Neural Computing and Applications
https://doi.org/10.1007/s00521-021-06086-4

S.I. : NCACVIP

Predicting image credibility in fake news over social media using multi-modal approach

Bhuvanesh Singh[1] · Dilip Kumar Sharma[1]

Received: 2 December 2020 / Accepted: 23 April 2021

Abstract
Social media are the main contributors to spreading fake images. Fake images are manipulated images altered through software or by other means to change the information they convey. Fake images propagated over microblogging platforms generate misrepresentation and stimulate polarization in the people. Detection of fake images shared over social platforms is extremely critical to mitigating its spread. Fake images are often associated with textual data. Hence, a multi-modal framework is employed utilizing visual and textual feature learning. However, few multi-modal frameworks are already proposed; they are further dependent on additional tasks to learn the correlation between modalities. In this paper, an efficient multi-modal approach is proposed, which detects fake images of microblogging platforms. No further additional subcomponents are required. The proposed framework utilizes explicit convolution neural network model EfficientNetB0 for images and sentence transformer for text analysis. The feature embedding from visual and text is passed through dense layers and later fused to predict fake images. To validate the effectiveness, the proposed model is tested upon a publicly available microblogging dataset, MediaEval (Twitter) and Weibo, where the accuracy prediction of 85.3% and 81.2% is observed, respectively. The model is also verified against the newly created latest Twitter dataset containing images based on India's significant events in 2020. The experimental results illustrate that the proposed model performs better than other state-of-art multi-modal frameworks.

Keywords EfficientNet · Fake images · Sentence transformer · Social media · Swish activation

1 Introduction

There is a paradigm shift in how people consume news today. They mostly look for a summarized version of news over the social media platforms to quickly gather more information [1]. This change is due to easy access to news readily available over social media platforms like Twitter and Facebook. Taking advantage of this inevitable dependency, people with malicious intent use this platform to spread fake images. Fake images are digitally manipulated images which undergo multiple altering. Morphed images are an excellent example of fake images where a person's face is replaced with another person's face. Nowadays, it is widely used to propagate a narrative or propaganda under the political arm. Norwegian Media Authority, Norway, [2] conducted a survey on fake information over coronavirus. The study's findings concluded that social media, mostly microblogging platforms, were the most significant contributor to spread false information.

Similarly, a survey conducted by CIGI-IPSOS and Internet Society [3] showed that Facebook and Twitter are the top two platforms in spreading fake news. Fake images and videos are the key material in the broadcast of fake news. Fake images gather more attention than text [4]. Some of the impacts of fake images and videos have led to grave impacts. Global tech giants like Adobe, Facebook, and Google are investing in developing artificial intelligence (AI) applications to counter the fake images and videos flooding the internet.

Figure 1a shows a digitally alerted fake image of a child with three eyes, and Fig. 1b displays the morphed image of

✉ Bhuvanesh Singh
bhuvanesh.singh_phdca18@gla.ac.in;
bhuvanesh_singh@yahoo.com

Dilip Kumar Sharma
dilip.sharma@gla.ac.in

[1] GLA University, Mathura, India

Published online: 24 May 2021

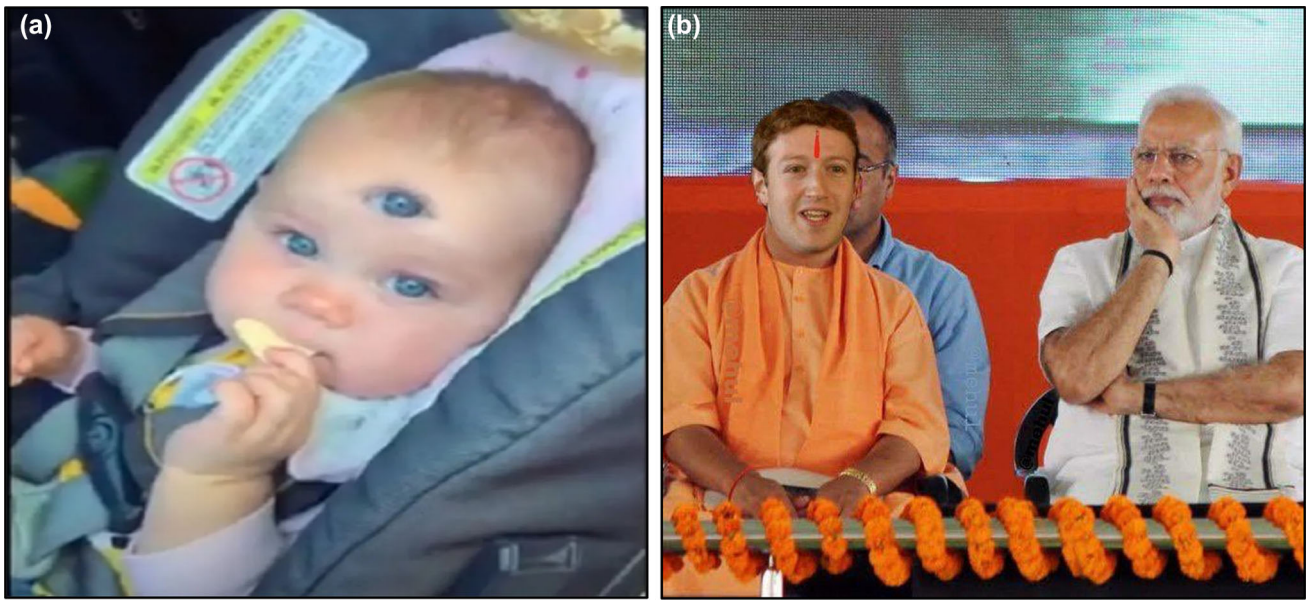

Fig. 1 Examples of fake images **a** child with three eyes—copy and move technique **b** Zuckerberg with Modi—image splicing technique (**a** and **b** are from boomlive)

