



# An image and text-based multimodal model for detecting fake news in OSN's

Santosh Kumar Uppada<sup>1</sup> · Parth Patel<sup>1</sup> · Sivaselvan B.<sup>1</sup>

Received: 22 July 2022 / Revised: 18 October 2022 / Accepted: 3 November 2022 /  
Published online: 30 November 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

Digital Mass Media has become the new paradigm of communication that revolves around online social networks. The increase in the utilization of online social networks (OSNs) as the primary source of information and the increase of online social platforms providing such news has increased the scope of spreading fake news. People spread fake news in multimedia formats like images, audio, and video. Visual-based news is prone to have a psychological impact on the users and is often misleading. Therefore, Multimodal frameworks for detecting fake posts have gained demand in recent times. This paper proposes a framework that flags fake posts with Visual data embedded with text. The proposed framework works on data derived from the Fakeddit dataset, with over 1 million samples containing text, image, metadata, and comments data gathered from a wide range of sources, and tries to exploit the unique features of fake and legitimate images. The proposed framework has different architectures to learn visual and linguistic models from the post individually. Image polarity datasets, derived from Flickr, are also considered for analysis, and the features extracted from these visual and text-based data helped in flagging news. The proposed fusion model has achieved an overall accuracy of 91.94%, Precision of 93.43%, Recall of 93.07%, and F1-score of 93%. The experimental results show that the proposed Multimodality model with Image and Text achieves better results than other state-of-art models working on a similar dataset.

**Keywords** Fake news detection · Xception · Bert+Dense · Fusion models · Click-baits · Fakeddit · Visual sentiment analysis · Error level analysis

---

Parth Patel and B. Sivaselvan contributed equally to this work.

✉ Santosh Kumar Uppada  
coe18d005@iiitdm.ac.in

Parth Patel  
coe17b020@iiitdm.ac.in

Sivaselvan B.  
sivaselvanb@iiitdm.ac.in

<sup>1</sup> Department of Computer Science and Engineering, IIITDM Kancheepuram, Melakottiyur, Chennai, 600127, Tamil Nadu, India

## 1 Introduction

The widespread use of the internet and online platforms has changed the perspective of information sharing, making it quick and straightforward to reach the masses. The rapid growth of various platforms resulted in news from various sources that lacked credibility. As news comes from different sources, it is always challenging to determine the credibility of news or posts. From the Hadson river plane crash incident to the prevailing pandemic situations, online media has become a prominent tool to get diversified news across the globe within no time. Online social media has become a buzz, and popularity in social networks has opened doors to the news patterns that are informative and target the users' emotions interacting with such news. In order to attract people, users started creating news that creates a sensation, which might be far enough from reality. It is estimated that a massive amount of unverified news is being generated by *WhatsApp Univeristy*. Fake news or Yellow journalism is a process of creating, spreading, and propagating news that could be biased or unreal (Campbell, 2001). Fake news is generally an intentional or unintentional spread of news that could be unreal. In general, Fake News can be biased, unreal, satirical, propaganda, clickbaits, satirical news, or disinformation to degrade the reputation or create hoaxes in the society. The intentional spread of fake news creates chaos and panic in society (Shu et al., 2017). Even though there are many fact-checking sites, it takes time to check the credibility of every news that populates in multiple platforms (Robertson et al., 2020). There is also the 'power law' observed in such social media that posts can spread more quickly and reach wider audiences if the posts target a few influential people in the online social network (Adamic & Huberman, 2000). Stance detection and reliability study of the Fake posts in online social networks is another challenging task. It deals with evaluation of reliability of statements and news articles. It relies on the Statement conflict data, with a prior assumption that the news article and statement relationships are already known (Zhang et al. 2022).

Users tend to believe in the news when it comes from trusted groups or sites, which is generally termed as Homophily (McPherson et al., 2001). Confirmation bias is the psychological aspect that makes users interpret and recall the news they hear and support one's beliefs and faith. Frequency Heuristic is another factor that makes users believe in the news they interact with, as the same news is spread from different sources or spread multiple times. Therefore unintentional spread will also have some psychological aspects related to the spread of news (Pennycook & Rand, 2021). If a user is biased to a specific domain, channel, person, or site, they blindly tend to believe news without even thinking about its credibility, hence will be termed as biased users. The echo chamber effect has a significant impact on users believing in the news coming from specific sources (Cinelli et al., 2021). When it comes to intentional spread, users tend to publish news targeting the users who have more followers. Targeting a more influenced person makes the news spread deeper into the network. Users even tend to create click-baits, tempting headings to draw the attention of the users, or by creating automated programs or bots, to make news reach with more intensity (Bazaco, 2019). As social posts with images have higher reachability than other posts, users tend to spread the news with more images to attract the masses. In spreading fake news, images spread can be either tampered with (edited) or used out of context (Uppada et al., 2022).

Tracking the tampered images is easy, but dealing with images used out-of-context is always tedious and time-consuming. Users even tend to spread the news as a combination of images and text. Here image-related or context-related captions are added to the images and published online, as it is always challenging to classify such posts. Fact-checking website Boom reports that the number of false/misleading claims and misinformation has a positive

correlation with the number of COVID-19 cases in the country (Chowdhury, 2020). Around 65% of COVID-19 related misinformation is shared using multimedia, mainly images and videos. For example, during the COVID-19 pandemic, a piece of news that got attention and circulated more was about Cocaine killing the coronavirus. Social networking sites like Facebook, Twitter, and Whatsapp are significant sources for spreading such news. Figure 1 depicts the fake news circulated the most in China about cocaine killing coronavirus. Users aim to spread the news in terms of image or video, as it draws more attention and has more reachability and retweet frequency than regular posts. Tweets with videos, for example, posts with images, receive 18 percent more clicks, 89% more likes, and 150 percent more retweets than those without videos (Cao et al., 2020). As fake news is coming in multimedia format, there is a need for multimodal detection systems for fake news. As Fake images can be either tampered with or used in a different context, traditional forensic techniques are inadequate to handle such diverse nature of fake images on social media. There is a need to develop a framework that can effectively learn useful features from the varied nature of images in fake news to distinguish them from those in actual posts. Such a framework could hugely benefit online social networks in their efforts to curb the proliferation of fake news on their platforms (Jin et al., 2016).

Fake images are often eye-catching and have a substantial emotional impact. Fake images are generally misleading and thus provokes the user's attention. Thus, it becomes necessary to map psychological triggers to the characteristics of the image. These psychological patterns are limited to visual appearance and beyond the standard object-level features. Hence traditional image sets are not suitable for this task of fake image classification (Jin et al., 2017). Gathering large labeled datasets containing posts with real and fake images is difficult because human verification and labeling of posts is time-taking and not fast enough to deal with the big data online (Jin et al., 2016). Images often get auto-compressed when shared using specific social networking sites. Most forensic techniques that aim to detect fake images rely on features retrieved from these compression factors; hence fail to work with these compressed images when uploaded and downloaded multiple times.

Metadata and sentiment of the images also play a vital role in detecting fake images (Luo et al., 2007). Users generally depend on specific third-party tools like Mechanical Turk for analyzing sentiment. The polarity of images will help in understanding the impact that image creates on users (Ragusa et al., 2022). Similar to the text polarity, images will



Fig. 1 Cocaine kills Coronavirus (Clever et al., 2020)

also have positive, negative, and neutral (Ragusa et al., 2019) polarity. Therefore, observing text-related and image-related features makes the system detect fake news more accurately. Image-based social posts often get combined with text related to the image (Image Caption) or context. Users get misled when textual data is added to the images, as clickbait to attract attention. It is often tricky to work with posts with more than one form (image combined with text data). Thus, multimodal models are getting attention from the research community to combine features of text and image for detecting fake posts very quickly with more accuracy (Shah & Kobti, 2020; Giachanou et al., 2020). As social media posts contain multimedia data, it is heterogeneous data that should be handled, and hence, multimedia-related frameworks are to be proposed that work on multiple modals. This process became interesting with the intended use of Machine Learning and Deep Learning models. There is a need to propose different methods that work on different feature combinations to detect Fake posts, mostly with misleading images (Galli et al., 2022). Therefore, multimodal analysis has become a prominent research objective in detecting Fake posts in Online Social Networks. The significant outcomes of the study can be summarized as follows.

- The proposed framework, given the input post containing Images/Images with Caption, extracts features such as the probability of image being manipulated, polarity of the image, and the probability of image caption being manipulated using different learners, treating each feature independently. Fusion models such as Maximum and Concatenate are used to learn the classifier model.
- A dense layer is added to the BERT model to enhance the learning capability of the model.
- Various models (algorithms) have been tested on the data, including ELA (Error Level Analysis). The learners with high accuracy and less loss have been chosen for the proposed framework.
- The proposed model helps detect fake posts on Social Networks, especially Images with embedded captions.
- The proposed ensemble framework has shown better results compared to the state-of-art models working on similar data. The proposed work outperforms existing models in terms of sample size and accuracy.

