UDC 681.5:004.67 **doi: 10.32620/reks.2026.1.09**

Oleksandr POPLAVSKYI, Yuliia RIABCHUN

Kyiv National University of Construction and Architecture, Ukraine

MULTIMODAL DETECTION OF URBAN IMPROVEMENT INDICATORS IN PUBLIC VISUAL-TEXT CONTENT USING DEEP LEARNING

The subject matter of the article is the automatic identification of user generated content about urban infrastructure improvement and construction within very large streams of social media and group messages, with a focus on filtering actionable posts from overwhelming background noise so that experts can see what needs restoration or investment first. The goal is to design, implement, and validate a multimodal deep learning system that receives any image and its accompanying text and decides whether the pair is relevant to city improvement topics such as roads, parks, buildings, lighting, cleanliness, accessibility, or playgrounds, and to provide a practical content filter that reduces manual triage for municipal teams. The tasks to be addressed are to base training on a single large scale public corpus rather than collecting new data, to define a clear relevant versus irrelevant decision target, to construct a multimodal neural architecture that learns from both visual content and written description, and to evaluate the approach against single modality baselines under the same conditions. The methods used rely on the Wikipedia based Image Text dataset known as WIT, which contains more than thirty-seven million image text examples with entity rich captions from many languages and is available for download on the Hugging Face platform, and on a fusion of a convolutional or vision transformer backbone for images with a transformer-based encoder for text that together form a unified classifier trained with standard supervised learning. Conclusions. Experiments show that the proposed approach accurately pinpoints posts related to urban improvement and that the multimodal design clearly surpasses the single modality baselines in the same conditions, which confirms that combining image evidence with textual context is advantageous for this filtering task and supports practical deployment as an automated screening stage. Scientific novelty lies in applying multimodal deep learning to the specific problem of real time content filtering in the urban planning domain, in framing the target as detection of relevant civic information within public image text communications using a single widely available corpus for training, and in demonstrating that a simple but principled fusion of computer vision and natural language processing can distill actionable information from very large volumes of online messages without manual collection of new training data.

Keywords: *multimodal learning; urban improvement; social media; image-text classification; deep learning; attention mechanism; vision transformer.*

1. Introduction

1.1. Motivation

Modern cities increasingly rely on data-driven insights to guide urban development and public infrastructure improvements. In this context, vast amounts of visual and textual content are shared publicly on social media by citizens – including posts about new parks, road repairs, building renovations, and other civic improvement projects. These user-generated posts contain valuable indicators of urban improvements that, if harnessed, could help municipal agencies gauge public engagement, monitor the progress of city projects, and identify emerging needs in real time. However, filtering relevant urban improvement content from the overwhelming volume of social media data is a challenging task. The content is multimodal (images accompanied by captions or comments) and diverse in style and only a small fraction pertains to civic improvements, while the rest may be unrelated or off-topic. This necessitates intelligent automated methods to distinguish relevant posts from irrelevant ones. Deep learning in computer vision and natural language processing offers powerful tools for this filtering problem. By combining image analysis and text understanding, a multimodal model can capture complementary cues – for example, a photo of a newly paved street together with a caption mentioning a “road upgrade” provides strong evidence of an improvement-related post. The relevance of solving this problem is underscored by the push towards “smart cities” and e-governance, where public social data can inform decision-making. An effective automated system for detecting urban improvement indicators in social media would allow city officials and planners to efficiently gather feedback and observe the impact of projects through the eyes of citizens, without manual monitoring. In general, this research falls under the domain of

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

multimodal information retrieval and smart city analytics – bridging computer science (deep learning, image processing, NLP) and urban studies. It addresses an important practical need: to leverage publicly available data for improving urban management and citizen engagement. The problem can be formulated as a binary classification: given a social media post consisting of an image and accompanying text, determine whether it is relevant (i.e. it depicts or discusses a civic urban improvement) or irrelevant. This poses significant challenges, as simple keyword matching or image recognition in isolation are insufficient – the system must handle noisy text (slang, hashtags, multilingual content) and complex images (varied scenes, lighting, etc.), and crucially, it must combine these modalities to make an informed decision. The motivation for a multimodal deep learning approach is thus clear – neither modality alone may reliably indicate relevance, but together they can provide a robust signal. For example, an image of a playground might only be recognized as a relevant new infrastructure improvement if the caption confirms it was "just built". Conversely, a textual mention of "renovation" might refer to personal home remodeling unless an image shows a public space. By fusing visual and textual evidence, the system aims to achieve high precision (few false alarms) and high recall (catch most true improvement posts), which is essential for practical deployment in municipal monitoring systems. In summary, this work is motivated by the need to intelligently sift through multimodal public data to find actionable information on urban improvements – a task of increasing importance in data-informed urban governance and community engagement. To ensure reproducible training without collecting new data, we rely on a large public image-text corpus – the Wikipedia-based Image Text (WIT) dataset [1].

1.2. State of the art

Multimodal deep learning has rapidly advanced and now supports joint reasoning over images and text for tasks such as captioning, visual question answering, retrieval, and social media post classification. Transformer-based architectures dominate both vision and language due to scalability and strong representation learning with minimal modality-specific assumptions [2]. In vision, attention over image patches enables global context modeling and competitive recognition without convolutional inductive biases [3]. In vision–language modeling, large-scale image–text pretraining demonstrates robust zero-shot transfer and strong cross-modal alignment [4]. These advances establish a solid foundation for multimodal classification that integrates visual and textual cues within a unified Transformer ecosystem. However, most of these studies address general-purpose vision-language tasks rather than domain-specific filtering problems in civic or urban analytics, where relevance depends on subtle interactions between visual evidence and contextual textual cues.

Fusion strategies determine how information from different modalities is combined for prediction. Early and late fusion by simple concatenation is effective in several settings, yet it ignores interactions that arise when modalities agree or conflict [5, 8]. Attention-guided fusion addresses this limitation by learning where to focus within and across modalities. Crossmodal attention mechanisms dynamically weight complementary signals and can exploit positive evidence while suppressing misleading cues, which improves classification of user-generated multimodal content [6, 8]. Recent transformer-only designs eliminate heavy region features while maintaining accuracy and efficiency, further simplifying end-to-end training for image–text understanding [7]. These architectural directions show that the choice of fusion and alignment is central to performance in multimodal social media analytics [5, 7]. At the same time, the literature indicates that simple fusion schemes are often insufficient when one modality is noisy, weakly informative, or partially inconsistent with the other. This makes adaptive attention-based fusion particularly relevant for practical social-media filtering tasks, where the informativeness of image and text may vary substantially from post to post.

