhibits the least confusion when classifying Positive instances, with only 16 such cases being misidentified.

For a more rigorous and standardized assessment, we derived key performance metrics from the confusion matrix, which are summarized in Table 1. The model achieved a high overall accuracy of 89.4%, confirming its general effectiveness. A closer examination of the per-class metrics reveals a more nuanced performance profile. The model attains an exceptionally high precision of 0.95 for the Negative class, indicating that when it predicts a sentiment as negative, it is very likely to be correct. Its recall for this class is 0.89, suggesting it successfully identifies the vast majority of all true negative instances. Conversely, the Positive class exhibits a very high recall of 0.93, meaning the model is

adept at finding positive feedback, but with a lower precision of 0.76, indicating it sometimes misattributes positive sentiment to neutral or negative posts. The Neutral class shows a balanced performance with a precision and recall of 0.88. The weighted average F1-Score, which accounts for class imbalance, stands at a robust 0.9, providing a single, comprehensive measure of the model's high predictive capability.

Table 1
Detailed Performance Metrics for the Multimodal Sentiment Classification Model

Sentiment Class	Precision	Recall	F1-Score	Support
Negative	0.95	0.89	0.92	892
Neutral	0.88	0.88	0.88	447
Positive	0.76	0.93	0.84	234
Overall				
Accuracy			0.894	1500
Weighted Avg	0.9	0.89	0.9	1500

3.1.2. Ablation Study

To validate our central hypothesis that a multimodal approach provides superior performance over unimodal methods, we conducted a comprehensive ablation study. This study systematically deconstructs our full model to isolate and measure the contribution of each modality. We evaluated three distinct model configurations: (1) a Text-Only baseline, (2) a bimodal model combining Text and Image data, and (3) our full Multimodal model incorporating Text, Image, and Geospatial data.

The results of this study are presented in Figure 6, which plots the weighted F1-score for each configuration. The findings provide clear and compelling evidence for the value of data fusion. The Text-Only model, serving as our baseline, achieved a respectable F1-score of 0.76. The introduction of the visual modality in the Text + Image configuration yielded a significant performance increase, elevating the F1-score to 0.83. This substantial improvement underscores the rich, and often essential, contextual information that images provide, which is entirely absent in a purely text-based analysis.

The final configuration, our full Multimodal model, which integrates the geospatial data via the Graph Neural Network encoder, achieved the highest F1-score of 0.9. This final incremental gain demonstrates that spatial context, while perhaps more subtle than visual or textual cues, provides a further layer of disambiguation that is critical for resolving complex cases and achieving peak performance. These results empirically validate our architectural design, proving that each modality offers a unique and complementary signal, and that their effective fusion is essential for building a highly accurate and reliable urban sentiment analysis system.

