Integrating Graph Learning and Multimodal Fusion for Detecting Misinformation on Social Platforms

Abstract

Nowadays, modern social networks allow the rapid sharing of news worldwide. Alas, these news are frequently unverified or shared on the basis of users' opinions or beliefs, which can cause confusion widespread, public trust erosion, and contribution to social and political instability. In this complex and evolving scenario, the early detection of fake news has become a critical issue. Multimodal approaches, which integrate various data types such as text, images, audio, video, and network structures, have shown promising results in addressing such a problem. The literature presents different fusion strategies, but there is no consensus on which one is the most effective. In this work, we propose *M3DUSA*, a modular multimodal framework able to combine different modalities to effectively detect malicious and misleading content. By using deep attention-based architectures, our framework discovers informative latent representations that can be combined using different early or late fusion strategies. Experiments conducted on a real-world dataset demonstrate the effectiveness of our solution. The achieved results highlight that while both early and late fusion approaches can effectively exploit the complementary contributions from different modalities, they can exhibit distinct advantages depending on the desired outcomes.

CCS Concepts

• **Information systems** → **Web mining**; **Social networks**; *Data mining*; • **Computing methodologies** → *Neural networks.*

Keywords

Multimodal Fake News Detection, Deep Fusion Methods, Social Networks, Heterogeneous Information Networks

ACM Reference Format:

. 2025. Integrating Graph Learning and Multimodal Fusion for Detecting Misinformation on Social Platforms. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Motivations and challenges. In today's world, social media has transformed how information is shared globally. Platforms like *Twitter, Facebook, and Instagram* are commonly used for sharing news worldwide. However, the information shared through these platforms often lacks of verification and is open to personal interpretation. In this complex and evolving scenario, early detecting fake news on social media has become a critical challenge [7], given its far-reaching societal impacts and the rapid dissemination of information across platforms.

In recent years, Artificial Intelligence (AI) has proven to be a valuable ally in mitigating the effects of this phenomenon. Various approaches based on Deep Learning (DL) techniques have been proposed for identifying malicious or unverified content on the Web [19]. In particular, there has been growing interest in developing multimodal approaches that can simultaneously leverage multiple types of data. Indeed, unimodal approaches can fail to capture the full complexity of fake news, as they rely on a single type of data or input limiting their ability to detect the complete range of deceptive signals [1].

Although multimodal approaches have shown promising results, the effective combination of multiple modalities with their different structures and dimensions remains an open challenge [4]. Both *early* and *late* fusion approaches have been proposed in the literature for integrating multiple modalities. In early fusion, data from different modalities are combined at the input level so that each modality can benefit from a joint representation; by contrast, in late fusion, the outputs of individual models are combined at the bottom of the overall architecture (typically, just before the classification layer). While the former can effectively capture cross-modal interactions, the latter allows for learning more specialized and effective models. As a result, there is no consensus on adopting one strategy over the other [4]. In addition, many approaches heavily rely on multimodal data combining text and images, audio and/or video; however, in many social media environments, frequently the posts don't include media content, making textual analysis and social network structure – providing critical context for understanding how information is propagated within the network – pivotal for tasks like claim classification. Finally, the imbalanced class distribution of data can further affect the performance of the learned detectors.

Contribution. To address the aforementioned issues, we proposed *M3DUSA* - a **M**odular **M**ulti-**M**odal **D**eep f**US**ion **A**rchitecture for Fake News Detection. *M3DUSA* enables the exploitation of multiple modalities, with a primary focus on text and social network structure. Our solution integrates Attention-based architectures and allows for the effective combination of latent representations extracted from Large Language Models (LLMs) belonging to the BERT (Bidirectional Encoder Representations) family [5] for text and from Graph Attention Networks (GATs) for the social network structure. The fusion mechanism is a configurable parameter of *M3DUSA*, integrating both early and late fusion strategies. In addition, to cope with the class imbalance problem, *M3DUSA* adopts a weighted version of the standard binary cross-entropy. The effectiveness of *M3DUSA* is assessed through extensive experimentation conducted against a real-world dataset named MuMiN [14], a recent comprehensive benchmark specifically created for the task of multi-modal fake news detection.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Organization of the paper. The remainder of this paper is structured as follows. Section 2 discusses recent works particularly related to ours. Section 3 introduces *M3DUSA* framework and describes its base components. Section 4 describes the dataset employed in our analysis. Section 5 discusses our experimental results. Finally, Section 6 concludes the work and provides pointers for future research.

