

Unmasking Deception: A Topic-Oriented Multimodal Approach to Uncover False Information on Social Media

Riccardo Cantini^{1*}, Cristian Cosentino¹, Irene Kilanioti²,
Fabrizio Marozzo¹, Domenico Talia¹

¹University of Calabria, Rende, Italy.

²National Technical University of Athens, Athens, Greece.

*Corresponding author(s). E-mail(s): rcantini@dimes.unical.it;

Contributing authors: ccosentino@dimes.unical.it;

eirinikoilanioti@mail.ntua.gr; fmarozzo@dimes.unical.it;

talia@dimes.unical.it;

Abstract

In the digital landscape, social media has emerged as a prevalent channel for global communication, connecting like-minded individuals worldwide. However, while facilitating information exchange, it is also susceptible to the dissemination of false information, posing a constant challenge to the reliability of online content. To address this issue, this paper introduces a novel methodology called *TM-FID* (*Topic-oriented Multimodal False Information Detection*), which combines false information detection and neural topic modeling within a semi-supervised multimodal approach. By jointly leveraging textual and visual information contained in online news, our approach provides insights into how false information influences specific discussion topics, thus enabling a comprehensive and fine-grained understanding of its spread and impact on social media conversation. Experimental evaluation carried out on a set of multimodal gossip-related news demonstrates the quality of the identified topics, assessed through a novel centroid-based metric, as well as the efficacy of the cross-attention mechanism used within *TM-FID* to accurately identify false information in multimodal news. Overall, the proposed methodology can enable effective strategies to counter the spread of false information, thereby fostering trust and confidence in the information shared on social media platforms.

Keywords: Social media, False information detection, Multimodal analysis, Topic Modeling, Natural Language Processing, Semi-supervised learning

1 Introduction

In the digital age, social media has become an integral part of our daily lives, transforming the way we communicate, access information, and engage with the world around us. The rapid dissemination of news, opinions, and multimedia content across platforms has empowered individuals to become more informed and aware of global events, fostering a sense of interconnectedness and shared experiences. Indeed, people can access news and opinions from various perspectives, connecting globally and participating in diverse conversations, fostering a deeper comprehension of complex issues [1, 2]. Nonetheless, the rapid proliferation of information on social media also comes with challenges and risks, such as the amplification of echo chambers and the pervasive spread of false information [3, 4].

False information, whether intentionally misleading or inadvertently shared, poses a significant threat, potentially harming our interconnected society. In particular, false information can mislead individuals, leading to incorrect beliefs, decisions, or actions, also harming the reputation of individuals, organizations, or businesses. The spread of false information related to emergencies, health issues, or security threats can also lead to panic, distrust, and compromise public safety. Furthermore, false information can be used to manipulate public opinion, create echo chambers, influence elections, and undermine democratic processes [5, 6]. Overall, the multifaceted threats highlighted by false information pose the need for robust methodologies to detect and combat its spread. Traditional methods of false information detection often fall short in the face of the dynamic and multimodal nature of content shared on social media. Indeed, the coexistence of textual and visual elements within social posts adds significant layers of complexity, emphasizing the limitations of unimodal approaches. Therefore, more advanced solutions are needed to enable a thorough understanding of the nuanced interplay between text and images in social media content, thereby effectively enhancing false information detection [7–10].

Building upon these considerations, this work proposes *TM-FID (Topic-oriented Multimodal False Information Detection)*, a novel methodology aimed at assessing the presence of false information in the main topics discussed by social users in a multimodal setting, fusing textual and visual information contained in social media posts. Following this holistic approach, *TM-FID* provides a comprehensive understanding that goes beyond individual modalities, effectively addressing the multifaceted nature of false information on social media and enhancing the overall effectiveness of strategies aimed at mitigating the spread of false information.

The proposed methodology adopts a semi-supervised approach that seamlessly integrates multimodal false information detection with neural topic modeling, enabling the topic-oriented investigation of the impact of false information within a single multimodal framework. Specifically, a limited amount of labeled data is employed to train a multimodal false information detection model, which harnesses both textual and visual content to effectively identify false information within social media posts. This model is trained through a two-step process: (i) an unimodal pre-fine-tuning step, where textual and visual classifiers are independently trained; (ii) a subsequent multimodal training phase that leverages cross-attention to fuse insights from both modalities filtering out noise due to negative information transfer. Following this, the

multimodal capabilities of BERTopic are leveraged to extract the main topics underlying the social discourse from a large set of unlabeled posts, identifying a topic-based clustering structure. Finally, the multimodal classifier is utilized to assess the presence of false information within the identified topics by classifying the sentences associated with them, enabling a fine-grained evaluation of how false information influences and shapes social discussions from a topical perspective.

Given the critical task of accurately evaluating the identified topics in our methodology, this study also introduces a centroid-based metric to assess the representativeness of topics identified through a multimodal neural approach. Indeed, while coherence metrics such as CV and Normalized Point-wise Mutual Information (NPMI) are widely accepted indicators of topic quality, it is challenging to generalize such assessments for neural-based approaches [11, 12]. Furthermore, these metrics generally leverage only textual information, which can lead to misleading evaluations in a multimodal setting.

This paper significantly extends our previous conference work [3] in the following main aspects:

- It introduces multimodal capabilities to effectively utilize textual and visual information within social posts, thereby fostering a comprehensive understanding of the multifaceted nature of false information from a topical perspective.
- It conducts a comprehensive experimental evaluation by comparing with state-of-the-art multimodal techniques that leverage both textual and visual content.
- It introduces a centroid-based metric to assess the quality of identified topics in a multimodal neural setting.
- It includes an ablation study to assess the benefits of incorporating different modalities and using the cross-attention layer for embedding fusion.

The remainder of the paper is organized as follows. Section 2 discusses related work on false information detection. Section 3 describes the proposed methodology. Section 4 presents the experimental results, and Section 5 concludes the paper.

2 Related Work

Social media plays a crucial role in information spreading and staying updated on current trends and discussions. However, the reliability of news circulating on social platforms is often questionable and susceptible to various biases, making it difficult to assess the reliability of online published content. This issue stems from the presence of false information, which can take various forms. Specifically, *misinformation* describes the unintentional spread of false information, whereas *disinformation* denotes purposely conveyed inaccurate or misleading information. Furthermore, the term *fake news* is frequently used to describe fabricated news aimed at deceiving public opinion, which can have a huge impact on socio-political issues.

2.1 Unimodal approaches

The multifaceted nature of false information and the threats it poses entail the need for robust methodologies to detect and counteract its spread. Among the major works

in the literature, deep learning-based approaches have been proposed, leveraging convolutional neural networks (CNNs), and recurrent neural networks (RNNs) [13, 14]. Moreover, Natural Language Processing (NLP) techniques have been increasingly used to effectively exploit the linguistic features and patterns of news articles or social media posts [15, 16]. In particular, Transformer-based architectures [17] such as BERT (Bidirectional Encoder Representation from Transformers) [18], have showcased remarkable effectiveness across various practical applications, leading to notable advancements in computational linguistics, in the field of natural language generation, processing, and understanding. Noteworthy examples in the field of false information detection include the combined use of BERT with CNNs, as in FakeBERT [19], and RNNs, as proposed in [20]. Additionally, beyond their conventional applications, Transformer-based architectures have been utilized for fact-checking purposes and explanatory endeavors. For instance, a recent study [21] outlines a two-stage system capable of not only assessing the veracity of COVID-19-related claims but also providing users with comprehensive textual explanations to aid in understanding the assessment process.

