Expert Systems With Applications 184 (2021) 115412

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection

Rina Kumari *, Asif Ekbal

Department of Computer Science and Engineering, Indian Institute of Technology Patna, India

ARTICLE INFO

Keywords:
Multimodal fake news detection
Deep learning
Attention mechanism
Multimodal Feature fusion
Multimodal Factorized Bilinear Pooling

ABSTRACT

Fake news is the information or stories that are intentionally created to deceive or mislead the readers. In recent times, Fake news detection has attracted the attention of researchers and practitioners due to its many-fold benefits, including bringing in preventive measures to tackle the dissemination of misinformation that could otherwise disturb the social fabrics. Social media in recent times are heavily loaded with multimedia news and information. People prefer online news reading and find it more informative and convenient if they have access to multimedia content in the forms of text, images, audio, and videos. In early studies, researchers have proposed several fake news detection mechanisms that mostly utilize the textual features and not proper to learn multimodal (textual + visual) shared representation.

To overcome these limitations, in this paper, we propose a multimodal fake news detection framework with appropriate multimodal feature fusion that leverages information from text and image and tries to maximize the correlation between them to get the efficient multimodal shared representation. We empirically show that text, when combined with the image, can improve the performance of the model. The model detects the post once it is introduced into the network in an early stage. At the early stage of a news post's introduction into the network, the model takes the text and image of the post as input and decides whether this is fake or genuine. Since this model only analyzes news contents, It does not require any prior information regarding the user and network details. This framework has four different sub-modules *viz.* **Attention Based Stacked Bidirectional Long Short Term Memory (ABS-BiLSTM)** for textual feature representation, **Attention Based Multilevel Convolutional Neural Network–Recurrent Neural Network (ABM-CNN–RNN)** for visual feature extraction, **multimodal Factorized Bilinear Pooling (MFB)** for feature fusion and finally **Multi-Layer Perceptron (MLP)** for the classification. We perform experiments on two publicly available datasets, *viz.* Twitter and Weibo. Evaluation results show the efficacy of our proposed approach that performs significantly better compared to the state-of-the-art models. It shows to outperform the current state-of-the-art by approximately 10 points for the Twitter dataset. In contrast, the Weibo dataset achieves an overall better performance with balanced F1-scores between fake and real classes. Furthermore, the complexity of our proposed model is significantly lower than the state-of-the-art.

1. Introduction

Misleading and fake news content on social media is one of the considerable challenges in our society. Fake news can be defined as the information created intentionally by manipulating text, images, audios, or videos. Some popular social media platforms, such as Twitter, Facebook, and blogs, play an inevitable role in the rapid dissemination of news stories. As the number of online news readers is continuously increasing, some people are taking advantage of spreading false information to mislead them. This can create a negative impact on society and even manipulate important public events. The US presidential election 2016 gives a better realization of it. During this election, fake news was generated to support either of the two candidates. Many people had believed and also shared more than 37 million times on Facebook (Wang et al., 2018).

Fig. 1 depicts a few examples of multimodal fake news from the Twitter dataset, which we use for evaluation (details are in subsequent sections). Each tweet contains a piece of text associated with an image. The image in the first tweet has been photo-shopped, but the text is real because the solar eclipse was a real event, but the view was different. The second tweet shows the actual image, but this is a picture of a

* Corresponding author.
E-mail addresses: rina_1921cs13@iitp.ac.in (R. Kumari), asif@iitp.ac.in (A. Ekbal).
URL: http://www.iitp.ac.in/~asif/ (A. Ekbal).

https://doi.org/10.1016/j.eswa.2021.115412
Received 12 August 2020; Received in revised form 11 March 2021; Accepted 9 June 2021
Available online 29 June 2021

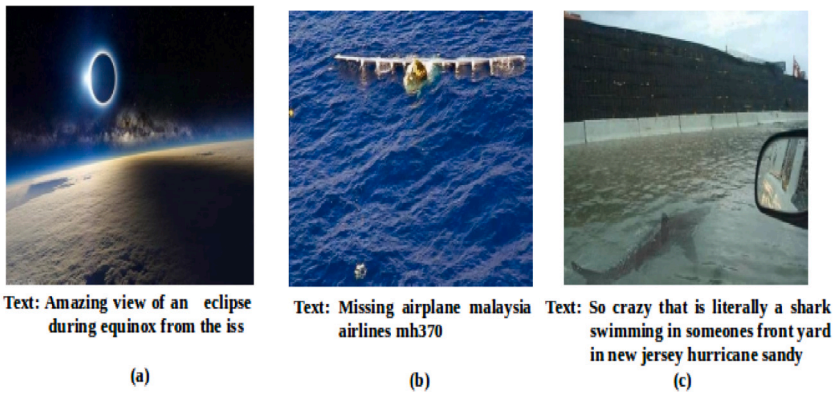

Fig. 1. Fake news example from the twitter data set.

plane crash in Sicily. In the right part of the image, an artificial shark has been created that did not exist during Hurricane Sandy. These tweets are potentially fake, and their primary purpose of diffusion is to mislead the population. It is a very challenging task to detect this type of information as fake or real.

Problem Definition: In our current work, we address the problem of fake news detection by leveraging information from multiple sources, such as text and image. In particular, we focus on utilizing and fusing textual and visual content to detect a multimedia post as fake or real. We formulate this problem as follows.

Suppose we are given a set of 'm' number of multimedia news posts $P = (p_1, p_2, \ldots, p_m)$. *Each post* (p_m) *contains texts* $T = (t_1, t_2, \ldots, t_m)$, *corresponding images* $I = (i_1, i_2, \ldots, i_m)$ *and labels* $Y = (y_1, y_2, \ldots, y_m)$. *Text content* (t_m) *of the news post is composed of either a single sentence or group of sentences. Label* (y_m) *is given for each post* (p_m). *A classifier 'C' is to be trained with the help of given set of the news posts. The classifier takes the whole text content along with the associated image of a news post*(p_m) *as input to classify the given post as fake* $(y_k = 0)$ *or real* $(y_k = 1)$, *i.e.* $(y_k = f(t_k, i_k))$. *Here,* y_k, t_k *and* i_k *are the predicted label, text instance and corresponding image of the news post, respectively and* $k \in [1...m]$.

Motivation and Contribution: Many traditional learning methods, and in recent times the deep learning models have been made available for fake news detection. However, the focus has been predominantly on text, e.g., Castillo, Mendoza, and Poblete (2011). Examples in Fig. 1 show that the fake news content cannot be correctly identified using only a single modality, i.e., either text or image modality information. The identification could be possible if we exploit the information from various sources (i.e., text and image), which gives rise to the concept of multimodal fake news detection.

Some of the prior works, such as Khattar, Goud, Gupta, and Varma (2019), Wang et al. (2018) have tried to learn the shared features among all the events and the correlation between text and image. These mechanisms provide equal importance to all parts of text and image, which may not be a correct approach to fuse the information. In Singhal, Shah, Chakraborty, Kumaraguru, and Satoh (2019) the author uses Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2018) and pre-trained VGG19 model (Shaha & Pawar, 2018) for textual and visual feature extraction, respectively. After extraction, they concatenated both the features that do not provide a good correlation between the text and image. It may not detect misleading news (news shown in the second example of Fig. 1) correctly.

In the literature, authors utilized mainly the pre-trained convolution neural network (CNN) like VGG19 (Shaha & Pawar, 2018) for visual feature extraction. The VGG19 was trained on the ImageNet dataset (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009), which is most suitable for object detection. As the ImageNet dataset is a general domain dataset, VGG19 may not capture the domain-specific semantic features of fake news images due to the lack of task-relevant information. So extraction of the inherent characteristics of fake news images is a very challenging task. Apart from feature extraction, feature fusion is also a very crucial task. Authors have just concatenated textual and visual features to obtain a joint multimodal feature representation in the existing works. Based on this, they have classified the news posts as fake or real.

Inspired by the lack of good feature extraction and effective feature fusion mechanisms, we propose a novel deep neural network-based framework for multimodal fake news detection in this work. This framework is named as "**A**ttention based **M**ultimodal **F**actorized **B**ilinear Pooling (AMFB)".

We summarize the contributions of this work as below:

(i). We propose an Attention Based Stacked Bi-directional LSTM (ABS-BiLSTM) network that captures textual information at different levels.

(ii). We propose an Attention Based Multi-level Convolutional Neural Network–Recurrent Neural Network (ABM-CNN–RNN) for visual feature extraction that extracts inherent features from the image.