2 Related work

In this section, we review the deep multimodal frameworks proposed in the literature for the Fake News Detection problem that most closely resemble our approach, highlighting the modalities utilized and the fusion strategies adopted. Interestingly, only a limited number of works focuses on the combination of text and social information. For instance, in [6] the authors employ BERT-based techniques to represent user historical posts and news information, and a Graph Neural Network (GNN) to build a joint user engagement embedding before the final concatenation. [13] encodes news content features using BERT, article's comments using a hierarchical attention network, and user-news interactions using a Feed-forward Neural Network (FNN); the final classification is performed after concatenation and an additional FNN. Both aforementioned approaches define a unique, specific late fusion strategy, and they rely on historical or sequential data, resp. A preliminary work [2] uses word embeddings to extract latent representations of news articles, captures their contextual similarities via a graph-based representation scheme, and performs the final classification leveraging an additional Graph Convolutional Network (GCN) or an attention-based GNN. In contrast, our approach relies on a single graph neural network architecture and explicitly integrates multiple modalities. [16] employs a hierarchical attention mechanism to perform node representation learning in a Heterogeneous Information Network (HIN) with multiple node and relations types, but focuses on a single modality and exploits an additional active learning framework to enhance learning performance. Another preliminary work [10] addresses social networks' misinformation checking from a modality-level explainable perspective, by concatenating the shallow metadata of the tweet-nodes and tweet's textual content to produce the final tweet-node encoding for classification. Like our approach, it employs a GAT for the HIN structure encoding leveraging multiple types of information. Compared to the above-discussed works, we mainly focus on designing a modular multimodal framework that is able to exploit interchangeably early and late fusion approaches. In particular, differently from [10] which is the closest study to us, we carry out a more extensive evaluation, investigating multiple data configurations and early/late fusion strategies.

3 *M3DUSA* framework

In this section, we provide a comprehensive description of the proposed DL-based framework for identifying malicious and misleading information by combining text data and social network structure. Our framework is designed to operate with both early and late fusion approaches: *(i)* in the early fusion approach, the node attributes of a heterogeneous graph are initialized with a combination of textual, image, numeric, and categorical information to effectively embed all modalities into a unified graph structure; *(ii)* in the late fusion approach, specialized models are trained on different modalities, and their outputs are combined using various techniques. In the following, we first illustrate the building blocks composing *M3DUSA*, and then present the fusion mechanisms integrated within the system. An overview of *M3DUSA* architecture is depicted in Figure 1.

3.1 Building blocks

As mentioned above, *M3DUSA* focuses on two modalities, i.e., textual content and social network structure, in a highly unbalanced dataset with multiple social relationships and numerous textual attributes associated with nodes. Below, we detail the base components to extract the encoding for text and social structure.

Text Encoding. For the text modality, two kinds of input are considered: claims and tweets discussing them. Both data are processed using BERT-like architectures [5]. Basically, it is a Transformer model used to address natural language processing tasks. The training process of BERT encompasses two key phases: *Word Masking* (WM) and *Next Sentence Prediction* (NSP). In the WM step, a percentage of words within a sentence are masked or randomly replaced. The goal is to predict the masked words. In the second step, the BERT model is fine-tuned to capture the relationships between two subsequent sentences. This process includes generating negative examples by replacing the second sentence with a random one. In particular, to encode our text data, we used two specialized instances of BERT: the content of tweets leverages TwHIN-BERT [18], a multi-lingual Tweet language trained model that integrates both text-based self-supervision and a social objective. The encoding of claims is obtained by using mp-net[1] [15], a sentence-transformer trained by adopting a Siamese architecture.

Notably, for both cases, the outputs are further fine-tuned by processing the encoded representation through an unsupervised autoencoder (AE) architecture over the specific input dataset.

Social Network Structure Encoding. For the social network modality, we employed a Graph Attention Network [3] on a heterogeneous graph, leveraging the adaptive weighting of neighboring nodes' contributions and of the relationship importance during the aggregation process to effectively capture the complex interactions within the network. Given the heterogeneous graph $G = \langle \mathcal{V}, \mathcal{E}, A, R, \phi, \varphi, \rangle$, where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges, A and R are the sets of node and relation types, with $|A| + |R| > 2$, $\phi : \mathcal{V} \rightarrow A$ and $\varphi : \mathcal{E} \rightarrow R$ are the node- and edge-type mapping functions, resp., the updated node representation $\mathbf{h}_i^{(l+1)}$ for each node $i \in \mathcal{V}$ at each layer l is computed by aggregating over all relation types:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{r \in R} \sum_{j \in \mathcal{N}_r(i)} \alpha_{ij,r}^{(l)} \mathbf{W}_r^{(l)} \mathbf{h}_j^{(l)}\right)$$

where $\mathcal{N}_r(i)$ denotes the neighborhood of node i under relation r, i.e., all nodes connected to i via an edge of type r; $\mathbf{W}_r^{(l)}$ is the learnable weight matrix for layer l specific to relation type $r \in R$;