2.2 Multimodal approaches

Current multimodal approaches for false information detection extract textual and visual features of news, combining them via embedding fusion mechanisms such as simple concatenation or sum, or more sophisticated approaches like cross-modal ambiguity learning [7]. Among these works, *SpotFake+* [9] utilizes the XLNet [22] Transformer to extract textual representations, which are fused through a dense layer with visual ones, extracted using the VGG16 deep convolutional neural network. Another similar work is represented by *SAFE* [10], which examines the relationship between textual and visual features to identify fake news, using a cross-modal similarity mechanism. Specifically, representations of textual and visual information, along with their relationship, are jointly learned and used to predict fake news, facilitating the detection of fake content based on either text or images, as well as the mismatch between them. *CAFE* [7] addresses cross-modal ambiguity through a cross-modal alignment module, which transforms unimodal embeddings into a shared space. By leveraging cross-modal ambiguity learning, *CAFE* can adaptively fuse the unimodal features and cross-modal correlations to improve fake news detection effectiveness. The *CMC* [8] methodology proposes a two-stage approach for fake news detection, which involves the training of textual and visual networks using cross-modal knowledge distillation. In this setting, the soft target from one network is used to guide the training of the other, thus capturing correlations between modalities. Finally, trained networks are utilized to extract features that are combined through a fusion mechanism.

Comparison

While unimodal false information detection techniques have shown substantial improvements, particularly due to the use of Transformer-based architectures, they often fall short in handling the multifaceted nature of social media content. Multimodal approaches have addressed this issue, achieving higher accuracy by combining textual and visual information. Despite these advancements, some key challenges and gaps remain. Several techniques employ simple embedding fusion mechanisms, such as

concatenation [9, 10], which may not fully capture the intricate relationships between different modalities. More sophisticated fusion techniques can potentially enhance detection performance further, such as cross-modal ambiguity learning [7], or cross-attention, as proposed in this study. Specifically, the latter effectively reduces noise in text-image pairs, thus minimizing negative information transfer across modalities [23]. Additionally, employing a two-stage learning process can enhance the representation of cross-modal content, conveying useful information for the downstream task of false information detection. Major approaches include cross-modal knowledge distillation and bilinear pooling [8], or cross-modal training after unimodal fine-tuning, as performed by *TM-FID*. Furthermore, existing techniques treat false information as a monolithic issue, failing to consider the specific topics around which it clusters. Therefore, integrating multimodal topic modeling, as proposed in *TM-FID*, can provide insights into the nuanced ways false information impacts different areas of discourse within the broader social media conversation, highlighting the underlying factors and dynamics that facilitate its spread.

3 Proposed Methodology

This section provides a comprehensive description of *TM-FID*, delving into its neural design, the learning process, and the integration of false information detection with topic modeling to thoroughly represent false information conveyed in multimodal news.

In *TM-FID*, a multimodal binary classifier is employed to identify false information within a given post effectively. This classifier is trained on a small set of annotated data via transfer learning, by integrating *BERTweet* [24] with a *Vision Transformer (ViT)* [25] through a cross-attention module [23]. Integrating visual and textual information is crucial for capturing the fine-grained details of multimodal posts and ensuring effective false information detection. Specifically, *BERTweet* can identify linguistic and stylistic patterns that may indicate false information, while *ViT* provides essential context that text alone cannot convey, such as the use of manipulated or misleading images. By combining both visual and textual data, the system can better understand the full context of a post and detect inconsistencies that may indicate false information. In particular, the cross-attention module enables *TM-FID* to correlate information across modalities and identify text-image discrepancies, a common trait of fabricated content.

A large set of unlabeled data is used to unveil the main discussion topics underlying social media conversation by using a multimodal topic modeling framework. Specifically, this step relies on the multimodal capabilities of *BERTopic* [12], one of the most used neural topic modeling methods in the literature. Finally, a false information score is computed for each identified topic, leveraging the multimodal false information classifier, enabling a topic-oriented assessment of the impact of false information on social media conversation. Overall, this approach allows for a quantitative evaluation of the main discussion subjects predominantly affected by false information, also facilitating the identification of specific instances of user-generated false information linked to these subjects. In the following, we provide a detailed description of the main steps making up *TM-FID*, whose execution flow is represented in Figure 1.

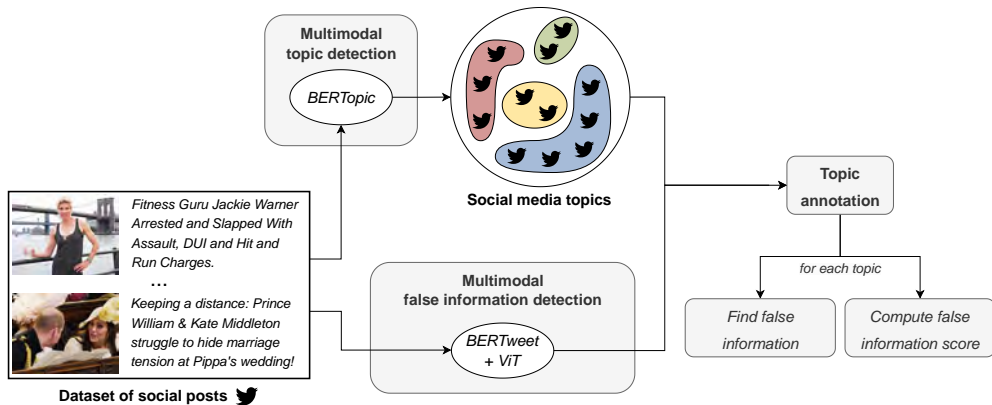


Fig. 1 Execution flow of *TM-FID* (Topic-oriented Multimodal False Information Detection).

3.1 Multimodal false information detection

In this step, the multimodal false information detection model is trained through a two-step process: (i) *unimodal fine-tuning*, where two Transformer-based models are individually fine-tuned for the downstream task of false information detection in each modality; and (ii) *cross-modal training*, where information from the two modalities, extracted from the fine-tuned models, is fused through cross-attention.

Unimodal fine-tuning

During the unimodal fine-tuning step, a BERTweet and a ViT model undergo individual training processes specific to their modalities (i.e., textual and visual, respectively). BERTweet is a large-scale language pre-trained on a vast corpus of English Tweets, following the RoBERTa [26] procedure, to effectively deal with features like hashtags, mentions, URLs, and emojis. ViT is a Transformer-based network encoder pre-trained on the ImageNet-21k image dataset, specially designed to effectively capture both local and global image features. Unlike traditional convolutional neural networks (CNNs) that rely on convolutions to capture local patterns in images, ViT divides an image into a sequence of smaller fixed-size patches, treats each patch as a token, and applies a transformer encoder to capture the global context and relationships between these patches. Both BERTweet and ViT models are fine-tuned by placing a classification head on top of the Transformer encoder, leveraging the embedding representation of the $[CLS]$ token, a special token added during pre-training for classification purposes. The classification head in this step is represented by a single fully connected layer with two output neurons and a softmax activation. The textual and visual models were configured using a sequence length of 128 for BERTweet and an input image size equal to 224×224 for ViT. They were fine-tuned individually for 8 epochs using a batch size of 32, the cross-entropy loss, and the Rectified Adam optimizer [27], designed to mitigate variance issues found in adaptive learning rate optimizers. Furthermore, we set the learning rate to $3 \cdot 10^{-5}$, in line with the standard order of magnitude usually employed for fine-tuning Transformer-based models [18, 28]. Specifically, such a small

value is crucial for subtly adjusting the weights of pre-trained models to suit the false information detection task, preventing large weight updates that can cause overfitting and catastrophic forgetting [29], which occurs when the model loses previously learned information.