Zuckerberg with prime minister Narendra Modi. These fake photos were viral on social media platforms and were circulated across the globe. Some fake images do not harm, but fake images like hurricane Sandy instilled fear among the citizens [5]. Thus, there is a need to build solutions to detect fake images over microblogging platforms.

1.1 Motivation

The utilization of fake images in fake news has increased, as its impact is more than text. There are psychological reasons that images change the way humans remember and consume information [4]. A similar observation was outlined by Adobe's 2015 state of content survey results, which showed that the interaction was three times more on the post with images than with post with just textual messages [6]. A survey conducted by the activist group Avaaz showed that Facebook causes most public health threats by sharing significant health misinformation [7]. Therefore, there is a critical need to create solutions to spot fake images over social media. The need for time is to check the proliferation of tampered images and mitigate its impact on the people.

Sharma & Sharma [8] described various methods in its detection. The conventional hand-crafted image forensic methods do not fair very well to identify manipulated images over social platforms. A multi-modal approach has recently been used, which uses multiple content and context type like text, visual, statistical, user profile, and network propagation to detect fake news. Out of these, a multi-modal framework using image and text has fared a little better than others [9, 10].

Our paper proposes a multi-modal approach that utilizes the new and upgraded models to detect fake images shared over social media platforms. Many well-known convolutional neural networks (CNN) models are available for image classification like ResNet, InceptionNet, and ImageNet. These pre-trained models are trained over millions of images. A new model EfficientNetB0 has shown better accuracy in image classification with fewer parameters and lower FLOPS than other CNN models like ResNet34, ResNet50, and InceptionNet-v2 [11]. Our model employs EfficientNetB0 for learning the inherent features of fake images. At times, in some cases of fake news, the images are authentic but out of context. Thus, text analyses are also required for fake image detection. The proposed model uses bidirectional encoder-based sentence transformer RoBERTa [12] for text analyses. Bidirectional Encoder Representations from Transformers (BERT) has been widely used in text classification. The creators of RoBERTa have proved that it has better results than BERT itself and RoBERTa tends to understand the context better [12]. Therefore, our proposed multi-modal framework utilizes EfficientNetB0 and RoBERTa for images and text analyses, respectively.

In summary, the critical points of this paper are as follows:

- Developing a practical multi-modal deep learning framework for the detection of fake images shared over social media platforms.

- The model applies error level analysis (ELA) images instead of regular images for image learning, which helps deep learning models to converge faster and have better accuracy.
- The model employs the novelty of using EfficientNet on the images and optimized Sentence transformer for text analysis within a multi-modal approach.
- Study and analyse the previous Twitter dataset changes by analysing the latest Twitter dataset containing images shared in India. [13]

The model can be used by fact-checking websites across the world to move towards automated marking of fake news, fake images for the posts shared over microblogging websites. Currently, a lot of manpower is required for doing the detection work. Secondly, as its automated more content can get generated over their websites as now only limited viral news/images are selected for fact checking. Another use case is of applying these models directly over the social media platforms in form of extensions over browser or apps in mobile.

The remainder of this paper is organized as follows. Section 2 reviews related work on detecting fake images using various techniques. Section 3 outlines the proposed model framework explaining all three components. Section 4 shares information about the datasets, experimental results, and comparative analysis with other models. Section 5 concludes and provides direction towards future work.

2 Related work

Digital alterations over images can be done in various ways. Image splicing, copy and move, resampling, and compression are majorly used techniques. There are numerous software tools available, like Photoshop, GIMP, Pixlr, and Paint.net, for altering the images.

Detection of manipulated images can be done either by hand-crafted extraction and learning forensic image features or by applying deep learning methods that learn the features by itself.

2.1 Forensic methods

For detecting copy and move tampering, the forensic approach primarily uses discrete cosine transformation (DCT) and discrete wavelet transform (DWT) coefficients [14–17]. Other novel methods like multiscale WLD histograms [18] and fractional Zernike moments (FrZms) [19] are also used. Similarly, for identifying image splicing CFA [20, 21], discrete octonion cosine transforms (DOCT) [22], and histograms gradients [23, 24] are applied. The problem with forensic techniques is that each technique is suitable for individual manipulation type. Various researches using forensic techniques resulted in high detection accuracy where a single tampering method was applied over an image for manipulation. However, when multiple tampering methods like rotating, resampling, mirroring, and compression were applied along with copy and move or image splicing over the same image, the accuracy was impacted ([17–19, 22, 24]). Fake images shared over social media platforms typically undergo multiple tampering. The quality of fake images is further deteriorated by adding noise. Thus, using forensic techniques is not an optimized option.