(iii). We combine the textual and visual feature representations using the MFB module and pass it through a Multi-Layer Perceptron (MLP) model with two hidden layers and one output layer with a sigmoid activation function for fake news detection.

(iv). We perform extensive experiments on two benchmark datasets to validate the performance of the proposed model. The evaluation shows that the proposed approach attains a new state-of-the-art performance.

The remainder of this paper is organized as follows. In Section 2, we present a brief survey of the related works. Section 3 explains the research objectives. Section 4 depicts the methodology in detail. In Section 5, we first describe the datasets used for our experiments and demonstrate the experiments performed, followed by detailed analysis. Section 6 concludes our work along with some road maps for future direction.

2. Related work

A news article contains various aspects such as source, author, catchy headlines, writing styles, images, videos, etc. Any changes made to these aspects yield misleading or fake news. Fake news detection task is somewhat similar to many other tasks, such as rumor detection (Jin, Cao, Jiang, & Zhang, 2014; Zhang, Fang, Qian, & Xu, 2019), spam detection (Liu, Pang et al., 2019; Shen et al., 2017), and satire detection (Rubin, Conroy, Chen, & Cornwell, 2016). Rumor detection is very similar to fake news detection because fake news is a type of false rumor.

The method proposed in Alkhodair, Ding, Fung, and Liu (2020) detects breaking news rumors rather than long-lasting rumors. The theory behind this mechanism is **the breaking news rumors spread rapidly**

compared to long-lasting rumors because of less awareness and time-consuming automatic fact verification algorithms. It demonstrates an approach that simultaneously learns word2vec (Church, 2017) word embeddings and recurrent neural networks with multiple objectives to automatically identify breaking news rumors. The Word2vec provides context-independent word vectors, and simple RNN only takes care of the sequential representation of the words in the text but does not extract the vital information from the text. So, the performance of this framework can be improved using some contextual word embedding like BERT and attention mechanism over RNN during implementation.

The work proposed in Liu, Jin et al. (2019) tries to capture dynamic changes of news contents, news spreaders, and diffusion structures. Authors in this paper have implemented a Short-Term Long Memory (LSTM) based early rumor detection model to identify rumors in the initial stage. It is a novel idea to capture the dynamic differences between the diffusion structures and spreaders of rumors and non-rumors. However, if the same message is written in a different word sequence, the model fails. More specifically, it gives incorrect predictions in the case of paraphrasing. In another work (Zubiaga et al., 2018), authors have shown that sequential classifier performance increases if it exploits the discourse features extracted from social media interactions.This paper also shows the LSTM to outperform the sequential classifier while using a reduced set of features. Since this is the feature-based mechanism, the attention over LSTM can improve the performance by selecting important information from the selected feature set. Researchers have developed various approaches to detect fake news and stop disseminating it on social media. Below we present a review of the existing works in two broad categories: (i). Fake news detection based on a single modality, i.e., unimodal; and (ii). Fake news detection based on multi-modality.

2.1. Fake news detection based on single modality

Fake news detection has been predominantly carried out for text only, but authors have also focused on visual and contextual information in recent studies. Textual features are generally the semantic or statistical features extracted from the text. The authors in Faustini and Covões (2020) proposed a mechanism to detect fake news using only text features. It generates source platform independent and language independent textual features and apply Naive Bayes (NB) (Rish et al., 2001), Random Forest (RF) (Gilda, 2017), K-Nearest Neighbor (KNN) (Zhang, 2016) and Support Vector Machine (SVM) (Huang et al., 2018) for fake news classification . This mechanism does a kind of feature engineering that cannot be a faster and automatic solution for fake news detection. This work can further be improved if it automatically learns the language and platform-independent feature representation and performs fake news detection using end-to-end deep learning algorithms. The work reported in Castillo et al. (2011) proposed automatic credibility assessment using classification techniques such as SVM and Decision Tree (DT) (Song & Ying, 2015) for identifying the credibility of the tweets based on user and tweet features, which are mainly statistical and semantic types. In this work, the performance can be improved by utilizing an automatic feature extraction model because the statistical features do not capture the dynamic changes in the propagation over time.

In Gravanis, Vakali, Diamantaras, and Karadais (2019), the authors have introduced a machine learning-based ensemble model that uses content-based features and machine learning algorithms for fake news detection. Machine learning algorithms work well, but it is challenging to obtain hand-crafted textual features for the traditional machine learning based fake news detection models due to lack of domain knowledge and expertise. (Kwon, Cha, Jung, Chen, & Wang, 2013; Rashkin, Choi, Jang, Volkova, & Choi, 2017) identified the rumors based on the three aspects of dissemination: structural, temporal, and linguistic. The author has demonstrated the linguistic and structural key differences during the diffusion of rumor and non-rumor posts. Proposed methods extract the linguistic features of the deceitful text. They pull the features by analyzing how the language pattern of real news is different from hoaxes, satire, and propaganda. Based on these linguistic features, the models classify the news post as true or false. The linguistic features are highly dependent on domain knowledge and specific events; the linguistic patterns are not yet well understood. So, the models can further be strengthened by improving the extracted linguistic features.

Authors of Potthast, Kiesel, Reinartz, Bevendorff, and Stein (2018) have reported a mechanism that analyzes the style of fake news and hyper-partisan (extremely biased) information. They show how mainstream and hyper-partisan news can be distinguished by style analysis.The writing styles of the same news make the fake news detection task difficult and make wrong predictions.

Apart from textual and linguistic features, social context also provides useful evidence for fake news detection. Recently, in Shu, Wang et al. (2019), authors have depicted how the social contexts are utilized in fake news detection. This paper proposed a framework for tri-relationship embedding named TriFN that simultaneously models user–news interactions and publisher–news relations for bogus news classification. Shu, Cui et al. (2019) explains why a particular piece of news or post is detected as fake. The authors have developed a sentence-comment co-attention sub-network that exploits news contents along with user comments to jointly learns and captures check-worthy sentences and user comments for explainable detection of fake news. Since the social context features are unstructured, very noisy, and labor-intensive to collect, it cannot provide sufficient and relevant information for newly emerged events. The work presented in Shu, Wang et al. (2019) and Shu, Cui et al. (2019) can be improved if the social context features and their selection mechanisms are strengthened.

Since all the above-discussed models are feature-based, it requires domain knowledge to derive a good feature set, which is time-consuming and requires many efforts. In contrast, deep learning-based models such as (Huang & Chen, 2020; Ma et al., 2016) can extract the features automatically from the data and hence does not need any in-depth knowledge. In Huang and Chen (2020), the authors have introduced an ensemble-based mechanism that combines different deep learning models for fake news detection. In Ruchansky, Seo, and Liu (2017), the author has proposed a recurrent neural network (RNN) based model that tries to find out the patterns of user activities on a given post and decides whether the post is fake or real based on the user activities. Although deep learning models extract the features automatically, they may very often contain noise and irrelevant features that degrade the model performance.

Recently, In Wu, Rao, Nazir, and Jin (2020), authors have explained how the extracted features through deep learning suffer from many noisy and irrelevant features that reduce the performance of the approaches. They have proposed a novel model based on Adversarial neural networks aiming to reduce irrelevant and redundant features from the extracted features for information credibility measure. Although the adversarial training of the model gives better features, but it is tough to train. Again the model is very complex to run with limited resources, and it also takes a long period to generate the relevant and noiseless features. Authors in Karimi, Roy, Saba-Sadiya, and Tang (2018) introduced an approach that combines information from multiple sources and proposed a framework named Multi-source Multi-class Fake news Detection (MMFD). It also differentiates between the different degrees of fakeness. Finally, This framework combines automated degrees of fakeness, automated feature extraction, and multi-source fusion into an interpretable and coherent model for fake news detection. This work only considers the news text from the different sources and perspectives but does not care for the comments and responses. Since the spread of fake news also depends on the evidence provided in the comments, the work further can be improved if it includes the comments on that news post.

There is very limited literature that utilizes visual features for verifying multimedia posts. Jin, Cao, Zhang, Zhou, and Tian (2016) have shown the importance of an image feature for automatic fake news verification on social media. It has demonstrated that real and fake news events contain different distribution patterns of images. (ping Tian et al., 2013) explained image feature extraction and the representation mechanism. Here, the author analyzed the performance of fake news detection models after fusing local and global features of the images. The work presented in Jin et al. (2016), ping Tian et al. (2013) are entirely based on the image features. However, these image features are still hand-crafted and can hardly represent complex distributions of visual contents. The researchers need to work on automatic image feature extraction by implementing some deep learning-based mechanisms.