The remainder of the paper is as follows. Section 2 describes the proposed models similar to the proposed models, finally comparison is derived with the proposed models. Section 3 has the methodology section that describes the overall framework. Further Section 4 is given with an introduction to the proposed Image Manipulation and Polarity based Fake News detection model. Sections 4.1, 4.2, and 4.3 has Image manipulation and polarity based fake posts detection models, including Error Level Analysis. For every modality different pre-trained models like VGG-16, Vgg-19, Xception, Inception-Resnet50 are tried. Even ELA based analysis is also performed on the images. In Section 4.4, Image caption data is analyzed. Section 5 has the proposed framework for Fake post detection. Section 5.1 has the summary of Fusion Models, Section 5.2 has Result Analysis, Section 5.3 has Error Analysis for the proposed model. Further Section 6 has conclusion and future scope discussion.

## 2 Related work

Anastasia et al. have proposed a multimodal Multi-image Fake News detection that works on the posts' textual, Visual, and Semantic Features. The BERT model is used for the

textual part, and VGG16 is used for the visual aspect of the posts. Tokens derived from the textual data and image tags are given to a similarity metric (cosine similarity of title and image tags), a semantic branch. These branches are fused using *Concatenate*, and finally, the attention layer is added. VGG16+BERT+fusion (attention) recorded an accuracy of 76.83%, VGG16+BERT+fusion (Concatenation) recorded 78.30% accuracy, and VGG16+BERT+similarity+fusion (attention) recorded an accuracy of 76.83% (Giachanou et al., 2020).

Kai Nakamura et al. proposed a multimodal model for Fake News Detection. For analysis, Reddit and Fakeddit datasets are considered, where the samples are classified into six classes. For combining the class labels from various models, fusion methods like Maximum, Concatenate, Add, and Average are considered. It is observed that BERT for Text and ResNet50 for image classification, combined with the fusion method as Maximum, has shown better results. BERT+ResNet50 achieved an accuracy of 89.29% for 2-way, 89.05% for 3-way, and 86% for 6-way classification with *Maximum* as fusion method (Nakamura et al., 2019).

Kirchknopf Armin et al. proposed a multimodal detection for Information Disorder in social media. The model proposed works on various combinations of Text data, Visual content, Comments for the visual content, and metadata. Classification results are fused using Sum, Concatenate, and maximum methods, and Fakeddit data is used for analysis. The combination of Visual content and Comments related to the visual content recorded a Validation and Testing accuracy of 88% and 88.1%, respectively. The model achieved better results when text, image content, comments, and metadata related to social posts were considered (Kirchknopf et al., 2021).

Priyanka Meel and Dinesh Kumar V have proposed an ensemble multimodal for Fake News detection that utilizes a Hierarchical Attention Network (HAN), Image Captioning, and Error Level Analysis. Max-voting is the fusion method employed to combine the models' results. The Text part of the dataset is analyzed using HAN, ELA, and Noise Variant Inconsistency for the images and Max fusion for attaining the Max vote class label for the Image with Text (caption and comments) content embedded. It is observed that the combined model outperformed individuals and other state-of-art models when working on the Fake News Samples dataset. The proposed ensemble model on the Fake News Sample dataset achieved an accuracy of 94.7% (Meel & Vishwakarma, 2021).

Yan Wu et al. have proposed multimodal Co-Attention networks based fusion network for Fake News detection. The model used BERT for working on the Textual aspects and VGG19 for working on the visual features of the data. Textual and Image related features from Twitter and Weibo datasets are used for analysis. Textual, spatial, and frequency domain aspects are fused to detect Fake posts. It is observed that the proposed MCAN model achieved an accuracy of 80.9% on the Twitter dataset and 89.9% on Weibo dataset (Wu et al., 2021).

Dhruv Khattar et al. proposed a multimodal variational autoencoder for detecting fake news. A bimodal variational autoencoder and a binary classifier were used for fake news classification. This model contains three components: an encoder that transforms data from text and images into latent vectors, a decoder that uses these latent vectors to re-construct the text and images, and latent vectors to detect fake news and images. Each encoder and decoder contains individual blocks for text and images. Twitter and Weibo datasets are used for processing. Proposed MVAE gave an accuracy of 74.5% on Twitter and 82.4% on Weibo datasets (Khattar et al., 2019).

Rina Kumari and Asif Ekbal proposed an Attention-based Multimodal factorized Bilinear Pooling model to detect Fake posts with Image and Textual data. The proposed

framework has four modules for textual features representation (Attention-based Stacked BiLSTM), image feature representation (Attention-based Multilevel CNN-RNN), a Multimodal Factorized Bilinear pooling for the fusion of textual and image features, and Multi-Layer Perceptron (MLP) for final classification. Data from Twitter and Weibo is used for analysis, and it is observed that the proposed model gave an accuracy of 88.3% on Twitter and 89.23% on the Weibo dataset (Kumari & Ekbal, 2021).

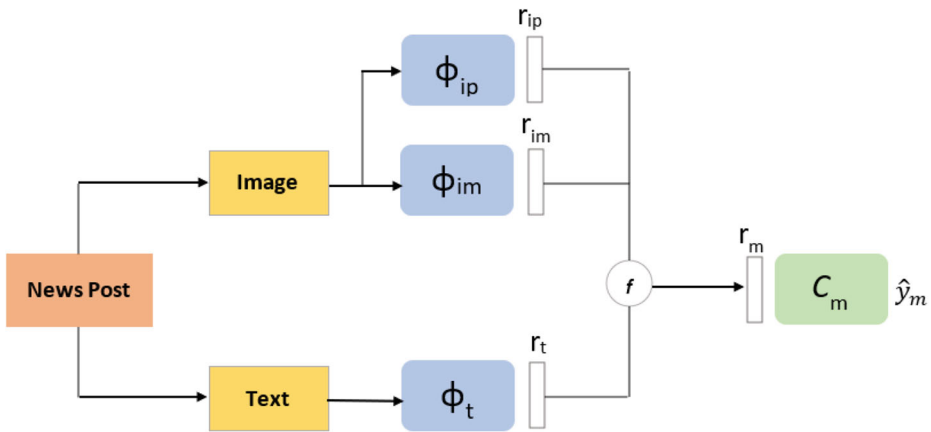
Peng Qi et al. proposed exploiting multi-domain visual information for fake news detection. The CNN-based network captured complex patterns and multi-branch CNN-RNN for extracting visual features at semantic levels. The MVNN model has three sub-models: Frequency domain sub-network, Discrete Cosine Transform, and Pixel domain sub-network. The fusion domain sub-network utilizes pixel and frequency domain sub-networks features to detect whether the images are fake or real. Verified data from Twitter and Weibo are considered for processing. When Attention-RNN (attRNN) was used, the proposed MVNN model gave an accuracy of 90.1%, Event Adversarial Neural Networks (EANN) with 89.7%, and MVAE around 89.1% (Qi et al., 2019).

Mathieu and Brahim proposed a multimodal sentiment analysis model for text and image-based posts and a fusion network to combine both modalities. Flickr Emotion VSO (Visual Sentiment Ontology) datasets are used in this model. The model proposed has achieved an accuracy of 91.17% on the Flickr Emotion dataset and 86.35% VSO dataset. As a part of the second experiment, multitask framework was enabled to be trained only with monomodal data. When trained with more images, Multitask model achieved an accuracy of 91.59% on the Flickr dataset. In addition, two auxiliary image and text-based classifiers are introduced to the traditional multimodal framework to handle missing modalities (Fortin & Chaib-Draa, 2019).

Jiangfeng et al. proposed a multimodal correlation model for detecting Fake News for epidemic emergencies using the deep correlations between text and images. The model has three phases; wherein the first phase, the image representatives are learned using a pre-trained VGG model and used for learning the textual representations using a hierarchical attention mechanism. In the second phase, multimodal representations are modeled to learn the fused text and image representations. In the third phase, image-enhanced text representations and the fused eigenvectors are combined to detect Fake News. It is observed that the proposed model achieved an accuracy of 83.4% (Zeng et al., 2021).

## 2.1 Motivation for the study

- Often Image analysis is carried out using Forensic techniques, which alone are insufficient to handle the problem of fake image detection. There is a need for a universal approach that can handle the scale and varied nature of fake news images.
- Often Image Caption data is neglected for analysis. Image Captions are being used as Clickbaits to enhance the reachability of the posts. Therefore, it is important to work on Image captions to detect clickbaits or misleading captions.
- There is also significant amount of work in using multiple modalities for sentiment analysis. However, different research works use varied datasets and hence comparison of performances of various frameworks developed is required.
- There is a need to develop a framework which focuses on identifying features unique to fake news images and its corresponding captions to aid in their identification.
- There is need to design a framework that works on Manipulated as well as sentiment cues from the images.



**Fig. 2** Overview of the proposed approach

### 3 Methodology

Social posts often contain a combination of images and text. Text embedded might be caption related to the image published or event that took place. Hence, it is always tedious to work with such social posts that combine image and text. Sophisticated methods are to be employed to work on such posts. Figure 2 depicts a model that works with image and text data separately and then combines the insights from both models.

The proposed framework has an image model that works on visual features,  $\Phi_{ip}$  and image polarity,  $\Phi_{im}$ . The textual network handles the textual portion or caption of the social post,  $\Phi_t$ . A multimodal classifier combines the features from these networks using a fusion method,  $C_m$ . When a social post is published, the framework works on both models independently and gives the combined classification result.