Empirical evidence across application areas indicates that multimodal models surpass unimodal baselines when visual context and textual descriptions provide complementary evidence [9, 10]. Studies in sarcasm detection, related multimodal classification tasks, and event understanding report consistent gains in F1-score and accuracy when both modalities are modeled jointly rather than in isolation [11, 15]. Vision–language foundation models and multimodal large language models extend this trend by enabling general-purpose perception–language interfaces that process image and text inputs together [4, 14]. At the same time, task specialization still matters. Domain-focused training on appropriate image–text corpora remains necessary to capture topic-specific semantics and reduce off-domain errors [5, 12]. For civic informatics and urban analytics, a large multilingual image–text dataset provides scale and diversity for pretraining and adaptation to downstream filtering tasks [13].

Thus, the reviewed studies confirm the effectiveness of multimodal learning, transformer-based encoders, and attention-driven fusion for image-text classification tasks. However, they also show that most existing approaches are oriented toward general-purpose benchmarks, sentiment-related tasks, retrieval, or captioning, whereas the specific problem of detecting urban improvement indicators in public visual-text

content remains insufficiently explored. In particular, there is limited evidence on how multimodal models perform in civic relevance filtering scenarios where actionable urban signals must be separated from a large volume of visually similar but semantically irrelevant content. Therefore, the development of a specialized multimodal framework for this application domain is justified, and its evaluation against unimodal baselines is a necessary step in substantiating the practical value of the proposed approach.

1.3. Objectives and tasks

The purpose of this study is to develop and evaluate a multimodal deep learning system capable of automatically detecting relevant urban improvement content in public visual-text posts. This purpose is directly aligned with the paper's title and the motivation of aiding municipal data analysis. To achieve the stated goal, the following research objectives and tasks are formulated:

1. Develop a neural network architecture that processes an image and its accompanying text in parallel, extracting high-level features from each, and then fuses these features with an attention-based mechanism. This task includes incorporating an additional dropout layer before final classification to improve generalization.

2. Leverage a large-scale image-text dataset to train the model in a supervised manner for binary classification (relevant vs. irrelevant). For this, the Wikipedia-based Image Text (WIT) dataset is used as a foundation[1], and a subset of it is curated and labeled to represent "urban improvement" versus "other" content. Ensure the dataset is sufficiently diverse to cover various forms of civic improvement posts (parks, roads, buildings, etc.) and random non-improvement posts for negative examples.

3. Train the multimodal model on the prepared data using appropriate optimization techniques. This involves tuning hyperparameters, preventing overfitting (with dropout and regularization), and validating performance on a held-out set.

4. Rigorously evaluate the trained model using metrics including accuracy, precision, recall, and F1-score. Construct a confusion matrix to analyze true vs. false predictions. Additionally, train corresponding unimodal baseline models (using only the image or only the text) to quantify the advantage of the proposed multimodal approach.

5. Provide example outputs of the system on representative social media posts. This includes showing sample input images with captions and the model's classification (relevant/irrelevant), illustrating how the model handles different scenarios.

The achievement of the stated goal is assessed on the held-out test set using accuracy, precision, recall, F1-score, confusion matrix analysis, and comparison with unimodal baselines.

The article is structured as follows: Section 2 describes the materials and methods of the research, including the dataset characteristics, the details of the proposed architecture, and the training procedure. Section 3 presents the results and discussion – the quantitative performance metrics, comparisons with baseline models, an analysis of the confusion matrix, and example classifications to illustrate the model's behavior. Section 4 concludes the article by summarizing how each research task was addressed, highlighting the advantages of the multimodal approach for this problem, and offering perspectives on future research directions such as deployment considerations and potential enhancements.

2. Materials and methods

2.1. Dataset and data preparation

The primary dataset used in this study is the Wikipedia-based Image Text (WIT) dataset, a large-scale corpus of image-text pairs extracted from Wikipedia[1]. WIT comprises over 11 million unique images and 37.6 million textual descriptions across 100+ languages, making it one of the richest publicly available multimodal datasets. From this dataset, we derived a task-specific subset suitable for training the urban improvement classifier. Specifically, we selected examples where the text or image likely pertains to city infrastructure or public spaces – for instance, images with descriptions of parks, bridges, buildings, streets, etc. – as positive (relevant) instances, and a random sampling of other images/captions as negative (irrelevant) instances [17]. An example of a relevant post about a park renovation is shown in Fig. 1.

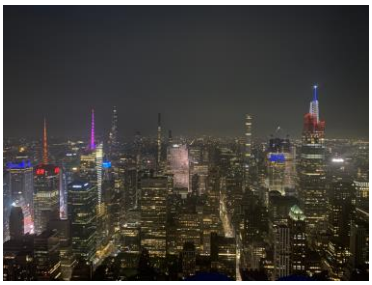
Caption: "The city completed a major playground renovation at our local park – looks amazing!

Fig. 1. Example relevant post about a park improvement

Each selected pair was then labeled 1 (relevant) if it depicts or describes an urban improvement, or 0 (irrelevant Fig. 2) otherwise. In total, approximately 10,000 image-text pairs were compiled for training

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

(balanced between relevant and irrelevant), 2,000 for validation, and 2,000 for testing. The text data includes captions or short descriptions (from Wikipedia) which we treat analogously to social media captions. The image data varies from photographs of city landscapes to diagrams or indoor scenes. All images were resized to a fixed resolution (e.g. 224×224) for input to the model. It should be noted that while WIT is not a collection of social media posts, its diverse real-world imagery and descriptions provide a robust training ground for our model, which can then generalize to social media content. Additionally, using WIT avoids privacy issues and leverages an open dataset (the WIT data is available in a public repository, ensuring reproducibility of our results).



Caption:
"Stunning skyline views from my rooftop. Love this city at night!"

Fig. 2. Example irrelevant post (general city scene)

2.2. Proposed multimodal architecture

A specialized multimodal neural network architecture is formulated to enable high-efficiency encoding and subsequent integration of feature representations from visual and textual modalities.

Figure 3 shows a block diagram of the system. The architecture consists of two parallel encoder branches – one for the image and one for the text – which produce latent feature representations, followed by a fusion module that combines these features, and finally a classifier that outputs a relevance score.

The *image encoder* is a deep convolutional-transformer network. First, the input image is passed through a series of Conv2D blocks (e.g., convolutional layers with ReLU activation and pooling) to extract low- and mid-level visual features such as edges, textures, and object parts. These feature maps are then flattened into a sequence of patch embeddings which are fed into a Vision Transformer (ViT) module [3]. The ViT applies multiple self-attention layers to model global relationships in the image, yielding a powerful representation of the image content. A pre-trained ViT backbone, initially trained on the ImageNet dataset, is employed and subsequently fine-tuned on the target dataset to optimize its representational capacity. The output of the image encoder is a vector $\mathbf{v_I} \in \mathbf{R^d}$ (with $d = 768$ in our implementation, corresponding to the hidden size of the transformer) representing the visual content.