2.2. Fake news detection based on multi-modality

In recent times, multimodal information analysis has attracted the attention of researchers and practitioners for solving several practical problems like sentiment analysis (Ghosal et al., 2018), emotion analysis (Chauhan, Akhtar et al., 2019), image captioning (Karpathy & Fei-Fei, 2015), visual question answering (Antol et al., 2015), and also fake news detection (Jin, Cao, Guo, Zhang, & Luo, 2017) etc.

In Jin et al. (2017), a deep learning-based model has been proposed to extract multimodal features along with social context features and combine them by an attention mechanism. Hence, the model extracts event-specific features and cannot be generalized to detect fake news on newly arrived events. The performance of this work can be improved by combining a domain and event independent feature extraction mechanism with the proposed model. To overcome the limitation found in Jin et al. (2017), Wang et al. (2018) proposed a deep learning-based model named as Event Adversarial Neural Network (EANN). The model generates event invariant feature representations with the help of an adversarial network (Goodfellow et al., 2014). This model is not capable of learning shared representation of multimodal posts. To avoid this problem of representation learning, Khattar et al. (2019) introduced multimodal variational auto-encoder (MVAE) for fake news detection. An additional sub-task (event discriminator in EANN (Wang et al., 2018) and decoder part of variational autoencoder in MVAE (Khattar et al., 2019)) have been introduced to obtain the outputs. In these models, results are highly dependent on the sub-task, the absence of which degrades the performance of the model. Finding shared multimodal representations in both (EANN and VAE) these works is another limitation. Fake news detection frameworks presented in both these papers concatenate the text and image feature representations, and thereby, obtained multimodal features that may not introduce the proper interaction and alignment between textual and visual features. The author in Singhal et al. (2019) has introduced a multimodal framework (Spotfake) that detects fake news without performing any subtask. In this work, the author has resolved the first limitations found in EANN and VAE, but it continues with another limitation of obtaining better shared representations.

As described in the above literature, in any multimodal learning framework, one of the crucial issues is investigating the appropriate fusion techniques to effectively combine the information of multiple sources. The work presented in Fukui, Park, Yang, Rohrbach, Darrell, and Rohrbach (2016) introduced multimodal Compact Bilinear Pooling (MCB) that generates a very high dimensional joint feature representation based on the outer product of two feature vectors. To overcome the problem of high dimensionality (Kim et al., 2016) presented the multimodal Low-rank Bilinear Pooling (MLB). This mechanism performs the element-wise multiplication (Hadamard product operation) of two feature vectors to get the joint feature representation in the shared space. Recently, Chauhan, Firdaus et al. (2019) and Yu, Yu, Fan, and Tao (2017) utilized the MFB module to obtain the fused representation of text and image features for Natural Language Generation (NLG) (Reiter & Dale, 1997) in the fashion domain.

The prior works with single modality majorly used handcrafted features for textual feature extraction. Some existing unimodal and multimodal works used RNN for textual feature extraction. For visual feature extraction in multimodal fake news, detection authors have used the pre-trained VGG19 network. Hand-crafted feature extraction needs domain expertise and is very time-consuming. RNN may not extract the high-level semantic features, and a pre-trained VGG19 network, trained on the general domain ImageNet dataset, may not extract domain-specific image features. In previous research works, authors have also not designed any multimodal feature fusion mechanism. We propose attention-based multi-level CNN–RNN architecture for visual feature extraction, and attention-based stacked Bi-LSTM for textual feature extraction to overcome these limitations in the focus of the above shortcomings. After that, we combine both the features using the MFB mechanism and pass the unified representation through a multilayer perceptron (MLP) to classify the multimedia news posts as real or fake. Evaluation of two benchmark datasets shows that our proposed approach attains state-of-the-art performance, possibly due to better feature extraction and future fusion mechanism.

3. Research objective

This section presents the specific objectives for multimodal fake news detection. The prime objective of our research is to analyze the content of a news post spread on social media and detect whether it is fake or real. Our current work presents an analysis of the extent to which the fusion of different modalities (e.g., text, image) of a news post boosts the classification performance. We set forth the following four primary research objectives:

RO 1. *Extraction of better and relevant features from text and image content of a multimedia news post.*

Our first research objective aims to extract useful and relevant features from the news content. Since we consider different modalities, we focus on the feature extraction from both texts and image contents of a news post. We do this using some deep learning based frameworks.

RO 2. *Propose a feature fusion mechanism to maximize the correlation between text and image features.*

After Feature extraction, another objective is to fuse the extracted features of text and image to maximize the correlation between them and generate a better shared representation. We achieve this using a Multimodal Factorized Bilinear-pooling (MFB) mechanism.

RO 3. *Design a novel multimodal fake news detection framework that classifies the news posts with high accuracy.*

The main objective of this research work is to design a novel fake news detection framework that leverages information from text and image content of the news, fuses that information, and decides whether this news is fake or real. We do this by implementing Multilayer Perceptron (MLP) over the shared representation.

RO 4. *Evaluate the consistency of the proposed model across different datasets.*

Our final aim is to build a multimodal fake news detector that generalizes to the different datasets. To attain this, we evaluate our proposed model on two different datasets. We perform a detailed analysis to show the effectiveness, strength, and weaknesses of the proposed fake news detection model.

4. Methodology

This section describes an intuition of the theoretical foundations and shows how to restate the problem to allow accurate, efficient, and fast computation of news item credibility. Findings (shown in Fig. 1 and discussed in the preceding paragraph) and the role of multiple modalities in fake news detection offer a theoretical foundation for further empirical examination of fake-news. This work lays the foundation

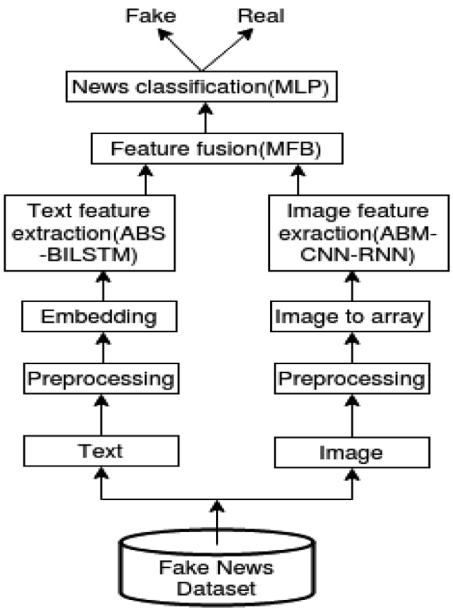

Fig. 2. Process diagram of the proposed multimodal fake news detection framework.

for building a repeatable end-to-end process to detect multimodal fake news spread on social media. These efforts to characterize multimedia news posts as fake or real will help researchers, journalists, scientists, and general people with misinformation. More specifically, the paper presents three essential foundations of fake news detection, *viz.*(i). Multimodal feature extraction: In this step, we extract the textual and visual features from multimedia news posts; (ii). Multimodal feature fusion: In this step, we combine the extracted textual and visual features to get a single shared representation (iii). Multimodal fake news detection: In this last step, we classify the news post using the shared representation. A high-level view of our technique is depicted in Fig. 3. We discuss all the theoretical foundations in details in this section.

Fig. 2 shows the overall process diagram of the proposed multimodal fake news detection framework. With the help of this diagram, we explain how the proposed methodology is implemented from the raw dataset to the final decision. We discuss these stages one by one in details.

Fake News Dataset: Initially, we take the publicly available multimodal fake news datasets. The dataset includes id, piece of text, associated image, context features, and ground truth labels.

Feature Selection: Our current model deals with the multimodal fake news detection, and we consider a piece of text, image, and the corresponding ground truth label attributes of the dataset. Textual and image attributes extracted from the neural network are also utilized.

Preprocessing: In this step we pre-process the text and image. Initially, we remove the instances only containing either text or image to prepare a complete multimodal dataset. Further, for text data, we tokenize the sentences, remove punctuation marks, and stop words. We make all the images of equal size and convert the images into a 3-dimensional array.

Embedding: We use pre-trained fasttext embedding of 300-dimensions to find the word vectors.

Feature extraction: After finding word embeddings of text we implement **ABS-BiLSTM** for the textual feature extraction. For visual feature extraction, we implement **ABM-CNN–RNN**.

Feature Fusion: After getting the textual and visual feature representations, we implement the feature fusion mechanism to obtain the shared representation. We show a detailed description of the feature fusion mechanism in the methodology section of this research work.