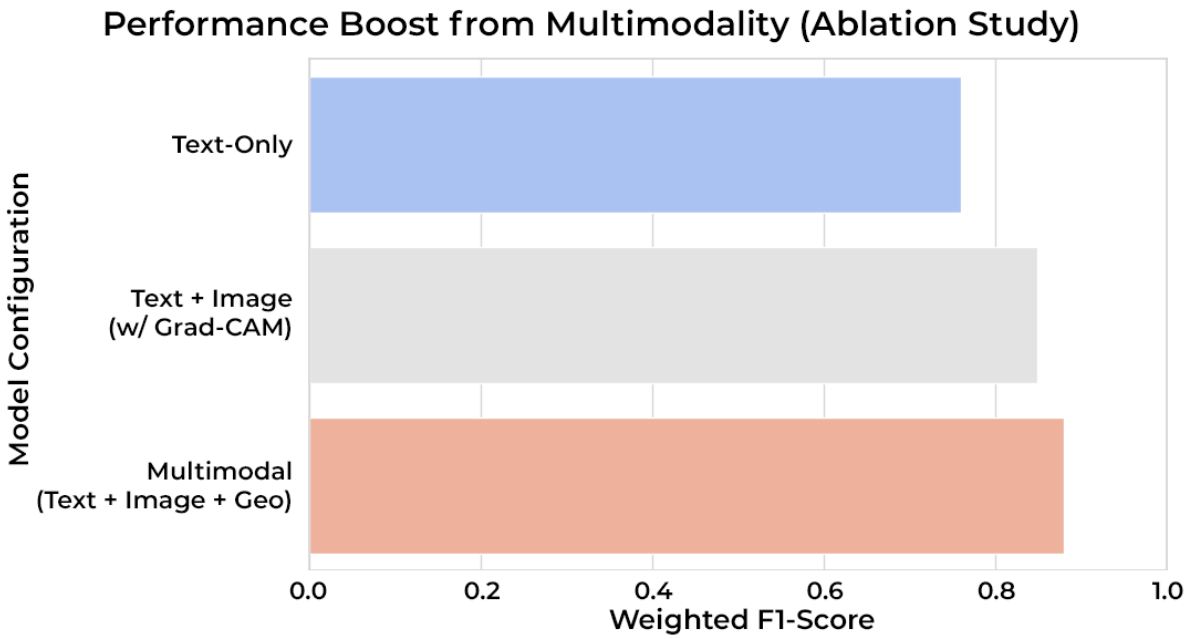

Figure 6: Results of the ablation study illustrating the performance contribution of each data modality (Figure by the authors).

3.2. Qualitative Analysis of Multimodal Explanations

While the quantitative metrics presented in the preceding section establish the high predictive accuracy of our model, they do not fully capture the primary contribution of this work: the generation of transparent and trustworthy explanations. To demonstrate the practical utility and interpretive power of our XAI Orchestrator, this section presents a qualitative analysis of its outputs through two distinct case studies. These examples, representative of common scenarios in urban monitoring, are designed to showcase the system's ability to deconstruct its predictions and present its reasoning in a human-intelligible format, thereby bridging the gap between a model's prediction and a stakeholder's understanding.

3.2.1. Case Study 1 – Deconstructing a "Negative" Sentiment Prediction from Urban Decay Imagery

The first case study examines a report containing an image of a building that suffered severe damage during a rocket attack, which is a clear instance of negative urban sentiment. Figure 7 provides a comprehensive breakdown of the visual explanation generated by our XAI Orchestrator for this example. Panel 1 shows the original input image, depicting the aftermath of a structural collapse. Panel 2 displays the raw Grad-CAM heatmap, where the warmer, red-hued regions indicate areas of high model activation and importance for the final prediction. Panel 3 presents the intuitive overlay, blending the heatmap with the original image to directly link the model's focus to specific visual features. Finally, Panel 4 utilizes the heatmap as a mask to isolate the most salient regions, providing an unambiguous view of the model's visual attention.

