







Multimodal Rumour Detection: Catching News that Never Transpired!

Raghvendra Kumar¹ , Ritika Sinha¹ , Sriparna Saha¹ ,
and Adam Jatowt² 

¹ Indian Institute of Technology Patna, Dayalpur Daulatpur, India
raghvendra.kumar1004@gmail.com, ritika16sinha@gmail.com

² University of Innsbruck, Innsbruck, Austria
adam.jatowt@uibk.ac.at

Abstract. The growth of unverified multimodal content on microblogging sites has emerged as a challenging problem in recent times. One major roadblock to this problem is the unavailability of automated tools for rumour detection. Previous work in this field mainly involves rumour detection for textual content only. As per recent studies, the incorporation of multiple modalities (text and image) is provably useful in many tasks since it enhances the understanding of the context. This paper introduces a novel multimodal architecture for rumour detection. It consists of two attention-based BiLSTM neural networks for the generation of text and image feature representations, fused using a cross-modal fusion block and ultimately passing through the rumour detection module. To establish the efficiency of the proposed approach, we extend the existing PHEME-2016 data set by collecting available images and in case of non-availability, additionally downloading new images from the Web. Experiments show that our proposed architecture outperforms state-of-the-art results by a large margin.

Keywords: Rumour Detection · Multimodality · Deep learning · PHEME Dataset · Twitter

1 Introduction

Considering the recent developments in technology, there is still insufficient control over the proliferation and dissemination of information transmitted through untrusted online sources like micro-blogging sites [27]. This leads to the propagation of unverified news, especially in the context of breaking news, which may further unfold rumours among the masses [19]. These rumours, if left unmitigated, can influence public opinion, or corrupt the understanding of the event for journalists. Manually fact-checking the news in real-time is a tremendously

R. Kumar and R. Sinha—These authors contributed equally to this work.

difficult task. So, there is a need to automate the process of rumour detection to promote credible information on online sources.

Additionally, information transfer in the modern era is increasingly becoming multimodal. *Oxford Reference*¹ defines multimodality as “The use of more than one semiotic mode in meaning-making, communication, and representation generally, or in a specific situation”. Information on micro-blogging sites is rarely present in the textual mode only. These days, they contain images, videos, and audio, among other modalities of information.



Fig. 1. A sample Twitter thread

Twitter, a well-known micro-blogging site, lets users exchange information via brief text messages, that may have accompanying multimedia. For the task of rumour detection, it is actually more effective to utilize the entire Twitter thread. Reply tweets are useful in this regard as the users often share their opinions, suggestions, criticism, and judgements on the contents of the source tweet. Also, one gets a better understanding of the context when provided with visual cues (images). As elaborated through an example shown in Fig. 1, a user (through reply tweet), has indicated that the image associated with the source tweet has wrongly stated the crash site. Other users can also interact by adding their views. These replies on the source tweet help in understanding the situation better.

¹ <https://www.oxfordreference.com>.

Recent research on Twitter rumour detection usually uses the PHEME or SemEval datasets and it mainly involves machine learning-based models [19, 28]. In these methods, authors extracted statistical or lexical features from textual data. These fail to capture dependencies between the features, and the semantics of the Twitter thread. To overcome this problem, deep learning-based methods were applied. These methods provide a more robust representation of the features. However, all these works are done on a textual content only.

In complex disciplines where a sole modality might not be able to offer sufficient information, multiple modalities could strengthen the overall system performance by combining the advantages of different modalities and offering a more thorough and reliable representation of the data. Also, multimodal systems are more natural to interact with as they mimic how humans use diverse senses to comprehend their surroundings. Just as importantly, multimodal systems provide a more engaging experience by using multiple modalities to create dynamic and interactive content. In this work, we put forward a deep learning-based approach that employs multimodality via an extended dataset that we have prepared to distinguish between reliable and unreliable content on micro-blogging websites such as Twitter. We also conduct thorough studies on the proposed model and the extended dataset. Attention-based RNN models have mainly been used for fusing feature representations of the two modalities [11]. Cheung *et al.* [5] have made significant progress towards extending the PHEME-2016 dataset to make it multimodal². However, the authors have performed this extension only for those tweet threads where images had already been uploaded by the users, so the resulting dataset is only partially multimodal, with only 46% of the tweets containing images. We further extend the dataset to be fully multimodal.