[1] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

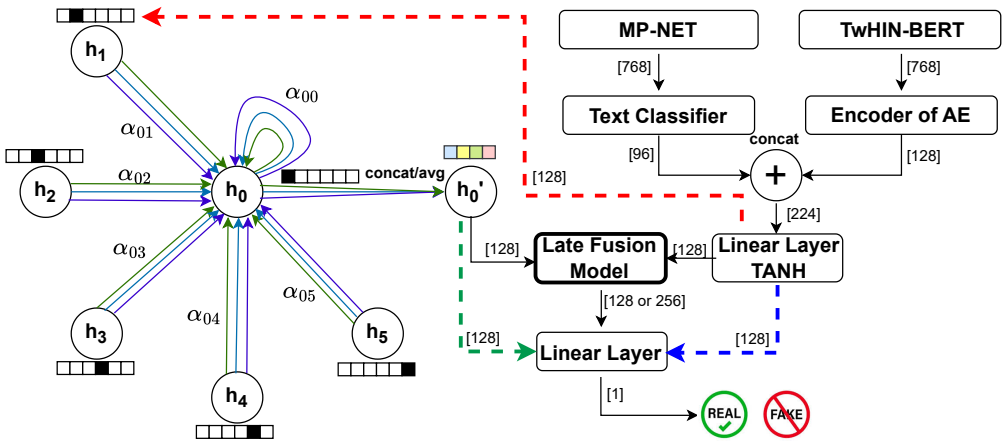

Figure 1: *M3DUSA* **framework learning flow. The base embeddings, yielded by BERT-like models and the GAT module, are combined using the Late Fusion. The dashed red arrow indicates the initialization of node embeddings with text latent representations in the Early Fusion. The dashed green and blue arrows indicate the ablation study w.r.t. individual modalities.**

$\alpha_{ij,r}^{(l)}$ is the relation-specific attention coefficient between i and j at layer l, calculated as follows:
$\alpha_{ij,r}^{(l)} = \text{softmax}_{j \in \mathcal{N}_r(i)} \left(\text{LeakyReLU} \left(\mathbf{a}_r^{(l)T} \left[\mathbf{W}^{(l)} \mathbf{h}_i^{(l)} \odot \mathbf{W}^{(l)} \mathbf{h}_j^{(l)} \right] \right) \right)$,
where $\mathbf{a}_r^{(l)}$ is the relation-specific learnable attention vector, and $\odot$ denotes element-wise multiplication.

Initial node features are generated using identity matrices so to ensure that each node's unique characteristics are preserved while focusing on the structural relations represented by the edges. In the initial stage, we performed our classification task based solely on the direct interactions between nodes. Hence, to capture more nuanced patterns and indirect relationships, we extended our approach to include *meta-paths* as additional relations between the terminal nodes of the sequences, representing higher-order connections within the graph.

3.2 Multimodal fusion

Here, we define how the learned embeddings are combined. In more detail, we denote the text embedding as $\mathbf{e_{text}} \in \mathbb{R}^{128}$ and the social network embedding as $\mathbf{e_{net}} \in \mathbb{R}^{128}$.

Late Fusion. Each late fusion technique combines the embeddings learned from each modality to create a joint representation, denoted as $\mathbf{e_f} \in \mathbb{R}^{\text{dim}}$, which is subsequently used to feed a simple linear classifier for fake news detection $\hat{y} = \mathbf{W}\mathbf{e_f} + \mathbf{b}$, where $\mathbf{W} \in \mathbb{R}^{C \times \text{dim}}$ and $\mathbf{b} \in \mathbb{R}^C$ are the weight matrix and bias vector of the classifier, respectively, $C = 2$ is the number of classes and *dim* is the dimension of the fused embedding.

The employed late fusion methods can be broadly categorized based on how they treat the individual modality embeddings into *Equal Contribution Methods* (ECM) and *Adaptive Contribution Methods* (ACM). The methods belonging to the first group – `Concatenation`, `Average Pooling`, and `Max Pooling` – handle both embeddings equally without assigning explicit importance weights. Basically, they assume an equal contribution from the different components. While they are computationally efficient, they can fail to capture the importance of each modality. Differently, the other fusion techniques, such as `Weighted Fusion`, `Attention-Based Fusion`, `Gated Fusion`, and `Bilinear Fusion`, employ mechanisms to modulate in an adaptive way the contributions of each modality. These strategies can achieve more effective performance at the cost of being more computationally expensive.