Fakeddit, a publicly available dataset, is used for analysis. It has 1 million large-scale multimodal fake news data containing text, image, metadata, and comments data gathered from a wide range of sources. The fake news articles in this dataset are scraped from Reddit, social news and discussion platform, where users can submit submissions on various subreddits. Data scrapped is between March 2008 to October 2019. Data samples have multiple labels- namely 2-way, 3-way, and 6-way. Here, 2-way classification states whether the news is authentic or fake; 3-way classification states if it is entirely true, completely fake, or either with fake text with correct sample and vice versa; 6-way classification that states if the samples come into categories like Satire, True, Fake, Misleading content, Manipulated content, False content or Imposter content (Nakamura et al., 2019). For analysis, samples with both text and images are only considered, and other samples are ignored. It is observed that around 64% of the samples have both image and text embedded with the image. The statistics of the dataset are as given in Table 1.

**Table 1** Statistics of fakeddit dataset

Samples	Training	Validation	Testing
Real	222081	23320	23507
Fake	341519	35979	35763
Total	563600	59299	59270



Fig. 3 Image manipulation examples

## 4 Image manipulation and image polarity based fake news detection

It is observed that tweets or social posts with images spread faster and have a high level of retweets and shares. Images generally target the emotions of the people, and hence users are targeted with image-based fake news as it not only catches the attention but also has high spread and interaction patterns. Fake images can be either tampered images or images used out of context. Visual information from the social posts is used to determine if they are manipulated. Figure 3 depicts the famous manipulated images that have created a buzz on online social media. The left image is the famous image manipulation example, the composite photo of Senator Millard Tydings and American Communist Party leader Earl Browder (Thakur & Rohilla, 2020). The right image is the most circulated image during Hurricane sandy that took place in 2012 (Boididou et al., 2015).

### 4.1 Fake news detection based on image manipulation

Pre-trained deep neural networks are used to detect images that are manipulated. Initially, images are trained on different neural networks and the model that gives best accuracy is chosen for construction of our model. Figure 4 depicts the design of image modality models.

#### 4.1.1 Inception-ResNet-V2

Inception-ResNet-V2 is a 164 layer deep network capable of identifying images into 1000 categories. This model is trained on an ImageNet dataset with more than 1 million images, and this model analyzes the images and returns a list of class probabilities (Szegedy et al., 2017). For analysis, images from the Fakeddit dataset are used to fine-tune Inception-ResNet-V2. The convolutional part of the model is instantiated, and pre-trained weights

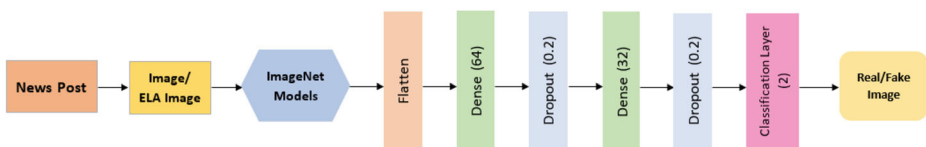


Fig. 4 Design of image modality models



from ImageNet are loaded. Images are scaled to 150 x 150, ReLu is the activation function used for every connected layer, and Softmax activation function is added on the top of the convolutional part. The model is run for a batch size of 256, with Adam optimizer and at a learning rate of 0.0005. The model is trained for 15 epochs, and it is observed that the model has the best validation loss at the 5th epoch. The model obtained validation accuracy of 80.49% and test accuracy of 80.6%.

#### 4.1.2 Xception

For analysis, images from the Fakeddit dataset are used to fine-tune Xception model. The convolutional part of the model is instantiated, and pre-trained weights from ImageNet are loaded (Chollet, 2017). Images are scaled to 150 x 150, ReLu is the activation function used for every connected layer, and Softmax activation function is added on the top of the convolutional part. The model is run for a batch size of 256, with Adam optimizer and at a learning rate of 0.0005. The model is trained for 15 epochs, and it is observed that the model has the best validation loss at the 2nd epoch. The model obtained validation accuracy of 82.07% and test accuracy of 82.32%.

### 4.2 Fake news based on image manipulation using error level analysis

The use of image editing tools has now made the manipulation of images very convenient. Visual content is an essential promoter for fake news propaganda as images offer a perception of *reality*, and hence users are often easily misled. Forensic techniques like Error Level Analysis (ELA) helps in identifying the digital alterations in the images, which analyses compression artifacts and helps identify regions in the image with different compression levels (Sudiatmika et al., 2019). ELA intentionally re-saves images at a compressed level and then computes the difference between these images (Abd Warif et al., 2015).

Figure 5 depicts the original image with its ELA and Fig. 6 depicts ELA for the modified image. Images clearly show that edited images have higher errors at the tampered regions. ELA images help identify digitally altered images since the error levels in such images are not uniform. Therefore, ELA for all the images is computed and these images are used to

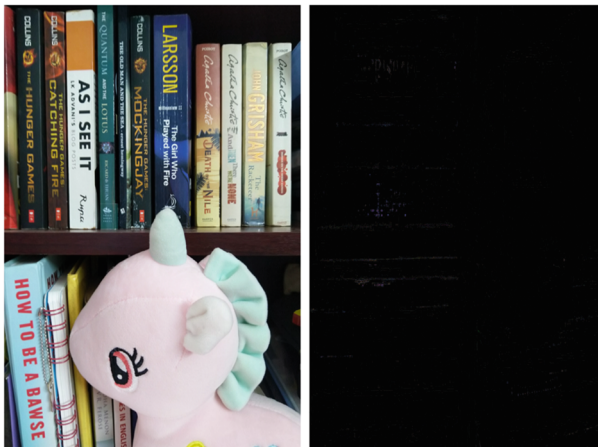


Fig. 5 Original Image and its ELA output



**Fig. 6** Edited Image and its ELA output

fine-tune CNN architectures to identify the altered images. Different CNN architectures are used to work on the ELA images.

#### 4.2.1 Inception-ResNet-V2 with ELA images

Inception-ResNet-V2 is a convolutional network trained on 1 million images from the ImageNet database. The network is 164 layers deep and can classify images into 1000 categories (Szegedy et al., 2017). Every fully connected layer in the model has a ReLU activation function, and a fully connected classifier with Softmax activation function is added on the top of the convolutional part. The entire model is trained on the Fakeddit dataset. Images from Fakeddit dataset is resized to 150X150, into a batch size of 256. The model is trained with Adam optimizer at a learning rate of 0.0005. The model is trained for 15 epochs. The best validation loss is obtained at the fifth epoch. The model obtained validation accuracy of 80.49% and test accuracy of 80.6%.

#### 4.2.2 ResNet50 with ELA images

ResNet50 has 48 deep convolutional layers. ELA computed images from the Fakeddit dataset are used to fine-tune the ResNet50 model for analysis (He et al., 2016a; Rezende et al., 2017). The convolutional part of the model is instantiated, and pre-trained weights from ImageNet are loaded. Images are scaled to 150 x 150, ReLU is the activation function used for every connected layer, and Softmax activation function is added on the top of the convolutional part. The model is run for a batch size of 256, with Adam optimizer and at a learning rate of 0.0005. The model is trained for 15 epochs, and it is observed that the model has the best validation loss at the fourth epoch. The model obtained validation accuracy of 79.01% and test accuracy of 79.58%.

#### 4.2.3 Xception with ELA images

ELA computed images from the Fakeddit dataset are used to fine-tune the Xception model for analysis. The convolutional part of the model is instantiated, and pre-trained weights

**Table 2** Validation and testing accuracy of image manipulation data

Text model	Validation accuracy	Test accuracy
VGG16 (Baseline)	73.55%	73.76%
EfficientNet (Baseline)	61.15%	60.87%
ResNet50 (Baseline)	80.43%	80.70%
Inception-ResNet-V2	80.49%	80.66%
Xception	<b>82.07%</b>	<b>82.32%</b>
Inception-ResNet-V2 with ELA	79.52%	79.70%
ResNet50 with ELA	79.01%	79.58%
Xception with ELA	79.58%	80.01%

Bold indicates models with better performance measures (here validation and Test accuracy)

from ImageNet are loaded (Chollet, 2017). Images are scaled to 150 x 150, ReLu is the activation function used for every connected layer, and Softmax activation function is added on the top of the convolutional part. The model is run for a batch size of 256, with Adam optimizer and at a learning rate of 0.0005. The model is trained for 15 epochs, and it is observed that the model has the best validation loss at the 2nd epoch. The model obtained validation accuracy of 79.58% and test accuracy of 80.01%.

Table 2 depicts the Performance measures of different Neural networks on Images from Fakeddit as well as for the ELA images. It is observed that the Xception model has better Validation and Testing accuracy. Therefore, the Xception model is used to construct the proposed model on the Fakeddit dataset.

### 4.3 Fake news based on visual sentiment of images

Online social posts with images often target the users' sentiment and induce strong visual impact. Analyzing the polarity of the images could help in detecting fake posts quickly. The proposed model is added with an additional branch that works on images with sentiment-related data. Transfer learning on CNN architecture is chosen for learning the features to analyze the sentiment of the images. A subset of the CrowdFlower dataset with positive and negative sentiment is chosen for analysis. Table 3 depicts the statistics of the CrowdFlower dataset.