Cross-modal training

Once the BERTweet and ViT models are fine-tuned, as described earlier, they are employed to extract latent representations from the two modalities, which are then fused via cross-attention as depicted in Figure 2. The use of this mechanism for embedding fusion can effectively address the presence of noise in text-image pairs, which can hinder the performance of multimodal tasks such as the classification of tweets, due to potential negative information transfer across modalities [23].

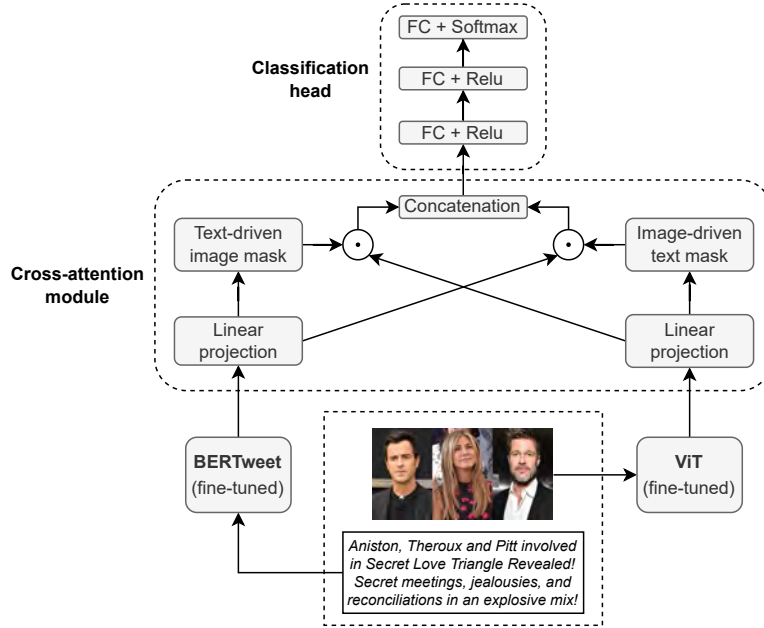


Fig. 2 Architecture of the multimodal false information detection model within *TM-FID*.

Specifically, let $x = \langle t, i \rangle$ denote a multimodal input instance, represented by a text-image pair. The latent representations e_t and e_i of the text t and image i are computed by the BERTweet and ViT encoders, respectively. Such representations are both represented by a d -dimensional vector (with $d = 768$) and are extracted from the hidden representation of the $[CLS]$ token of the corresponding encoder. Afterward, for each modality, a linear projection of the corresponding embedding in a k -dimensional space is computed, with k equal to 512:

$$h_t = W_t^\top e_t, \quad h_i = W_i^\top e_i \quad (1)$$

where $h_t \in \mathbb{R}^d$ and $h_i \in \mathbb{R}^d$ are the 512-dimensional linear projections of e_t and e_i , while $W_t \in \mathbb{R}^{d \times k}$ and $W_i \in \mathbb{R}^{d \times k}$ are learnable matrices.

Before combining the projected text and image embeddings, the *cross-attention module* enables each modality to selectively block features from the other, guided by its confidence in the relevance of its input. This cross-modal selective blocking is achieved by computing two mask vectors, where the mask for one modality is determined exclusively by the other. These masks are computed by feeding the corresponding projected embedding to a fully-connected layer with a sigmoid activation:

$$\mu_{t \rightarrow i} = \sigma(W_{t \rightarrow i}^\top h_t), \quad \mu_{i \rightarrow t} = \sigma(W_{i \rightarrow t}^\top h_i) \quad (2)$$

where $W_{t \rightarrow i} \in \mathbb{R}^{k \times k}$ and $W_{i \rightarrow t} \in \mathbb{R}^{k \times k}$ are learnable weight matrices. The $W_{t \rightarrow i}$ matrix is used to compute the mask for filtering image features based on the text embedding h_t , while $W_{i \rightarrow t}$ is used for the mask that filters text features based on the image embedding h_i . The sigmoid function σ is finally applied element-wise to squash the output values to the range $(0, 1)$, allowing to selectively attenuate or amplify features. As a result, the two masks are obtained, specifically:

- $\mu_{t \rightarrow i} \in \mathbb{R}^k$, i.e., the *text-driven image mask*, derived from the projected text embedding and employed to filter features from the image embedding.
- $\mu_{i \rightarrow t} \in \mathbb{R}^k$, i.e., the *image-driven text mask*, derived from the projected image embedding, and employed to filter features from the text embedding.

This masking approach allows each modality to influence features from the other modality that are considered important, enhancing the mutual understanding and interaction between the text and image representations before they are combined, thus addressing negative information transfer issues. Finally, given the mask vectors, projected embeddings are fused, to obtain a single multimodal representation ω :

$$\begin{aligned} \tilde{h}_t &= \mu_{i \rightarrow t} \odot h_t, & \tilde{h}_i &= \mu_{t \rightarrow i} \odot h_i \\ \omega &= \tilde{h}_t \parallel \tilde{h}_i \end{aligned} \quad (3)$$

where \tilde{h}_t and \tilde{h}_i are the masked text and image embeddings, \odot and \parallel are the Hadamard product and the concatenation operator, and $\omega \in \mathbb{R}^{2k}$ is the multimodal representation of the input pair $x = \langle t, i \rangle$. This representation is finally fed to the *classification head*, composed of two hidden fully-connected ReLU layers and an output softmax layer that computes the class probabilities. The model is trained for 5 epochs, by keeping the fine-tuned encoders frozen. We used a batch size of 32, a cross-entropy loss, and the Rectified Adam optimizer, initialized with a learning rate equal to 10^{-3} .

3.2 Multimodal topic detection

In this phase, the main topics underlying social media conversations are extracted using multimodal topic modeling. Unlike the false information detection step (Section 3.1), which involves training a multimodal classifier on a curated set of labeled data, discussion topics are extracted in an unsupervised manner. Consequently, a vast

amount of unlabeled posts can be used, which is crucial for extracting a large set of diverse and coherent topics that thoroughly represent online user discussions.

Among all topic modeling techniques in the literature, we chose *BERTopic* [12], which relies on semantically-rich data representations achieved through Transformer-based pre-trained models. As highlighted in recent studies, neural topic modeling techniques like BERTopic can enhance performance in terms of coherence and diversity, especially in the case of data gathered from social media platforms [12, 30, 31], and have been widely applied across various fields, ranging from healthcare, to political and social sciences [32–34]. Furthermore, unlike other ready-to-use neural topic modeling techniques that handle only textual data, BERTopic easily supports topic modeling in multimodal settings thanks to its modular design.¹

In the case of multimodal data, BERTopic adopts a pre-trained embedding model based on *CLIP (Contrastive Language-Image Pre-training)* [35]. This model from OpenAI leverages contrastive learning to generate latent representations of both text and images within a unified embedding space, which enables the use of natural language to reference learned visual concepts enabling zero-shot transfer of the model to several downstream tasks. Within BERTopic, the embedded representation of both text and images are extracted from CLIP, fused together, and projected into a low-dimensional space using UMAP (Uniform Manifold Approximation and Projection). Then, reduced representations are clustered into semantically-related groups through HDBSCAN, a density-based clustering algorithm capable of handling varying density scenarios. Finally, for each cluster, a class-based version of tf-idf is employed to compute topic representations. The modular approach implemented by BERTopic ensures the extraction of coherent and diverse discussion topics from the social media conversation, allowing to capture semantic relationships across different modalities through the use of large-scale multimodal pre-trained encoders.