2.2 Single modality

Another approach is using deep learning frameworks using a single modality. Here, single content type is used to predict the fake or real classification of the information over the social platform like image, text, context, and user profile. Huang et al. [25] proposed the spatial–temporal structural neural network framework to model the message spread from temporal and spatial perspectives for rumour detection. It worked fine for rumours, but the propagation of fake images was not considered. A single SRM-CNN-based model was suggested by Rao and Ni [26] for the detection of fake images. Other hybrid CNN models were proposed later [27–29]. Mangal and Sharma [30] used the cosine similarity index between text over images and headline text to identify fake images. The model used the CNN-LSTM framework. Singh and Sharma [31] used custom CNN model with high-pass filters for fake image detection over social platforms. Johnston et al. [32] proposed a CNN model to spot and localize tampered regions in manipulated videos. The model used CNN to estimate a quantization parameter, intra/inter mode, and deblock setting of pixels patch up in videos to identify and mark the tampered regions in videos. Ghanem et al. [33] proposed using the suspicious account's semantic and stylistic features to detect the fake credibility of the news generated from these accounts. On the contrary, Vishwakarma et al. [34] proposed web scrapping and image reverse search for fake image detection. Kaliyar et al. [35] used text-based modality. Wang and Chen [36] used the information credibility model and suggested a solution that uses an online social network credibility evaluation behaviour model based totally on the SOR framework.

2.3 Multi-modal methods

Recently, research has been done using multi modalities which perform better than single modalities [9, 10]. Jin et al. [37] integrated multiple content types and suggested

solution using a recurrent neural network (RNN) having an attention mechanism for combining features of the visual, textual, and social context. Text and social context were initially combined with an LSTM network for a fused representation. The resultant representation was then bonded with image features which were mined from deep CNN. Wang et al. [38] proposed EANN [event adversarial neural networks] to detect fake news, that obtain event-invariant characteristics, and assist fake detection on newly emerged events. The architecture comprises three major modules: first, the multi-modal feature extractor, second, the fake news detector, and at last, the event discriminator. The main work of the multi-modal feature extractor is generating the visual and textual features from posts. The work of event discriminator is to eliminate event-specific features and keep event invariant features among events. Gupta et al. [39] proposed MVAE (multimodal variational autoencoder), an end-to-end network. The main task was to build an autoencoder model. The proposed model has three primary modules: encoder, decoder, and classifier module. The model uses two streams—text and visual—where their respective features are learned in the encoder component. It uses bidirectional LSTM for producing text features and VGG19 for image features. Cui et al. [40] presented a novel method SAME [sentiment-aware multi-modal embedding] incorporating users hidden opinions from users' comments into a unified deep multi-modal embedding framework for detecting forged news. Different networks are used to handle the heterogeneous data, like text, image, user profile, and publisher. In the next phase, the adversarial mechanism is adopted to learn semantically meaningful spaces per data modality. The model characterizes a unique regularization loss in the last phase to bring embeddings of relevant pairs closer. Zhou et al. [41] proposed the SAFE (similarity aware fake) framework. The model computes the probability of false reports by text and visual learnings separately. Later, it considers both these probabilities along with the calculated similarity index between the text and visual content to classify it as fake or not. Another prominent multi-modal framework proposed by other researchers is Spotfake[42].

However, in the models mentioned above, which employ text and images, there are certain drawbacks. First, they have low accuracy over social platforms datasets ([38, 39, 42]). We hypothesize this is because the deep learning model learns the main features of the image and subside manipulated features. The second drawback is that they use sub-activities like learning correlation across modalities or using sub-tasks like event discriminator and domain classifier ([38–41]). This paper suggests an explicit multi-modal approach using text and visual content. It uses two streams, each for text and image. The intrinsic features of the image and text are learnt separately and are fused for the final classification. The proposed model has better accuracy than the above stated state-of-art models. Table 1 illustrates the studies mentioned above and shows the features and techniques used, datasets, and resulting performance evaluation.

During our research work, authors have attempted to overcome the drawbacks of the problems mentioned above by using the following: first, we employ EfficientNetB0 model. EfficientNet utilizes inverted residual networks and is very optimized for image classification. Second, to inflate manipulated features in an image, we use ELA images instead of regular images. Third, to improve the text analysis, a fine-tuned sentence transformer RoBERTa is employed, which shows better results in a similar text by understanding context better. Last, the model is not dependent on any sub-activities for prediction.

3 System design

The paper proposes an efficient approach to tackle the problem of fake image detection using a multi-modal framework. Text modality is also considered to fill the gap where the image is authentic, but out of context to the news shared. The proposed model considers both text and image modalities from the social media platform and passes it to their respective feature extraction channels.

The comprehensive architecture of the recommended model is illustrated in Fig. 2. It comprises of 3 components:

- Image feature learning—It learns the intrinsic features from the fake images.
- Text feature learning—This layer learns the latent text features provided along with the fake images.
- Classifier—Softmax is used as a classifier which classifies the image using the fused features.