News Classification: After obtaining the combined representation, the final stage is the classification of the news. We implement the Multilayer Perceptron (MLP) for the fake news classification and put the description again in the methodology section. This last stage decides whether the information is fake or real. We extend this process diagram to design and explain our proposed model, which we discuss later in this section.

In this paper, we devise a deep learning based technique for multimodal fake news detection. It utilizes both textual and visual information to decide whether a multimedia news post is fake or real. We define a post instance $I_P = \{T,V\}$ as a tuple representing two different modalities of a news post, where T and V represent textual and visual content, respectively. The proposed model extracts textual feature (F_T) and visual feature (F_V) for the given instance I_P. The overall architecture of the proposed model is depicted in Fig. 3. It includes four components: (a). Attention based Stacked Bi-LSTM (ABS-BiLSTM); (b). Attention based Multi-level CNN–RNN (ABM-CNN–RNN); (c). Multimodal Factorized Bi-linear pooling (MFB) and (d). Multi Layer Perceptron (MLP).

We discuss each of these modules below:

4.1. Attention based stacked BiLSTM(ABS-BiLSTM)

Different modalities yield different aspects of a multimedia news post. This part of the proposed architecture extracts the textual features from the multimedia news posts. It captures the best contextual and semantic feature representations of the words. The stacking of two BiLSTM layers forms this sub-network. We give the output of the first BiLSTM layer as the input to the second BiLSTM layer. Here, the second BiLSTM layer extracts more intricate features and some different patterns over the extracted features. Mathematically, we formulated it as:

Given a news post P_x, two stacked BiLSTMs are employed to encode each word into the hidden vectors $h_{P_{x,i}}$. Words are represented by 300-dimensional embedding. Here, P_x is the x^th news post, $w_{x,i}$ is the i^th word of x^th news post, $w_{x,i}, i \in (1, \ldots, n), x \in (1, \ldots, m)$ and $k \in [1, 2]$.

$$\begin{aligned} \overrightarrow{h^k_{P_{x,i}}} &= LSTM_{p,f}(w_{x,i}, \overrightarrow{h^k_{P_{x,i-1}}}) \\ \overleftarrow{h^k_{P_{x,i}}} &= LSTM_{p,b}(w_{x,i}, \overleftarrow{h^k_{P_{x,i-1}}}) \\ h_{P_{x,i}} &= [\overrightarrow{h^k_{P_{x,i}}}, \overleftarrow{h^k_{P_{x,i}}}] \end{aligned} \tag{1}$$

All the words do not contribute equally to the meaningful text representation. We apply an attention mechanism over the hidden state obtained from the second BiLSTM layer to extract important words and aggregate the representation of those informative words to form the final text representation vector, similar to Yang et al. (2016). We formulate this attention mechanism using Eqs. (2)–(4).

$$h_{it} = tanh((W_w h_{P_{x,i}}) + b_w) \tag{2}$$

$$\alpha_{it} = \frac{exp((h_{it}^T h_w))}{\Sigma_t exp((h_{it}^T h_w))} \tag{3}$$

$$s_i = \Sigma_t \alpha_{it} h_{it} \tag{4}$$

Again we pass the output of the second BiLSTM layer ($h_{P_{n,i}}$) to a fully connected layer that produces the hidden representation (h_{it}) of a word. We pass this hidden representation to a softmax function to measure the importance of the word and to obtain a normalized importance weight (α_{it}). Finally, the weighted sum (s_i) of the word representation yields the attended textual feature representation. Here, W_w, b_w, and h_w are randomly initialized vectors that learn jointly during the training process.

To make the final dimension length 32, the output of the attention layer is passed through a fully connected layer as shown in Eq. (5) and R_T yields the attended representation of the text.

$$R_T = relu((W_{s_i} s_i) + b_{s_i}) \tag{5}$$

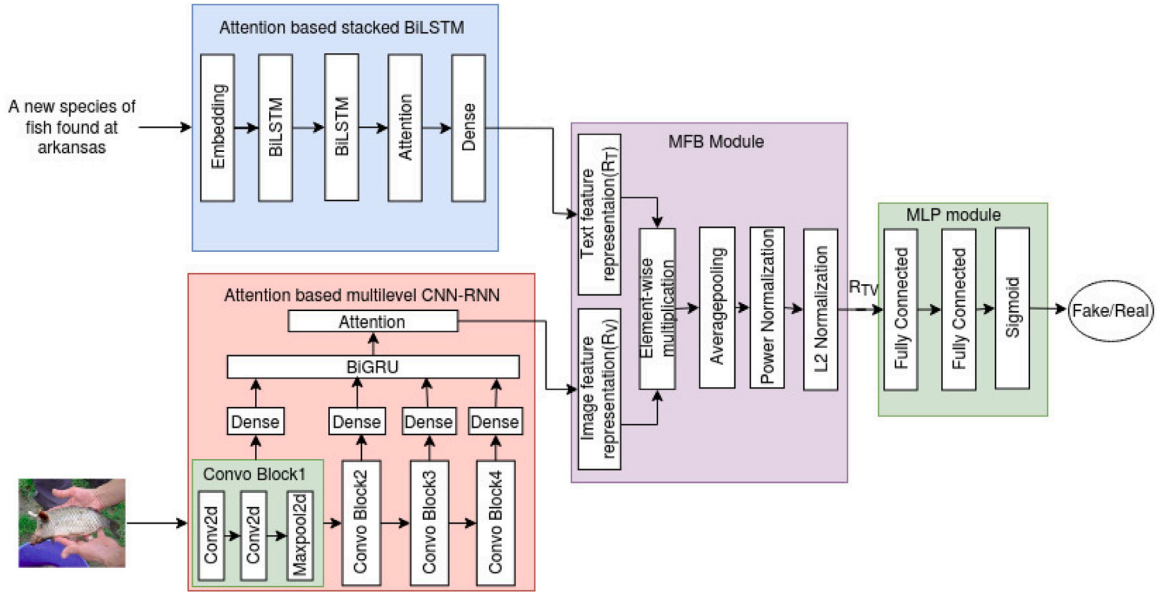

Fig. 3. AMFB: Attention based multimodalFactorized Bilinear Pooling for multimodalFake News Detection. In this model **Attention Based Stacked Bi-LSTM** extracts the textual feature representation, **Attention Based Multi-level CNN–RNN** extracts visual feature representation, **Multimodal Factorized Bi-linear pooling** combines both textual and visual features to give shared representation and **Multi Layer Perceptron** finally classifies the news post as fake or real.

4.2. Attention based multi-level CNN–RNN(ABM-CNN–RNN)

In general, people are tempted more towards visual content as it is faster and easier to capture than textual content. Most of the news posts nowadays contain some visual content, such as images or videos. Hence, an accurate visual feature extraction model should play an important role in implementing a multimodal fake news detector. We build attention-based multi-level CNN–RNN for visual feature extraction. Generally, CNN learns high-level features or semantic features (features used by humans to describe images) through layer-by-layer abstraction. Initially, it retains low-level elements such as shape, line, color, etc. Later, it learns high-level features by focusing on objects, actions, etc., of the image. Semantic features of an image are highly dependent on low-level features. The intermediate layer on CNN gives complementary information for the upper layer. Low-level features are often overshadowed and sometimes lost due to the high-level semantic feature representation. Existing literature on image emotion classification (Yang, Sun, Liang, Yang, & Cheng, 2018) and salient object detection (Zhang, Wang, Lu, Wang, & Ruan, 2017) have proven that integration of low level and high (semantic) level features provide better visual feature representations than only using high-level features.

Fake news images also show emotional provocation and some visual impacts. Motivating by this idea, we design an attention-based multi-level CNN–RNN subnetwork that captures better semantic visual feature representations by integrating low and high-level image features. Initially, the image with a size (224,224) is given as input to a CNN network. We consider the two 2-dimensional convolution layers with a ReLu activation function and one 2-dimensional max-pooling layer with pool size (2,2) as a CNN block. Upon each block output, a fully connected layer is added to reduce the feature-length to 32. Now we consider the output of these blocks as a sequence and pass it to a bidirectional Gated Recurrent Unit (BiGRU) to find internal and sequential dependencies among the features in both the directions, i.e., from high level to low level and from low level to high level. Like textual features, the whole extracted visual features also cannot have equal importance, so we design an attention mechanism over the obtained sequence from BiGRU. This is then passed through a fully connected layer to produce the final visual representation (R_V). Mathematically, it can be formulated using Eq. (6) to Eq. (10):

$$Cb_j = Maxpool(Conv2d(Conv2d(Cb_j - 1))) \tag{6}$$

$$v_j = relu((W_{Cb_j} Cb_j) + b_{Cb_j}) \tag{7}$$

$$\begin{aligned} \overrightarrow{v_f} &= \overrightarrow{\mathrm{GRU}(v_j)} \\ \overleftarrow{v_b} &= \overleftarrow{\mathrm{GRU}(v_j)} \\ v &= [\overrightarrow{v_f}, \overleftarrow{v_b}] \\ j &\in (1..4) \end{aligned} \tag{8}$$

$$s_i = att(v) \tag{9}$$

$$R_V = relu((W_{s_i} s_i) + b_{s_i}) \tag{10}$$

Here, $W_{Cb_j}, W_{s_i}, b_{Cb_j}$ and b_{s_i} are the learnable parameters. In summary, our analysis reveals that **ABS-BiLSTM** extracts the high-level implicit textual features, while **ABM-CNN–RNN** extracts the multi-level visual features. The proposed feature extractors are better than the existing feature extraction mechanisms, and also we are the first to utilize a novel feature extractor instead of some pre-trained methods. This analysis solves our first research objective.