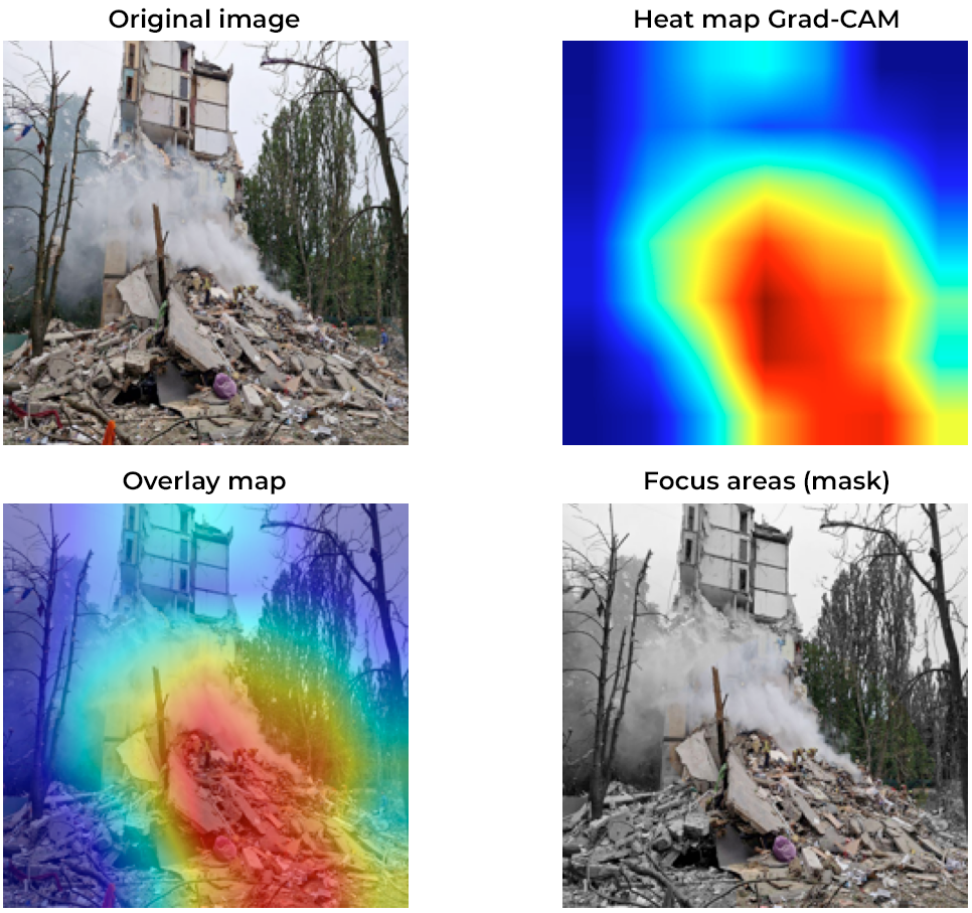

Figure 7: A complete qualitative analysis of a "Negative" sentiment prediction using Grad-CAM (Figure by the authors, the photograph is sourced from publicly available media and social network posts documenting real-world events).

The analysis clearly reveals that the model's prediction was overwhelmingly driven by the large pile of rubble and debris at the base of the damaged structure. However, a critical challenge arises from the nature of the pre-trained image encoder. Trained on the general-purpose ImageNet dataset, the encoder lacks the specific vocabulary and context of urban planning. Consequently, its raw, class-based prediction for this salient region is technically correct but contextually meaningless for a planner; it might classify the object as "ruin" or, more abstractly, "breakwater" due to visual similarities in texture and form.

This is the juncture at which the XAI Orchestrator's LLM Explanation Composer becomes critical. The composer receives a structured set of inputs: the final sentiment prediction ("Negative"), the primary visual focus (the rubble), and the raw object classification ("breakwater"). The LLM then synthesizes this information, re-contextualizing the naive technical label into a domain-relevant explanation. A representative output from the composer would be: "The system has classified this submission with a Negative sentiment. The decision was primarily based on the visual content of the image, with the model's attention concentrated on the extensive rubble and structural collapse at the center of the frame, which it identifies as a clear indicator of urban decay and destruction." This composed explanation effectively translates the opaque internal logic of the model into an actionable and trustworthy insight for a human expert.

3.2.2. Case Study 2 – Synthesizing a "Neutral" Sentiment Prediction from Reconstruction Activity

The second case study explores a more nuanced scenario: an image depicting a building undergoing reconstruction, featuring a prominent construction crane. This scenario could be interpreted as neutral (objective reporting of an activity) or positive (a sign of recovery). Figure 8 presents the visual explanation for this case. The Grad-CAM analysis (Panels 2, 3, and 4) clearly indicates that the model's visual attention is localized almost exclusively on the construction crane and its immediate operational area.