Presuming that we have that N training pairs then model $M = \{\mathrm{FS}_k, G_k\}_{k=1}^{N}$, the FS_k is the feature set from text and image embeddings, and G is the correct label of the data. As this is multi-modal, the features from both modalities are taken.

$$\mathrm{FS}_k = \mathrm{FS}_t + \mathrm{FS}_i \tag{1}$$

For extracting the latent features of the images, we have used the latest lightweight CNN model called EfficientNetB0 [11]. EfficientNetB0 is a highly optimized variation of CNN. In the pre-processing phase, their ELA-generated images are used despite using regular images in the dataset. ELA highlights the compression features within an image. It is noted that applying any image processing filter helps in improving the generalization ability and expedite the convergence of deep learning networks

Table 1 Summary of former studies on fake image detection

Study	Methodology	Technique	Dataset	Performance
[17]	Forensic (Copy&Move)	DWT + Surf	50 Images from MICC-F2000	Acc 95%
[18]	Forensic (Copy&Move) MWLD	Multiscale Weber's Law Descriptor	CASIA 1.0 and CASIA 2.0	Acc 92.62 and 96.52%
[19]	Forensic (Copy&Move)	FrQZMs	GRIP, FAU (Factor 1.2)	Pixel level F-measure of 0.8848 over GRIP and 0.9296 over FAU
[22]	Forensic(image splicing)	Markov features of DOCT domain	CASIA 1.0 and CASIA 2.0	Acc 98.77 and 97.59%, respectively
[43]	Forensic (image splicing)	SVD + DCT + PCA	Columbia DVMM	Acc 80.79 (No PCA) and 98.78% (PCA)
[24]	Machine learning (image splicing)	Logistic Regression using DWT + HOG + LBP	CASIA 1.0, CASIA 2.0, Columbia	Accuracy of 98.3, 99.5 & 98.8%, respectively
[25]	Graph convolutional networks	Spatial & Temporal Structures	MediaEval 15/16	Acc 75.2 and 77.3%
[34]	Web retrieval	Image reverse search	BuzzFeed, PolitiFact, and BuzzFeed election	Acc 85.3, 88.0 and 86%
[35]	Deep neural networks—text-based	Text embedding with CNN	Kaggle dataset	Acc 98.36%
[27]	Deep neural networks	Prediction error filters + CNN	Self-generated images from 12 digital cameras	Acc 98.40% (original images)
[28]	Deep neural networks	CNN with random mini-batches	CASIA-FASD, replay-attack	HTER 4.59 and 5.74 (Intra database)
[29]	Deep neural networks	Fusion network with binary classification and residual loss branch	Columbia	mAP 0.99
[38]	Multi-modal—EANN	VGG19 for Image, Text CNN for text	Twitter, Weibo	Acc 0.648 and 0.795 (EANN-)
[39]	Multi-modal—MVAE	VGG19 for Image, RNN + LSTM for text	Twitter, Weibo	Acc 0.745 & 0.824
[40]	Multi-modal—SAME	Image, text, and user profile. Text and user profile analysis using adversarial loss and image through CNN	PolitiFact, GossipCop	Micro F1 Score of 76.31 and 81.58
[41]	Multi-modal—SAFE	Text CNN for text and Image2Sentence for images. Cross-modal similarity	PolitiFact, GossipCop	Acc 8.74 and 0.838
[42]	Multi-modal—SpotFake	VGG19 for Image, BERT for text	Twitter, Weibo	Acc 0.77 and 0.892

Acc, Accuracy; AUROC area under ROC curve, F1 Score—it is harmonic mean of Precision and Recall, half total error rate (HTER)—average of FPR and FNR, *mPA*, mean average Precision the average of AP; *BERT*, bidirectional encoder representations from transformers; *CNN,* convolutional neural network; *RNN,* recurrent neural network; *LSTM*, long short-term memory

[26]. The ELA images are passed through EfficientNetB0 pre-trained model and transfer learning of EfficientNetB0 is used to generate the image embeddings from the output of its third to the last layer. The image embeddings are forwarded to two layers of the fully connected dense layers to learn the image features. The image features are represented as FS_i. After preprocessing the text, it is passed through sentence transformer RoBERTa for generating the text embeddings for learning the text features. The text embeddings are forwarded to two layers of the fully connected dense layers. The text features are represented as FS_t. After normalization, the image and text feature sets are concatenated and passed through two fully connected networks (FCN) layer. Here, feature vectors from both the modalities are learned and they are passed through the final classifier Softmax for classification. The Softmax predicts the probability of fake images. The learning of the model can be represented as below.