4.3. Multimodal Factorized Bilinear Pooling (MFB)

We implement this component of the proposed model to support our second objective. After obtaining the final textual (R_T) and visual (R_V) feature representations from both the modalities, we fuse them using the MFB module. Here, we prefer MFB over the standard concatenation because of the following reasons: (a). It is challenging to determine the boundary of the extracted features obtained from the different modalities in the standard concatenation.

(b). After concatenation, features are stacked one after another, and hence it may not discover the correlation between the image and text feature representations.

These two problems can be efficiently solved using the MFB module. This fusion mechanism maximizes the correlation between textual and visual feature representations. Let us assume that the textual feature vector is denoted as $(R_T) \in R_m$ for text, and visual feature vector as $(R_V) \in R_n$ for an image. The basic multimodal bilinear model is then defined according to following Eq. (11).

$$R_{TV} = {R_T}^T W_i R_V \tag{11}$$

Where $W_i \in R^{m*n}$ is a projection matrix. R_{TV} is the output of the bilinear model. Bilinear pooling effectively captures the pairwise interactions between the feature dimensions and simultaneously introduces a huge number of parameters that accompany high computational cost and risk of over-fitting. To reduce the number of parameters, W_i in Eq. (11) is factorized as two low-rank matrices:

$$R_{TV} = {R_T}^T U_i V_i^T R_V = \sum_{d=1}^{k} {R_T}^T u_d v_d^T R_V \tag{12}$$

$$R_{TV} = 1^T (U_i^T R_T \circ V_i^T R_V) \tag{13}$$

where k is the latent dimensionality of the factorized matrices $U_i = [u_1, \ldots, u_k] \in R^{m*k}$ and $V_i = [v_1, \ldots, v_k] \in R^{n*k}$, $\circ$ is the element-wise multiplication of two vectors, $1 \in R_k$ is an all-one vector. To obtain the output feature R_T by Eq. (13), we need to learn two three order tensors, $U = [U_1, \ldots .U_o] \in R_{m*k*o}$ and $V = [V_1, \ldots .V_d] \in R_{n*k*o}$ as weights for o output dimension. It can be further reformulated as two dimensional matrices, $U' \in R^{m*ko}$ and $V' \in R^{n*ko}$ and can be rewritten as follows:

$$R_{TV} = AveragePooling({U'}^T R_T \circ {V'}^T R_V) \tag{14}$$

$$R_{TV} = sign(R_{TV}) |R_{TV}|^{0.5} \tag{15}$$

$$R_{TV} = R_{TV}^T / \|R_{TV}\| \tag{16}$$

In summary, we can say that our feature fusion mechanism is different from the existing fusion methods. Most of the existing works focused on concatenating textual and visual features in order to obtain the shared representation that shows a very limited performance for fake news detection. Following the second research objective, we design a novel feature fusion mechanism that maximizes the correlation between text and image features and provides proper alignment.

4.4. Multi-layer Perceptron (MLP)

We design a multi-layer perceptron sub-network with two hidden layers and an output layer with a sigmoid activation function. This multi-layer perceptron network takes fused features as input. It projects them into the target space of two classes to produce the final prediction probability that decides whether a multimedia news post is fake or real. In AMFB we define binary cross-entropy loss between the original and predicted labels as the objective function. This is mathematically formulated as shown in Eq. (17) :

$$L = -\Sigma[ylogp + (1-y)log(1-p)] \tag{17}$$

where y is the original class and p is the predicted class of the news post.

All the above discussed four components of the proposed methodology together leads us to solve our third research objective. First of all, we give text and image of a news post as input, ABS-BiLSTM extracts textual features, ABM-CNN–RNN extracts visual features, MFB combines these two features, and finally, MLP determines whether the news is fake or real.

5. Dataset and experiments

For our experiments, we use two benchmark datasets, *viz.* Twitter and Weibo. These are the two datasets that researchers have used for designing high-quality multimodal fake news detection systems. Hence, to compare with the prior works, we train our model on these two publicly available real-world benchmark datasets, i.e., Twitter.[1] and Weibo[2]

[1] https://github.com/MKLab-ITI/image-verification-corpus.

[2] https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view?usp=sharing.

Table 1
Data distribution of different dataset.

Data set	Train		Test		Image
	Fake	Real	Fake	Real	
Twitter	6841	5009	2564	1217	410
Weibo	3748	3783	1000	996	13274

Twitter Dataset: Twitter dataset was released by Boididou et al. (2015) as a part of Verifying Multimedia Use at MediaEval challenge. This dataset consists of two parts as a training set and a test set. Tweets in this dataset contain text, associated images, and contextual information. Tweets in the training and test set have been collected from the different events, and hence there is no overlapping of events between the training and test sets. The training set consists of 14,483 tweets labeled with three different classes: fake, real, and humor. Out of the total, 6848 tweets are fake, 5001 tweets are real, and 2634 are humor class. The test set consists of 3781 tweets. Since the test set does not contain humor class instances, so we ignore these for our experiments. The Twitter dataset contains 360 training set images and 50 test set images. We use 20% of the training set as the validation set.

Weibo Dataset: Weibo is a multimodal Chinese dataset. Weibo is a micro-blogging website in China that encourages users to inform suspicious tweets on Weibo. This is then verified as fake or real by the committee of reputed users. Fake news of the Weibo dataset has been collected from Weibo from May 2012 to June 2016. All the fake tweets, crawled during this time and later on verified by the Xinhua News Agency, are considered real ones. The Xinhua News Agency is an authoritative news agency in China. According to Jin et al. (2017), very small and duplicate images have been removed for maintaining the quality of the dataset. Posts without images have also been removed to make it completely multimodal in nature. The Weibo dataset consists of a total of 9527 news posts. We use 7531 as the training data and 1996 as the test data. Training data is split as 90:10 for training and validation of the model. This dataset also consists of 7954 fake and 5320 real images.

Table 1 shows the complete data distribution for both the datsets.

The above-discussed datasets are the benchmark datasets and publicly available for fake news detection research. Initially, we have downloaded these datasets from the respective repositories. We have discussed in the introduction section, and we only consider the news contents, i.e., text and image, along with the individual labels to perform experiments for our model. We give text and image attributes as input to the model and keep the labels as ground truth. First of all, we preprocess the text and image before putting them as input into the network. For text, we tokenize the sentence, remove punctuation and stop words, and make all the images of equal size. We convert the text into vector format using fasttext embedding and images into vectors using image-to-array. We pass the text vectors into the ABS-BiLSTM (discussed in the methodology section) component of the proposed model to get the textual feature representation. Similarly, we give the image vector into two vectors as input into ABM-CNN–RNN (discussed in the methodology section) to get the visual feature representation. After getting these two representations, we pass them into the MFB (described in the methodology section) module of the proposed work to get the shared multimodal feature representation. Finally, we give this shared representation as input into the MLP (described in the methodology section) component of the model to provide the final output as fake or real. Now the ground truth labels of the datasets are compared with the predicted labels to compute the loss. The training set of the dataset is used to train the model, and the test set is used to validate the model's performance. The loss is optimized during the training process, and the model comes up with the best hypothesis.

Table 2
Hyper parameters used for training the proposed model.