The following is a summary of our work’s significant contributions:

- We propose a dual-branch Cross-fusion Attention-based Multimodal Rumour Detection (CAMRD) framework for effectively capturing multimodal interdependencies between the text and image modalities for the detection of rumours on micro-blogging sites.
- The PHEME-2016 dataset comprises textual data only. We extend the dataset³ by collecting images using the associated metadata and by the means of web scraping. We make use of cross-modal fusion of CAMRD to effectively capture the interdependencies between the two modalities.
- We perform extensive studies for selecting the best image amongst multiple images using various heuristics.

2 Related Works

Over the past few years, there have been significant research efforts focused on identifying rumours and misinformation. This section provides an overview of

² Unfortunately, that dataset was not made public.

³ <https://drive.google.com/file/d/1XR7g6UL8.4yqvo12alQn2iqmWvHb6iKr/view?usp=sharing>.

previous works in this area and delves into specific studies related to cross-modal learning which is related to the proposed framework for detecting rumours.

Previous methods for detecting rumours heavily relied on statistical analysis. One of the earliest works examined the spread of rumours on Twitter by analyzing the frequency of certain words in data sets [25]. Despite being constrained to specific events and unable to generalize to new scenarios, these techniques became the basis for later developments in machine learning-based approaches to detect rumours. Kwon et al. [12] were the first to use machine learning techniques for this task including support vector machines, decision trees and random forests, accompanied with linguistic features for rumour detection.

Others have used Recurrent Neural Networks (RNNs) [15] and Convolutional Neural Networks (CNNs) [4], as such neural network-based models are able to effectively uncover and learn the underlying patterns within a data set. These techniques are often combined with pre-trained non-contextual word embeddings, such as GloVe [20], which yield a unique vector for each word without taking note of the context. Contextual word embeddings like BERT [8] have also been used as the vector representation, as the generated embeddings convey the semantic meaning of each word. In this study, we have used BERTweet [18] which is a publicly available, large-scale language model that has been pre-trained specifically for understanding and processing English tweets. It is trained based on the RoBERTa [14] pre-training procedure to generate contextual embeddings.

The ensemble graph convolutional neural network (GCNN) technique proposed in [1] uses a two-branch approach, a graph neural network (GNN) and a convolutional neural network (CNN) to process the features of a node in relation to its neighbouring nodes and to obtain feature representations from weighted word embedding vectors, respectively. The weighted combination of these two branches' features is considered the final feature vector. This method results in a poor representation of the conversation as it does not fully utilize the relationship between global and local information. The method outlined in [22] suggested using a recurrent neural network (RNN) for rumour detection in conversations but it suffered the same limitation of not considering the relative importance of each message in the conversation. To overcome this limitation, we have incorporated an attention-weighted average module in order to achieve a more precise representation of tweets and images.

Visual computing techniques are also useful for the analysis of social media contents, as shown in [3]. In recent times, processes which earlier had unimodal input and output such as emotion and opinion analysis, fake-news identification, and hate-speech detection have now expanded into the multimodal domain [9] which is the integration of multiple modalities of information. Another paradigm, namely, knowledge distillation which is a method used to transfer knowledge from a multimodal model to a single-modal model, has been applied in various fields like computer vision and speech processing [2, 13, 23].

3 Dataset

3.1 Original Dataset

As mentioned earlier, we used the PHEME-2016 dataset [28] for our experiments. It is a collection of Twitter threads consisting of 3,830 non-rumour and 1,972 rumour tweets posted during 5 breaking news, namely Charlie Hebdo, Ferguson, Germanwings crash, Ottawa Shooting, and Sydney Siege. It contains metadata regarding source tweets and reply tweets for each tweet thread.