3.3 Topic annotation

In *TM-FID*, a topic-oriented approach is employed to thoroughly examine the impact of false information within social media conversations, quantitatively assessing its prevalence on the different subjects underlying the online discussion. Let \mathcal{C} be the set of clusters composing the topic-based clustering structure identified by BERTopic. Each topical cluster $c \in \mathcal{C}$ is a set of text-image pairs $x = \langle t, i \rangle$ characterized by a probability p_x^c indicating the degree of membership of x to the cluster c .

Instance-level false information probability

For each cluster $c \in \mathcal{C}$, we utilize a multimodal classification model to assess the likelihood of false information in each instance $x = \langle t, i \rangle \in c$. Particularly, the classification model, trained as described in Section 3.1, performs the following operations:

- *Feature extraction*: textual and visual features are extracted from t and i using BERTweet and ViT, respectively.
- *Feature fusion*: the extracted modality-specific features are fused by the cross-attention layer, to achieve a joint representation.

¹<https://maartengr.github.io/BERTopic/>

- *Probability computation*: the achieved cross-modal embedded representation is fed through the classification head of the model to determine the probability p_x^{fi} , which measures the likelihood that the instance x contains false information.

Cluster-level false information score

Given the probabilities p_x^{fi} for each instance x in a cluster c , a false information score $\mathcal{S}(c)$ is computed. This score is the average false information probability of the instances within that cluster, weighted by the degree of membership of those instances to that cluster. Formally, the false information score for a cluster $c \in \mathcal{C}$ is defined as:

$$\mathcal{S}(c) = \frac{\sum_{x \in c} p_x^c \cdot p_x^{fi}}{\sum_{x \in c} p_x^c}, \text{ where } x = \langle t, i \rangle, c \in \mathcal{C} \quad (4)$$

where p_x^{fi} is the probability that instance x contains false information, while p_x^c is the degree of membership of instance x to cluster c . This weighted average ensures that instances with a higher degree of membership to a cluster have a greater influence on the false information score of that cluster, providing a more accurate reflection of the prevalence of false information within it. By employing this approach, *TM-FID* effectively quantifies the impact of false information across different topics in social media conversations, facilitating a deeper understanding of how misinformation propagates within various subject areas.

4 Experimental Results

As the Internet becomes a primary source of information, distinguishing between genuine news and fabricated stories has become a complex challenge. This is particularly true in the realm of social media, where information can spread rapidly and without verification. The spread of false information can have serious consequences, influencing public opinion, harming individuals' reputations, and even affecting political and social outcomes. For instance, during election cycles, fake news can sway voter perceptions and decisions, potentially altering the results [6]. Similarly, in the field of public health, misinformation about vaccines or treatments can lead to dangerous health behaviors and undermine public trust in medical institutions [3]. Furthermore, businesses can suffer severe financial and reputational damage from false rumors about their products, while celebrities often fall victim to false information on the Internet due to their public visibility and the continuous interest from the media and the public.

In this section, we discuss the results achieved by applying the *TM-FID* methodology to a large set of celebrity-related Twitter news, provided by the *FakeNewsNet* dataset [36]. It is an extensive collection of news gathered from reliable fact-checking websites such as *PolitiFact*², which provides detailed evaluations of political news, and *GossipCop*³, which focuses on entertainment stories. In addition to the news content, the dataset includes extensive social context information sourced from social media

²<https://www.politifact.com/>

³<https://www.gossipcop.com/>

platforms, including retweets, likes, and replies, which provide valuable insights into how fake news spreads and how users react to it. Moreover, the dataset contains spatiotemporal information, including geographical locations inferred from user profiles and timestamps of social media interactions, enabling the analysis of how fake news propagates over time and across different regions.

Among the large set of multimodal news provided by the dataset described, we maintained all pairs where both image and text are well formatted. The resulting dataset is composed of 9500 news divided as follows: (i) 66% of news, partitioned into train, validation, and test sets, are labeled as false information or trustworthy content; (ii) 33% of remaining news are unlabeled, and utilized to uncover the main discussion topics underlying the online discourse regarding celebrities on Twitter.

In the following sections, we present our main findings, specifically focusing on the following aspects: (i) a comprehensive experimental evaluation with state-of-the-art techniques for multimodal fake news detection; (ii) an ablation study, highlighting the individual contributions of textual and visual modalities, as well as their combined impact through cross-attention; (iii) a discussion of the main topics identified that drove the celebrity-related discussions on Twitter; (iv) a topic-oriented analysis of false information related to celebrity gossip.

4.1 State-of-the-art comparison

In this section we assess the performance of the false information detection model used within the proposed *TM-FID* methodology, by comparing it to the most commonly used techniques in the literature, detailed in Section 2, which include:

- *SAFE*, which examines the relationship between textual and visual features to identify fake news via cross-modal similarity.
- *SpotFake+*, which uses a dense layer to fuse visual and textual information extracted from a VGG16 convolutional neural network and an XLNET model, respectively.
- *CAFE*, which exploits cross-modal ambiguity learning to adaptively fuse unimodal features, and cross-modal correlations to improve fake news detection effectiveness.
- *CMC*, which uses cross-modal knowledge distillation to train a textual and a visual network, then combining the produced features through a fusion mechanism.

Figure 3 shows the results achieved by the aforementioned methods, in terms of accuracy, precision, recall, and F1-score. Apart from accuracy, the macro average is used for the other metrics. This comparison highlights the superior performance achieved by the multimodal model used within *TM-FID* compared to the other approaches in the literature. We argue that this improvement may be related to the use of BERTweet for the textual modality, which allows for effectively grasping semantically-rich aspects of analyzed data, due to the Twitter-specific nature of this model. Moreover, the two-step training process outlined in Section 3.1 allows for accurate extraction and information fusion from the two modalities. Indeed, the encoders are firstly fine-tuned separately, allowing for easier convergence, and then the extracted representations, already tailored to the downstream task under consideration, are selectively fused via cross-attention, effectively addressing negative information transfer across modalities.

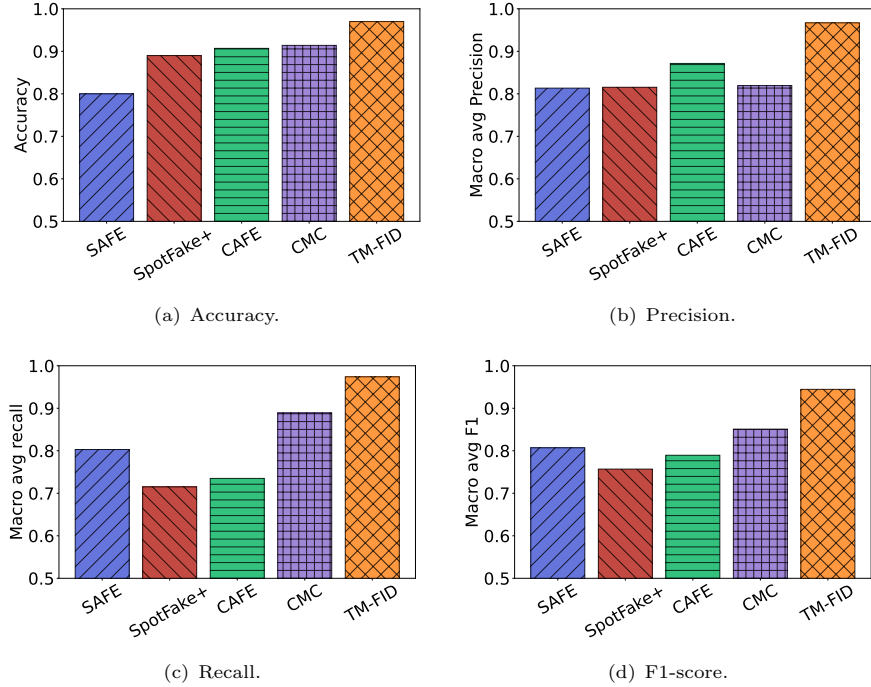


Fig. 3 Comparison with state-of-the-art multimodal false information detection models.