#### 4.3.1 Transfer learning with VGG19

The pre-trained VGG19 model with transfer learning detects the images with sentiment data. Pretrained weights from ImageNet data are used for analysis (Simonyan & Zisserman, 2014; Rajinikanth et al., 2020). The model is run in two phases with varying learning rates and epochs. In the first phase, all layers are frozen with preloaded ImageNet weights. Adam optimizer, with a learning rate of 0.00001 and binary cross-entropy loss function, is used. The model is run with a batch size of 100 and 40 epochs. The model is run in two phases. In the first phase, it is observed that the model gave the best accuracy in the 37th epoch with a

**Table 3** Statistics of CrowdFlower dataset with Positive and Negative sentiment

Samples	Training	Validation	Testing
Positive sentiment	1000	112	250
Negative sentiment	1000	113	250

validation accuracy of 67.56%. In the second phase, 18 layers are frozen for the base model with pre-trained ImageNet weights with Adam optimizer, a learning rate of 0.00001, and a binary cross-entropy loss function. The model is run with a batch size of 100 and 40 epochs. It is observed that the model gave the best accuracy in the 31st epoch with a validation accuracy of 74.22%.

#### 4.3.2 Transfer learning with ResNet50

ResNet50 has 48 convolutional layers with one max pool and an average pool layer. For fine-tuning ResNet50, the base model is instantiated with pre-trained weights from ImageNet (Rezende et al., 2017). The model is run in two phases with varying learning rates and epochs. In the first phase, 170 layers are frozen for base-model with preloaded ImageNet weights. Adam optimizer, with a learning rate of 0.00001 and binary cross-entropy loss function, is used. The model is run with a batch size of 100 and 40 epochs. It is observed that the model gave the best accuracy in the 15th epoch with a validation accuracy of 68.44%. In the second phase, 165 layers are frozen for the base model with pre-trained ImageNet weights with Adam optimizer, learning rate of 0.00001, and binary cross-entropy loss function. The model is run with a batch size of 100 and 30 epochs. It is observed that the model gave the best accuracy in the fifth epoch with a validation accuracy of 71.11%.

#### 4.3.3 Transfer learning with ResNet50V2

In order to fine-tune ResNet50V2, the base model is instantiated with input size (150,150,3), and pre-trained weights from ImageNet are loaded and run in two phases (He et al., 2016b; Siegfried, 2020). In phase-1, the first 170 layers are frozen for the base model with preloaded ImageNet weights. Adam optimizer, with a learning rate of 0.00001 and binary cross-entropy loss function, is used. The model is run with a batch size of 100 and 40 epochs. It is observed that the model gave the best accuracy in the 15th epoch with a validation accuracy of 68.44%. In the second phase, 180 layers are frozen for the base model with pre-trained ImageNet weights with Adam optimizer, a learning rate of 0.00001, and a binary cross-entropy loss function. The model is run with a batch size of 100 and 30 epochs. It is observed that the model gave the best accuracy in the fifth epoch with a validation accuracy of 72.89%.

#### 4.3.4 Transfer learning with InceptionV3

InceptionV3 is a 48 layers deep convolutional network. In order to fine-tune InceptionV3, the base model is instantiated with input size (150,150,3), and pre-trained weights from ImageNet are loaded and run in three phases (Szegedy et al., 2016). In the first phase, 290 layers are frozen for the base model with preloaded ImageNet weights. Adam optimizer, with a learning rate of 0.00001 and binary cross-entropy loss function, is used. The model is run with a batch size of 100 and 30 epochs. It is observed that the model gave the best accuracy in the 16th epoch with a validation accuracy of 70.67%. In the second phase, 250 layers are frozen for the base model with preloaded ImageNet weights. Adam optimizer, with a learning rate of 0.00001 and binary cross-entropy loss function, is used. The model is run with a batch size of 100 and 40 epochs. It is observed that the model gave the best accuracy in the 16th epoch with a validation accuracy of 72.44%. In phase-3, all layers are frozen with Adam optimizer, learning rate of 0.00001, and binary cross-entropy loss

**Table 4** Validation and testing accuracy for visual sentiment data

	Text model	Validation accuracy	Test accuracy
	VGG19	74.22%	68%
	ResNet50	71.11%	66.44%
	ResNet50V2	72.89%	68.89%
Bold indicates models with better performance measures (here validation and Test accuracy)	InceptionV3	73.33%	68.89%
	Xception	<b>75.11%</b>	<b>70%</b>

function. The model is run with a batch size of 100 and 30 epochs. It is observed that the model gave the best accuracy in the third epoch with a validation accuracy of 73.33%.

#### 4.3.5 Transfer learning with Xception

The Xception model is fine-tuned by instantiating the base model with input size (150,150,3). Pre-trained weights from ImageNet are loaded and run in three phases (Chollet, 2017). In phase-1, all layers are frozen for the base model with preloaded ImageNet weights. With a learning rate of 0.0001 and binary cross-entropy loss function, Adam optimizer is used. The model is run with a batch size of 100 and 30 epochs. It is observed that the model gave the best accuracy in the seventh epoch with a validation accuracy of 75.11%. In the second phase, 115 layers are frozen for the base model with preloaded ImageNet weights. With a learning rate of 0.0001 and binary cross-entropy loss function, Adam optimizer is used. The model is run with a batch size of 100 and 40 epochs. It is observed that the model gave the best accuracy in the 16th epoch with a validation accuracy of 72%. In phase-3, all layers are frozen with Adam optimizer, learning rate of 0.00001, and binary cross-entropy loss function. The model is run with a batch size of 100 and 30 epochs. It is observed that the model gave the best accuracy in the third epoch with a validation accuracy of 74.67%. Table 4 depicts the validation and Testing accuracy of different models on the visual sentiment data. It is observed that the testing accuracy is higher for the Xception model. Therefore, the Xception model is used to construct the proposed model on the Fakeddit dataset.

### 4.4 Fake news detection based on image caption

Fake news potentially differs from the truth in writing style and quality, word count, and sentiment expressed. As a result, it is fair to identify fake news using linguistic features that capture various writing styles and sensational headlines. Various text modality models were implemented and evaluated. The models were fine-tuned on the Fakeddit dataset on the textual information (image caption) in posts to determine whether they were fake or not. Various pre-trained neural networks are employed to classify the image captions into fake and real categories.

#### 4.4.1 LSTM + CNN

LSTM, added with a layer of one-dimensional CNN with max pool layer, is used to learn the spatial features of the image caption. A dense layer with softmax as an activation function is finally used to classify the captions (Xia et al., 2020). Adam optimizer with binary cross-entropy loss and three callback functions (CSV logger, Tensorboard, and Model check) is

used, and the model is run for 20 epochs with a batch size of 1024. The best validation loss was obtained at the fourth epoch with validation accuracy of 85.51% and test accuracy of 85.25%.

#### 4.4.2 BiGRU + CapsuleNet

This model uses word embeddings from pre-trained Glove and Paragram, combined with meta-embedding. A bidirectional GRU layer added with the Capsule network is used to classify the data (Deng et al., 2020). Adam optimizer with binary cross-entropy loss and three callback functions (CSV logger, Tensorboard, and Model check) is used, and the model is run for 20 epochs with a batch size of 1024. The best validation loss was obtained at the third epoch with validation accuracy of 85.98% and test accuracy of 86.18%.

#### 4.4.3 BiLSTM+BiGRU+attention

BiLSTM helps understand the context of the sentences by using the words before and after the current word, thereby offering better predictions (Zhou & Bian, 2019). Word embeddings are used from Glove and fasttext. The attention layer is added after BiGRU, and the output is passed through the Global max Pooling layer. Adam optimizer with binary cross-entropy loss and three callback functions (CSV logger, Tensorboard, and Model check) is used, and the model is run for 15 epochs with a batch size of 512. The best validation loss was obtained at the fourth epoch with validation accuracy of 87.89% and test accuracy of 87.90%.

#### 4.4.4 2D CNN

In the 2D CNN model, the word is embedded from pre-trained Glove and Fasttext. The concatenated word embeddings are reshaped, and 2D CNN is applied with different filter sizes (Zhao et al., 2019). Adam optimizer with binary cross-entropy loss and three callback functions (CSV logger, Tensorboard, and Model check) is used, and the model is run for 15 epochs with a batch size of 512. The best validation loss was obtained at the second epoch with validation accuracy of 86.52% and test accuracy of 86.77%.

#### 4.4.5 BERT+Dense

The BERT model helps understand and process ambiguous text by learning the context of the sentence (Wang et al., 2021). A dense output layer with softmax as an activation function is added to the BERT model. Adam optimizer with binary cross-entropy loss and one callback function (Model check) is used, and the model is run for ten epochs with a batch size of 144. The best validation loss was obtained at the second epoch with validation accuracy of 89.34% and test accuracy of 89.46%.