In Table 1, we summarized how the fusion strategies work. In particular, $[\cdot;\cdot]$ denotes the concatenation operation, and w, $\mathbf{b}$ and $\mathbf{W}$ are learnable parameters. For the attention-based strategy, α_{text} and α_{net} are defined as follows:

$$\alpha_{text} = \frac{\exp(\mathbf{W_a}^\top \mathbf{e_{text}})}{\exp(\mathbf{W_a}^\top \mathbf{e_{text}}) + \exp(\mathbf{W_a}^\top \mathbf{e_{net}})}$$

$$\alpha_{net} = \frac{\exp(\mathbf{W_a}^\top \mathbf{e_{net}})}{\exp(\mathbf{W_a}^\top \mathbf{e_{text}}) + \exp(\mathbf{W_a}^\top \mathbf{e_{net}})}$$

while the gate component g is defined as: $g = \sigma\left(\mathbf{W_g}[\mathbf{e_{text}}; \mathbf{e_{net}}] + \mathbf{b_g}\right)$.

Early Fusion. In *M3DUSA*, the GNN module is used to implement the early fusion by initializing the nodes with the embeddings learned according to the other modalities. Here, text encodings are embedded as additional node features.

Moreover, different preprocessing procedures are performed for the other feature types (e.g., MinMax feature scaling for numeric values and one-hot encoding for categorical ones) to make them suitable for the learning phase. SentenceBERT [15] is used to handle short text and sparse categorical textual attributes (e.g., keywords or hashtag tags). Hashtags are further processed by performing the clustering algorithm HDBSCAN [11] on the tag encodings to highlight semantic similarities. As regards image features associated with some tweets, they are processed by a ResNet architecture[8], a convolutional neural network including skip connections to boost performance and mitigate the vanishing gradient problem.

The initial embedding of each node i is thus obtained by concatenating K_a encodings, with K_a equal to the number of attribute

Table 1: Late Fusion strategies and their formal definitions.

Type	Strategy	Formal Definition
ECM	Concatenation	$\mathbf{e_f} = [\mathbf{e_{text}}; \mathbf{e_{net}}] \in \mathbb{R}^{256}$
	Average Pooling	$\mathbf{e_f} = 1/2\,(\mathbf{e_{text}} + \mathbf{e_{net}}) \in \mathbb{R}^{128}$
	Max Pooling	$\mathbf{e_f}[i] = \max(\mathbf{e_{text}}[i], \mathbf{e_{net}}[i]),\ \forall i \in \{0, \ldots, 127\}$
ACM	Weighted Fusion	$\mathbf{e_f} = w_{text} \cdot \mathbf{e_{text}} + w_{net} \cdot \mathbf{e_{net}} \in \mathbb{R}^{128}$
	Attention-Based Fusion	$\mathbf{e_f} = \alpha_{text} \cdot \mathbf{e_{text}} + \alpha_{net} \cdot \mathbf{e_{net}} \in \mathbb{R}^{128}$
	Gated Fusion	$\mathbf{e_f} = g \odot \mathbf{e_{text}} + (1-g) \odot \mathbf{e_{net}} \in \mathbb{R}^{128}$
	Bilinear Fusion	$\mathbf{e_f} = \mathbf{W_b}(\mathbf{e_{text}} \otimes \mathbf{e_{net}}) \in \mathbb{R}^{128}$

Table 2: Dataset statistics, in terms of no. of nodes, edges and meta-paths.

# Nodes	Claim (C)	2168
	Tweet (T)	4340
	Reply (R)	195459
	User (U)	153168
	Hashtag (H)	28091
	Image (I)	1020
	Article (A)	1453

# Meta-paths	C-T-U-T-C	28867
	C-T-H-T-C	21577
	C-T-R-T-C_r	2859
	C-T-R-T-C_q	3042

# Edges	T *discusses* C	5081
	R *reply_to* T	90196
	R *quote_of* T	101216
	T *has_hashtag* H	2289
	T *has_article* A	1898
	T *has_image* I	1028
	T *mentions* U	1119
	U *posted* T	4091
	U *posted* R	179247
	U *retweeted* T	13402
	U *follows* U	18379
	U *mentions* U	2797
	U *has_hashtag* H	50451

of node type $a \in A$: $\mathbf{h}_i^0 = \left[\mathbf{e}_i^{(1)} \| \mathbf{e}_i^{(2)} \| \ldots \| \mathbf{e}_i^{(K_a)}\right]$, where $\mathbf{e}_i^{(k)}$ represents the k-th attribute encoding of node i. Here, PCA can be additionally used to ensure that the embeddings across all node types have the same size.