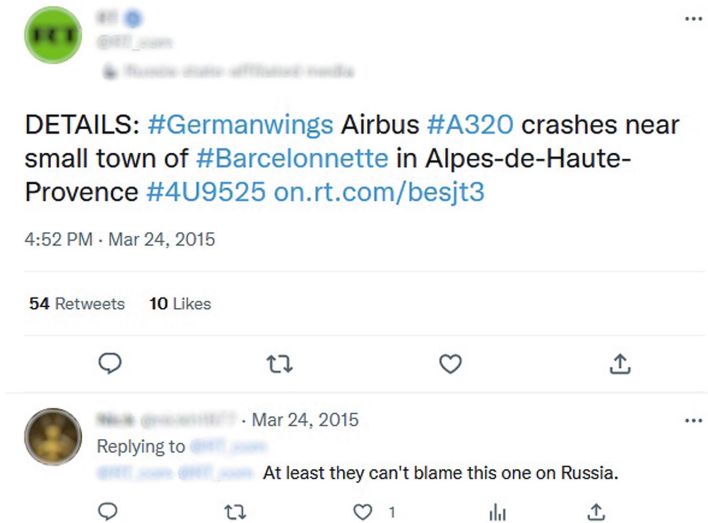


Fig. 2. Example of a tweet thread with no image

3.2 Dataset Extension

The original PHEME-2016 dataset initially did not contain any images. So the tweet threads which already had images uploaded by the users were collected using the image information specified in the metadata, with the distribution illustrated in Table 1. For the remaining tweet threads, we augmented images through web scraping⁴, as shown by an example via Fig. 2 and Fig. 3. Our criteria for downloading images included only the source tweet and not the reply tweets because reply tweets have higher chances of not being relevant to the tweet’s content, relevance, and appropriateness. Also, the rationale behind making the dataset multimodal was that even though the textual data can provide valuable information, they are often limited in their ability to convey complex or detailed information. Images can provide supplementary information and help to provide

⁴ <https://github.com/Joeclinton1/google-images-download>.



Fig. 3. Web scraped image for the above tweet thread showing rescuers examining the situation at the crash site

a comprehensive understanding of a situation. For example, in the case of the Ottawa shooting news topic in our dataset, the tweet “Penguins will also have the Canadian National Anthem before tonight’s game. A thoughtful gesture by the Pens. Sing loud, Pittsburgh #Ottawa”. was not able to fully illustrate the impact of the event, as the user may not have known that the “Penguins” here referred to a Canadian ice hockey team. However, the downloaded image of the sports team helps to provide additional context and clarify the meaning of the tweet.

Table 1. Distribution of tweet threads in rumour and non-rumour classes in PHEME-2016 Dataset, and count of images that were present originally on these threads.

News Event	Tweets		Images	
	Rumour	Non-rumour	Rumour	Non-rumour
Charlie Hebdo	458	1,621	234	1,010
Ferguson	284	859	72	414
Germanwings Crash	238	231	82	141
Ottawa Shooting	470	420	172	134
Sydney Siege	522	699	241	310

4 Problem Statement

Our research aims to create an end-to-end multimodal framework for rumour detection given an array of tweets $\{TUI\}$ where $T = \{T_1, T_2, \dots, T_n\}$ and each T_i represents the text content (further, $T_i = \{s_i, r_{i1}, r_{i2}, \dots, r_{ik}\}$ where s_i represents source tweet and reply tweets are denoted by r_{ik}), and $I = \{I_1, I_2, \dots, I_n\}$ where each I_i denotes the image related to the i^{th} tweet thread. The task is to train an end-to-end rumor detector $f : \{TUI\} \rightarrow \{Rumour, Non - Rumour\}$ by inter-relating the dependence of text and images.

5 Proposed Methodology

Our model primarily consists of three modules: (i) Attention-based feature extraction sub-modules of text and image modalities (ii) Cross-modal fusion module and (iii) Rumour Classification module (RC) for rumour detection, which are elaborated in detail in the following subsections. Figure 4 represents the architecture of our proposed model.

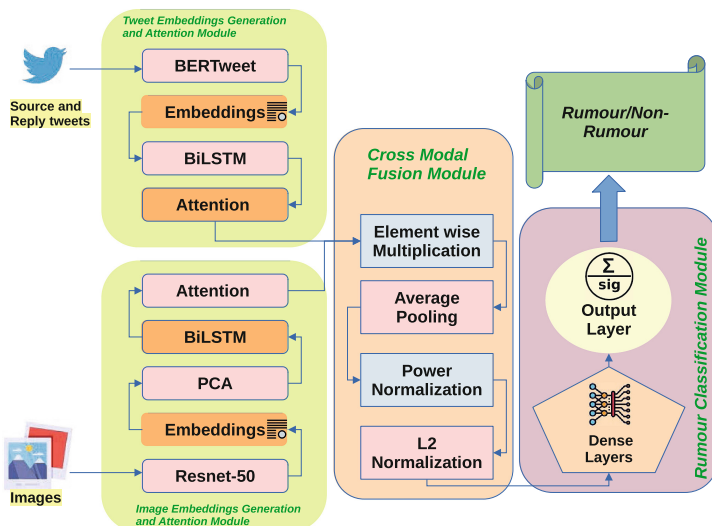


Fig. 4. Proposed dual-branch Cross-fusion Attention-based Multimodal Rumour Detection (CAMRD) model architecture