#### 4.4.6 RoBERTa+Dense

RoBERTa is the model built based on the BERT model that helps predict the unintentionally hidden sections of the text. RoBERTa is pre-trained on Books Corpus and English Wikipedia; in addition to this dataset, RoBERTa is trained on CommonCrawl, Web text corpus, and stories from Common Crawl datasets. A dense output layer with softmax as an activation function is added to the RoBERTa model (Kalyan & Sangeetha, 2020). Adam

**Table 5** Performance measures of various text models

	Text model	Validation accuracy	Test accuracy
	BERT (Baseline)	86.54%	86.44%
	InferSent (Baseline)	86.34%	86.31%
	LSTM+CNN	85.51%	85.25%
	BiGRU+Capsule	85.98%	86.18%
	BiLSTM+BiGRU+attention	87.89%	87.90%
	2D CNN	86.52%	86.77%
Bold indicates models with better performance measures (here validation and Test accuracy)	BERT+Dense	<b>89.34%</b>	<b>89.46%</b>
	RoBERTa+Dense	88.52%	88.62%

optimizer with binary cross-entropy loss and one callback function (Model checkpoint) is used, and the model is run for ten epochs with a batch size of 144. The best validation loss was obtained at the sixth epoch with validation accuracy of 88.52% and test accuracy of 87.90%.

Table 5 depicts the performance measures of various pre-trained models on the image caption data. As Validation and Testing accuracy is high for BERT+Dense model, this model is used to construct the proposed model.

#### 4.5 Inferences from image manipulation, visual sentiment and image caption

After analyzing the data using different Deep Learning models on the manipulated and visual sentiment of the data, the models with better accuracy are considered further. Different models like VGG16, EfficientNet, ResNet50, Inception-ResNet-V2, Xception; Inception-ResNet-V2, ResNet50, and Xception with ELA images are used on manipulated data. Further VGG19, ResNet50, ResNetV2, ResNetV3, and Xception on the visual sentiment, and finally, the models with better accuracies are considered for developing the proposed framework. The following are inferences drawn

- Xception (Extreme Inception), with 71 layers, has shown better accuracy when compared to other models. Instead of partitioning input data into several compressed chunks, the Xception model tries to map the spatial correlations for each output channel separately and then perform a 1x1 depthwise convolution to capture cross-channel correlations. Xception combines the advantage of the Inception module with the Residual feature, making Xception give good results combined with other models (A Dense layer, as in the proposed model).
- Xception also reduces the effect of the vanishing gradient problem, which makes this model better in classifying Fake and Real images. Xception, contrary to the models like Inception, has no non-linearity module (No intermediate ReLU non-linearity).
- ELA (Error Level Analysis), a forensic technique, is not observed to give better results than other models. Therefore, ELA models are not considered for further analysis.
- Visual sentiment analysis from the CrowdFlower dataset is chosen for analysis. Transfer learning (Retraining the Pre-trained model) is used to work on the image polarity data. As the dataset is around 2000 samples, it is observed that the accuracy level is less when compared to the manipulated data.
- As BERT and RoBERTa has contextual embedding and are trained on a larger dataset, BERT and RoBERTa outperformed other models like BiGRU, LSTM, and InferSent.

As a deep layer is added to the BERT/RobERTa model, it has shown a significant improvisation in terms of accuracy. For the dataset crawled for the current analysis, BERT has shown to be a bit better than RobERTa (which might not be the case all the time, as RobERTa, is said to be an Optimized version of BERT and worked on mode data points, and BERT model without NSP (Next Sentence Prediction)).

- Upon analyzing the data, BERT+Dense is considered for Textual data, and Xception is used for Image Manipulation and Visual Polarity data while creating the multi-modal framework.

## 5 Proposed ensemble model for fake news detection

An ensemble model was designed to improve the identification of fake news with

- Xception model to help in identifying images with high digital alterations (Chollet, 2017).
- BERT to learn contextual knowledge.
- Visual sentiment analysis to learn features that distinguish an image with negative sentiment from that which induces positive emotions, thereby identifying misleading and tampered fake images with high confidence.

---

**Input :** Images (I)  
Image Captions ( $I_t$ )  
Visual Sentiment (Image Polarity) data ( $I_{vsd}$ )

**Output:** Classification labels for the Images  
Class Labels,  $C_i \in \{Fake, Real\}$

**procedure:** Visual Sentiment based Image classification

$\Phi_{ip} \leftarrow Xception(I_{vsd})$

Classifying the Manipulated Images

$\Phi_{im} \leftarrow Xception(I)$

Classifying the embedded Text (Image Caption)

$\Phi_t \leftarrow (BERT + Dense)(I_t)$

Fusing the classifying labels generated from Image Polarity ( $\Phi_{ip}$ ) and Image Manipulation data ( $\Phi_{im}$ )

$\Phi_i \leftarrow \Phi_{im} \oplus \Phi_{ip}$

Classifying results of the Proposed Multimodality model

$C_i \leftarrow MC(\Phi_i, \Phi_t)$

where,  $MC(\bullet)$  is defined as Maximum or Concatenate

---

**Algorithm 1** Proposed multimodality model.

The design of the ensemble construction is depicted in Fig. 7. The leftmost vertical branch consists of layers from the BERT model fine-tuned on image captions, the middle branch is composed of layers from the Xception model fine-tuned on Fakeddit dataset images, and the rightmost vertical branch is comprised of layers from the Xception model fine-tuned on Sentiment dataset of images. This proposed ensemble model includes all layers except the last classification layers. Finally, the last module is a multimodal fusion module that combines representations from various modalities (such as text and image) to



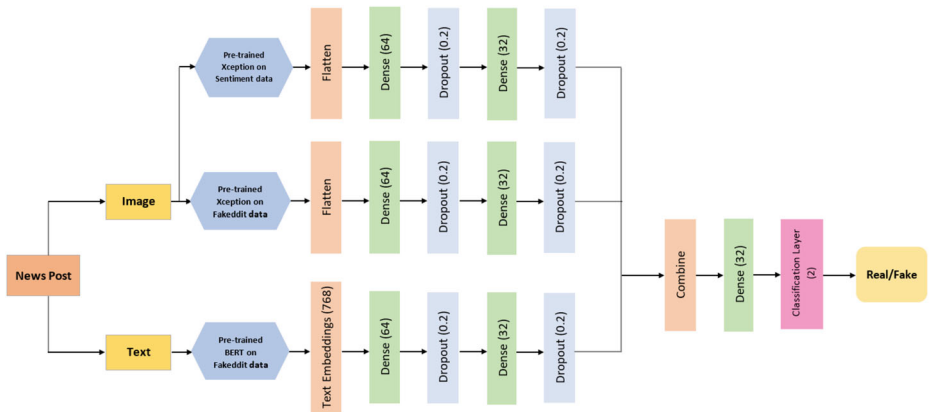


Fig. 7 Proposed model design

form a new feature vector. This news representation is fed into a completely connected neural network with softmax activation for fake news classification.

### 5.1 Fusion models

Social Media consists of multimedia posts which can normally be Text, Images, Audio, or Video. Data from various modals should be analyzed, and the prediction probabilities from different modals should be grouped to predict a final class label for the post. Multi-modal Fusion, therefore, acts as a process of combining features from various modalities to perform a prediction. Fusion can be either Early, Late or Intermediate (Kiela et al., 2018).

#### Early fusion

- Early Fusion tends to concatenate features from various modals into a single feature vector and is fed into the model to obtain the prediction.
- It becomes tedious to work with features with higher granularity and can be very highly dimensional (due to the Fusion of pre-processed features from different modals).
- Figure 8 denotes Early Fusion.

#### Late fusion

- Late Fusion or decision-level Fusion aims at aggregating decisions from multiple modalities, each trained separately.

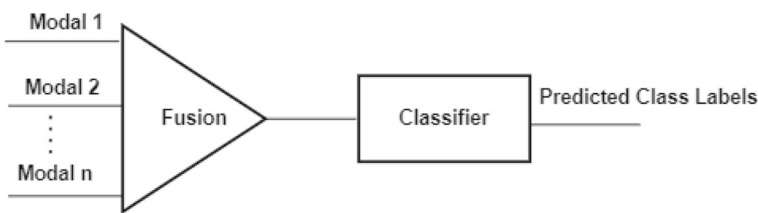


Fig. 8 Early fusion

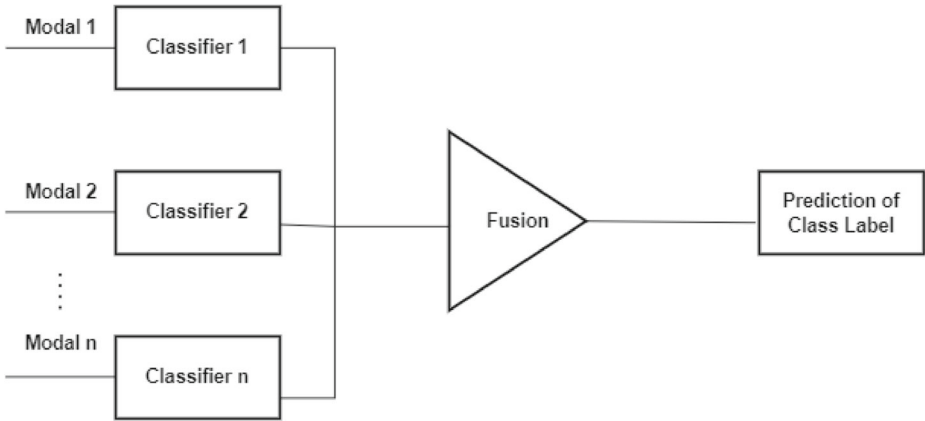


Fig. 9 Late fusion

- The method is feature independent, and any errors from multiple modals tend to be uncorrelated.
- Figure 9 denotes Late Fusion.