The GAT attention mechanism enables the combination of all modalities.

4 Case study: MuMiN Dataset

We conducted our experiments on the MuMiN-small dataset, modeled as a Heterogeneous Information Network, with multiple node and edge types and external information associated with nodes available as a set of attributes. A fully description of the dataset is provided in [14].

Classification is performed w.r.t. nodes of type Claim (C), of which 93% are labeled as *misinformation*, while only 7% are labeled as *factual*. To enhance classification performance by capturing richer structural and semantic relationships, we build 4 different *meta-paths* toward the Claim node type, which can be seen as additional high-order relationships connecting nodes of that type: C-T-U-T-C, C-T-H-T-C, C-T-R-T-C_r and C-T-R-T-C_q, i.e., we are interested in pairs of claims discussed by tweets written by the same user, having the same hashtag, belonging to the same conversation thread as replies or quotes, respectively. Table 2 shows the dataset statistics in terms of number of nodes, number of edges and number of meta-path instances for each type.

Each node type is associated with categorical, numeric, and/or textual attributes. Claim attributes include an original embedding, the language, keywords extracted from the text, the membership in specific clusters (topics) along with associated keywords, and detailed information about the review/categorization process. For Tweet and Replies, the language of the content, the creation date, the source platform, the textual content itself and various popularity measures like number of retweets, replies and quote tweets (crucial for understanding the spread and engagement of content on social media) are provided. User nodes are enriched with attributes like a textual description, the location, the date the profile was created, and indicators of the status (e.g., whether they are verified or have a protected account); additionally, popularity metrics such as the number of followers, followees, posted tweets, and listings are included. Hashtag nodes are provided with the corresponding tag, Image nodes with their width, height and pixel vectors, Article nodes with title and content.

5 Experimental results

In this section, we assess the effectiveness of *M3DUSA*. First we define the experimental setting and then we discuss the achieved results.

In particular, we pose the following set of interrelated research questions:

RQ1. How informative are the different modalities, i.e, the text of social media posts discussing claims and the underlying social network structure?

RQ2. Do the learned embeddings effectively encode the semantics, separating the classes correctly?

RQ3. Do early and late fusion strategies exhibit different behaviors? Do they help avoiding biases towards the majority class in a unbalanced setting?

RQ4. To what extent the choice of training data impact on the classification task, especially considering the class imbalance and the topics variety?

5.1 Experimental setup and evaluation metrics

We conducted our experiments on a NVidia DGX Station featuring 4 GPU V100 32GB. For both modalities, we trained the model using the Adam optimization algorithm [9] with full batch size for 5 independent runs, differing in the train-val-test split, and we set the optimal hyperparameters via grid search algorithm. For the text modality, we employed a binary cross entropy and a mean squared error loss function for the text classifier and the autoencoder, resp., and trained both models for 20 epochs with learning rate set to 0.0001. The text classifier is a neural network consisting of three linear layers interleaved by ReLu, pretrained by using GossipCop[2] and CoAid[3]. For the network modality, we employed a 3-layer

[2]https://github.com/KaiDMML/FakeNewsNet
[3]https://github.com/cuilimeng/CoAID

Table 3: Performance comparison of different models. The best results are highlighted in bold, the second best are underlined.

Approach	Model	F1-micro	F1-macro	ROC-AUC	Precision_0	Recall_0
Baseline	MuMiN	0.837±0.014	0.573±0.017	0.709±0.044	0.184±0.025	0.358±0.098
	MuMiN_{128}	0.749±0.019	0.526±0.016	0.667±0.033	0.133±0.019	0.428±0.055
Unimodal	Text	0.899±0.012	0.699±0.076	0.838±0.083	0.367±0.087	0.617±0.259
	Net	0.917±0.038	0.780±0.064	0.850±0.032	0.562±0.120	0.580±0.049
	Net+mps	0.954±0.007	0.788±0.117	0.864±0.116	0.721±0.139	0.647±0.191
Early Fusion	EF	0.961±0.003	0.856±0.011	0.854±0.021	0.740±0.039	0.729±0.046
	EF_{256}	0.962±0.002	**0.860±0.007**	0.860±0.015	**0.745±0.031**	0.739±0.032
Late Fusion	Concat	0.959±0.011	0.824±0.136	0.922±0.083	0.674±0.103	0.725±0.339
	Avg pool	0.961±0.011	0.833±0.125	0.927±0.078	0.700±0.090	0.733±0.321
	Max pool	0.961±0.012	0.833±0.126	0.930±0.050	0.709±0.090	0.721±0.314
	Weigh LF	0.958±0.010	0.825±0.121	0.926±0.080	0.678±0.074	0.727±0.318
	Attn LF	0.956±0.014	0.810±0.133	0.920±0.052	0.666±0.116	0.679±0.327
	Gated LF	**0.964±0.014**	0.836±0.145	0.910±0.067	0.704±0.159	0.718±0.335
	Bil LF	0.961±0.009	0.856±0.062	**0.932±0.035**	0.727±0.072	**0.775±0.211**