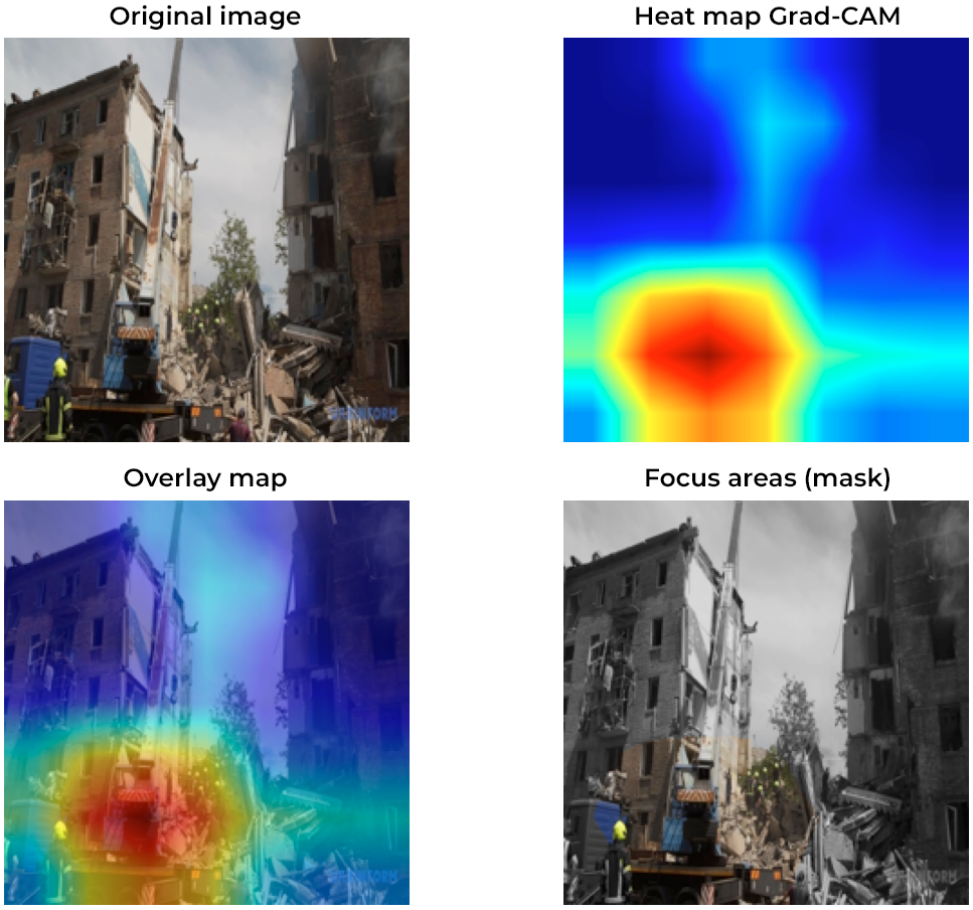

Figure 8: A complete qualitative analysis of a "Neutral" sentiment prediction (Figure by the authors, the photograph is sourced from publicly available media and social network posts documenting real-world events).

In this instance, the power of the XAI Orchestrator lies in its ability to synthesize multiple sources of evidence into a coherent narrative. Let us assume the accompanying text for this image was a simple, factual statement: "Work is underway to restore the building on Main Street." The XAI Orchestrator would receive the following inputs for its LLM composer: (1) a "Neutral" sentiment prediction, (2) a primary visual focus on the "crane" from Grad-CAM, and (3) the textual content. The LLM Composer then fuses these elements to produce the final, multimodal explanation:

"A Neutral sentiment was predicted for this report. The model's decision was primarily influenced by the visual presence of a construction crane, which it identified as the key object of interest. This visual focus on reconstruction activity is consistent with the factual, non-emotive language used in the accompanying text."

These two case studies demonstrate the core functionality of the XAI Orchestrator. It not only provides accurate, low-level attributions through methods like Grad-CAM but, more importantly,

leverages a Large Language Model to elevate these technical outputs into high-level, contextually aware, and directly interpretable explanations. This process is essential for building stakeholder trust and ensuring that the insights generated by the AI system are truly actionable.

Conclusion

The increasing integration of artificial intelligence into the domain of urban planning presents a profound paradox: while the predictive power of complex, data-driven models offers unprecedented opportunities for responsive governance, their inherent opacity creates a critical barrier to stakeholder trust and practical adoption. This "black box" problem is particularly acute in the context of urban revitalization, where decisions directly impact communities and demand the highest standards of transparency and accountability. The inability of decision-makers to understand the reasoning behind an AI-generated insight renders that insight unactionable, perpetuating a gap between technological capability and real-world utility.

In this paper, we addressed this fundamental challenge by proposing and validating a novel, end-to-end framework for explainable multimodal sentiment analysis. Our solution is built upon a high-performance multimodal analysis model that effectively fuses textual, visual, and geospatial data to achieve a nuanced understanding of citizen sentiment. The core innovation, however, is the XAI Orchestrator, a dedicated module that couples this predictive engine with a suite of state-of-the-art interpretability techniques. By synergizing modality-specific explainers – namely Grad-CAM for visual saliency and SHAP for textual attribution – and synthesizing their outputs through a Large Language Model composer, our framework deconstructs complex predictions into clear, human-intelligible justifications.