**Intermediate fusion**

- Intermediate Fusion aims at creating representation layers (typically a single shared layer) that merge the units from multiple modality-specific paths.
- The representation layer created can be either a single layer that maps multiple channels or can be a combination of layers fusing different sets of modals at different levels.

**5.1.1 Available fusion techniques**

The fusion Model aims at merging independent features into unique features. If  $f_n^t$  is the normalized text feature representation and  $f_n^v$  is the visual feature representation; and U, V, and W are weight matrices or kernel that helps in combining different features, The following are some of the major fusion models in practice

- **Element-wise Sum:** It is also termed Component-wise Sum, which combines the features from multiple modalities. The features from the combining modalities should be similar in terms of their shape. The Additive/ Element-wise sum is given by

$$W(U f_n^t + V f_n^t)$$

The Element-wise sum is generally disordered. This method does not account for being best when working with relatively large data or where the ordering is prominent.

- **Attention/ Gated:** This Fusion method is used if one modality needs to be given attention or importance over the other. Attention is a hyper-parameter. Sigmoid non-linearity is used to gate one modality over the other. The Attention/Gate is given by

$$W(\sigma(U f_n^t) * V f_n^t)$$

or

$$W(U f_n^t * \sigma(V f_n^t))$$

- **Maximum:** It is also called Max-Pooling, generally used when combining features from multiple modalities. This function computes component-wise maximum, given by

$$W(\max(U f_n^t, V f_n^t))$$

This helps attain the features with maximum weight when comparing different modalities like Text and Image.

- **Concatenate:** In general, when combining features computed from the same algorithm, either Element-wise addition or attention can be used. If the features are generated from different algorithms (each modality learned by a different algorithm), then Concatenate can be used to improve the performance of the combined model. Combining features from different modalities computed by different algorithms using attention or gated mechanisms is given by.

$$W(U f_n^t V f_n^t)$$

Other fusion models used include Average, Element-wise Product, and Polling. The fusion model to be chosen depends on what modals are being combined and the weightage to be given to each modal. In case no attention is granted to a particular modal, fusion models like Sum, Average, and Concatenate can be used. One might also come across fusion techniques which combine features from different layers and then combine all these features, which comes under intermediate fusion models (Boulahia et al., 2021; Baltrušaitis et al., 2018).

### 5.1.2 Proposed fusion model for image and text modality data

As the proposed framework works with Textual and Visual content (Visual Manipulated and Visual Polarity related data), independently learning features using different models (algorithms), hence Maximum and Concatenate are the fusion models considered for analysis. In the proposed framework, the textual data is analyzed using pre-trained BERT, Visual data, which has two characteristics related to Image Polarity and Image manipulation is analyzed using pre-trained Xception models. Finally, the features obtained from these branches are combined using fusion models like Concatenate and Maximum to finally classify the social posts as Fake/Real.

**Maximum fusion model** While working with Maximum Fusion model, all the layers from visual and Textual branches are frozen (i.e, the layer weights of the trained model are not changed). Freezing helps in retaining weights from its pretrained phase (ImageNet for Xception and Wikipedia/Brown Corpus for BERT). All the layers in the Xception branch are frozen until the Flatten layer, and all layers in the BERT branch are frozen until the Text embeddings. The Image Polarity-related data was frozen till the Merge layer. Both modalities' 32-dimensional vectors are combined using the maximum fusion method and fed into a fully connected neural network classifier with a 32-layer hidden and a 2-layer classification layers with softmax activation with batch size 256 and Adam optimizer with 0.0005 learning rate. The model is run for 20 epochs. The Maximum fusion model proposed is shown in Fig. 10. The best validation loss was obtained at the 12th epoch. The classification scores are shown in Table 6

**Concatenate fusion model** While working with Concatenate Fusion model, all the layers from visual and Textual branches are frozen (i.e, the layer weights of the trained model are not changed). Freezing helps in retaining weights from its pretrained phase (ImageNet for Xception and Wikipedia/Brown Corpus for BERT). All the layers in the Xception branch

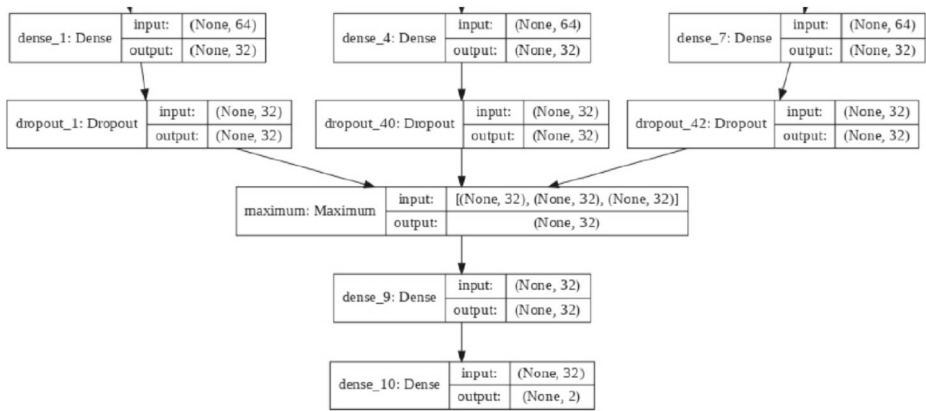


Fig. 10 Proposed framework with maximum as fusion method

are frozen until the Flatten layer, and all layers in the BERT branch are frozen until the Text embeddings. The Image Polarity-related data was frozen till the Merge layer. Both modalities’ 32-dimensional vectors are combined using the concatenation fusion method and fed into a fully connected neural network classifier with a 32-layer hidden, and a 2-layer classification layers with softmax activation with batch size 256, and an Adam optimizer with 0.0005 learning rate. The model is run for 20 epochs. The Maximum fusion model proposed is shown in Fig. 11. The best validation loss was obtained at the 13th epoch. The classification scores are shown in Table 7

### 5.1.3 Coupling

Coupling generally aids the fusion models in integrated feature representation and feature fusion mechanisms. Coupling Layers may exist at different layers or the fusion step, depending on the type of fusion model considered for analysis. When a vast dataset is analyzed as batches using the same algorithm, or when Intermediate coupling is employed where features at various levels are to be combined, a strong coupling is generally used. Loose coupling is generally employed when different modalities are being coupled using various algorithms. In the proposed algorithm, as the Textual features, Image features (Image Manipulation and Image Polarity related features) are independently analyzed using BERT and Xception models and are coupled to classify the posts; loose coupling is employed (Song et al., 2021).

## 5.2 Result analysis

The proposed ensemble model is loaded with the best weights obtained from 3 models trained independently - fine-tuning Xception on Fakeddit dataset images, fine-tuning BERT

Table 6 Classification scores of max fusion

Epoch	Accuracy	Loss	Val_Accuracy	Val_Loss	Test_Accuracy
12	94.43%	13.76%	91.68%	21.95%	91.94%

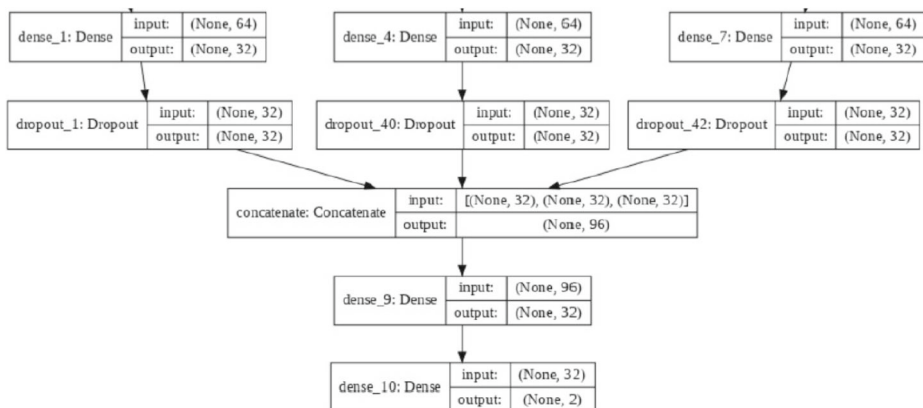


Fig. 11 Proposed framework with concatenate as fusion method

on image captions(i.e., text), and fine-tuning Xception network for sentiment analysis. All layers in the Xception branch are made untrainable until the Flatten layer, and all layers in the BERT branch are made untrainable until the Text embeddings. The whole sentiment branch was made untrainable till the Merge/combine layer. The ensemble model was fine-tuned again on Fakeddit dataset samples with both images and text. The padded and tokenized text was passed into the BERT model to receive word vectors of dimension 768. The images were rescaled to 150x150 pixels before being passed into the models. As depicted in Fig. 7, both modalities’ 32-dimensional vectors are combined using *Maximum* and *Concatenate* fusion method (Atrey et al., 2010) and fed into a fully connected neural network classifier with a 32-layer hidden layer and a 2-layer classification layer with softmax activation. The model was trained for 20 epochs, a batch size of 256, and an Adam optimizer with a learning rate of 0.0005. Better validation accuracy for *Maximum* as fusion method is observed at 12th epoch (91.68%) and for *Concatenate*, it is observed at 13th epoch (91.94%).