5.1 Embedding Generation

The following is the process used for the generation of embeddings across multiple modalities:

Tweet Embedding: BERTweet [18], a pre-trained language model for English tweets, was used for generating word-wise text embeddings. The *bertweet-large* model in particular (max-length = 512, Parameters = 355, pre-trained on 873M English tweets (cased)) was utilized. The normalized textual content of the source tweet s_i and reply tweets r_{ij} , where j is the number of reply tweets for a source tweet s_i were concatenated and passed through BERTweet:

$$t_i = s_i \oplus r_{ij} \quad (1)$$

$$Embed_{tweet} = BERTweet(t_i) \quad (2)$$

to get the required array of $[Embed_{tweet}]^{n*1024}$ dimensional embeddings, where n is the total count of tweet threads.

Image Embedding: The image features I_i for each tweet thread t_i are extracted using ResNet-50 [10] as indicated below,

$$Embed_{image} = ResNet50(I_i) \quad (3)$$

ResNet-50 is a convolutional neural network, 50 layers deep. A pre-trained version of the network trained on the ImageNet [7] database was used for our task. The images were first resized and normalized into a fixed dimension of $224 \times 224 \times 3$. The final array of $[Embed_{image}]^{n*2048}$ dimensional embeddings were obtained by extracting the features from the fully connected layer of the model. Next, in order to match the dimensions of the image vector to that of the tweet vector, we performed the Principal Component Analysis (PCA) on the image vector:

$$PCA([Embed_{image}]^{n*2048}) = [Embed_{image}^{PCA}]^{n*1024} \quad (4)$$

The vectors obtained after PCA showed that more than 95% variance of the data was retained. Lastly, we passed the embeddings to the BiLSTM layer, depicted below in Eq. 5.

$$Embed_{image}^{BiLSTM} = BiLSTM(Embed_{image}^{PCA}) \quad (5)$$

5.2 Attention-based Textual and Image Feature Extraction Module

Different modalities contribute to our design to the overall feature generation. In this module, we extract the textual and image feature representations.

The most suitable semantic representation of the text embeddings (explained above) is captured through the first branch of this module. The text embeddings are fed to this branch as represented in Eq. 6. It consists of a BiLSTM layer for capturing the long-term dependencies of textual embeddings in both forward and backward directions.

$$Embed_{tweet}^{BiLSTM} = BiLSTM(Embed_{tweet}) \quad (6)$$

Some words are more important than others, contributing more toward meaningful textual representation. Hence, the output of BiLSTM is passed through an attention layer, shown in Eq. 7, for extracting those words that are crucial to the tweet's meaning, and forming the final embedding vectors out of those descriptive words [26].

$$Embed_{tweet}^{final} = Attention(Embed_{tweet}^{BiLSTM}) \quad (7)$$

The second branch of this module focuses on the critical parts of the image for understanding which aspect of the picture makes it more likely to be categorized as rumour or non-rumour. Thus, we pass the output vector of the image from the previous module to an attention layer, represented in Eq. 8.

$$Embed_{image}^{final} = Attention(Embed_{image}^{BiLSTM}) \quad (8)$$

5.3 Cross-modal Fusion Module (CMF)

We use the cross-modal fusion module to merge the final textual and image feature vectors in lieu of a plain concatenation because the CMF module magnifies the association between the vectors of both modalities. Additionally, it overcomes the necessity of choosing the appropriate lengths of the extracted vectors which poses a challenge in plain concatenation. The first step of the CMF module involves element-wise multiplication of the vectors of the two modalities which adequately encapsulates the relationship between them.

$$Embed_{mul} = Embed_{tweet}^{final} * Embed_{image}^{final} \quad (9)$$

The $*$ in Eq.9 denotes element-wise multiplication. Next, average pooling is performed as it retains more information often than not when compared to max or min pooling.

$$Embed_{pooled} = Avg.Pooling(Embed_{mul}) \quad (10)$$

Then power normalization is carried out to reduce the chances of over-fitting.