GAT [3] architecture with hidden channels dimension set to 128 and out channel dimension set to 2 as the number of classes. We employed a weighted cross entropy loss function, with weights inversely proportional to the class frequency. We trained the model for 200 epochs, with learning rate equal to 0.005, weight decay equal to 0.0005 and dropout set to 0.4. The initial node features are initialized with the identity matrix for each node type. For the early fusion approaches, we used the same (hyper-)parameters as the network modality, with node features initialized using the attribute vectors; when using PCA, the dimension is set to 256 for all node types. For a fair comparison, we used the same final classifier in all experiments, defined as a simple linear layer to map the embeddings to class scores and trained independently for 50 epochs using Adam with batch size equal to 32 and learning rate set to 0.01. The dimension of all the learned embeddings is equal to 128, except for the concatenation strategy which is double. The threshold of the HDBSCAN algorithm is 0.75. For the UMAP projection, we set *nearest neighbors* and *minimum distance* hyperparameters equal to 5 and 0.1, resp., to control the balance between local and global structure. To assess the effectiveness of *M3DUSA*, we adopted well-known performance metrics, i.e., F1-micro, F1-macro, and ROC-AUC [17], particularly suitable for imbalanced scenarios, and delve into Precision and Recall of the minority class.

5.2 Research findings

Our experimental results are summarized in Table 3. For each experiment, we report the average and standard deviation of 5 independent runs. For each run, we randomly extracted three data samples – training, validation, and test sets – respectively corresponding to 60%, 15%, and 25% of the whole dataset.

RQ1. First, we compared the results achieved by *M3DUSA* in the unimodal setting with the baseline, i.e., the original MuMiN claim embeddings. The proposed models demonstrate improvements across all evaluation metrics, highlighting the quality of the learned embeddings for both text and network modalities. The poor baseline performance worsens after applying dimensionality reduction with PCA to match the dimension of our learned embeddings. In the unimodal setting, text-based classification led to poor precision for the minority class, with many false positives, mislabeling factual claims as misinformation. Network topology-based classification improved overall performance, especially for minority class precision, with fewer misclassifications. The inclusion of meta-paths further enhanced these results, indicating their contribution in improving accuracy. However, the network modality was more sensitive to the training data used, as shown by a higher standard deviation in performance metrics. Text encoding is faster than network encoding; adding meta-paths does not significantly impact execution time.

RQ2. We performed a qualitative analysis of the learned embeddings by visualizing their relative proximities using Uniform Manifold Approximation and Projection (UMAP) [12] algorithm in order to better understand the model capabilities in separating classes. Figure 2 shows the two-dimensional UMAP visualization of the initial embeddings as provided by the MuMiN dataset (Fig. 2a), of the intermediate embeddings learned under text (Fig. 2b) and network modality (Fig. 2c), and of the final embeddings learned under early (Fig. 2d) and late fusion (Fig. 2e). In the initial representation (Fig. 2a), all claims are grouped closely together regardless of their class, resulting in a cluttered representation. Nonetheless, Figs. 2b and 2c show how even in the unimodal setting the learned embeddings allow UMAP to better separate entities of different classes; the separation is more pronounced when looking at the fused embeddings, as shown in Figs. 2d and 2e.

RQ3. As regards RQ3, we investigated the benefits of early and late fusion strategies in the multimodal setting. Early fusion strategies, which initialize node embeddings with meaningful features, improved performance, and enhanced stability, especially when PCA was applied. Late fusion strategies exhibit good performances, with adaptive techniques outperforming equal contributions, though the Attention-based fusion underperformed due to the small dataset size. Although not included in the main table due to space constraints, we observed consistently high precision and recall scores