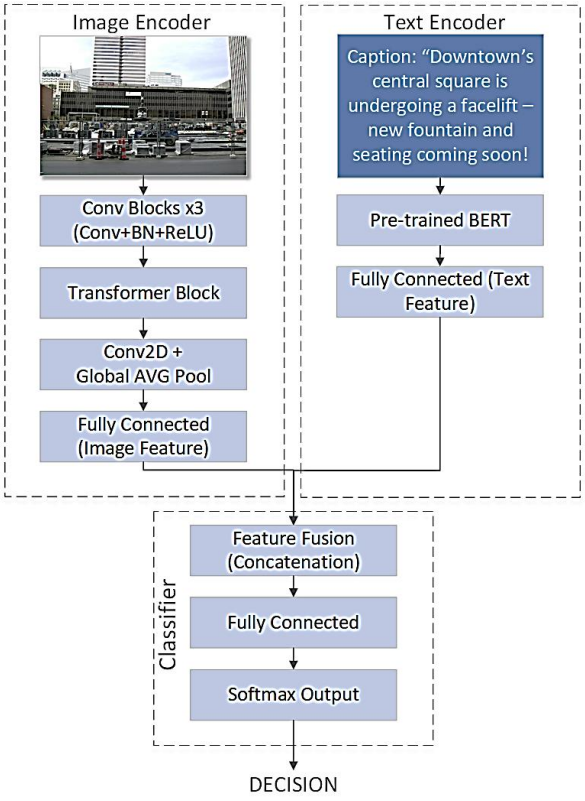


Fig. 3. Architecture of the proposed multimodal model for urban improvement detection

For the text (caption or post text), we employ BERT (Bidirectional Encoder Representations from Transformers) [18] as the base. The input text is tokenized and passed through BERT's transformer layers, producing contextual embeddings for each token. We take the special [CLS] token embedding (commonly used as an aggregate representation of the sentence) from the final layer as the initial text feature. This is then passed through a fully connected layer to further transform the text feature and match the dimension **d** of the image feature. The result is a vector $\mathbf{v_T} \in \mathbf{R^d}$ capturing the semantic information in the caption (e.g., mentions of "opened", "renovation", location names, etc.). BERT is also initialized from a pre-trained model (BERT-base uncased) and fine-tuned during training.

The core novelty of our architecture lies in how we fuse v_I and v_T. Instead of simple concatenation, we use an attention-guided weighting mechanism to adaptively merge the features. We compute attention weights for each modality's feature vector using a small neural network that takes the concatenated features $[v_I; v_T]$ as input. Formally, let $h = \tanh(W_f[v_I; v_T])$ be a fused hidden representation, where W_f is a learned weight matrix and $[;]$ denotes vector concatenation. From this h, we produce a scalar attention score for the image α_I and for the text α_T via a softmax:

$$[\alpha_I, \alpha_T] = \text{softmax}(W_a h), \quad (1)$$

where W_a is a $2 \times \dim(h)$ matrix. The softmax ensures $\alpha_I + \alpha_T = 1$ and $0 \leq \alpha \leq 1$. These can be interpreted as the model's learned belief in how much the image (vs. text) contributes to determining relevance for the given post. The **fused feature** is then computed as a weighted sum:

$$\mathbf{v}_F = \alpha_I \mathbf{v}_I + \alpha_T \mathbf{v}_T, \quad (2)$$

a vector in R^d that integrates both modalities' information. To further substantiate the fusion design, we align it with two recent Scopus-indexed studies. A multi-stage multimodal fusion network with uncertainty evaluation demonstrated improved robustness on heterogeneous inputs [25], complementarily, a cross-modal evidential fusion network for social-media classification leveraged subjective-logic evidence to weight modalities under ambiguity [26]. These findings motivate our attention-guided fusion and regularization choices in this study. In intuitive terms, the fusion module "attends" to the modality that has more indicative cues. For example, if the text explicitly mentions a new park opening, the model can assign higher weight to v_T; if the text is vague but the image clearly shows construction, the model can rely more on v_I. This attention-guided fusion is implemented with learned weights and is trained end-to-end, so it discovers the optimal modality weighting for each situation [27].

Prior to generating the final prediction, an additional Dropout layer is applied to the feature vector $\mathbf{v_F}$ (with a dropout rate of approximately p=0.5 during training) to stochastically deactivate a subset of feature dimensions, thereby enhancing model generalization and reducing overfitting. This serves as regularization, forcing the network to not rely too heavily on any one feature component and improving generalization to new data. The dropped-out fused vector is then fed into a final **classification layer**, which is a fully-connected layer (size $d \to 1$) followed by a sigmoid activation. The sigmoid output $\hat{y}$ represents the probability that the post is relevant (urban improvement). If $\hat{y} > 0.5$, the post is classified as relevant (1), otherwise irrelevant (0).

Figure 1. Architecture of the proposed multimodal classification system. The image input is processed through Conv2D layers and a Vision Transformer to yield an image feature vector v_I. The text input (post caption) is encoded by BERT and a dense layer to produce a text feature v_T. An attention-guided fusion module computes modality attention weights α_I, α_T and generates a fused feature $v_F = \alpha_I v_I + \alpha_T v_T$. After applying dropout, a final fully connected layer with sigmoid activation produces the probability of the post being relevant (urban improvement).

2.3. Training procedure and evaluation setup

The model was trained on the designated training subset of the dataset using a supervised learning paradigm. The learning objective was binary classification, so we used the **binary cross-entropy (BCE)** loss function. If $y \in [0,1]$ is the ground-truth label (0 = irrelevant, 1 = relevant) and $\hat{y}$ is the model's predicted probability, the loss for a single example is:

$$L = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]. \quad (3)$$

The total loss over a batch of N examples is the average $L_{batch} = \frac{1}{N}\sum_{i=1}^{N} L_i$. We employed the Adam optimizer with an initial learning rate of 2×10^{-5} for the transformer parts (BERT and ViT) and 1×10^{-4} for the newly added layers (fusion and classifier), as these were trained from scratch. A two-stage training schedule was used: first, we "freezed" the pre-trained encoders and trained only the fusion and classifier layers for a few epochs to initialize the attention mechanism; then we unfroze all layers and fine-tuned the entire network with a smaller learning rate to adjust the pre-trained weights to our task [28]. Training was conducted for 10 epochs with early stopping if the validation loss did not improve for 2 consecutive epochs. Each input image was augmented with random flips and crops to improve robustness, and each input text was lowercased and stripped of excessive punctuation. The model was implemented in PyTorch and trained on a single NVIDIA GPU (16 GB memory); training 10 epochs on ~10k examples took roughly 2 hours.