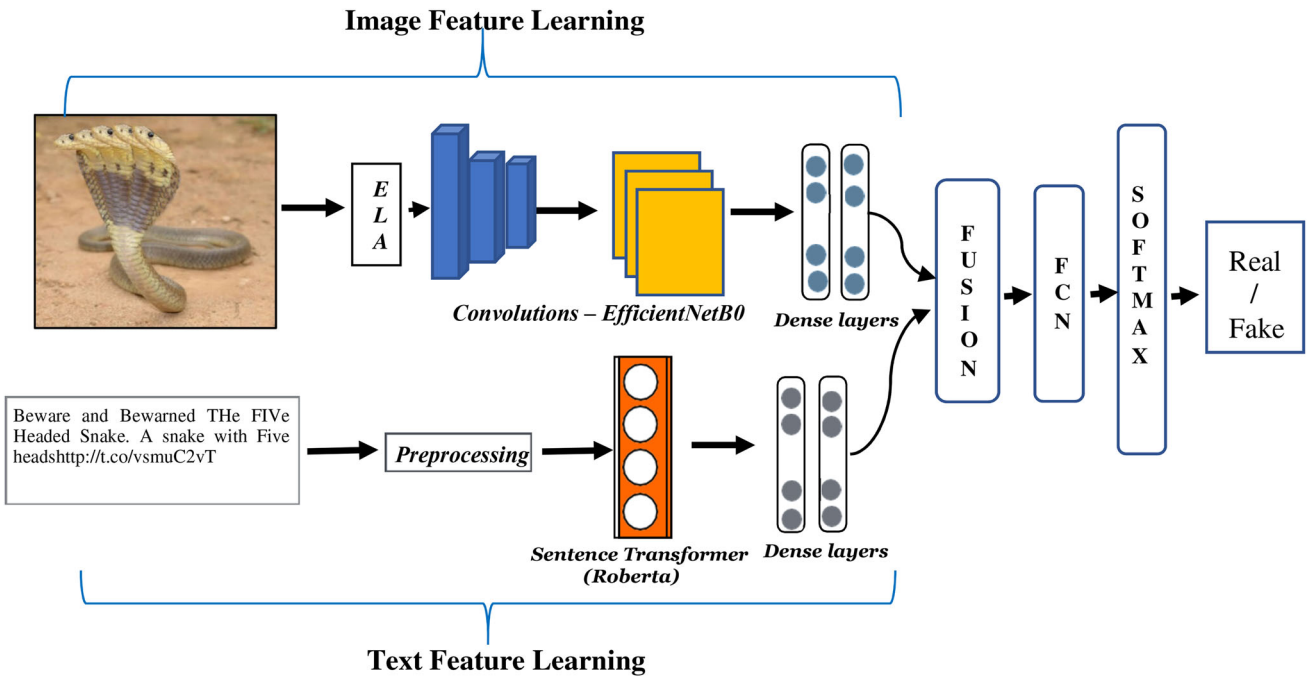

Fig. 2 Architecture of the proposed system

3.1 Image feature learning

At the pre-processing level besides resizing, all the images are passed through the error level analysis (ELA) process. ELA is a forensic method to highlight the compression differences in an image. The fundamental concept behind ELA is that if an image is tampered with and compressed, then there will not be uniformity in the compression levels within an image. A significant difference in compression levels will be observed. The ELA images prove beneficial as they subside an image's main features and bloat the manipulation features. ELA-type forensic technique supports the neural network to learn and converge faster [26]. For learning the latent features of images, EfficientNetB0, a variant of deep convolution neural network is utilized. Deep networks get saturated, and the output accuracy is at par with their shallow networks at a lot of computation cost. Hence, EfficientNet originated from the Google Brain gave the compounding scaling formula for DNN and designed EfficientNet. They verified their multiple variations of EfficientNet frameworks from EfficientNetB0 to EffiecientNetB7 and proved them to be more efficient than other well known DNN's like ResNet-152, Inception-ResNet-V2, and NASNet-A.

Their basic model EfficientNetB0 outperforms many DNN's by having better accuracy with very few parameters and FLOPS in image classification [11]. Our proposed framework experimented with multiple variations from EfficientNetB0 to EfficientNetB5. The highest accuracy was achieved on EfficientNetB0.

The key architecture of EfficientNet are:

- Swish activation—It is a multiplication of a linear and a sigmoid function. It has been proved that the Swish activation function matches or outperforms the rectified linear unit (ReLU), especially in image classification [44]. Figure 3 illustrates the comparison graph between ReLU and Swish activation functions. The swish's advantages are primarily because it is bounded below and unbounded above and is also non-monotonic. These attributes help it to outperform ReLU in deep networks and avoid dead neurons in the neural network.
- Inverted residual block (MBConv block)—These form a shortcut between the beginning and end of a convolutional block. A traditional residual block has a

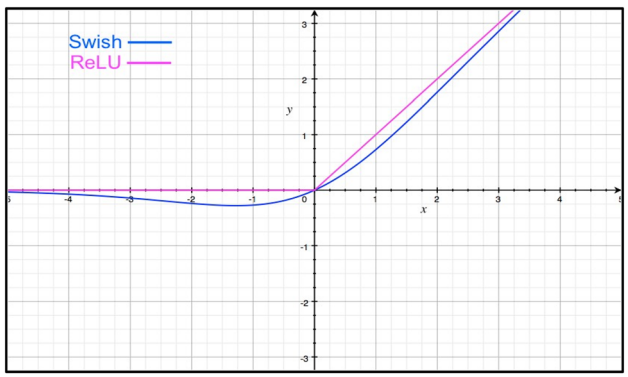

Fig. 3 Comparison of graphs between ReLU and Swish activation function [45]

wide—> narrow—> wide structure with several channels. There are a large number of channels at the input layer, which are compressed with a 1 × 1 convolution. The number of channels then increases again with a 1 × 1 convolutions so input and output can be added. In contrast, an inverted residual block follows a narrow—> wide—> narrow approach, hence the inversion. We first widen with a 1 × 1 convolution, then use a 3 × 3 depthwise convolution, then we use a 1 × 1 convolution to lower the number of channels so input and output can be added.

- Squeeze and excitation block—It is a way to give weightage to each channel instead of treating them all equally.

Figure 4 illustrates the complete architecture of EfficientNetB0. The other variations like B1 to B7 also have similar architecture but with prescribed scaling as per the formulae suggested.

To extract the image embeddings, all the preprocessed images are passed through EfficientNet-B0, and the output from the third last layer is extracted out. The output from this layer has the image feature vectors.