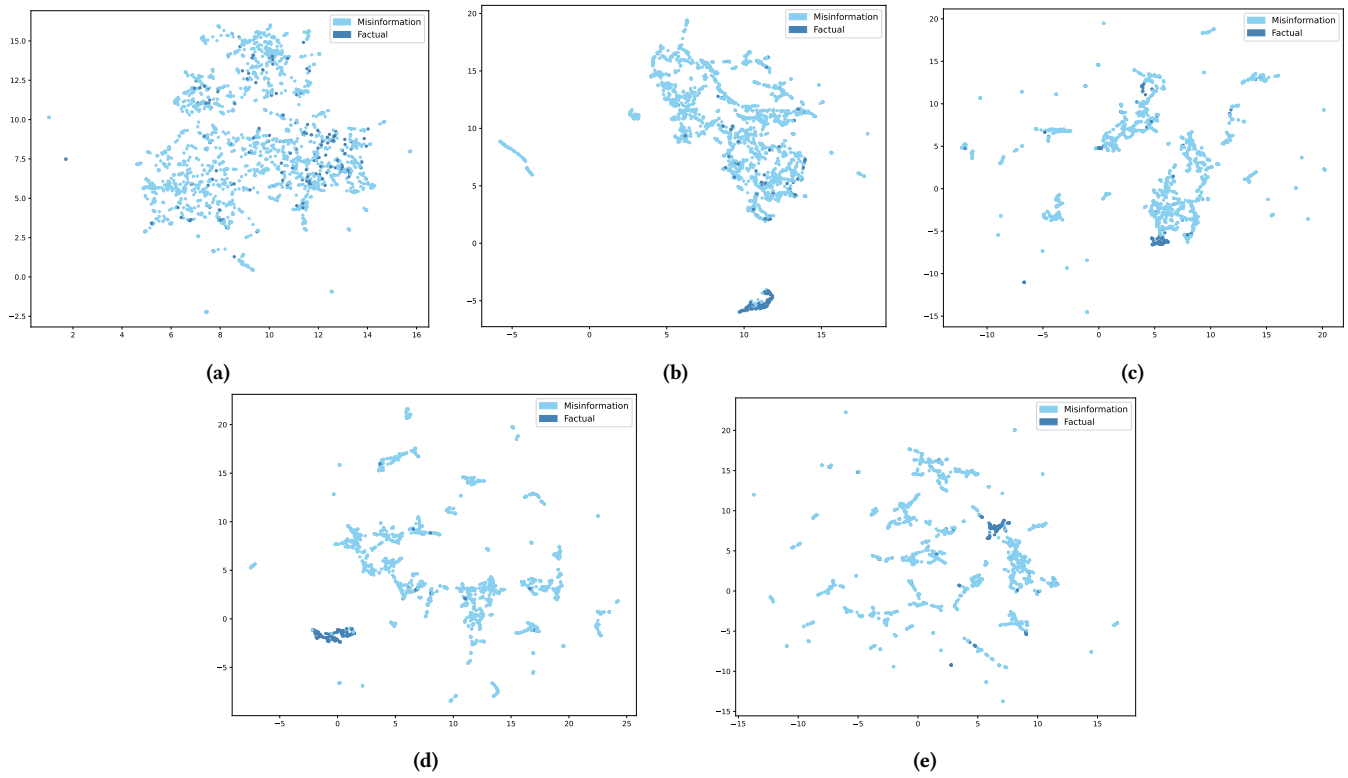

Figure 2: UMAP 2D visualization of claim embbeddings (baseline (a), only text (b), only network (c), early fusion (d) and late fusion (e)), with dark and light blue points indicating the verdict of the claim is factual or misinformation, resp.

for misinformation class across all fusion strategies, with the lowest values for the Attention-based model being 0.9771±0.0182 and 0.9756±0.0117, resp. These results also confirm the effectiveness of the adopted strategies for this unbalanced environment.

In more detail, the best results came from early fusion strategies and advanced late fusion models (specifically, Gated and Bilinear), with the first offering higher stability and the latter achieving the best performance on single experiments. Advanced late fusion techniques may thus excel when training data can be manually selected, but early fusion is recommended for more robust and generalizable models. We recall that, for a fair comparison, the early fusion approach also incorporates the (limited) number of images, hashtags, and articles associated with tweets discussing claims.

RQ4. Finally, we analyzed whether variations in class and topic (cluster) distribution across random splits can affect the detector's performance. As shown in Figure 3, there is no significant variation as the split data changes, suggesting a strong dependence on the specific data rather than its distribution.

6 Conclusion

Mitigating the diffusion of malicious or unverified content through social media is a critical challenge. Effectively combining different types of information is crucial for developing accurate and reliable fake news detection systems. In this work, we presented *M3DUSA*, a framework that enables the combination of text and social data by leveraging various Deep Fusion methods. The results achieved on a real-world dataset demonstrate the effectiveness of the proposed solution.

In the future, we are interested in detecting crucial subgraph structures and node features influencing the predictions via GNN explainers, and in investigating how explanation methods could be integrated into a late fusion setting to support decision-making by operators. Additionally, the use of Adversarial Learning techniques could help extract more robust (cross-domain) features.