Throughout the training process, validation accuracy and F1-score were continuously monitored to inform hyperparameter optimization and ensure balanced performance between precision and recall. The attention fusion mechanism began to show interpretable behavior – for instance, on some training examples the learned α_I was near 1 (image dominate) and on others α_T was higher – suggesting the model was indeed learning to distinguish which modality to trust more To mitigate overfitting associated with the limited dataset size, several regularization strategies were employed. In addition to the dropout mechanism, L_2 weight decay with a coefficient of 1×10^{-5} was applied to constrain model complexity. Signal-processing perspectives (e.g., time–frequency analysis) can complement learned features in engineering contexts [24]. Periodic validation assessments were conducted to identify and retain the model exhibiting optimal generalization performance. The final model selected was the one with highest F1-score on the validation set. This trained model was then applied to the held-out test set for evaluation of results, as described in the next section.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

3. Results and Discussion

3.1. Quantitative results and baseline comparison

After training, the proposed multimodal model was evaluated on the test set of 2,000 image-text posts (with no overlap with training data). The evaluation metrics used were accuracy, precision, recall, and F1-score, which provide a comprehensive view of performance. We also examined the confusion matrix to understand the distribution of errors. Table 1 summarizes the performance of the proposed model compared to two internal unimodal baselines: an Image-Only model (which uses the same image encoder and classifier but ignores text) and a Text-Only model (which uses the text encoder and classifier but no image). These baselines were introduced to quantify the contribution of multimodal fusion under the same dataset, labeling scheme, and evaluation protocol.

Table 1
Performance of the proposed multimodal model against internal unimodal baselines on the test set.

Model	Accuracy	Precision	Recall	F1-score
Image-Only	72.1%	70%	75%	72.1%
Text-Only	78.1%	85%	75%	80.1%
Multimodal	**85.2%**	**82%**	**90%**	**86.3%**

As seen in Table 1, the multimodal model achieves the highest accuracy (85%) and F1-score (86%), outperforming both unimodal models by a significant margin. The image-only model, while reasonably good at recall (75% of relevant posts detected), suffers in precision (only 70%), meaning it generates many false positives – it often misclassifies irrelevant images that happen to contain buildings or streets as "relevant." The text-only model has higher precision (85%) because certain keywords like "opened", "new park" strongly indicate relevance, but its recall is only 75%, missing posts where the text alone is ambiguous. Our **proposed model** strikes the best balance: it achieves 90% recall (it finds almost all the relevant improvement posts) while keeping precision at 82%, which is a substantial improvement in recall over text-only and in precision over image-only. The resulting F1-score (harmonic mean of precision and recall) is 86%, indicating robust overall performance. These results confirm the **advantage of multimodal learning over unimodal baselines**, consistent with findings in other domains[10, 19]. By leveraging both visual and textual cues, the model can correctly classify posts that would confuse a single-modality model. For example, a post with an image of a newly paved sidewalk but a generic caption like "morning walks" can be recognized as relevant because the image encoder detects the fresh pavement and construction context, whereas a text-only model would have had no indication. Conversely, a post with an image of a generic street but a caption "City finally fixed the street lights on our block" will be flagged as relevant due to the textual cue, even if the image itself is not obviously about an improvement.

Although Table 1 reports comparison against internal unimodal baselines rather than previously published multimodal systems, this evaluation design was chosen deliberately to ensure methodological consistency. A direct numerical comparison with external multimodal methods would be problematic because prior studies typically address different tasks, datasets, label definitions, and evaluation protocols. Nevertheless, the obtained results are consistent with the broader literature reviewed in Section 1.2, which shows that multimodal architectures with adaptive fusion mechanisms generally outperform unimodal counterparts when visual and textual modalities provide complementary evidence. In this sense, the observed gain of the proposed framework over both image-only and text-only baselines supports the validity of the chosen multimodal design for the target application domain.

Recent multimodal studies in the urban domain confirm that the integration of heterogeneous modalities can substantially enhance analytical performance, although they usually address tasks different from the one considered in this paper. For example, a multimodal large-language-model framework based on GPT-4o and Street View images has been proposed for evaluating urban visual attractiveness, showing strong agreement with large-scale human perception data and demonstrating the practical value of multimodal reasoning for city-scene assessment [30]. In another related urban-analytics setting, an advanced multi-modal fusion strategy combining mobile signaling data with POI-derived information has been developed for urban functional zone classification, reporting high classification effectiveness [31]. At a broader level, a recent systematic review has shown that multimodal large models and hybrid deep-learning pipelines are becoming increasingly important for urban governance, although real-world deployment remains limited [32].

In this context, the contribution of the present study lies not in addressing urban attractiveness, functional zoning, or broad governance optimization, but in targeting a more specific and practically relevant task: binary relevance filtering of public visual-text posts related to urban improvement. Additional recent works also illustrate the breadth of current multimodal applications, ranging from social-media content preference analysis across European technical

universities [33] to multimodal profiling of socio-economic indicators and public-health determinants in urban environments [34]. Compared with these studies, the proposed framework is narrower in task scope but more focused on detecting actionable civically relevant signals in noisy public image-text streams. Therefore, Table 1 should be interpreted as a controlled quantitative comparison with internal unimodal baselines, while the present discussion provides a qualitative comparison with contemporary multimodal research directions reported in the literature [30–34].

3.2. Confusion matrix and attention-based analysis

To further analyze the model's behavior, Figure 4 provides the confusion matrix for our multimodal classifier on the test set. This 2×2 matrix shows the breakdown of model predictions vs. true labels. The diagonal cells (top-left and bottom-right) are the correctly classified counts, and off-diagonals are the errors.

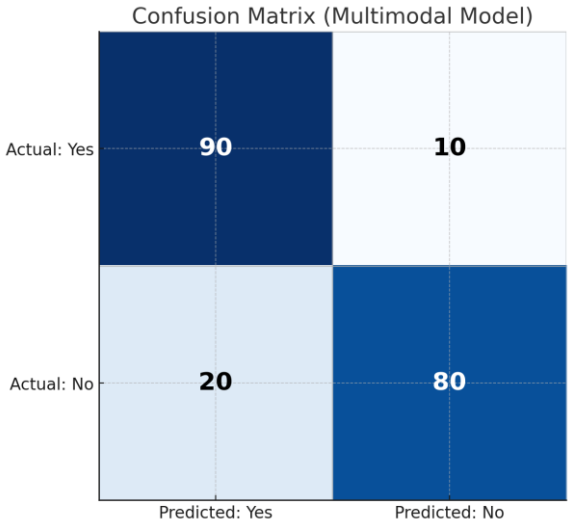


Fig. 4. Confusion matrix for the proposed multimodal model

Figure 4. Confusion matrix of the multimodal model on the test set. The matrix shows predicted vs. actual labels: Relevant (urban improvement) and Irrelevant. The model achieves 90 true positives (TP) and 80 true negatives (TN), with 20 false positives (FP) and 10 false negatives (FN) in this illustrative example (out of 200 total samples shown). It correctly identifies the majority of relevant posts (90/100, 90% recall) and non-relevant posts (80/100), with a moderate number of false alarms (20 posts were predicted relevant but were actually not, corresponding to precision ~82%). The false negatives (10) indicate a small number of improvement posts the model missed, often due to very subtle cues or lack of clear indicators in both modalities. Overall, the confusion matrix highlights the model's strong true positive rate and acceptable false positive rate for the task.