Table 8 shows the result of multi-modality models trained on the Fakeddit dataset. Table 9 displays both the baseline results and the proposed method result on Fakeddit dataset.

In terms of accuracy, precision, recall, and F1 score, the proposed method outperforms the current methods overall. It is evident from the tables that multimodal models outperform unimodal models. These results further validate that multi-modality helps learn better distinguishing features between fake and real news. When it comes to assessing the accuracy of the news, data from various sources complement each other. A good recall score is crucial in this context of fake news identification since we would not want to miss flagging a fake news post. At the same time, we also need to be reasonably precise with predictions. The proposed method has a high recall, precision, and an F1-score of ~ 93% as depicted in Table 8.

Table 7 Classification scores of concatenate fusion

Epoch	Accuracy	Loss	Val_Accuracy	Val_Loss	Test_Accuracy
13	94.48%	13.66%	91.70%	22.31%	91.87%

**Table 8** Performance measures of the text + image models on fakeddit dataset

Text+Image model	Fusion method	Validation accuracy	Test accuracy	Precision	Recall	F1-score
BERT+Xception	Maximum	91.61%	91.87%	93.43%	93.07%	93.25%
BERT+Xception	Concatenate	91.67%	91.88%	93.31%	93.22%	93.26%
(BERT+Dense)+Xception	Maximum	91.68%	<b>91.94%</b>	<b>93.76%</b>	92.83%	93.29%
(BERT+Dense)+Xception	Concatenate	<b>91.70%</b>	91.87%	93.39%	<b>93.29%</b>	93.25%

Bold indicates models with better performance measures (here validation and Test accuracy)

As the proposed framework works on Manipulated data and Visual sentiment data, there is more scope for analyzing the social posts with high polarity images. Often Fake News is spread with higher sentiment to grab the users' attention by targeting the psychological aspects of the users interacting with such posts. As text analysis is added to work with the embedded image captions, the proposed framework helps detect the clickbaits, which is one of the major aspects that helps in the easy propagation of the posts. Further, the error analysis is discussed in the next Section 5.3.

### 5.3 Error analysis

The proposed model incorporates various concepts relating to the visual and linguistic components of the post (Image caption). Several combinations are attempted for the visual and textual features of the social posts, and models with higher accuracy are fused for analysis. Fake images online tend to have a substantial visual impact and induce high sentiment.

The visual sentiment branch suffers from a higher error rate when compared to others, as the sentiment of the data can often be misleading. Hence, Image polarity detection is often challenging (identifying whether an image reflects positive or negative emotion). It is also difficult to gather a large sample set for visual sentiment data. In most cases, there is either human power to explicitly tag image sentiment after crawling the data or third-party tools

**Table 9** Performance of the proposed model in comparison to the baseline models




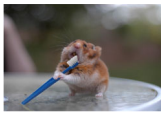



Multimodality Models	Fusion method	Validation accuracy	Test accuracy
InferSent+VGG16 (Baseline)	Maximum	86.55%	86.58%
InferSent+EfficientNet (Baseline)	Maximum	83.28%	83.39%
InferSent+ResNet50 (Baseline)	Maximum	88.88%	88.91%
BERT+VGG16 (Baseline)	Maximum	86.94%	86.99%
BERT+EfficientNet (Baseline)	Maximum	83.34%	83.18%
BERT+ResNet50 (Baseline)	Maximum	89.29%	89.09%
BERT+ResNet50 (Baseline)	Concatenate	85.64%	85.68%
BERT+Xception (Proposed)	Maximum	91.61%	91.87%
BERT+Xception (proposed)	Concatenate	91.67%	91.88%
(BERT+Dense)+Xception (proposed)	Maximum	91.68%	<b>91.94%</b>
(BERT+Dense)+Xception (proposed)	Concatenate	<b>91.94%</b>	91.87%

Bold indicates models with better performance measures (here validation and Test accuracy)


**Table 10** Error rate of different modalities

Model	Error rate
Image modality (Xception)	0.18
Visual sentiment (Xception)	0.30
Text modality (BERT+Dense)	0.11
Fusion (Concatenate)	0.08

**Table 11** Samples of error analysis- true (class related fake news), and false (class related to real news)

Actual label			Predicted label			Percentage	Image	Caption
Multi	Text	Image	Multi	Text	Image			
Real	Real	Real	Fake	Fake	Fake	10.51		Penguin battle
Fake	Fake	Fake	Fake	Real	Real	3.41		Exotic plant
Real	Real	Real	Fake	Real	Fake	10.45		The rottness monster
Fake	Fake	Fake	Fake	Fake	Real	21.92		Hamsters care about dental hygiene too
Real	Real	Real	Real	Fake	Fake	0.41		Outlet with usbc port
Real	Real	Real	Real	Real	Fake	6.99		There is a chicken sitting on my car
Fake	Fake	Fake	Real	Fake	Real	17.15		A yawning seal

**Table 11** (continued)

Actual label			Predicted label			Percentage	Image	Caption
Multi	Text	Image	Multi	Text	Image			
Fake	Fake	Fake	Real	Real	Real	29.13		A tortoise near a grove of mushrooms

like Amazon Mechanical Turk (crowd-sourcing marketplace) to the tag sentiment of the images. For image captions, it is observed that most of the captions are intentionally written to grab attention, and Click-baits are common while spreading Fake News. The error rate is determined for each modality and the combination of modalities. The error rate for several models is shown in Table 10.

It is observed that using either text or an image alone might not be sufficient for detecting falsification. However, in the multimodality framework, Image caption is observed to have a high impact on correctly classifying Fake News. Table 11 depicts examples of the error analysis for Image, Text, and Multimodality.

## 6 Conclusion and future scope

This paper proposed a framework that combines visual and textual features to detect fake news. Posts are crawled from the Fakeddit dataset, with an image and its caption, and Fakeddit has around 1 million images crawled from Reddit. Fine-tuned BERT is implemented on textual information in posts to determine whether they are fake or real. Fine-tuned BERT achieved an accuracy of 89.31%. On the other hand, fine-tuned Xception network is used on the visual content of the posts, and it showed an accuracy of 82.32%. The fusion of models is considered, and unlike the traditional image forensic methods, a framework is proposed to identify both tampered and images that are not altered. The proposal model achieved an accuracy of 91.94% and an F1-score of 93%. It is evident that the textual (caption of the image) part of the social post, followed by the visual component of the image, plays a vital role in detecting fake posts.

As part of the future scope, the plan is to use metadata and comments of the posts and combine these with the user-related data to track the user's credibility in the interactions. More samples is planned to be collected with visual sentiment (polarity) to enhance the capability of the visual sentiment branch. A cross-domain generalization model is planned to be implemented on social posts across domains, topics, websites, and languages. The user engagement patterns can also be combined to the models that helps in attaining generalization across domains.

**Author Contributions** Santosh Kumar Uppada: Conceptualization, Mathematical Modeling, Methodology, Writing the draft, Validation.

Parth Patel: Formal Analysis, Visualization, Data Curation. Dr. Sivaselvan B: Draft Review and Editing, Supervision, Investigation.



**Funding** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Data Availability** The paper does not include any supporting data.

## Declarations

**Consent for Publication** There is no content that requires permission of any third-party organizations or persons to publish the above manuscript.