$$Embed_{p-norm} = sgn(Embed_{pooled}) * \sqrt{|Embed_{pooled}|} \quad (11)$$

where $sgn(x)$ is the signum function, described in Eq. 12.

$$\begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \quad (12)$$

The last step in the module carries out L2 normalization, to minimize the overall error and lower the computational cost.

$$Embed_{L2-norm} = \|Embed_{p-norm}\|_2 = \sqrt{\sum (Embed_{p-norm})^2} \quad (13)$$

To recap, our CMF module aligns the tweets and image features by boosting their association and is novel compared to existing works that fixate on the plain concatenation of feature representations of different modalities. The mathematical formulation of the CMF module is explained using Eq.9 to Eq. 13.

5.4 Rumour Classification Module

Our final rumour classification module consists of five hidden layers with rectified linear activation function as depicted by

$$ReLU(z) = \max(0, z) \quad (14)$$

and an output layer with a logistic function,

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (15)$$

as an activation function. This module intakes the L2 normalized vectors, output by the CMF module, and extrapolates them into the objective array of two classes to yield the final expectancy probability that determines whether the multimodal tweet is rumour or non-rumour. The loss function used for our model is binary cross-entropy which is calculated as

$$- \sum (y \log(p) + (1 - y) \log(1 - p)) \quad (16)$$

where y is the binary indicator (0 or 1) and p is the predicted probability observation. The optimizer used is the Adam optimizer.

6 Experiments, Results, and Analysis

In this segment, we present and analyze the various experimental configurations we have used and their respective outcomes.

6.1 Experimental Setting and Evaluation Metrics

This section describes the process of extracting embeddings, the pre-processing of tweets and images, various hyperparameters used, and all the technical implementation details. Python libraries NumPy, Pandas, Keras, and Sklearn, were used on the TensorFlow framework to execute the tests. The system’s performance was assessed on the basis of parameters like accuracy, precision, recall, and F1-score.

The experimental protocol followed was random shuffling of the tweet threads, and then splitting into an 80–10–10 train-validation-test set. Fine-tuning was performed on the validation dataset. As raw tweets contain mostly unorganized and dispensable information, we preprocess the tweets using RegEx in Python by removing URLs, usernames, punctuation, special characters, numbers, and hash(#) characters in hashtags. In addition, character normalization was also done. Pre-processed tweets were then passed to the BERTweet model to generate embeddings of 1,024 dimensions. Next, a BiLSTM layer with 512 units intakes these feature vectors and passes them on to the attention module. In parallel, we re-scale images to the size of (224,224,3) and feed them to the ResNet-50 module, which produces a 2,048-dimensional feature vector that in turn is fed to PCA to reduce it to 1,024 dimensions. Next, we pass it through the attention module. Following that, both the vectors are then advanced to the CMF module which fuses them and then feeds them to the rumor classification module that has five hidden layers with 512, 256, 128, 64, and 16 units, respectively, and the output layer with 1 unit. The activation functions for the hidden layers and the output layer are the ReLU and Sigmoid, respectively. Our model is trained with a batch size of 32 and an epoch number of 100. Table 2 summarizes the details of all the hyperparameters used for training our model. The hyperparameters shown in Table 2 were concluded after performing a parameter sensitivity study in which we analyzed the effect of variations in individual parameters over a range and observed how it affected the output of the system.

Table 2. Hyperparameters utilized for training the presented model.

Parameters	Values
Tweet length	512
Image size	(224, 224, 3)
Optimizer	Adam
Learning rate	0.001
Batch Size	32
Epochs	100
Filter size	(3, 3)
Strides	(1, 1)
Padding	'same'

6.2 Results and Discussion

In addition to the augmented PHEME-2016 dataset which contains a single image per tweet thread, i.e., 5,802 tweet-image pairs in total, we also created a further expanded dataset for experimental purposes which consists of multiple images, with a majority of tweets containing two images and few of the tweets with several images, altogether, 10,035 images. Various heuristics were applied to select the best image from the multiple images, which are as follows:

- H1: Manual selection of a single image per tweet.
- H2: Selection of the first image present in the dataset per tweet.
- H3: Random selection of an image per tweet.
- H4: Selection of the image with the largest dimension present in the dataset per tweet.