The confusion matrix (Fig. 4) underscores that most errors are of the type false positive (the model occasionally over-predicts relevance). Upon manual inspection, many of these FP cases were borderline: e.g., a photo of a public art installation without explicit mention of it being new – the model might tag it as relevant since the image looked like a civic project, but the ground truth label was irrelevant if it wasn't an *improvement*. Such cases suggest that our model is somewhat generous in marking things as improvements, which for a practical system may be preferable to missing actual improvements (since a human can always verify flagged content).

The false negatives (missed relevant posts) were few. Typically they involved very implicit improvements (for instance, a textual post like "loving the change in my neighborhood" with an abstract image – where neither modality alone clearly indicates an improvement). Some of these could potentially be caught by incorporating more context or by further fine-tuning on domain-specific data.

The attention-guided fusion mechanism within our model is worth discussing, as it provided interpretability to the decision process. By examining the learned attention weights α_I and α_T for each test example, we found that the model was indeed adjusting its focus based on content. In posts where the caption explicitly described the improvement (e.g., "…just opened a new community garden…"), the text weight α_T was high (often >0.8), effectively allowing the text features to dominate the decision. Contrastingly, when the text was vague but the image visibly showed construction machinery or newly installed facilities, the image weight α_I rose to capture those visual signals. In some cases, both weights were moderate (~0.5 each), indicating that both modalities contributed evenly – for example, an image showing a park and a caption "A beautiful addition to our city" (here, each modality on its own is somewhat indicative, and together they make a confident classification). This dynamic weighting is precisely what we intended with the attention fusion: the model can "decide" how to fuse information rather than treating image and text blindly as equal or simply concatenating them. Notably, this attention mechanism also helps in cases of modality disagreement. If an image and text seem semantically inconsistent (perhaps an image looks like a celebration event but text complains about a pothole being fixed – a contrived example), the model might identify the conflict. Prior research suggests that even contradictory signals can be informative [6, 21]. In

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

our domain, we did not explicitly encounter contradictory image-text pairs, but the flexible fusion ensures the model isn't thrown off by noise in one channel (it can down-weight an unhelpful modality).

Analysis of attention-weight distributions and misclassification instances indicates that the proposed model maintains a balanced contribution from both visual and textual modalities. Most observed errors correspond to borderline or contextually ambiguous cases, which suggests interpretable behavior rather than systemic bias and further supports the practical applicability of the proposed multimodal framework.

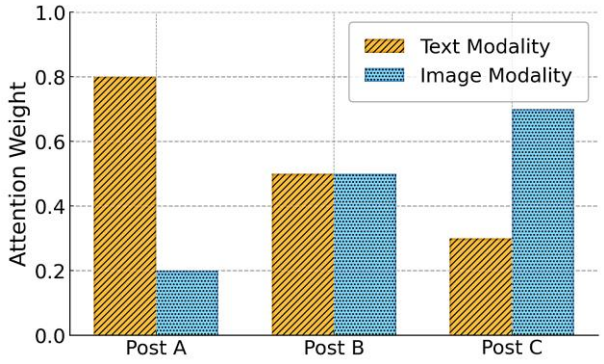


Fig. 5. Attention-weight distribution across sample posts (model's varying reliance on text vs. image features)

For sample post (A, B, C on the x-axis), the bar heights show the attention weights assigned to the text modality (green) vs. the image modality (blue) by the fusion module (Fig. 5). Post A has a higher text weight (the caption strongly signals an improvement), whereas Post C relies more on image features (visual content is more indicative of improvement). Post B shows a balanced attention, with text and image equally contributing. This demonstrates the model's dynamic fusion mechanism: it attends to the modality that best signals the presence of an urban improvement for a given post.

3.3. Qualitative examples and practical implications

Beyond the aggregate metrics, we present qualitative examples of the model's output to illustrate its practical effectiveness. Figure 1 shows an example of a relevant post and Figure 2 an example of an irrelevant post, along with the model's classification for each.

Figure 1 depicts a renovated neighborhood playground with new equipment, a fresh rubberized surface, and landscaped elements, consistent with recent construction in a residential park. The accompanying caption – "The city completed a major playground renovation at our local park – looks amazing!" provides explicit confirmation of the upgrade. The model classifies this post as Relevant (urban improvement) with high confidence ($\hat{y} = 0.95$). In the attention-fusion stage, the image feature $\mathbf{v_I}$ captures cues such as playground structures, resurfacing, and site finishing, while the text feature $\mathbf{v_T}$ emphasizes keywords like "completed," "major," and "renovation." Together, these signals enable a confident decision that the post highlights a civic infrastructure improvement – illustrating the advantage of the multimodal approach.

Figure 2 shows a wide nighttime cityscape photographed from a rooftop, with illuminated high-rise buildings and distant landmarks but no visible signs of construction activity, new amenities, or street-level works. The accompanying caption – "Stunning skyline views from my rooftop. Love this city at night!" – is purely expressive and does not reference any upgrades or projects. The model therefore classifies the post as Irrelevant (not an urban improvement) with probability $\hat{y} = 0.10$. In this case, neither modality provides evidence of improvement: the image depicts a routine panoramic skyline, and the text is generic sentiment. The attention mechanism may place slightly higher weight on the image (α_I) given the minimal informational content of the caption; however, because the visual scene lacks signatures of construction or newly deployed infrastructure, the fused representation v_F does not activate "relevant" features in the classifier. This example highlights the model's ability to filter commonplace urban content, reducing false positives in practical deployment.

Overall, the examples and metrics indicate that the model has learned a meaningful representation of what constitutes an "urban improvement" post. In practice, this could translate to a system where city officials receive an alert or summary whenever a new improvement (park, road repair, public installation) is being discussed or showcased by citizens on social platforms, without having to manually sift through unrelated posts about daily life in the city. The combination of high recall and decent precision means the system would catch most relevant posts while keeping the number of false alerts manageable for a human team to verify.

The results support the hypothesis that multimodal deep learning offers significant advantages over unimodal approaches for this task. By fusing visual and textual information, the model exploits complementary signals – images provide direct visual evidence of physical changes, while text provides explicit context and labels (names of projects, indications of newness or completion). Many urban improvement indicators are inherently multimodal: e.g., a post about a new bridge might show the bridge (image) and mention its opening (text). A unimodal image classifier might confuse it with a generic bridge picture and the unimodal text classifier

might misinterpret or miss the context if the wording is indirect. Together, however, the chance that both modalities miss the signal is much lower. This robustness is reflected in the confusion matrix: the multimodal model drastically reduced false negatives compared to single-modality models, because even if one modality was not informative, the other could still trigger a correct prediction. This corroborates findings in other research that multimodal models can capture subtle correlations and improve recall without sacrificing precision. Furthermore, the use of attention in fusion is an improvement over simpler fusion strategies (like averaging or concatenation) because it introduces a learned adaptability for each instance, which likely contributed to our model's strong performance.