**Competing interests** The Authors does not have any competing interests.

## References

- Abd Warif, N.B., Idris, M.Y.I., Wahab, A.W.A., & et al. (2015). An evaluation of error level analysis in image forensics. In *2015 5th IEEE international conference on system engineering and technology (ICSET)* (pp. 23–28). IEEE. <https://doi.org/10.1109/ICSEngT.2015.7412439>.
- Adamic, L.A., & Huberman, B.A. (2000). Power-law distribution of the world wide web. *Science*, 287(5461), 2115–2115. <https://doi.org/10.1126/science.287.5461.2115a>.
- Atrey, P.K., Hossain, M.A., El Saddik, A., & et al. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6), 345–379. <https://doi.org/10.1007/s00530-010-0182-0>.
- Baltrušaitis, T., Ahuja, C., & Morency, L.P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>.
- Bazaco, A. (2019). Clickbait as a strategy of viral journalism: Conceptualisation and methods. <https://doi.org/10.4185/RLCS-2018-1323en>.
- Boididou, C., Andreadou, K., Papadopoulos, S., & et al. (2015). Verifying multimedia use at mediaeval 2015. *MediaEval*, 3(3), 7. <https://doi.org/10.1145/1235>.
- Boulahia, S.Y., Amamra, A., Madi, M.R., & et al. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6), 1–18. <https://doi.org/10.1080/24725854.2021.1987593>.
- Campbell, W.J. (2001). *Yellow journalism: Puncturing the myths, defining the legacies*. United States: Greenwood Publishing Group. <https://doi.org/10.1002/9781118841570.iejs0159>.
- Cao, J., Qi, P., Sheng, Q., & et al. (2020). Exploring the role of visual content in fake news detection. In *Disinformation, misinformation, and fake news in social media* (pp. 141–161). Cham: Springer. [https://doi.org/10.1007/978-3-030-42699-6\\_8](https://doi.org/10.1007/978-3-030-42699-6_8).
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258). <https://doi.org/10.1109/CVPR.2017.195>.
- Chowdhury, A. (2020). Fake news in the time of coronavirus: A BOOM study. BOOM.
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., & et al. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), e2023301118. <https://doi.org/10.1073/pnas.202330111>.
- Clever, L., Assenmacher, D., Müller, K., & et al. (2020). FakeYou!- A gamified approach for building and evaluating resilience against fake news. In *Multidisciplinary international symposium on disinformation in open online media* (pp. 218–232). Cham: Springer. [https://doi.org/10.1007/978-3-030-61841-4\\_15](https://doi.org/10.1007/978-3-030-61841-4_15).
- Deng, J., Cheng, L., & Wang, Z. (2020). Self-attention-based BiGRU and capsule network for named entity recognition. arXiv:2002.00735. <https://doi.org/10.48550/arXiv.2002.00735>.
- Fortin, M.P., & Chaib-Draa, B. (2019). Multimodal sentiment analysis: A multitask learning approach. In *ICPRAM* (pp. 368–376). <https://doi.org/10.5220/0007313503680376>.
- Galli, A., Masciarì, E., Moscato, V., & et al. (2022). A comprehensive Benchmark for fake news detection. *Journal of Intelligent Information Systems*, 59, 237–261. <https://doi.org/10.1007/s10844-021-00646-9>.
- Giachanou, A., Zhang, G., & Rosso, P. (2020). Multimodal multi-image fake news detection. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)* (pp. 647–654). IEEE. <https://doi.org/10.1109/DSAA49011.2020.00091>.
- He, K., Zhang, X., Ren, S., & et al. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>.

- He, K., Zhang, X., Ren, S., & et al. (2016b). Identity mappings in deep residual networks. In *European conference on computer vision* (pp. 630–645). Cham: Springer. <https://doi.org/10.48550/arXiv.1603.05027>.
- Jin, Z., Cao, J., Luo, J., & et al. (2016). Image credibility analysis with effective domain transferred deep networks. arXiv:1611.05328.
- Jin, Z., Cao, J., Guo, H., & et al. (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). <https://doi.org/10.1145/3123266.3123454>.
- Kalyan, K.S., & Sangeetha, S. (2020). Social media medical concept normalization using roberta in ontology enriched text similarity framework. In *Proceedings of knowledgeable NLP: The first workshop on integrating structured knowledge and neural networks for NLP* (pp. 21–26).
- Khattar, D., Goud, J.S., Gupta, M., & et al. (2019). Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference* (pp. 2915–2921). <https://doi.org/10.1145/3308558.3313552>.
- Kiela, D., Grave, E., Joulin, A., & et al. (2018). Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, (Vol. 32, no. 1, pp. 5198–5204). <https://doi.org/10.48550/arXiv.1802.02892>.
- Kirchknopf, A., Slijepećević, D., & Zeppelzauer, M. (2021). Multimodal detection of information disorder from social media. In *2021 International conference on content-based multimedia indexing (CBMI)* (pp. 1–4). IEEE. <https://doi.org/10.1109/CBMI50038.2021.9461898>.
- Kumari, R., & Ekbal, A. (2021). Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*, 184, 115412. <https://doi.org/10.1016/j.eswa.2021.115412>.
- Luo, W., Qu, Z., Pan, F., & et al. (2007). A survey of passive technology for digital image forensics. *Frontiers of Computer Science in China*, 1(2), 166–179. <https://doi.org/10.1007/s11704-007-0017-0>.
- McPherson, M., Smith-Lovin, L., & Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415–444. <https://doi.org/10.1146/annurev.soc.27.1.415>.
- Meel, P., & Vishwakarma, D.K. (2021). HAN, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences*, 567, 23–41. <https://doi.org/10.1016/j.ins.2021.03.037>.
- Nakamura, K., Levy, S., & Wang, W.Y. (2019). rf/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. arXiv:1911.03854.
- Pennycook, G., & Rand, D.G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>.
- Qi, P., Cao, J., Yang, T., & et al. (2019). Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)* (pp. 518–527). IEEE. <https://doi.org/10.1109/ICDM.2019.00062>.
- Ragusa, E., Cambria, E., Zunino, R., & et al. (2019). A survey on deep learning in image polarity detection: Balancing generalization performances and computational costs. *Electronics*, 8(7), 783–811. <https://doi.org/10.3390/electronics8070783>.
- Ragusa, E., Apicella, T., Gianoglio, C., & et al. (2022). Design and deployment of an image polarity detector with visual attention. *Cognitive Computation*, 14(1), 261–273. <https://doi.org/10.1007/s12559-021-09829-6>.
- Rajinikanth, V., Joseph Raj, A.N., Thanaraj, K.P., & et al. (2020). A customized VGG19 network with concatenation of deep and handcrafted features for brain tumor detection. *Applied Sciences*, 10(10), 3429. <https://doi.org/10.3390/app10103429>.
- Rezende, E., Ruppert, G., Carvalho, T., & et al. (2017). Malicious software classification using transfer learning of resnet-50 deep neural network. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1011–1014). <https://doi.org/10.1109/ICMLA.2017.00-19>.
- Robertson, C.T., Mourão, R.R., & Thorson, E (2020). Who uses fact-checking sites? The impact of demographics, political antecedents, and media use on fact-checking site awareness, attitudes, and behavior. *The International Journal of Press/Politics*, 25(2), 217–237. <https://doi.org/10.1177/1940161219898055>.
- Shah, P., & Kobti, Z. (2020). Multimodal fake news detection using a cultural algorithm with situational and normative knowledge. In *2020 IEEE congress on evolutionary computation (CEC)* (pp. 1–7). IEEE. <https://doi.org/10.1109/CEC48606.2020.9185643>.
- Shu, K., Sliva, A., Wang, S., & et al. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36. <https://doi.org/10.1145/3137597.3137600>.
- Siegfried, I.M. (2020). Comparative study of deep learning methods in detection face mask utilization. OSF Preprints. <https://doi.org/10.31219/osf.io/3gph4>.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

- Song, K., Zhou, L., & Wang, H. (2021). Deep coupling recurrent auto-encoder with multi-modal EEG and EOG for vigilance estimation. *Entropy*, 23(10), 1316. <https://doi.org/10.3390/e23101316>.
- Sudiatmika, I.B.K., Rahman, F., Trisno, T., & et al. (2019). Image forgery detection using error level analysis and deep learning. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 17(2), 653–659. <https://doi.org/10.12928/telkomnika.v17i2.8976>.
- Szegedy, C., Vanhoucke, V., Ioffe, S., & et al. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826). <https://doi.org/10.1109/CVPR.2016.308>.
- Szegedy, C., Ioffe, S., Vanhoucke, V., & et al. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence* (pp. 4278–4284). <https://doi.org/10.5555/3298023.3298188>.
- Thakur, R., & Rohilla, R. (2020). Recent advances in digital image manipulation detection techniques: A brief review. *Forensic Science International*, 312, 110311. <https://doi.org/10.1016/j.forsciint.2020.110311>.
- Uppada, S.K., Manasa, K., Vidhathi, B., & et al. (2022). Novel approaches to fake news and fake account detection in OSNs: User social engagement and visual content centric model. *Social Network Analysis and Mining*, 12(1), 1–19. <https://doi.org/10.1007/s13278-022-00878-9>.
- Wang, S., Zhuang, S., & Zuccon, G. (2021). Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval* (pp. 317–324). <https://doi.org/10.1145/3471158.3472233>.
- Wu, Y., Zhan, P., Zhang, Y., & et al. (2021). Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 2560–2569). <https://doi.org/10.18653/v1/2021.findings-acl.226>.
- Xia, K., Huang, J., & Wang, H. (2020). LSTM-CNN architecture for human activity recognition. *IEEE Access*, 8, 56855–56866. <https://doi.org/10.1109/ACCESS.2020.2982225>.
- Zeng, J., Zhang, Y., & Ma, X. (2021). Fake news detection for epidemic emergencies via deep correlations between text and images. *Sustainable Cities and Society*, 66, 102652. <https://doi.org/10.1016/j.scs.2020.102652>.
- Zhang, D., Xu, J., Zadorozhny, V., et al. (2022). Fake news detection based on statement conflict. *Journal of Intelligent Information Systems*, 59, 173–192. <https://doi.org/10.1007/s10844-021-00678-1>.
- Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47, 312–323. <https://doi.org/10.1016/j.bspc.2018.08.035>.
- Zhou, L., & Bian, X. (2019). Improved text sentiment classification method based on BiGRU-attention. In *Journal of physics: Conference series*. (Vol. 1345, No. 3, pp. 032097). IOP Publishing. <https://doi.org/10.1088/1742-6596/1345/3/032097>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.