Parameters	Twitter	Weibo
Text length	33	95
Image size	(224,224,3)	(224,224,3)
Batch size	32	32
Optimizer	Adam (lr = 0.00001)	Adam (lr = 0.00005)
Epochs	100	100
Regularizer	L2(0.5)	L2(0.5)
Dropout	0.2	0.2
Filter size	(3,3)	(3,3)
Strides	(1,1)	(1,1)

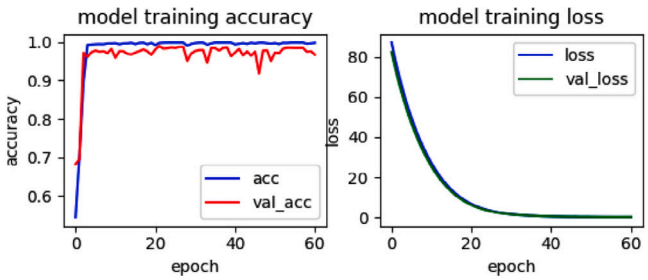

Fig. 4. Learning curves for twitter data.

5.1. Experimental setup

This section discusses the embedding mechanism, text and image pre-processing, different hyperparameters, and implementation details. All the experiments are carried out in a python environment. We utilize Keras, NLTK, Numpy, Pandas, and Sklearn libraries of python for conducting the experiments. We evaluate the performance of the system in terms of accuracy, precision, recall, and F-score.

We remove the Twitter handle, punctuation marks, numbers, special characters, and short words for Twitter text cleaning. For Weibo data, we use the Jieba python module for data segmentation. For textual feature extraction, we use fastText pre-trained embedding of 300 dimensions. We pass these vectors of 300 dimensions to the stacked BiLSTM layer, and then output is given to an attention layer following a dense layer. The first and second BiLSTM layers contain 256 and 128 units, respectively, and the dense layer is 32.

We resize all the images of size (224,224,3), then pass them to attention-based multi-level CNN–RNN. Here, we use Conv2d with 64 units, (3,3) size filter, and (1,1) strides. Every CNN block consists of two CNN layers with ReLu activation and one Maxpool 2d layer with pool size (2,2). The output of each block is passed through a fully connected layer of size 32. The output of all four blocks is then concatenated and passed to a bidirectional GRU layer of size 32. Attention is applied to this extracted feature, and the attended visual features are then passed to a fully connected layer of dimension 32 with the ReLu activation function to make its size equal to the textual features. We use the MFB module for feature fusion and pass it to a multilayer perceptron with two hidden layers of 32 and 16 with a ReLu activation function and an output layer of size one with a Sigmoid activation function. Here, we use binary cross-entropy as the loss function and Adam optimizer. We train the model for 100 epochs with 32 batch size and early-stopping callbacks. All the hyper-parameters used for training the proposed model are listed in Table 2.

We implement a few baseline models based on uni-modal and multimodal sources of information for validating the effectiveness of our proposed model.

Uni-modal baselines. We define the following uni-modal baselines for comparisons: (a). Textual: For textual feature-based baseline models, we use pre-trained fasttext embedding of 300 dimensions to obtain the textual feature vector and pass it to bidirectional LSTM layers of size 128. These extracted textual feature representations are again given to an MLP with two hidden layers with 64 and 32. The output layer of size 2 with the Sigmoid activation function detects the news post as fake or real. (b). Visual: To implement the visual model, we extract the visual features with a pre-trained VGG19 model. The extracted visual feature vectors are of sizes 4096 that are again fed into a fully connected layer with hidden size 32 and one output layer with Sigmoid activation function for the prediction.

Multimodal baselines. For multi-modality, we define the following baselines:

(a). att-RNN: att-RNN (Jin et al., 2017) uses visual attention to multimodal fake news detection. The author of this paper originally used concatenation for the joint text and image representation. We use element-wise multiplication for obtaining the joint multimodal feature representation. For embedding, we use a fasttext of 300 dimensions instead of word2vec embedding. The hyper-parameters are the same, as mentioned in the original paper.

(b). EANN*: Event Adversarial Neural Network (EANN) (Wang et al., 2018) has three components: the feature extractor, event discriminator, and fake news detector. In this model, the feature extractor extracts the event invariant textual and visual features with the help of an event discriminator. It then classifies the news posts as fake or real using these features. We propose a variation of this model for the performance comparison, which has only two components: feature extractor and fake news detector. All the parameters used for training this model are the same as the original model.

(c). MVAE*: multimodalVariational Autoencoder (Khattar et al., 2019) trains three sub-networks for fake news detection. Here, a variational autoencoder was trained for obtaining better textual and visual joint feature representation. The shared latent representation was further used for the classification. Here, we build the model with encoder and fake news detector parts to make a fair comparison. We train the model with the same hyper-parameters used for training the original model.

5.2. Results and analysis

The comparative analysis of the existing models and the proposed model on two different datasets are shown in Table 3. We report the accuracy, precision, recall, and F1-score of AMFB for fake and real classes. The results show that our proposed approach yields better performance than the existing state-of-the-art and baseline models. It is evident that the visual model yields better performance compared to the textual model. This might be because texts may sometimes contain noisy and unstructured information, but the image shows better evidence. It can be concluded from the results that incorporating image to text is beneficial as it attains higher performance compared to only image or text. Hence, this analysis proves to have achieved our final research objective.

Learning curves for training and validation show the loss and accuracy of the proposed (AMFB) model for each epoch. Figs. 4 and 5 depict the learning curves of the proposed model for Twitter and Weibo datasets, respectively. In both the learning curves, the loss continuously decreases to an equilibrium position that shows the model learns appropriately. The validation accuracy curve of the Weibo dataset is seen to fluctuate more than the Twitter data because the size of the Weibo validation data is lesser compared to the Twitter validation data.

To analyze how the proposed model (AMFB) discriminates fake news from real news, we employ dimensionality reduction using T-distributed Stochastic Neighbor Embedding (t-SNE) (Zhong, Li, Ma, Jiang, & Zhao, 2017). Figs. 6 and 7 show the projection of the feature representations learned by the proposed model in the 2-dimensional plane for both Twitter and Weibo dataset. We observe that our model obtains good separability for both datasets. Overlapping of instances for

Table 3
Classification results of existing and proposed model on Twitter and Weibo datasets.

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F-Score	Precision	Recall	F-Score
Twitter	Textual	0.538	0.43	0.71	0.53	0.72	0.43	0.54
	Visual	0.645	0.52	0.59	0.55	0.74	0.68	0.71
	VQA	0.631	0.765	0.509	0.611	0.550	**0.794**	0.650
	Neural Talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN*	0.741	0.69	0.55	0.61	0.76	0.85	0.81
	EANN	0.715	NA	NA	NA	NA	NA	NA
	MVAE*	0.724	0.62	0.64	0.63	0.79	0.77	0.78
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	Spotfake	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	AMFB	**0.883**	**0.89**	**0.95**	**0.92**	**0.87**	0.76	**0.81**
Weibo	Textual	0.593	0.62	0.50	0.55	0.58	0.69	0.63
	Visual	0.608	0.620	0.604	0.607	0.607	0.611	0.609
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	Neural Talk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN*	0.791	0.84	0.72	0.78	0.76	0.86	0.80
	EANN	0.827	NA	NA	NA	NA	NA	NA
	MVAE*	0.70	0.67	0.80	0.73	0.75	0.60	0.67
	MVAE	0.824	0.854	0.769	0.809	0.802	**0.875**	0.837
	Spotfake	**0.8923**	**0.902**	**0.964**	**0.932**	0.847	0.656	0.739
	AMFB	0.832	0.82	0.86	0.84	**0.85**	0.81	**0.83**

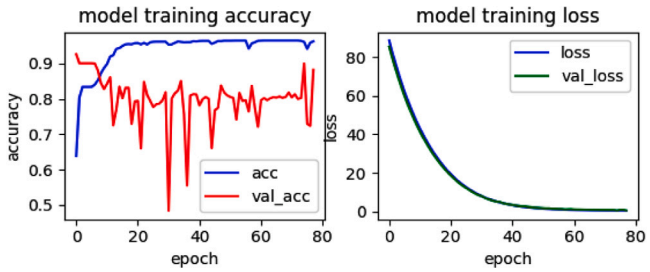

Fig. 5. Learning curves for Weibo data.