The latent features of images can be modelled as:

$$\mathrm{FS}_i = \varnothing(W_{\mathrm{if}}\mathrm{FSi}_{\mathrm{effb0}}) \tag{2}$$

Here, activation function is represented by Θ and W_{if} weights of the third last layer of EfficientNet-B0 and $\mathrm{FSi}_{\mathrm{effb0}}$ is the output from the previous layer.

3.2 Text feature learning

For text analysis, the text data is pre-processed where the NLP libraries are used to remove the stopwords and translate the text to English if in any other language. After the pre-processing, the text is passed to the sentence transformer RoBERTa. The usage of sentence transformer resolves the problem of vanishing and exploding gradients in RNN. RoBERTa is a fine-tuned and lighter version of the BERT-base. BERT designed and proposed by Google is an innovative self-supervised pretraining method that learns to forecast deliberately hidden (masked) sections of text. It has shown remarkable results in text classification, especially its use on Twitter tweet analyses.

RoBERTa designed by Facebook [46] uses 50 K subwords as compared to BERT's 30 k subwords. There are two main differences in RoBERTa from BERT. Firstly, it uses dynamic masking instead of static masking in BERT. Secondly, it works without NPS. The results achieved without NPS are better than that with NPS.

The sentence embeddings vectors obtained from the RoBERTa are passed through the two stacked dense fully connected layers. This is done as these features will be concatenated with the image embeddings in the next phase.

The textual feature learning can be modelled as:

$$\mathrm{FS}_t = \varnothing(W_{\mathrm{tf}}\mathrm{FSt}_{\mathrm{st}}) \tag{3}$$

Here, activation function is denoted by Θ and W_{tf} weights of the last dense layer, and $\mathrm{FSt}_{\mathrm{st}}$ is the output from the sentence transformer stacked layer.

3.3 Classification

Before the classification, we need to fuse the feature vectors obtained from the dense layers of image and text streams. The two distinct features set, i.e., FS_t * FS_i are fused into a vector of dimensionality $2p$, this can be denoted as $\mathrm{FS}_k \in \mathrm{FS}^{2p}$. Moreover, we can denote the multi-modal feature extractor as FE(IP; Θ_{fe}), where IP denotes the vectorized input data, and Θ_{fe} represents the set of parameters for the multi-modal extractor, and FE represents the overall mapping function. It is made sure that the dimensions from both the channels are in the same dimension and the batch normalization is applied. The final feature set after concatenating both the modalities is represented as below:

$\mathrm{IP}_k = \mathrm{FS}_t * \mathrm{FS}_i$ (the combination of both features sets)

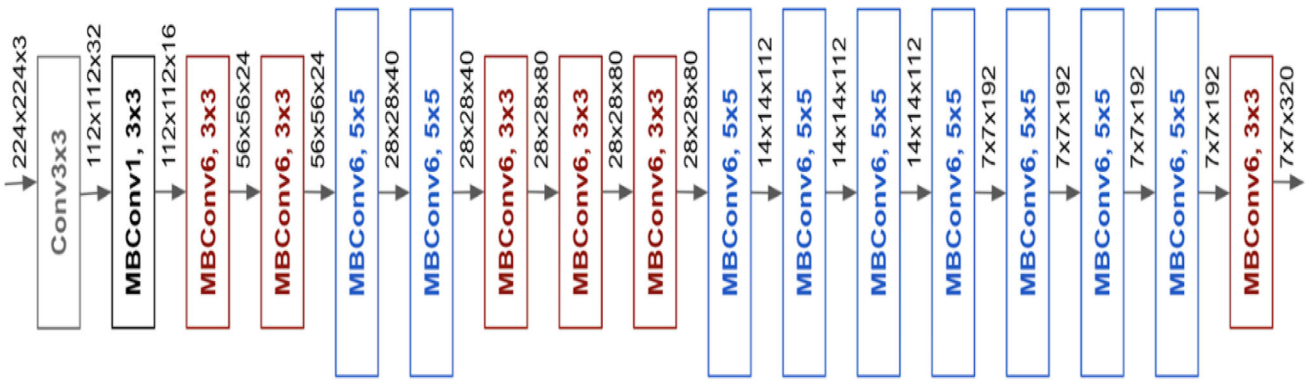

Fig. 4 The architecture for baseline network EfficientNet-B0 as provided by the authors [11]

$$FS_k = FE(IP; \varnothing_{fe}) \quad (4)$$

After fusion, two dense layers are added for learning the combined feature vectors. The activation function used in dense layers is tanh. The output from the dense layers is passed on to the Softmax layer for classification. We represent the predictor of the fake image from Softmax as $PR(FS_k; \Theta_{pr})$. Here, Θ_{pr} represents the parameter set of predictors, and PR represents the mapping function. Adam optimizer is used to optimize learning. The output from the predictor $\hat{y}$ for the multi-modal event IP^j represents the probability of the event and can be represented as:

$$\hat{y} = PR\left(FE\left(IP^j; \varnothing_{fe}\right); \varnothing_{pr}\right) \quad (5)$$