The first heuristic means that a human is manually selecting the best image for a tweet thread and the instructions followed by the human to manually select the image included:

- Relevance: The image should be relevant to the content of the tweet and help to convey the message or meaning of the tweet visually.
- Coherent to the hashtags: The image should align with the hashtags and help to convey the intended message or meaning.

Table 3 illustrates that amidst the various heuristics used, manually selecting the images produces the optimal result in terms of all four evaluation parameters with our proposed model. The proposed model produces the poorest accuracy when no heuristics were applied and multiple images were used. After the establishment of the best heuristic, we conducted experiments with various combinations of image and tweet embeddings including ResNet-50, ResNet-101 [10], ResNet-152 [10], VGG-19 [21], VGG-16 [21], InceptionV3 [24], and Xception [6] and BERTweet and OpenAI, respectively. Table 4 shows that the embeddings

generated by the combination of BERTweet and ResNet-50 produce the optimal outcomes in terms of accuracy, recall, and F1-score which demonstrate that it is the best-performing embedding-generating combination. The conjunction of BERTweet and ResNet-101 performs exceptionally well in terms of precision.

Table 3. Results obtained on the proposed CAMRD model using various heuristics.

Heuristic Used	Accuracy	Precision	Recall	F1-Score
H1	0.893	0.913	0.917	0.915
H2	0.877	0.911	0.906	0.909
H3	0.841	0.850	0.915	0.881
H4	0.844	0.881	0.875	0.878
None	0.832	0.858	0.897	0.877

Table 4. Results obtained after using different tweet and image embeddings on the proposed CAMRD model and the heuristics of manually selected images.

Tweet Embd.	Image Embd.	Accuracy	Precision	Recall	F1-Score
BERTweet	ResNet-50	0.893	0.913	0.917	0.915
BERTweet	ResNet-101	0.875	0.914	0.898	0.906
BERTweet	ResNet-152	0.868	0.888	0.917	0.902
BERTweet	VGG-19	0.860	0.897	0.890	0.894
BERTweet	VGG-16	0.862	0.887	0.908	0.897
BERTweet	InceptionV3	0.863	0.900	0.893	0.896
BERTweet	Xception	0.859	0.884	0.900	0.892
OpenAI	ResNet-50	0.886	0.897	0.891	0.894
OpenAI	ResNet-101	0.872	0.897	0.885	0.891
OpenAI	ResNet-152	0.871	0.872	0.913	0.892
OpenAI	VGG-19	0.882	0.909	0.888	0.898
OpenAI	VGG-16	0.877	0.887	0.904	0.895
OpenAI	InceptionV3	0.864	0.896	0.875	0.880
OpenAI	Xception	0.893	0.910	0.904	0.907

6.3 Comparison with State of the Art and Other Techniques

In this part, we compare our CAMRD model with the existing State-of-the-Art (SOTA) and various other techniques. In Multi-Task Framework To Obtain Trustworthy Summaries (MTLTS) [17], the authors have proposed an end-to-end solution that jointly verifies and summarizes large volumes of disaster-related

tweets, which is the current SOTA that we use for comparison. The other techniques that we compare our model with are SMIL: Multimodal Learning with Severely Missing Modality [16], Gradient Boosting classifier which attains the best results amongst different machine learning classifiers that we experimented upon and the final technique in which we passed tweets to both the branches of our proposed model. The techniques, namely, our proposed CAMRD model and gradient boosting classifier, when used with a single image per tweet and multiple images per tweet have been represented in Table 5 with subscript *SI* and *MI*, respectively. The values of the main parameters for the Gradient Boosting technique were as follows, learning_rate = 0.1, loss = ‘log_loss’, min_samples_split = 2, max_depth = 3. These values were selected after parameter sensitivity testing.

Table 5 demonstrates that our proposed model when operated with a single image per tweet outperforms the SOTA and other techniques in terms of accuracy and precision, while gradient boosting classifier when used with multiple images per tweet shows superior results in terms of recall and f1-score, however, this technique seriously falters in the case of a single image per tweet.