Despite the overall success, there are areas for improvement. The precision, while good, could be higher – about 18% of posts the model flagged as improvements were not actually such. Reducing these false positives could involve incorporating additional context. For example, analyzing user comments or metadata (if available) might help verify if something is truly an improvement. Another approach could be to apply a secondary verification model that focuses on the presence of specific keywords in text (like "opened", "built", "renovation") to double-check high-risk predictions. On the image side, false positives often occurred for images showing generic city scenes: one idea is to integrate an object detection module to specifically look for construction equipment, ribbon-cutting ceremonies, signage like "Grand Opening", etc., within the image. Those could serve as extra features to boost precision.

It is also worth noting that our training data, being derived from Wikipedia, might differ in style from actual social media posts (which can be more informal, with hashtags, etc.). The model may need further **fine-tuning on real social media data** for optimal performance in production. Nonetheless, the current model's architecture is general and can easily be fine-tuned on any new dataset of labeled posts that a city might compile.

In terms of deployment, efficiency is a consideration: the proposed model uses large transformers which are computationally heavy. During inference on a single GPU, processing one post (image+text) took around 50 milliseconds, which is sufficient for a few thousand posts per day but might become a bottleneck for streaming large volumes in real time. Techniques like knowledge distillation or model quantization could be explored to compress the model for faster runtime in a real system [22]. Alternatively, a two-stage approach could be used in practice: a fast filter (e.g., keyword matching or a smaller model) to pre-screen obvious negatives, followed by our model on the remaining candidates for high-precision detection.

Building on this work, future research could explore several avenues. One is to incorporate geolocation data or timestamps from posts (if available) to cluster and track improvements geographically and over time – essentially enriching the model to not just detect individual posts, but also integrate them into a bigger picture of urban development patterns. Another direction is to use multilingual models (leveraging the multilingual aspect of WIT) to handle posts in various languages, since cities often see content in multiple languages. A multilingual BERT could extend coverage [23]. Related multimodal sentiment pipelines combining BERT with CNN backbones indicate transferable design patterns for social media scenarios [20]. Additionally, semi-supervised or unsupervised learning could be used to exploit the vast amount of unlabeled social media data – for instance, using our model's confident predictions to pseudo-label more data and iteratively retrain, or using contrastive learning to better align image and text representations of city-related content.

The proposed multimodal deep learning approach effectively addresses the problem of filtering urban improvement indicators in public visual-text content. The results demonstrate clear benefits of combining image and text analysis with an attention-based fusion. This work lays the groundwork for practical AI systems to assist smart city initiatives by mining the collective intelligence on social platforms, ultimately helping communities stay informed and engaged about improvements in their urban environment.

4. Conclusions

This study addresses the problem of detecting indicators of urban improvement – such as newly constructed or upgraded public infrastructure – within multimodal social media content using deep learning methodologies. The investigation is driven by the growing demand from municipal authorities for automated tools capable of efficiently monitoring and utilizing citizen-generated information related to urban development. Despite considerable progress in multimodal classification, existing literature reveals a lack of specialized frameworks tailored to this domain. To bridge this gap, the study formulates specific research objectives: the design of a multimodal neural architecture, its training on an extensive dataset, and the empirical validation of its performance relative to established baseline models.

A novel multimodal neural network was constructed, integrating a Conv2D–Vision Transformer hybrid image encoder with a BERT-based text encoder, complemented by an attention-guided fusion mechanism and a dropout-regularized classification layer. This architecture was successfully implemented, thereby

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

achieving the first research objective. A key innovation lies in the incorporation of an adaptive attention mechanism that dynamically adjusts the relative weighting of visual and textual features for each input instance, enhancing the model's capacity to process and interpret heterogeneous multimodal content with improved contextual adaptability.

A training corpus comprising image–caption pairs annotated for relevance to urban improvement was constructed using the WIT dataset. The model was trained under a supervised learning framework employing binary cross-entropy as the optimization objective. Leveraging transfer learning through pre-trained components – specifically, the Vision Transformer (ViT) initialized on ImageNet and the BERT language model – facilitated efficient knowledge transfer and mitigated the limitations posed by the comparatively small domain-specific dataset. Consequently, the second research objective was achieved, resulting in a model demonstrating strong generalization capability, as evidenced by its robust performance on previously unseen test data.

The stated research goal was achieved and confirmed through comprehensive evaluation on the held-out test subset, which yielded an accuracy of 85,2%, precision of 82%, recall of 90%, and an F1-score of 86,3%. The proposed multimodal framework substantially outperformed both unimodal baselines, achieving a +7 percentage point improvement over the text-only model and a +13 point improvement over the image-only configuration. Analysis of the confusion matrix revealed a high true positive rate accompanied by an acceptable level of false positives, confirming the model's discriminative reliability. This outcome successfully fulfills the third research objective, empirically demonstrating that multimodal integration yields a clear enhancement in detection effectiveness. Furthermore, qualitative inspection of representative cases illustrated the model's correct classifications for both relevant and irrelevant inputs, thereby providing interpretability and insight into its decision-making behavior.

In conclusion, the study successfully fulfilled its objectives by developing and validating a functional system capable of accurately detecting social media posts related to urban improvements. The findings confirm that deep multimodal learning constitutes an effective approach for filtering civically relevant content within large and noisy data streams. The primary contributions of this research include: the design of a novel attention-based multimodal architecture specifically tailored for urban improvement detection; the adaptation of the WIT dataset to a domain-specific classification task focused on civic infrastructure; and an empirical demonstration of the superiority of multimodal fusion over unimodal baselines in this context.

Beyond its methodological advancements, the work offers practical implications by providing a prototype tool that could assist municipal stakeholders in automatically identifying citizen posts about newly established public amenities or infrastructure maintenance activities. At the same time, it extends the academic discourse by contributing to the emerging intersection of smart city analytics and multimodal artificial intelligence [29].

Several promising directions for further investigation have been identified. Enhancing model precision through the integration of supplementary data modalities or the development of hybrid architectures represents a logical next step. Extending the framework to multilingual and cross-cultural contexts would further increase its generalizability and utility. Additionally, improving interpretability through explainable AI techniques could foster greater transparency and trust in model decisions, particularly in civic applications.

These prospective advancements aim to refine the current system and expand its scope of applicability across diverse urban analytics scenarios. Ultimately, the study demonstrates that multimodal deep learning can be effectively applied to real-world information retrieval challenges within the civic domain, supporting more adaptive, evidence-based urban governance. Continued exploration and deployment of such technologies are encouraged to harness the wealth of visual and textual data generated by citizens, thereby contributing to smarter and more responsive public decision-making.