A noteworthy aspect is related to the dataset imbalance, with fewer instances of false information ($\approx 27\%$) compared to real and trustworthy news. Consequently, we analyzed how this imbalance impacts classification performance, assessing the classification effectiveness for both real and false classes separately. As shown in Table 1, the multimodal information detection model utilized by *TM-FID* outperforms all other approaches for both classes, exhibiting notably greater robustness, particularly for the minority class.

Table 1 Performance comparison with per-class evaluation metrics.

Model	Fake News			Real News		
	Precision	Recall	F1-score	Precision	Recall	F1-score
SAFE [10]	0.764	0.669	0.730	0.863	0.907	0.884
SpotFake+ [9]	0.736	0.539	0.622	0.895	0.891	0.892
CAFE [7]	0.822	0.549	0.658	0.827	0.921	0.921
CMC [8]	0.703	0.842	0.766	0.967	0.936	0.936
TM-FID	0.960	0.927	0.943	0.974	0.986	0.980

4.2 Ablation study

To perform a more in-depth analysis of the model performance, this section presents an ablation study investigating the contribution and interaction of individual modalities in false information detection.

As reported in Table 2, by focusing on single-modal models, particularly those fine-tuned using either textual or visual news data, we discovered that the textual model, leveraging BERTweet, achieved significantly higher performance compared to its visual counterpart, i.e., ViT. Moreover, combining textual and visual information in a multimodal approach resulted in further improvement. Specifically, leveraging a cross-attention mechanism proved to be more effective than naive fusion techniques such as concatenation and summation, which yielded comparable results in our experiments. These results underscore the nuanced nature of textual information, where linguistic cues may carry more weight in discerning falsehoods than visual cues alone. They also highlight the benefits of a holistic approach that integrates insights from different modalities and emphasizes the role of cross-attention in avoiding negative information transfer and ensuring effective fusion.

Table 2 Ablation study comparing the multimodal model with unimodal models using either text or image, alongside various embedding fusion mechanisms.





Model	Accuracy	Precision	Recall	F1-score
Image-only (ViT)	0.868	0.838	0.807	0.820
Text-only (BERTweet)	0.952	0.938	0.950	0.944
Multimodal (concat/sum)	0.966	0.959	0.955	0.957
Multimodal (cross-attention)	0.970	0.967	0.956	0.962

To gain a deeper understanding of the role of each modality in the classification, Table 3 shows examples of test posts containing false information, along with the corresponding false information probabilities computed by *TM-FID* and unimodal models using either text (BERTweet) or image (ViT).

The first test instance, p_1 , is correctly labeled as false information by ViT, while BERTweet fails to detect signs of deception based on text analysis alone. Overall, the *TM-FID* model accurately classifies the post as false information, drawing more clues from visual data. In fact, compared to the other post images in Table 3, the image of p_1 appears more fabricated due to the juxtaposition of the two celebrities in a single frame, which is a common trait in fake news. In contrast, the classification of post p_2 shows the text as more informative for detecting false information using *TM-FID*. Specifically, while ViT classifies it as real, BERTweet identifies a high likelihood of false information, mainly due to the style of the post. The text indeed exhibits distinct linguistic and stylistic patterns that may be indicative of false information, mainly related to sentence syntax and exaggerated tone.

An interesting case is post p_3 , labeled as real by both BERTweet and ViT. Although the unimodal models fail to detect the falsehood of p_3 when examining either the text or the image alone, their joint use allows *TM-FID* to correctly classify it by detecting a discrepancy between the text and image, which is a typical feature of false information.

Table 3 Examples of test posts containing false information and corresponding false information probabilities computed by ViT (image-only), BERTweet (text-only), and *TM-FID* (multimodal). Probabilities that result in a correct classification as false information are highlighted in bold.

Id	Post text	Post image	False information probability p_x^{fi}		
			ViT	BERTweet	TM-FID
p_1	Taylor Swift turned her apartments into five-star restaurants.		0.71 (False)	0.12 (Real)	0.99 (False)
p_2	CONFIRMED: Brad Pitt and Jennifer Aniston - They're back on!		0.13 (Real)	0.99 (False)	0.99 (False)
p_3	Matt Lauer spotted spending time with his wife Annette Roque amid divorce rumors.		0.07 (Real)	0.36 (Real)	0.97 (False)
p_4	Stan Lee feeling great after brief hospitalization.		0.26 (Real)	0.01 (Real)	0.27 (Real)

Lastly, p_4 presents an example where the model fails to recognize the post as false information, given the low likelihood from the unimodal models and the absence of a clear misalignment as seen in p_3 , since the image in p_4 aligns more closely with the accompanying text.

The presented findings underscore the critical importance of carefully considering the interplay between different modalities, each offering valuable insights, such as the detection of misleading images and the identification of specific linguistic or stylistic patterns. The effective fusion of these modalities via cross-attention also enables a comprehensive understanding of the post and is key to the identification of text-image misalignments. Overall, the multimodal approach used in *TM-FID* effectively captures the subtle ways in which textual and visual information interact, resulting in a more robust and reliable detection mechanism.

4.3 Gossip-related detected topics

In this section, we present the main topics identified by applying BERTopic to the *FakeNewsNet* dataset of celebrity-related multimodal Twitter news. As shown in Figure 4 and described below, a wide range of topics emerged, spanning from entertainment and politics to the personal lives of celebrities.