The learning loss is calculated using categorical crossentropy. The categorical crossentropy loss is calculated as below. If n = number of samples, m represents the number of categories. Hence for binary classification

$$\begin{aligned} Loss_{pr}\left(\varnothing_{fe}, \varnothing_{pr}\right) &= \sum_{i=1}^{n} y'_{i1} \log y_{i1} + y'_{i2} \log y_{i2} \\ &\quad + \ldots + y'_{im} \log y_{im} \\ &= \partial loss \,/\, \partial y_{in} = \sum_{i=1}^{n} Y'_{im}/Y_{im} \\ &= \sum_{i=1}^{n} Y'_{i2}/Y_{i2} \end{aligned}$$

For optimization of parameters $\varnothing_{fe}$ and $\varnothing_{pr}$, we need to minimize the crossentropy classification loss, which is represented as below:

$$\left(\varnothing^*_{fe}, \varnothing^*_{pr}\right) = \min_{\varnothing fe, \varnothing pr} Loss_{pr} \quad (6)$$

The summarized algorithm for the working of the proposed model is provided in *Algorithm 1*. The IS_k represents the input Set, FS_k denotes feature Set. FS_t and FS_i represent a feature set of text and images, respectively. Here, the algorithm illustrates the steps followed by the model. The text and visual features are taken, respectively, on different channels FS_{tk} and FS_{ik} for each of the dataset image and text combination. The fusion of features is optimized as per the loss until a good accuracy is achieved.

4 Experiment results and analysis

In this section, we present the experiments' results to evaluate the proposed multi-modal model's effectiveness empirically. This section covers the information about datasets, results compared with other multi-modal frameworks and our study on the latest Twitter dataset. Three evaluation metrics were considered to evaluate the experimental results: accuracy, area under the ROC (receiver operating characteristic) curve (AUC) and F-score. Accuracy measures how accurately model classifies correctly. AUC represents the degree or measure of separability. It represents how much the model can distinguish between classes. F-score is the measures of the harmonic mean of Precision and Recall.

4.1 Experimental setup

Images were resized to 300 × 300 size. The model was implemented using Keras library over google TensorFlow framework using a computer system with 32 GB RAM and Nvidia GEFORCE RTX 2080 8 GB GPU. For selecting the right combination of hyperparameters, multiple iterations were required employing different batch sizes and with different dropout probabilities to get the correct hyperparameter values. In each iteration, the number of possible combinations is reduced based on the previous iteration performance. For conducting random search and evaluating parameters in a random search, Talos library is used. Talos was developed for automated hyperparameter tuning and model evaluation of deep learning networks. The optimum results were achieved in 300 epochs having a batch size of 128. Adam optimizer was used with a learning rate of 10^{-4}. Figure 5 illustrates the plot diagram of the proposed model. We performed each experiment by randomly dividing our dataset into 75% training, 10% testing, and 15% validation subsets. The final results were obtained when the highest accuracy was reached. Accuracy metric was selected to stop the network.

4.2 Datasets

The experiment was conducted over three publicly available datasets. CASIA 2.0 [47] is image only dataset. MediaEval [48] and the Chinese Weibo [39] datasets are social media datasets consisting of images and text. MediaEval is the Twitter dataset, and Weibo is from the Weibo microblogging platform of China.

4.2.1 CASIA 2.0

The dataset has 12,616 images. There are 7492 authentic images and 5124 tampered images. The images are altered by applying copy-move and image splicing techniques. Cropping and resizing is also done while applying tampering over the images.

4.2.2 MediaEval

The dataset of social media has 193 cases of real images, 218 cases of fake images, and two altered videos. It has about 6000 rumours and 5000 non-rumour tweets from 11

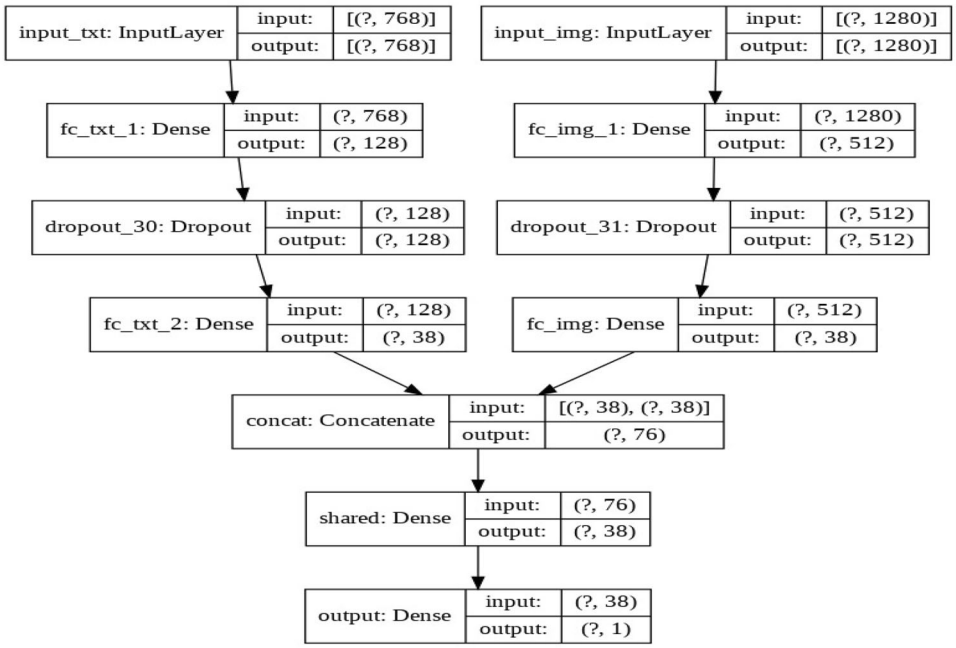

Fig. 5 Plot diagram of the proposed architecture with dimension information

events. We observed that tweets were in different languages, so we translated them using the google translate library. Though we found that specific tweets had problems translating using google trans API, those few tweets have been ignored.