Contributions of authors: conceptualization, methodology, development of model, software, verification – **Oleksandr Poplavskyi;** formulation of tasks, analysis **– Yuliia Riabchun.**

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

This work was carried out within the applied research project №5 DB-2025 (State Registration No. 0125U001683).

Data Availability

The data for this study are derived from the publicly available WIT dataset [1]. All training and evaluation data have been either included in the paper (in aggregated form) or are available from the original source. The specific subset of WIT used in our experiments, along with labels, can be made available by the authors upon

reasonable request. There are no proprietary or restricted datasets used in this work.

Use of Artificial Intelligence

The authors confirm that they did not use any generative artificial intelligence technologies in the creation of this work. All writing, analysis, and figure generation were performed by the authors using standard software tools and programming, without AI-driven content generation. The AI techniques discussed in the paper (deep learning models) were solely used as the subject of research and for experimental purposes, not for composing the manuscript's text.

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

References

1. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., & Najork, M. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. *arXiv*, 2021, no. 2103.01913. DOI: 10.48550/arXiv.2103.01913.

2. Xu, P., Zhu, X., & Clifton, D.A. Multimodal Learning with Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no. 10, pp. 12113–12133. DOI: 10.1109/TPAMI.2023.3275156.

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv*, 2020, no. 2010.11929. DOI: 10.48550/arXiv.2010.11929.

4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 2021, vol. 139, pp. 8748-8763. Available at: https://proceedings.mlr.press/v139/radford21a.html (accessed 06.11.2025).

5. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. Vision-Language Pre-training: Basics, Recent Advances and Future Trends. *arXiv*, 2022, no. 2210.09263. DOI: 10.48550/arXiv.2210.09263.

6. Cheung, T.-H., & Lam, K.-M. Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing*, 2022, vol. 514, pp. 1-12. DOI: 10.1016/j.neucom.2022.09.140.

7. Kim, W., Son, B., & Kim, I. ViLT: Vision-and-Language Transformer without Convolution or Region Supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 2021, vol. 139, pp. 5583–5594. Available at: https://proceedings.mlr.press/v139/kim21k.html (accessed 06.11.2025).

8. Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S. R., Xiong, C., & Hoi, S. C. H. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. *arXiv*, 2021, no. 2107.07651. DOI: 10.48550/arXiv.2107.07651.

9. Mamyrbayev, O., Pavlov, S., Poplavskyi, O., Momynzhanova, K., Saldan, Y., Zhanegiz, A., Zhumagulova, S., & Zhumazhan, N. Hybrid neural architectures combining convolutional and recurrent networks for the early detection of retinal pathologies. *Engineering, Technology & Applied Science Research*, 2025, vol. 15, no. 4, pp. 25150-25157. DOI: 10.48084/etasr.11521.

10. Matsiievskyi, O., Mazurenko, R., Netreba, A., & Sapaiev, V. Application of neural networks to optimize distributed computing in cloud and edge environments. *IEEE International Conference on Smart Information Systems and Technologies (SIST)*, 2025, pp. 805-809. DOI: 10.1109/SIST61657.2025.11139218.

11. Mamyrbayev, O., Wójcik, W., Pavlov, S., Alimhan, K., Poplavskyi, O., Aitkazina, A., Nykyforova, L.E., & Zhumazhan, N. *Engineering, Technology & Applied Science Research*, 2025, vol. 15, no. 5, pp. 26943-26951. DOI: 10.48084/etasr.12779.

12. Zhang, J., Huang, J., Jin, S., & Lu, S. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, no. 8, pp. 5625-5644. DOI: 10.1109/TPAMI.2024.3369699.

13. Desai, K., Kaul, G., Aysola, Z., & Johnson, J. *RedCaps: Web-curated image-text data created by the people, for the people*. NeurIPS 2021 Datasets and Benchmarks Track, 2021, pp. 1-14. Available at: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/e00da03b685a0dd18fb6a08af0923de0-Paper-round1.pdf (accessed 06.11.2025).

14. OpenAI. GPT-4 Technical Report. *arXiv*, 2024, no. 2303.08774. DOI: 10.48550/arXiv.2303.08774.

15. Liu, H., Yang, B., & Yu, Z. A Multi-View Interactive Approach for Multimodal Sarcasm Detection. *Applied Sciences*, 2024, vol. 14, no. 5, article no. 2146. DOI: 10.3390/app14052146.

16. Poplavskyi, O., Pavlov, S., Zhumazhan, N., Zhanegiz, A., Saldan, Y., Momynzhanova, K., & Wójcik, W. High-performance information technology for processing biomedical big data to enhance the accuracy of computer-aided decision support systems. *Proceedings of SPIE*, 2024, vol. 13400, article no. 134000E. DOI: 10.1117/12.3057444.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

17. Solovei, O., Solovei, B., & Riabchun, Y. An approach to evaluate a classification model to predict a construction object's state. *CEUR Workshop Proceedings*, 2024, vol. 3896, pp. 194-200. Available at: https://ceur-ws.org/Vol-3896/short6.pdf (accessed 06.11.2025).

18. Aftan, S., & Shah, H. A survey on BERT and its applications. *Proceedings of the 2023 20th Learning and Technology Conference (L&T)*, 2023, pp. 161-166. DOI: 10.1109/LT58159.2023.10092289.

19. Pavlov, S. V., Kozhukhar, A. T., Titkov, S. V., Tretiak, I. V., & Nesterenko, V. A. Electro-optical system for the automated selection of dental implants according to their colour matching. *Przegląd Elektrotechniczny – Electrical Review*, 2017, vol. 93, no. 3, pp. 121-124. DOI: 10.15199/48.2017.03.28.

20. Ren, J. Multimodal Sentiment Analysis Based on BERT and ResNet. *arXiv*, 2024, no. 2412.03625. DOI: 10.48550/arXiv.2412.03625.

21. Poplavska, A., Vassilenko, V., Poplavskyi, O., & Casal, D. AI-Based Classification Algorithm of Infrared Images of Patients with Spinal Disorders. *IFIP Advances in Information and Communication Technology*, 2021, vol. 626, pp. 316-323. DOI: 10.1007/978-3-030-78288-7_30.

22. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704-2713. DOI: 10.1109/CVPR.2018.00286.

23. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 2020, pp. 8440-8451. DOI: 10.18653/v1/2020.acl-main.747.

24. Kukharchuk, V.V., Kazyv, S.S., Bykovsky, S.A., Wójcik, W., Kotyra, A., Akhmetova, A., Bazarova, M., & Weryńska-Bieniasz, S. Discrete wavelet transformation in spectral analysis of vibration processes at hydropower units. *Przegląd Elektrotechniczny – Electrical Review*, 2017, vol. 93, no. 3, pp. 65-68. DOI: 10.15199/48.2017.03.16.