References

[1] Sara Abdali, Sina shaham, and Bhaskar Krishnamachari. 2024. Multimodal Misinformation Detection: Approaches, Challenges and Opportunities. arXiv:2203.13883 [cs.LG]
[2] Adrien Benamira, Benjamin Devillers, Etienne Lesot, Ayush K Ray, Manal Saadi, and Fragkiskos D Malliaros. 2019. Semi-supervised learning and graph neural networks for fake news detection. In *Proc. of IEEE/ACM Intl. Conf. on Adv. in Soc. Net. Anal. and Min.* 568–569.
[3] Shaked Brody, Uri Alon, and Eran Yahav. 2021. How Attentive are Graph Attention Networks? *CoRR* abs/2105.14491 (2021). arXiv:2105.14491
[4] Carmela Comito, Luciano Caroprese, and Ester Zumpano. 2023. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining* 13, 1 (2023), 101.
[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186. doi:10.18653/v1/N19-1423
[6] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proc. of the 44th Intl. ACM SIGIR Conf. on research and development in inf. retr.* 2051–2055.
[7] Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. 2023. A review of fake news detection approaches: A critical analysis of relevant

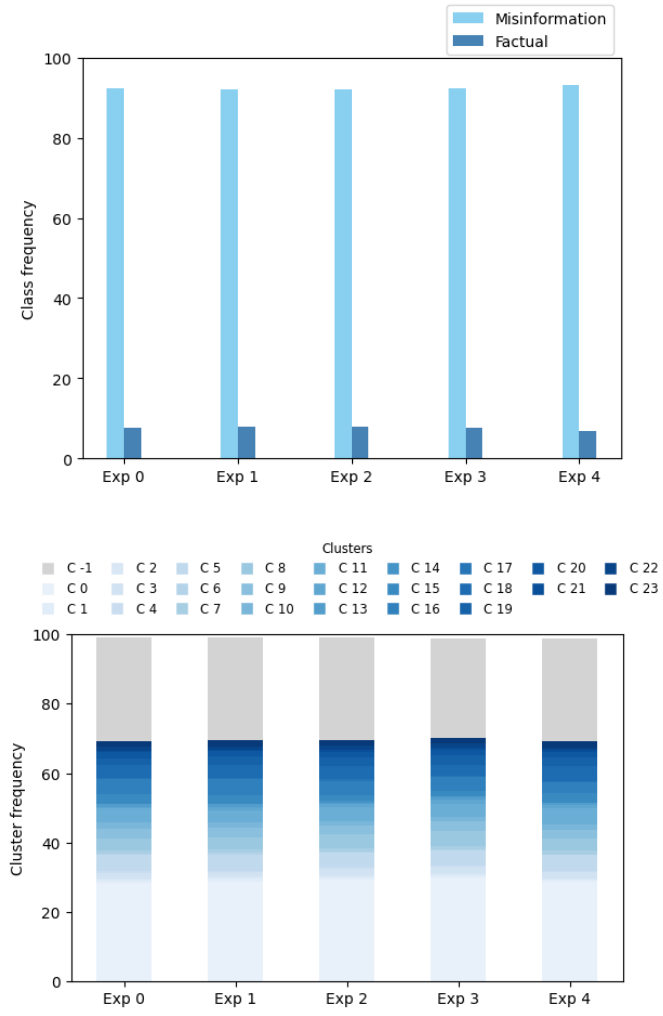

Figure 3: Class (top) and cluster (bottom) frequencies for different training set splits.

studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon* 9, 10 (2023), e20382. doi:10.1016/j.heliyon.2023.e20382

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conf. on comp. vis. and patt. rec.* 770–778.

[9] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2017).

[10] Vítor Lourenço and Aline Paes. 2022. A modality-level explainable framework for misinformation checking in social networks. *arXiv preprint arXiv:2212.04272* (2022).

[11] Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE Intl. Conf. on data mining workshops (ICDMW)*. IEEE, 33–42.

[12] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. doi:10.48550/ARXIV.1802.03426

[13] Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proc. of the ACM Web Conf. 2022*. 3632–3640.

[14] Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proc. of the 45th Intl. ACM SIGIR Conf. on research and development in inf. retr.* 3141–3153.

[15] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conf. on Empirical Methods in Natural Language Processing*.

[16] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. 2020. Adversarial active learning based heterogeneous graph neural network for fake news detection. In *2020 IEEE Intl. Conf. on Data Mining (ICDM)*. IEEE, 452–461.

[17] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45, 4 (2009), 427–437.

[18] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. *arXiv preprint arXiv:2209.07562* (2022).

[19] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.* 53, 5 (2020), 1–40.