Our empirical evaluation demonstrated the dual efficacy of this approach. The quantitative results confirmed the high predictive accuracy of the underlying model, achieving a weighted F1-score of 0.9. A rigorous ablation study further validated our multimodal design, showing a clear and significant performance improvement with the integration of each data modality. Beyond predictive power, our qualitative case studies revealed the interpretive strength of the XAI Orchestrator, showcasing its ability to generate composite explanations that are both visually intuitive and narratively coherent. The system successfully translated opaque model activations into clear, actionable insights, directly linking sentiment predictions to specific visual and textual evidence.

While this work establishes a viable framework for explainable urban analysis, it also illuminates several promising avenues for future research. A logical next step is the development of domain-specific encoders, fine-tuned on large-scale urban datasets, which could enhance the model's ability to recognize contextually relevant features beyond what is possible with general-purpose pre-training. Further investigation into more advanced feature fusion mechanisms, such as attention-based co-learning across modalities, may yield additional performance gains. Perhaps the most critical future direction lies in the development of robust, quantitative metrics to formally evaluate the quality, faithfulness, and utility of the generated explanations, moving the field beyond purely qualitative assessment. Finally, the framework's modular design invites the future integration of additional data modalities, such as audio from public forums or sensor data, to create an even more holistic understanding of the urban environment.

In conclusion, this research presents more than just a novel architecture; it offers a methodological blueprint for designing and deploying AI systems that are not only accurate but also transparent and trustworthy. By prioritizing explanation alongside prediction, our work contributes to the critical effort of transforming AI from a "black box" tool into a collaborative partner for urban planners, policymakers, and the public. Ultimately, it is through such human-centric, explainable systems that we can hope to build the smarter cities of the future – cities that are not only technologically advanced but also more accountable, equitable, and responsive to the citizens they serve.

Acknowledgements

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). The project was executed under the national priority research direction of "Information and Communication Technologies".

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 2.5 Pro, Claude 4.0 Sonnet, and GPT-5 in order to: accelerate the software code generation procedure and to improve the readability and correctness of the English language. 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.

References

[1] K. Honzák, S. Schmidt, B. Resch, and P. Ruthensteiner, P, "Contextual Enrichment of Crowds from Mobile Phone Data through Multimodal Geo-Social Media Analysis," ISPRS Int. J. Geo Inf., vol. 13, no. 10, pp. 350, October 2024. https://doi.org/10.3390/ijgi13100350.

[2] D. Hanny, and B. Resch, "Multimodal Geo-Information Extraction from Social Media for Supporting Decision-Making in Disaster Management," AGILE GIScience Ser., vol. 5, no. 28, pp. 1–8, September 2024. https://doi.org/10.5194/agile-giss-5-28-2024.

[3] D. Hanny, and B. Resch, "Multimodal GeoAI: An integrated spatio-temporal topic-sentiment model for the analysis of geo-social media posts for disaster management," Int. J. Appl. Earth Obs. Geoinformation, vol. 139, pp. 104540, May 2025. https://doi.org/10.1016/j.jag.2025.104540.

[4] T. Honcharenko, G. Ryzhakova, Y. Borodavka, D. Ryzhakov, V. Savenko, and O. Polosenko, "Method for representing spatial information of topological relations based on a multidimensional data model," ARPN Journal of Engineering and Applied Sciences, vol. 16, no. 7, pp. 802–809, April 2021.

[5] Y. Riabchun, T. Honcharenko, V. Honta, K. Chupryna, and O. Fedusenko, "Methods and means of evaluation and development for prospective students' spatial awareness," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 11, pp. 4050–4058, August 2019. https://doi.org/10.35940/ijitee.k1532.0981119.

[6] D. Hazarika, R. Zimmermann, and S. Poria, :MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis," Proceedings of the 28th ACM International Conference on Multimedia, pp. 1122–1131, October 2020. https://doi.org/10.1145/3394171.3413678.

[7] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," Conference on Empirical Methods in Natural Language Processing, pp. 1103–1114, September 2017. https://doi.org/10.18653/v1%2FD17-1115.