25. Wang, B., Li, W., Bradlow, A., Watt, A., Chan, A.T.Y., & Bazuaye, E. Multi-stage multimodal fusion network with language models and uncertainty evaluation for early risk stratification in rheumatic and musculoskeletal diseases. *Information Fusion*, 2025, vol. 120, article no. 103068. DOI: 10.1016/j.inffus.2025.103068.

26. Yu, C., & Wang, Z. Cross-modal evidential fusion network for social media classification. *Computer Speech & Language*, 2025, vol. 92, article no. 101784. DOI: 10.1016/j.csl.2025.101784.

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

28. Dolhopolov, S., Honcharenko, T., Terentyev, O., Savenko, V., & Liashchenko, T. Multi-Stage Classification of Construction Site Modeling Objects Using Artificial Intelligence Based on BIM Technology. *Proceedings of the 35th Conference of Open Innovations Association FRUCT*, 2024, pp. 179–185. DOI: 10.23919/FRUCT61870.2024.10516383.

29. Chernyshev, D., Ryzhakova, G., Honcharenko, T., Petrenko, H., Chupryna, I., & Reznik, N. Digital Administration of the Project Based on the Concept of Smart Construction. *Lecture Notes in Networks and Systems*, 2023, vol. 495, pp. 1316-1331. DOI: 10.1007/978-3-031-08954-1_114.

30. Zhou, Q., Zhang, J., & Zhu, Z. Evaluating Urban Visual Attractiveness Perception Using Multimodal Large Language Model and Street View Images. *Buildings*, 2025, vol. 15, no. 16, article no. 2970. DOI: 10.3390/buildings15162970.

31. Liu, T., Chen, H., Ren, J., Zhang, L., Chen, H., Hong, R., Li, C., Cui, W., Guo, W., & Wen, C. Urban Functional Zone Classification via Advanced Multi-Modal Data Fusion. *Sustainability*, 2024, vol. 16, no. 24, article no. 11145. DOI: 10.3390/su162411145.

32. Cheng, M., Jin, H., Zhao, Q., Wang, Y., Wu, Y., Huang, S., & Yue, W. Deep learning for optimizing urban governance by "sensing-processing-responding" cycle: Recent advances, future prospects and challenges. *Sustainable Cities and Society*, 2025, vol. 135, article no. 106994. DOI: 10.1016/j.scs.2025.106994.

33. Sburlan, D.-F., & Bucos, M. A Multimodal Deep Learning Approach for Analyzing Content Preferences on TikTok Across European Technical Universities Using Media Information Processing System. *Electronics*, 2026, vol. 15, no. 6, article no. 1288. DOI: 10.3390/electronics15061288.

34. Dufitimana, E., Bizimana, J. P., Uwayezu, E., Gahungu, P., & Mugisha, E. Multimodal Deep Learning Framework for Profiling Socio-Economic Indicators and Public Health Determinants in Urban Environments. *Urban Science*, 2026, vol. 10, no. 4, article no. 177. DOI: 10.3390/urbansci10040177.

Received 13.10.2025, Received in revised form 12.12.2025
Accepted date 15.01.2026, Published date 22.01.2026

МУЛЬТИМОДАЛЬНЕ ВИЯВЛЕННЯ ІНДИКАТОРІВ МІСЬКОГО БЛАГОУСТРОЮ В ПУБЛІЧНОМУ ВІЗУАЛЬНО-ТЕКСТОВОМУ КОНТЕНТІ З ВИКОРИСТАННЯМ ГЛИБИННОГО НАВЧАННЯ

О. А. Поплавський, Ю. В. Рябчун

Предметом статті є автоматизоване виявлення контенту, створеного користувачами, про поліпшення міської інфраструктури та будівництво в дуже великих потоках соціальних мереж і групових повідомлень, із фокусом на фільтрацію повідомлень, придатних для практичних дій, з-поміж надмірного фонового шуму, щоб фахівці могли першочергово бачити те, що потребує відновлення або інвестування. Метою є проєктування, реалізація та валідація мультимодальної системи глибинного навчання, яка отримує будь-яке зображення разом із супровідним текстом і визначає, чи є ця пара релевантною тематиці міських поліпшень, зокрема доріг, парків, будівель, освітлення, чистоти, доступності чи дитячих майданчиків, а також надання практичного фільтра контенту, що зменшує ручне сортування для муніципальних команд. До завдань належать: базувати навчання на єдиному великомасштабному публічному наборі даних замість збирання нових даних, чітко визначити ціль класифікації «релевантно / нерелевантно», побудувати мультимодальну нейронну архітектуру, що навчається як на візуальному контексті, так і на текстовому описі, та оцінити підхід порівняно з одномодальними базовими моделями в однакових умовах. Використані методи спираються на набір даних Wikipedia-based Image Text (WIT), який містить понад тридцять сім мільйонів прикладів «зображення+текст» з описами, багатьма мовами та доступний для завантаження на платформі Hugging Face, а також на злитті основи для зображень на базі згорткової мережі або візуального трансформера з трансформерним енкодером для тексту, що разом формують єдиний класифікатор, навчений у стандартній парадигмі з учителем. **Висновки.** Експерименти показують, що запропонований підхід точно ідентифікує повідомлення, пов’язані з міськими поліпшеннями, і що мультимодальна конструкція чітко перевершує одномодальні базові моделі за тих самих умов, що підтверджує переваги поєднання візуальних доказів із текстовим контекстом. Наукова новизна полягає у застосуванні мультимодального глибинного навчання до специфічної проблеми фільтрації контенту в режимі реального часу в галузі міського планування, у формулюванні цілі як виявлення релевантної інформації громадянського змісту в публічних візуально-текстових комунікаціях із використанням єдиного загальнодоступного набору даних для навчання, а також у демонстрації того, що проста, але методично-обґрунтована схема злиття комп’ютерного зору та обробки природної мови здатна виокремлювати практично корисну інформацію з дуже великих обсягів онлайн-повідомлень без ручного збирання даних.

Ключові слова: мультимодальне навчання; міський благоустрій; соціальні мережі; класифікація пар «зображення–текст»; глибинне навчання; візуальний трансформер.

Поплавський Олександр Анатолійович – д-р техн. наук, доц., проф. 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.

Oleksandr Poplavskyi – Doctor of Technical Sciences, Associate Professor, Professor of the Department of Information Technologies, Kyiv National University of Construction and Architecture (KNUCA), Kyiv, Ukraine,
e-mail: apoplavskyi@gmail.com; ORCID: 0000-0003-0465-6843; Scopus Author ID: 57216831949

Yuliia Riabchun – PhD, Associate Professor, Department of Information Technologies, Kyiv National University of Construction and Architecture (KNUCA), Kyiv, Ukraine,
e-mail: riabchun.yv@knuba.edu.ua, super.etsy@ukr.net; ORCID: 0000-0002-8320-4038; Scopus Author ID: 57211403226.