Table 5. Comparative analysis of SOTA and other methods

Approach	Accuracy	Precision	Recall	F1-Score
<i>CAMRD_{SI}</i>	0.893	0.913	0.917	0.915
<i>CAMRD_{MI}</i>	0.832	0.858	0.897	0.877
SMIL [16]	0.815	0.823	0.835	0.829
<i>GB_{SI}</i>	0.741	0.765	0.873	0.816
<i>GB_{MI}</i>	0.884	0.884	0.951	0.917
MTLTS [17]	0.786	0.77	0.766	0.768
Text-only	0.733	0.803	0.791	0.797

6.4 Classification Examples and Error Analysis

In this section we show few examples of tweets and explain the causes for both cases, when the tweets were correctly classified, and error analysis when the tweets were wrongly classified. The first case covers the true positive and true negative scenario where tweets from each class, rumour and non-rumour were rightly categorized as depicted in Fig. 5. The left-hand side tweet thread has been correctly classified as a rumour, with the subsequent reply tweets on the source tweet questioning the authenticity of the latter. The right-hand side tweet has been correctly classified as non-rumour as the image supports the tweet.

The second case covers the false positive as shown in Fig. 6 and false negative as shown in Fig. 7 where tweets from each class, rumour and non-rumour were wrongly categorized.

Here, the image shown in Fig. 6 represents disrespectful behaviour as it shows abusive symbols in public, and the text says that it is alright to rage against what happened, so the model might be biased against public display of resentment and this may be the reason for this tweet getting misclassified. The entire Twitter thread as shown in Fig. 7 seems quite convincing, as well the image also represents a man being pointed fingers at, which is quite in line with the word ‘punishable’. This may be the reason for this tweet getting misclassified.

In general, if there are discrepancies between text and image, the CAMRD model may misclassify them, which is reflected via examples (Fig. 6 and Fig. 7).

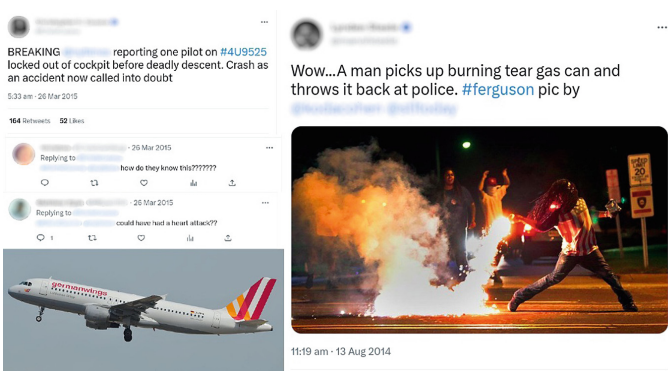


Fig. 5. Example of tweets correctly classified, left-hand side tweet thread and corresponding picture belongs to a rumour class and the right-hand side tweet and image belongs to non-rumour class



Fig. 6. Example of a non-rumour tweet wrongly classified as rumour



Fig. 7. Example of a rumour tweet wrongly classified as non-rumour

7 Conclusions and Future Works

In this paper, we have introduced a novel deep-learning based model named Cross-fusion Attention-based Multimodal Rumour Detection (CAMRD), which takes text and image(s) as input and then categorizes it as either rumour or non-rumour. Attention-based textual and visual features were extracted. Instead of a plain concatenation of both features, we have used the cross-modal fusion module which then passes it to the Multilayer Perceptron segment that classifies the data. Additionally, the PHEME 2016 dataset has been expanded and made fully multimodal by means of image augmentation. The diverse experiments conducted on the expanded dataset show that our approach is effective for multimodal rumour detection. Most social media platforms these days have seen a huge rise in textual and visual content being uploaded, in the form of short videos, commonly known as ‘reels’, which certainly opens our work to be extended in the direction of including videos and audio as well.

Acknowledgements. Raghvendra Kumar would like to express his heartfelt gratitude to the Technology Innovation Hub (TIH), Vishlesan I-Hub Foundation, IIT Patna for providing the Chanakya Fellowship, which has been instrumental in supporting his research endeavours. Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

Author contributions. Raghvendra Kumar, Ritika Sinha : These authors contributed equally to this work.