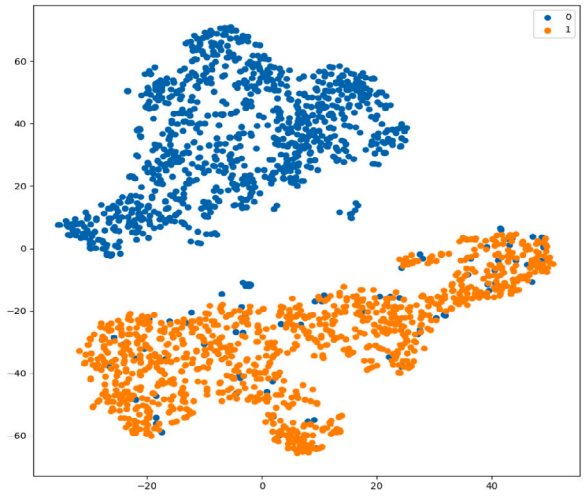

Fig. 6. Projection of feature representations for twitter data.

the Weibo dataset is more than the Twitter dataset because of the two reasons: (i). Most of the images are more involved in the Weibo dataset (ii). Weibo is a Chinese dataset, and after segmentation, the length of some sentences becomes greater than the sentence length of the Twitter dataset. The model does not extract better semantic textual features for complex or very long sentences. These drawbacks have been better explained in the limitation section.

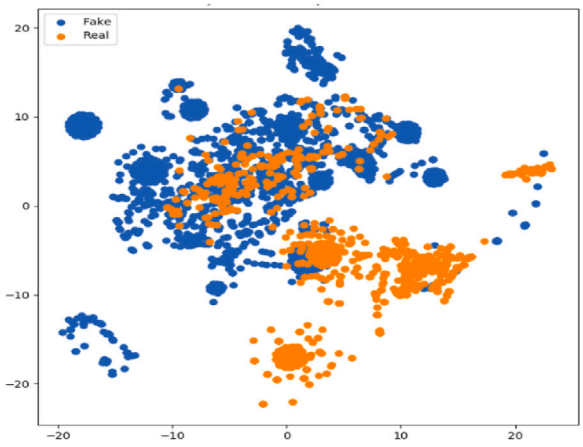

Fig. 7. Projection of feature representation for Weibo data.

Qualitative analysis: We closely analyze the outputs of the classifiers to get better insights. Fig. 8 shows two different tweets taken from the Twitter dataset. The text content of the first tweet seems real and is classified as real by the textual model, but it has tampering in the image. In the second tweet, text and image both are real, and it is classified as real by both textual and visual models. But the image in the second tweet has not been taken during the Nepal earthquake as described by the text. All the tweets are shown in Fig. 8 are fake, and our proposed model (AMFB) predicts as fake. This indicates that the proposed model effectively extracts both visual and textual features and decides whether it is fake or real. Fig. 9 shows some real tweets taken from the Twitter dataset. These tweets are correctly classified by the proposed model, whereas baseline models such as textual, visual, and existing models like att-RNN and EANN yield the wrong prediction. Therefore, the above observation shows that our proposed model outperforms textual, visual, att-RNN, and EANN. The state-of-the-art system, Spotfake, also fails to classify the first tweet directed in Fig. 8, i.e., the misleading tweet.

Comparison to the state-of-the-arts: This section describes the comparative analysis of the proposed model and the current state-of-the-art (Singhal et al., 2019). Details of the results are shown in Table 3. In Spotfake (Singhal et al., 2019), BERT has been used for textual

(a) new species of fish found at arkansas

(b) A 4 year old attempts to protect his sister from the devastation of NepalEarthquake

Fig. 8. Fake tweets correctly classified by AMFB but misclassified by single modality model.

(a) Great picture thierry legault captures iss transit of the sun during eclipse

(b) NepalEarthquake: Death Toll Rises To 758, Emergency Declared. North, Eastern India Also Hit

Fig. 9. Some real tweets correctly classified by AMFB but mis-classified by existing model.

feature extraction, that undoubtedly a powerful mechanism. We use fasttext embedding, and attention-based stacked LSTM instead of the stacked encoder of BERT for feature extraction, and hence our model is less complex. Twitter and Weibo data are very noisy, and sentences are not syntactically and semantically correct. Most of the words are also not complete in the sentences, yielding to the vocabulary's outliers. BERT performs better than fasttext; however, fasttext can even handle it in a better way. Hence, due to less complexity, we use a fasttext mechanism for embedding. Spotfake uses pre-trained VGG19 architecture for image feature extraction that extracts general image features and may not extract the invariant image features. The state-of-the-art obtains the joint feature representation only by concatenating textual and visual features that do not give a good correlation between image and text. So the misleading news (news in which text content of the tweet describes the image captured during another event as shown in the second example of Fig. 1) cannot be classified correctly. To maximize the correlation between textual and visual feature representation, our proposed model uses the MFB module, which gives better joint representation and helps the model classify the multimedia posts correctly.

Our proposed model performs better by taking appropriate measures to the problems identified by the model, Spotfake. Although the Twitter dataset is imbalanced and more than half of the instances belong to a particular event only, we obtain approximately 9.8% accuracy gain. Spotfake reported high accuracy for the Weibo dataset. Even for a balanced dataset like Weibo, Spotfake reports relatively low precision, recall, and F1-score for the real class. The images in the Weibo dataset are more complex, and pre-trained VGG19 cannot capture the high-level domain-specific image inherent features that are a cause of the biased result. Although our model shows inferior accuracy compared to the Spotfake model, we obtain more balanced precision, recall, and F1-score values for both classes. Moreover, our model is not as complex as Spotfake.

Table 4
Number of parameters for AMFB and Spotfake.

Model	Number of parameters
AMFB	50,003,909
Spotfake	2,61,582,168

For the complexity analysis of our proposed model compared to the Spotfake, we analyze the number of parameters for each model. A large number of model parameters represent the more complex model. Table 4 shows that the number of parameters for the proposed model (AMFB) is significantly less than the state-of-the-art (Spotfake). So, by the above observation, it is shown that the proposed model performs better even with less complexity.

Implications False, misleading, or fake news on social media can have significant adverse societal effects. Fake news detection on social media has recently attracted the attention of researchers and practitioners. The investigations of this study imply both theoretical and practical implications. Most of the previous works explored unimodal sources (i.e., text) for fake news detection. Although there exists a fair amount of prior research for detecting fake news on social media, these methods are still not fully capable of detecting these in the early stages or prevent after spreading. Some of the existing techniques have tried to overcome these limitations by utilizing only the news posts' text content. As multimodal news posts create much attraction and attention that can quickly compel the population to believe, social media is full of multimodal news posts nowadays. This provides the right motivation for detecting fake news by leveraging information from multiple modalities.

Theoretically, this study describes how to extract the features from different modals and fuse those features to obtain the shared representation that finally classifies news as fake if the image or text content of a news post contains a piece of misinformation. We have empirically tested and validated the role of images in fake news detection. Second, these findings have uncovered new facets of fake-news sharing and detection, proposing a deep learning based end-to-end framework. These revelations contribute to the theoretical knowledge in the domain.

Practically, we can detect the fake news at two stages: (i). immediately after introducing the fake news in the network, (ii). when the fake news is propagated in the network. Our main purpose is to detect the fake news in the early stage means immediately after the introduction into the network and stop it from spreading here only, but also our model monitors the network regularly for fake news detection. At each stage, it only takes the text and image as input, and based on these contents, it predicts the truthness of the news. Suppose the news is identified as real in the initial stage, and it is actually true, but somebody has made any changes in that particular news afterwards. Since the model monitors the network regularly, it takes the text and image of that tampered news as a new input and detects it as fake. But, if the news is identified as real in the initial stage and it is actually false, then the model will not detect it as false later in the network until somebody has made changes in the news content during propagation. We can conclude that if there is no tampering or changes in the news content at any stage of the propagation, then there will not be any difference in the fake news detection model's decision at different stages. Our model extracts the better syntactic and semantic textual and visual features and predicts the fake news with high accuracy, so this is a rare case.

Our findings show that adopting the strategy of multimodality improves the performance of fake news detection. We have demonstrated the practical implementation of deep learning based multimodal fake news detection framework in this study. No, any existing literature focuses on multimodal feature fusion for fake news detection. This can

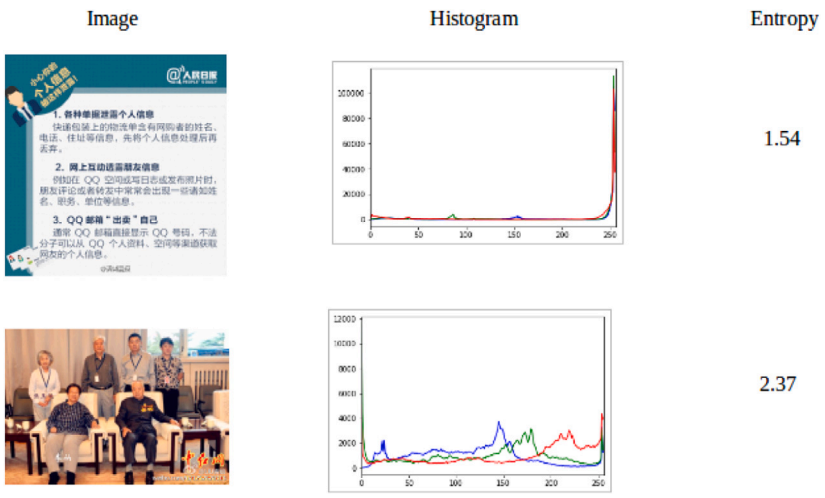

Fig. 10. Image complexity analysis.

go a long way in mitigating the spread and detection of fake news, especially in the news with audio and video content. The implementation results show that our proposed model performs better compared to the state-of-the-art. Regulators, researchers, scientists, journalists, and organizations could benefit from these findings and utilize them to prevent and detect the spread of fake news.