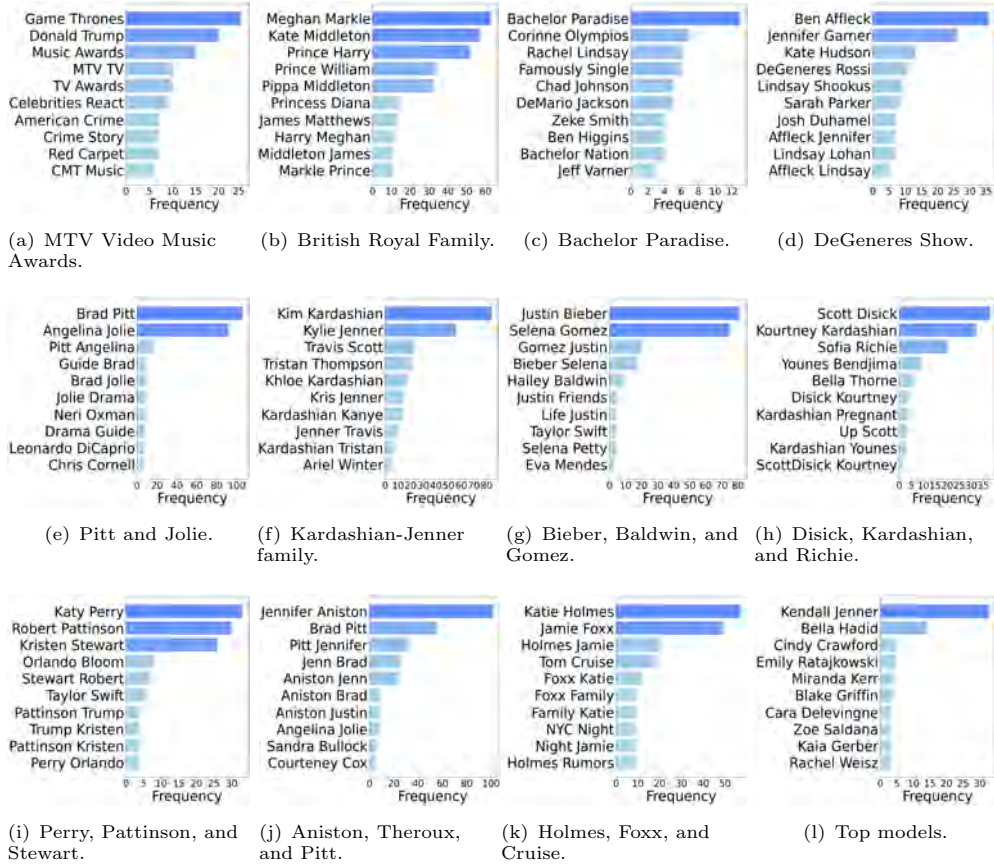


Fig. 4 Frequency of top-10 bigrams for each identified topic.

One prominent topic focuses on the *2017 MTV Video Music Awards*⁴ (Figure 4(a)), an event organized by the American television network MTV, where the best music videos and songs of the past 12 months are honored. The *British Royal Family* (Figure 4(b)) and royal weddings captivated the public’s interest, sparking conversations about traditions, protocols, and scandals. Other topics center around television shows such as *Bachelor Paradise*⁵ (Figure 4(c)), and the show hosted by *Ellen DeGeneres*⁶ (Figure 4(d)), where celebrity statements frequently foster discussions and gossip among users. Celebrities like *Brad Pitt and Angelina Jolie* (Figure 4(e)), and the members of the *Kardashian-Jenner* family (Figure 4(f)), were also at the center of rumors and

⁴Words like *Game*, *Throne*, *hbo*, and *nomination* collectively refer to the nomination of *Game of Thrones*, an HBO TV series, for several award categories, including the *show of the year*. The bigram *Donald Trump* may refer to the speech Paris Jackson gave before presenting the *Best Pop Video* award, where she criticized Trump for fueling racism, condemning the violent events of Charlottesville (August 2017).

⁵*Bachelor Paradise* is an American reality television elimination show. It is a spin-off of the *Bachelor* and *Bachelorette* programs, hence the presence of these terms. Other bigrams refer to contestants of the show, like *Rachel Lindsay* and *Corinne Olympios*.

⁶Among celebrities featured at the *DeGeneres* show, the topic focuses on *Ben Affleck*, with his *divorce* from *Jennifer Garner*, finalized in 2017, and rumors about his new relationship with *Lindsay Shookus*.

scandals, which ignited debates about their personal lives. Romantic entanglements, secrets, and rumors about love triangles involving celebrities also drew considerable attention from users, contributing to the proliferation of rumors and even false information. Particularly, online conversation focused on: *Justin Bieber, Selena Gomez, and Hailey Baldwin* (Figure 4(g)); *Scott Disick, Kourtney Kardashian, and Sofia Richie* (Figure 4(h)); *Katy Perry, Robert Pattinson, and Kristen Stewart* (Figure 4(i)); *Jennifer Aniston, Justin Theroux, and Brad Pitt* (Figure 4(j)); *Jamie Foxx, Katie Holmes, and Tom Cruise*⁷ (Figure 4(k)). Furthermore, discussions also revolved around top models such as *Bella Hadid, Kendall Jenner, and Emily Ratajkowski* (Figure 4(l)), who often feature in advertising campaigns and fashion events, television personalities such as *Luann de Lesseps*, as well as famous singers like *Taylor Swift, Rihanna, Drake, Mariah Carey, and Jennifer Lopez*.

Topic quality evaluation

As stated in [37], coherence metrics, such as CV and Normalized Pointwise Mutual Information (NPMI), serve as a valuable indicator of topic quality, including human judgment and interpretability. However, recent literature highlighted the challenge of generalizing such judgments, particularly in the case of neural-based approaches [11, 12]. Therefore, such metrics may offer only partial insights when employed to evaluate the output of neural topic modeling techniques like BERTopic. Moreover, these metrics typically evaluate the efficacy of topical words in conveying a specific theme or concept, thus not thoroughly considering the whole spectrum of available information in a multimodal setting.

Starting from these considerations, we introduced a new centroid-based metric, namely *SIL-CB*, which borrows ideas from internal metrics for clustering evaluation, specifically the Silhouette coefficient. In particular, we derived the *SIL-CB* metric by adapting the Silhouette coefficient to the evaluation of the topic-based multimodal clustering structure (\mathcal{C}) identified by BERTopic, thus assessing both intra-cluster cohesion and inter-cluster separateness as a joint measure of topic quality. In clustering evaluation, Silhouette coefficient \mathcal{S} is defined as follows:

$$\mathcal{S} = \frac{\delta_{inter} - \delta_{intra}}{\max\{\delta_{intra}, \delta_{inter}\}} \quad (5)$$

where δ_{intra} represents the average intra-cluster distance, calculated for each point relative to its cluster, while δ_{inter} denotes the average distance computed between different clusters. A Silhouette value near 1 indicates that the data point is far away from the neighboring clusters, which suggests a well-formed clustering structure; a score around 0 indicates that the data point is close to the decision boundary between two neighboring clusters; a score close to -1 indicates that the data point may have been assigned to the wrong cluster.

In *SIL-CB*, δ_{inter} and δ_{intra} are computed as follows. Let \mathcal{R}_c be the set of representative documents for the cluster $c \in \mathcal{C}$. This is computed, in BERTopic, by selecting the subset of documents that are most representative of their topic, based on the cosine

⁷This topic refers to the American actress *Katie Holmes* and her alleged secret meetings with actor *Jamie Foxx* following her divorce from *Tom Cruise*.

similarity with the corresponding c-TF-IDF topical representation [12]. In addition, let p_x^c be the degree of membership of the multimodal instance $x = \langle t, i \rangle$ to the cluster c . For each cluster $c \in \mathcal{C}$, the centroid γ_c is computed by averaging over the instances in \mathcal{R}_c , with each instance weighted by its membership degree:

$$\gamma_c = \sum_{x \in \mathcal{R}_c} p_x^c \cdot emb(x) \quad (6)$$

where $emb(x)$ represents the multimodal embedded representation of the text-image instance x , computed by BERTopic as the average between textual and visual embeddings extracted from CLIP. Afterward, for a given cluster c , inter-cluster dispersion is computed as the average cosine distance between each instance x in \mathcal{R}_c and its corresponding centroid γ_c . This calculation is then averaged across all clusters to obtain the overall δ_{inter} . Additionally, δ_{intra} is obtained as the average cosine distance between the centroids of each possible pair of clusters. Formally:

$$\delta_{inter} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{R}_c|} \sum_{x \in \mathcal{R}_c} d(x, \gamma_c), \quad \delta_{intra} = \frac{1}{|\mathcal{C}|(|\mathcal{C}| - 1)} \sum_{c', c'' \in \mathcal{C}} d(\gamma_{c'}, \gamma_{c''}) \quad (7)$$

where $d(a, b) = 1 - \frac{a^\top b}{\|a\| \|b\|}$ represents the cosine distance between two given vectors a and b . Hence, given the definition of δ_{intra} and δ_{inter} , the *SIL-CB* metric is computed using equation 5. By averaging across 10 different runs, the set of multimodal topics extracted by BERTopic achieved a mean *SIL-CB* score of 0.879, indicating the quality of the multimodal clustering structure identified by the algorithm, and, in turn, the meaningfulness of the corresponding topics.