References

1. Bai, N., Meng, F., Rui, X., Wang, Z.: Rumour detection based on graph convolutional neural net. *IEEE Access* 1 (2021). <https://doi.org/10.1109/ACCESS.2021.3050563>
2. Bai, Y., Yi, J., Tao, J., Tian, Z., Wen, Z., Zhang, S.: Fast end-to-end speech recognition via non-autoregressive models and cross-modal knowledge transferring from bert. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* **29**, 1897–1911 (2021). <https://doi.org/10.1109/TASLP.2021.3082299>
3. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*, pp. 223–232. Association for Computing Machinery, New York, NY, USA (2013). <https://doi.org/10.1145/2502081.2502282>
4. Chen, Y., Sui, J., Hu, L., Gong, W.: Attention-residual network with CNN for rumor detection. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*, pp. 1121–1130. Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3357384.3357950>
5. Cheung, T.H., Lam, K.M.: Transformer-graph neural network with global-local attention for multimodal rumour detection with knowledge distillation (2022). <https://doi.org/10.48550/ARXIV.2206.04832>, <https://arxiv.org/abs/2206.04832>
6. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807. IEEE Computer Society, Los Alamitos, CA, USA, July 2017. <https://doi.org/10.1109/CVPR.2017.195>, <https://doi.org/ieeecomputersociety.org/10.1109/CVPR.2017.195>
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. <https://doi.org/10.18653/v1/N19-1423>, <https://aclanthology.org/N19-1423>

9. Ghani, N.A., Hamid, S., Targio Hashem, I.A., Ahmed, E.: Social media big data analytics: a survey. *Computers in Human Behavior* **101**, 417–428 (2019). <https://doi.org/10.1016/j.chb.2018.08.039>, <https://www.sciencedirect.com/science/article/pii/S074756321830414X>
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016
11. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: *Proceedings of the 25th ACM International Conference on Multimedia, MM 2017*, pp. 795–816. Association for Computing Machinery, New York, NY, USA (2017). <https://doi.org/10.1145/3123266.3123454>, <https://doi.org/10.1145/3123266.3123454>
12. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: *2013 IEEE 13th International Conference on Data Mining*, pp. 1103–1108 (2013). <https://doi.org/10.1109/ICDM.2013.61>
13. Liu, T., Lam, K., Zhao, R., Qiu, G.: Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Trans. Circ. Syst. Video Technol.* **32**(1), 315–329 (2022). <https://doi.org/10.1109/TCSVT.2021.3060162>
14. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019)
15. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*, pp. 3818–3824. AAAI Press (2016)
16. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: multimodal learning with severely missing modality (2021). <https://doi.org/10.48550/ARXIV.2103.05677>, <https://arxiv.org/abs/2103.05677>
17. Mukherjee, R., et al.: MTLTS: a multi-task framework to obtain trustworthy summaries from crisis-related microblogs. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, ACM*, February 2022. <https://doi.org/10.1145/3488560.3498536>
18. Nguyen, D.Q., Vu, T., Nguyen, A.T.: BERTweet: a pre-trained language model for English Tweets. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 9–14 (2020)
19. Pathak, A.R., Mahajan, A., Singh, K., Patil, A., Nair, A.: Analysis of techniques for rumor detection in social media. *Procedia Comput. Sci.* **167**, 2286–2296 (2020). <https://doi.org/10.1016/j.procs.2020.03.281>, <https://www.sciencedirect.com/science/article/pii/S187705092030747X>, international Conference on Computational Intelligence and Data Science
20. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar, October 2014. <https://doi.org/10.3115/v1/D14-1162>, <https://aclanthology.org/D14-1162>
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations* (2015)
22. Song, C., Yang, C., Chen, H., Tu, C., Liu, Z., Sun, M.: Ced: credible early detection of social media rumors. *IEEE Trans. Knowl. Data Eng.* **33**(8), 3035–3047 (2021). <https://doi.org/10.1109/TKDE.2019.2961675>

23. Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for BERT model compression. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4323–4332. Association for Computational Linguistics, Hong Kong, China, November 2019. <https://doi.org/10.18653/v1/D19-1441>, <https://aclanthology.org/D19-1441>
24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016). <https://doi.org/10.1109/CVPR.2016.308>
25. Takahashi, T., Igata, N.: Rumor detection on twitter. In: The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems, pp. 452–457 (2012)
26. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489. Association for Computational Linguistics, San Diego, California, June 2016. <https://doi.org/10.18653/v1/N16-1174>, <https://aclanthology.org/N16-1174>
27. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and resolution of rumours in social media. *ACM Comput. Surv.* **51**(2), 1–36 (2018). <https://doi.org/10.1145/3161603>
28. Zubiaga, A., Liakata, M., Procter, R.: Exploiting context for rumour detection in social media. In: Ciampaglia, G.L., Mashhadi, A., Yasseri, T. (eds.) *SocInfo 2017*. LNCS, vol. 10539, pp. 109–123. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67217-5_8