[8] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis," AAAI Conference on Artificial Intelligence, vol. 35, no. 12, pp. 10790–10797, 2021. https://doi.org/10.1609/aaai.v35i12.17289.

[9] H. Yang, Y. Zhao, Y. Wu, S. Wang, T. Zheng, H. Zhang, W. Che, and B. Qin, "Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey," ArXiv, pp. 1–40, June 2021. https://doi.org/10.48550/arXiv.2406.08068.

[10] W. Wang, L. Ding, L. Shen, Y. Luo, H. Hu, and D. Tao, "WisdoM: Improving Multimodal Sentiment Analysis by Fusing Contextual World Knowledge," Proceedings of the 32nd ACM

International Conference on Multimedia, pp. 2282–2291, October 2024. https://doi.org/10.1145/3664647.3681403.
[11] D. Yang, M. Li, D. Xiao, Y. Liu, K. Yang, Z. Chen, Y. Wang, P. Zhai, K. Li, and L. Zhang, “Towards Multimodal Sentiment Analysis Debiasing via Bias Purification,” ArXiv, pp. 1–18, March 2024. https://doi.org/10.48550/arXiv.2403.05023.
[12] O. Matsiievskyi, T. Honcharenko, O. Solovei, T. Liashchenko, I. Achkasov, and V. Golenkov, “Using Artificial Intelligence to Convert Code to Another Programming Language,” 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST), pp. 379–385, May 2024. https://doi.org/10.1109/SIST61555.2024.10629305.
[13] O. Matsiievskyi, I. Achkasov, Y. Borodavka, and R. Mazurenko, “Behavioral model of autonomous robotic systems using reinforcement learning methods,” CEUR Workshop Proceedings, vol. 3896, pp. 560–568, October 2024. https://ceur-ws.org/Vol-3896/short14.pdf.
[14] S. Dolhopolov, T. Honcharenko, O. Terentyev, V. Savenko, A. Rosynskyi, N. Bodnar, and E. Alzidi, “Multi-Stage Classification of Construction Site Modeling Objects Using Artificial Intelligence Based on BIM Technology,” 2024 35th Conference of Open Innovations Association (FRUCT), pp. 179–185, April 2024. https://doi.org/10.23919/fruct61870.2024.10516383.
[15] A. Neftissov, A. Biloshchytskyi, I. Kazambayev, S. Dolhopolov, and T. Honcharenko, “An Advanced Ensemble Machine Learning Framework for Estimating Long-Term Average Discharge at Hydrological Stations Using Global Metadata,” Water, vol. 17, no. 14, pp. 2097, July 2025. https://doi.org/10.3390/w17142097.
[16] P. Linardatos, V. Papastefanopoulos, and S. B. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” Entropy, vol. 23, no. 1, pp. 18, December 2020. https://doi.org/10.3390/e23010018.
[17] D. A. Sharma, and K. S. NagendraKumar, ”Explainable AI: Scene Classification and GradCam Visualization,” International Journal of Advanced Research in Science, Communication and Technology, vol. 4, no. 4, pp. 1–5, June 2024. https://doi.org/10.48175/ijarsct-19006.
[18] V. Dhore, A. Bhat, V. Nerlekar, K. Chavhan, and A. Umare, “Enhancing Explainable AI: A Hybrid Approach Combining GradCAM and LRP for CNN Interpretability,” ArXiv, pp. 1–10, May 2024. https://doi.org/10.48550/arXiv.2405.12175.
[19] Y. Yang, S. Wang, D. Li, S. Sun, and Q. Wu, “GeoLocator: A Location-Integrated Large Multimodal Model (LMM) for Inferring Geo-Privacy,” Applied Sciences, vol. 14, no. 16, pp. 7091, August 2024. https://doi.org/10.3390/app14167091.
[20] Y. Wang, T. Zhang, X. Guo, and Z. Shen, “Gradient based Feature Attribution in Explainable AI: A Technical Review,” ArXiv, vol. 1, no. 1, pp. 1–25, March 2024. https://doi.org/10.48550/arXiv.2403.10415.
[21] M. Bachhawat, “Generalizing GradCAM for Embedding Networks,” ArXiv, pp. 1–7, February 2024. https://doi.org/10.48550/arXiv.2402.00909.