Limitations: We have tried to introduce and solve the problems of the state-of-art; however, our model also has some limitations such as *i)* The model does not extract very good image invariant features of a complex image. *ii)* If the length of text in the news post is very large, then sometimes it does not make semantic attention to the latter part of the sentence. *iii)* Our proposed modal provides better correlation than the state-of-the-art, but it still does not captures the semantic correlation between text and image.

In Fig. 10, we analyze the complexity of some images of the Weibo dataset using Shannon entropy (Khanzadi, Majidi, & Akhtarkavan, 2017) and also plot the histogram to show the frequency of pixels intensity values. Our model does not perform well for the images having low entropy and non-uniform distribution of the frequency of pixels intensity values. For the first example in Fig. 10, the model does not extract useful invariant features, whereas the second example model performs well.

Fig. 11 shows some examples of Twitter and Weibo datasets, where our proposed model fails. Both the tweets show the attended text and its corresponding image. In the first example, the text's length is very large, so the model fails to discover the attention on the latter part of the sentence. This tweet is fake, but the proposed model classifies it as real. The second tweet example in Fig. 11 better explains the third limitation. In this example, the model gives the high attention on *colapses* and *earthquake* but in image there is no any direct object as *colapses* and *earthquake*. In such cases, our model cannot find the semantic correlations and misclassify them as fake.

We can address these limitations to create a large sustainable impact to resist the fake news into the network. We can also incorporate the user characteristics, user behavior, propagation knowledge, etc., that might impact fake news detection significantly. However, it demands creating an appropriate dataset and the experimental framework.

6. Conclusion and future work

In this paper, we have proposed an end-to-end deep learning framework that takes the image and text of a multimedia post as input and detects whether this post is fake or real. To extract the textual feature, we have designed an Attention Based Stacked BiLSTM (ABS-BiLSTM). For visual feature extraction, we propose Attention Based Multi-level

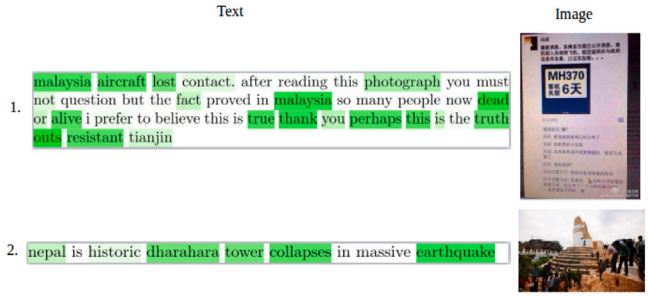

Fig. 11. Some example where our model fails.

CNN–RNN. Joint representation is obtained through the MFB module and fed to a Multi-Layer Perceptron (MLP) to detect a multimedia news post as fake or real. Extensive experiments have been conducted on two publicly available multimodal datasets. Experimental results show that the proposed model performs better for detecting fake news compared to the existing models.

There are many possibilities to investigate and extend this work in the future. Here, we list two possible directions of our future research: (a) Semantic alignments between text and image can further be investigated for better fusion mechanisms; (b). Nowadays, many news posts contain videos. Hence a possible extension would be to include audios and videos.

CRediT authorship contribution statement

Rina Kumari: Conceptualization, Methodology, Writing manuscript. **Asif Ekbal:** Supervision, Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors gratefully acknowledge the project "HELIOS - Hate, Hyperpartisan, and Hyperpluralism Elicitation and Observer System", sponsored by Wipro.

References

Alkhodair, S. A., Ding, S. H., Fung, B. C., & Liu, J. (2020). Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management*, *57*(2), Article 102018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., et al. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425–2433).

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., et al. (2015). Verifying multimedia use at mediaeval 2015.. *MediaEval*, *3*(3), 7.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684). ACM.

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 5646–5656).

Chauhan, H., Firdaus, M., Ekbal, A., & Bhattacharyya, P. (2019). Ordinal and attribute aware response generation in a multimodal dialogue system. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5437–5447).

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, *23*(1), 155–162.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv:1810.04805.

Faustini, P. H. A., & Covões, T. F. (2020). Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, Article 113503.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing* (pp. 457–468). ACL.

Ghosal, D., Akhtar, M. S., Chauhan, D. S., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3454–3466).

Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. In *2017 IEEE 15th student conference on research and development (SCOReD)* (pp. 110–115). IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Gravanis, G., Vakali, A., Diamantaras, K., & Karadais, P. (2019). Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, *128*, 201–213.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics*, *15*(1), 41–51.

Huang, Y.-F., & Chen, P.-H. (2020). Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, Article 113584.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 795–816). ACM.

Jin, Z., Cao, J., Jiang, Y.-G., & Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE international conference on data mining* (pp. 230–239). IEEE.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, *19*(3), 598–608.

Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1546–1557).

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128–3137).

Khanzadi, P., Majidi, B., & Akhtarkavan, E. (2017). A novel metric for digital image quality assessment using entropy-based image complexity. In *2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI)* (pp. 0440–0445). IEEE.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *The world wide web conference* (pp. 2915–2921). ACM.

Kim, J.-H., On, K.-W., Lim, W., Kim, J., Ha, J.-W., & Zhang, B.-T. (2016). Hadamard product for low-rank bilinear pooling. arXiv, arXiv–1610.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining* (pp. 1103–1108). IEEE.

Liu, Y., Jin, X., & Shen, H. (2019). Towards early identification of online rumors based on long short-term memory networks. *Information Processing & Management*, *56*(4), 1457–1467.

Liu, Y., Pang, B., & Wang, X. (2019). Opinion spam detection by incorporating multi-modal embedded representation into a probabilistic review graph. *Neurocomputing*, *366*, 276–283.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., et al. (2016). Detecting rumors from microblogs with recurrent neural networks.. In *Ijcai* (pp. 3818–3824).

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *ACL (1)*.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937).

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87.

Rish, I., et al. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (pp. 41–46).

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7–17).

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 797–806). ACM.

Shaha, M., & Pawar, M. (2018). Transfer learning for image classification. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 656–660). IEEE.

Shen, H., Ma, F., Zhang, X., Zong, L., Liu, X., & Liang, W. (2017). Discovering social spammers from multiple views. *Neurocomputing*, *225*, 49–57.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). Defend: Explainable fake news detection. In *25th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2019* (pp. 395–405). Association for Computing Machinery.

Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In *12th ACM international conference on web search and data mining, WSDM 2019* (pp. 312–320). Association for Computing Machinery, Inc.

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)* (pp. 39–47). IEEE.

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130.

ping Tian, D., et al. (2013). A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, *8*(4), 385–396.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., et al. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th Acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857). ACM.

Wu, L., Rao, Y., Nazir, A., & Jin, H. (2020). Discovering differential features: Adversarial learning for information credibility evaluation. *Information Sciences*, *516*, 453–473.

Yang, J., Sun, Y., Liang, J., Yang, Y.-L., & Cheng, M.-M. (2018). Understanding image impressiveness inspired by instantaneous human perceptual cues. In *Thirty-second AAAI conference on artificial intelligence*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). 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).

Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 1821–1830).

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, *4*(11).

Zhang, H., Fang, Q., Qian, S., & Xu, C. (2019). Multi-modal knowledge-aware event memory network for social media rumor detection. *Proceedings of the 27th ACM international conference on multimedia* (pp. 1942–1951).

Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 202–211).

Zhong, Z., Li, J., Ma, L., Jiang, H., & Zhao, H. (2017). Deep residual networks for hyperspectral image classification. In *2017 IEEE international geoscience and remote sensing symposium (IGARSS)* (pp. 1824–1827). IEEE.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., Bontcheva, K., et al. (2018). Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, *54*(2), 273–290.