4.4 Topic-oriented false information detected in gossip discussions

As described in Section 3, *TM-FID* enables a fine-grained analysis of the impact of false information on Twitter discourse, by taking a topical perspective. Figure 5 shows the discussion topics sorted according to their estimated level of false information. Specifically, for each topic $c \in \mathcal{C}$ a false information score $\mathcal{S}(c)$ is calculated, as defined in Section 3.3, which assesses the extent of false information within topic c . By analyzing Figure 5, a diversity emerges in the level of false information among the identified topics. In this regard, we report three of them, highlighting the minimum, medium, and maximum value of false information, based on the $\mathcal{S}(c)$ score distribution.

- *Bachelor Paradise*: this topic refers to Bachelor in Paradise, an American reality TV show where contestants compete with each other and form relationships.
- *DeGeneres Show*: this topic is centered around the television show hosted by Ellen DeGeneres, where well-known celebrities have been featured. Their statements during the show have sparked discussions on social media, giving rise to gossip about their personal lives and romantic relationships.

- *Holmes, Foxx, and Cruise*: this topic concerns the private life of Katie Holmes, specifically focusing on an agreement between the actress and her ex-husband Tom Cruise, in which she was supposed to maintain the secrecy of her new relationship for the following five years after their divorce. Viral images capturing her in the company of Jamie Foxx ignited gossip and theories among social media users, while the media produced sensationalist headlines further amplifying the discussion.

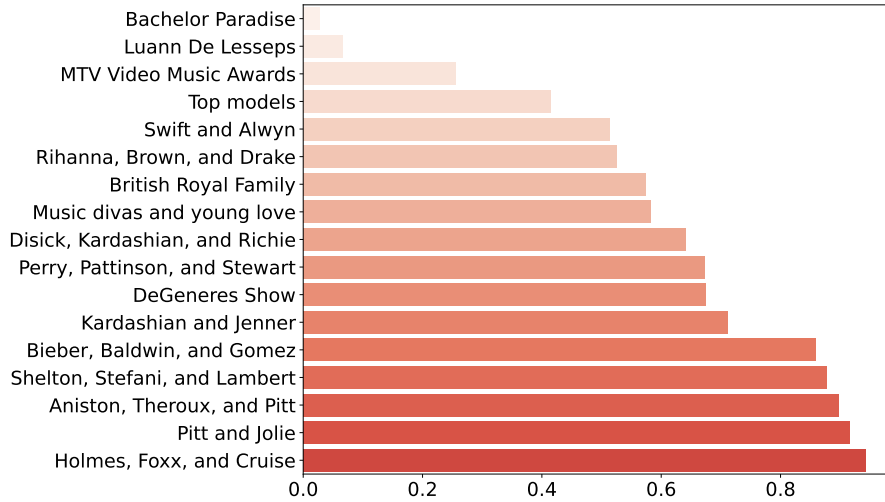


Fig. 5 Identified topic sorted by increasing false information score.

For each identified topic, we examined the distribution of the output probability generated by the multimodal false information detection model leveraged by *TM-FID*. This model, as explained in Section 3, calculates a probability, denoted as p_x^{fi} , indicating to what extent a given multimodal instance $x = \langle t, i \rangle$ is false information.

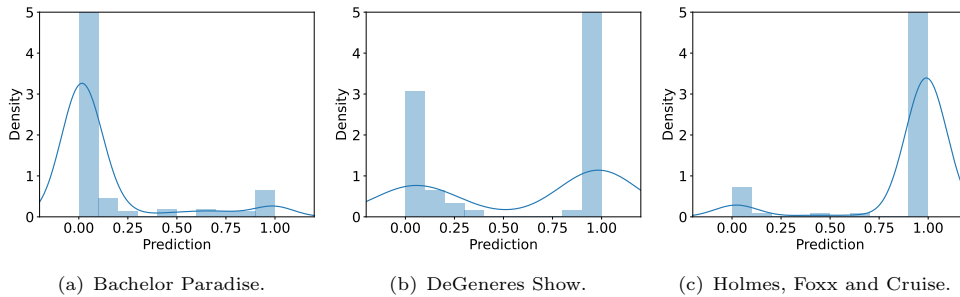


Fig. 6 Examples of p_x^{fi} distribution for topics with low, medium, and high false information scores.




The results obtained, shown in Figure 6, are consistent with the false information scores calculated earlier. In particular, Figure 6(a), related to the *Bachelor Paradise* topic, shows a distribution whose values are mainly concentrated toward 0, indicating a higher prevalence of non-false information. An opposite behavior is observed in Figure 6(c), related to the topic with the highest score of false information (i.e., *Holmes, Foxx, and Cruise*), in which a high concentration around 1 is present, indicating a marked presence of user-generated content identified by *TM-FID* as false information. Differently, in Figure 6(b), related to *DeGeneres Show*'s topic, an almost bimodal distribution is visible, which translates into a nearly equal distribution between false and non-false content.

As a final analysis, we focused on the top three topics with the highest false information score:

- *Holmes, Foxx, and Cruise*, described above, with a false information score of 0.94.
- *Pitt and Jolie*: this topic refers to scandals and gossip about Brad Pitt and Angelina Jolie, with a false information score of 0.91.
- *Aniston, Theroux, and Pitt*: this topic refers to gossip about Jennifer Aniston, her ex-husbands Brad Pitt, and Justin Theroux, with a false information score of 0.89.

Table 4 provides an example of false information detected by *TM-FID* for each of the three topics highlighted above. The reported tweets express untrustworthy theories and gossip about: (i) a love triangle involving Aniston, Theroux, and Pitt; (ii) Pitt and Jolie's separation; and (iii) romantic meetings between Holmes and Foxx. For each example, the related topic, the text of the tweet, the associated image, the false information probability, and the topic-level false information score are reported.

Table 4 Example test posts, identified as false information, for each of the top-3 topics per false information score.

Topic	Post text	Post image	p_x^{fi}	p_x^c
<i>Aniston, Theroux, and Pitt</i>	Aniston, Theroux and Pitt involved in Secret Love Triangle Revealed! Secret meetings, jealousies, and reconciliations in an explosive mix!		0.97	1.00
<i>Pitt and Jolie</i>	Brangelina Scandals - Separation orchestrated for the secret project?		0.98	0.99
<i>Holmes, Foxx, and Cruise</i>	Romantic evening with dinner for Katie Holmes and Jamie Foxx! Rumors about a possible happy event!		0.96	0.99

5 Conclusion

In the current digital age, social media has permeated every aspect of our daily lives, making it crucial to assess both the benefits and challenges of their rapid and widespread dissemination. While social media offers access to diverse information and global conversations, allowing for a deeper understanding of complex issues, the spread of false information poses a significant risk, undermining trust and distorting our perception of the world. This paper focuses on analyzing Twitter news to identify and address false information circulating online by effectively leveraging both textual and visual data. Specifically, we propose a novel methodology, namely *TM-FID* (*Topic-oriented Multimodal False Information Detection*), which combines false information detection and neural topic modeling within a unified semi-supervised multimodal framework. Similarly to state-of-the-art approaches, this study tackles the inherently multimodal nature of social media content to enhance the detection of false information, leveraging Transformer-based models and cross-modal embedding fusion mechanisms. However, unlike other approaches that generally treat the false information issue within the vast landscape of social media discourse as a singular entity, *TM-FID* offers a more nuanced analysis by adopting a topical perspective. Employing such an approach fosters a deeper understanding of how false information shapes and influences discussions focused on specific topics, highlighting the underlying factors and dynamics that facilitate its dissemination. This understanding is critical for developing targeted interventions and effective strategies to counter the spread of false information, helping to strengthen trustworthiness and confidence in information shared on social media. Future work could focus on integrating more sophisticated embedding fusion mechanisms, also assessing model performance across diverse scenarios. Another area for improvement is related to the high computational demand associated with processing and analyzing large volumes of multimodal data through Transformer-based models. To enhance scalability, the use of compression techniques such as model distillation and quantization will be explored, alongside distributed computing solutions. Finally, we will investigate how the proposed methodology can be applied in more challenging situations, where prior classification of deception-related information is limited or absent.

Declarations

Funding. This work has been partially supported by the “FAIR – Future Artificial Intelligence Research” project - CUP H23C22000860006, and Next Generation EU - Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of “Innovation Ecosystems”, building “Territorial R&D Leaders” (Directorial Decree n. 2021/3277) - project Tech4You - Technologies for climate change adaptation and quality of life improvement, n. ECS0000009.

Conflict of interest. The authors declare that they have no conflict of interest.

Ethics approval. Not applicable.

Consent for publication. Not applicable.

Data availability. Data employed in our experiments are publicly available at the following link: <https://github.com/KaiDMML/FakeNewsNet/tree/master>.

Materials availability. Not applicable.

Code availability. Source code is publicly available at the following link: <https://github.com/SCA labUnical/TM-FID>.

Author contribution. All authors conceived the presented idea and contributed to the structure of this paper, helping to shape the research. R.C. and C.C. designed the deep learning methodology, carried out the experiments, and wrote the manuscript. R.C., I.K., F.M., and D.T. reviewed the manuscript, while D.T. supervised the whole process, from the assessment of achieved results to the organization of the manuscript. All authors have read and agreed to the published version of the paper.

References

- [1] Belcastro, L., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Learning political polarization on social media using neural networks. *IEEE Access* **8**, 47177–47187 (2020)
- [2] Cantini, R., Marozzo, F., Bruno, G., Trunfio, P.: Learning sentence-to-hashtags semantic mapping for hashtag recommendation on microblogs. *ACM Trans. Knowl. Discov. Data (TKDD)* **16**(2), 1–26 (2021)
- [3] Cantini, R., Cosentino, C., Kilanioti, I., Marozzo, F., Talia, D.: Unmasking covid-19 false information on twitter: A topic-based approach with bert. In: *International Conference on Discovery Science*, pp. 126–140 (2023). Springer
- [4] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* **8**(3), 171–188 (2020)
- [5] Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., Luo, J.: Detection and analysis of 2016 us presidential election related rumors on twitter. In: *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, pp. 14–24 (2017). Springer
- [6] Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Analyzing political polarization on social media by deleting bot spamming. *Big Data Cogn. Comput.* **6**(1), 3 (2022)
- [7] Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., Shang, L.: Cross-modal ambiguity learning for multimodal fake news detection. In: *Proceedings of the ACM Web Conference 2022*, pp. 2897–2905 (2022)
- [8] Wei, Z., Pan, H., Qiao, L., Niu, X., Dong, P., Li, D.: Cross-modal knowledge distillation in multi-modal fake news detection. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4733–4737 (2022)

- [9] Singhal, S., Kabra, A., Sharma, M., Shah, R.R., Chakraborty, T., Kumaraguru, P.: Spotfake+: A multimodal framework for fake news detection via transfer learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13915–13916 (2020)
- [10] Zhou, X., Wu, J., Zafarani, R.: Safe: Similarity-aware multi-modal fake news detection. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 354–367 (2020). Springer
- [11] Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., Resnik, P.: Is automated topic model evaluation broken? the incoherence of coherence. *Adv. Neural Inf. Process.* **34**, 2018–2033 (2021)
- [12] Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
- [13] Nasir, J.A., Khan, O.S., Varlamis, I.: Fake news detection: A hybrid cnn-rnn based deep learning approach. *Int. J. Inf. Manag. Data Insights* **1**(1), 100007 (2021)
- [14] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S.: Ti-cnn: Convolutional neural networks for fake news detection. arXiv preprint arXiv:1806.00749 (2018)
- [15] Oliveira, N.R., Pisa, P.S., Lopez, M.A., Medeiros, D.S.V., Mattos, D.M.: Identifying fake news on social networks based on natural language processing: trends and challenges. *Information* **12**(1), 38 (2021)
- [16] Jarrahi, A., Safari, L.: Evaluating the effectiveness of publishers’ features in fake news detection on social media. *Multimed. Tools Appl.* **82**(2), 2913–2939 (2023)
- [17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process.* **30** (2017)
- [18] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
- [19] Kaliyar, R.K., Goswami, A., Narang, P.: Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimed. Tools Appl.* **80**(8), 11765–11788 (2021)
- [20] Kula, S., Choraś, M., Kozik, R.: Application of the bert-based architecture in fake news detection. In: 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12, pp. 239–249 (2021)
- [21] Vijjali, R., Potluri, P., Kumar, S., Teki, S.: Two stage transformer model for covid-19 fake news detection and fact checking. arXiv preprint arXiv:2011.13253

(2020)

- [22] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process.* **32** (2019)
- [23] Abavisani, M., Wu, L., Hu, S., Tetreault, J., Jaimes, A.: Multimodal categorization of crisis events in social media. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14679–14689 (2020)
- [24] Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200* (2020)
- [25] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
- [26] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019)
- [27] Liu, L., et al.: On the variance of the adaptive learning rate and beyond. *arxiv 2019*. *arXiv preprint arXiv:1908.03265* (2019)
- [28] Mosbach, M., Andriushchenko, M., Klakow, D.: On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884* (2020)
- [29] Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pp. 194–206 (2019). Springer
- [30] Egger, R., Yu, J.: A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology* **7** (2022)
- [31] Udupa, A., Adarsh, K., Aravinda, A., Godihal, N.H., Kayarvizhy, N.: An exploratory analysis of gsdmm and bertopic on short text topic modelling. In: *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, pp. 1–9 (2022). IEEE
- [32] Gabarron, E., Dorrnzoro, E., Reichenpfader, D., Denecke, K.: What do autistic people discuss on twitter? an approach using bertopic modelling. In: *Caring Is Sharing—Exploiting the Value in Data for Health and Innovation*, pp. 403–407 (2023)

- [33] Mendonça, M., Figueira, Á.: Topic extraction: Bertopic’s insight into the 117th congress’s twitterverse. In: Informatics, vol. 11, p. 8 (2024)
- [34] Gokcimen, T., Das, B.: Exploring climate change discourse on social media and blogs using a topic modeling analysis. *Heliyon* (2024)
- [35] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR
- [36] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
- [37] Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 530–539 (2014)