4.2.3 Weibo

Weibo dataset consists of data collected from Xinhua News Agency, an authoritative news source of China, and Weibo, a Chinese microblogging website. The fake images and text collected from Weibo are collected in time ranging from 2012 to 2016. Weibo's official rumour debunking system verifies the dataset. The system encourages everyday users to report suspicious tweets on Weibo, which are then examined by a reputable committee that classifies the suspicious posts as false or real. The posts were totally in Chinese, so all were first translated to English. Some of the posts were too long for sentence transformer, and those few have been ignored.

4.2.4 Twitter Indian dataset v.01

To study the changing trends over the social platform Twitter, a new dataset is created from an Indian perspective [13]. We have collected fake, and authentic images shared over Twitter. Authors searched specifically for morphed/forged image news over fact-checking websites of India. Then, those corresponding news articles were searched over Twitter platform and images and tweets were collected from Twitter. The events covered are mainly from politics and religion arena as they are the most targeted area for fake images in India. The data has been reviewed in two phases. First, all the collected news are verified from the various well-known fact-checking websites active in India, namely boomlive, Alt news and India Today. Peer reviewers have also done manual annotations in the second round. The manual reviewers reviewed the images by going over to the Twitter platform and cross-checking them. Dataset has a total of 110 such images. 61 images are fake, and 49 are valid. All the events covered are from November 2019 to November 2020, shared over Twitter in India.

4.3 Experiment results

The initial level of the experiment was conducted for selecting the EfficientNet variation for the proposed problem. The experiment was conducted with different variations of EfficientNet from B0-B5. We got the best results with the initial model EfficientNetB0. We observed that with the limited dataset and the low resolution of the images present in the datasets ($\sim$ 300/400 pixels) the EfficientNetB0 gave better results than other variations of EfficientNet. Using more scaled variations of EfficientNet beyond B0 leads to over learning and reduced accuracy. Table 2 shows the accuracy values over various EfficientNet variations.

We conducted the first experiment over CASIA 2.0 dataset. As this dataset has only images, image channel having EfficientNetB0 was employed. The experimental results showed an accuracy of 87.13% over the CASIA

Table 2 Experiment results on different variations of EfficientNet

Dataset /Accuracy	EfficientNet B0	EfficientNet B1	EfficientNet B2	EfficientNet B3	EfficientNet B4	EfficientNet B5
MediaEval	0.853	0.793	0.778	0.795	0.753	0.692
Weibo	0.812	0.809	0.809	0.801	0.798	0.793

Table 3 Performance metrics of the proposed model

Dataset	Accuracy	F1-Score	F2-Score
CASIA 2.0	87.13%	0.87	0.877
MediaEval	85.3%	0.85	0.843
Weibo	81.2%	0.80	0.80
Indian dataset	58.3%	0.54	0.565

dataset. Over the social media datasets MediaEval and Weibo, the accuracies were 85.3 and 81.2%, respectively. Table 3 provides information on performance metrics results on all the datasets. The accuracy over CASIA dataset is more, as CASIA dataset has manipulated images with only a single manipulation type. The tampered images are manipulated with either image splicing or copy and move. Also, no noise and compressions are applied. The images in MediaEval and Weibo are images which are morphed using multiple tampering including noise and compression. Also MediaEval and Weibo have more on human faces and buildings, and CASIA has more of nature images. Figure 6 illustrates the accuracies comparison of each modality over social media datasets. We got more accuracy with images than with text modality.

Table 4 illustrates the comparison of results with other benchmarked multi-modal methods. Among the benchmarking multi-modal framework, MVAE [39] and SpotFake [42] models have good accuracy over social media datasets. Both use text and visual content like our proposed model. MVAE uses an additional component of variational autoencoder to learn the similarities between both modalities, which gives it an edge over other previous models like EANN. SpotFake, on the other hand, uses VGG19 for extracting image feature vector and BERT transformer for text feature extraction. Our results surpass the MVAE by 10.8% and SpotFake by 7.6% over MediaEval dataset. The better results are attributed to the following reasons. First, we have employed EfficientNet for images that work better than other models like VGG19 and ResNet over the smaller dataset [11]. Second, we have used error level analysis (ELA)-generated images rather than simple images. ELA images bloat the latent features of compression, which is typically applied in social media images. ELA process shows the disparities in edges of images due to different compression levels. Highlighting tampering features and subsiding the main image features leads to faster and better learning in deep learning models. On the text side, RoBERTa also supports learning the context of short tweets better than other simple LSTM or BERT used in MVAE and SpotFake. As the tweets are small texts and similar words are used, word frequency plays a vital role in this detection. Here, we have used optimized sentence transformer for text-embedding generation, making it more advantageous than other methods used in the other state-of-art models.

Over the Weibo dataset, the accuracy is on the little higher side with most models and at par with few of them.

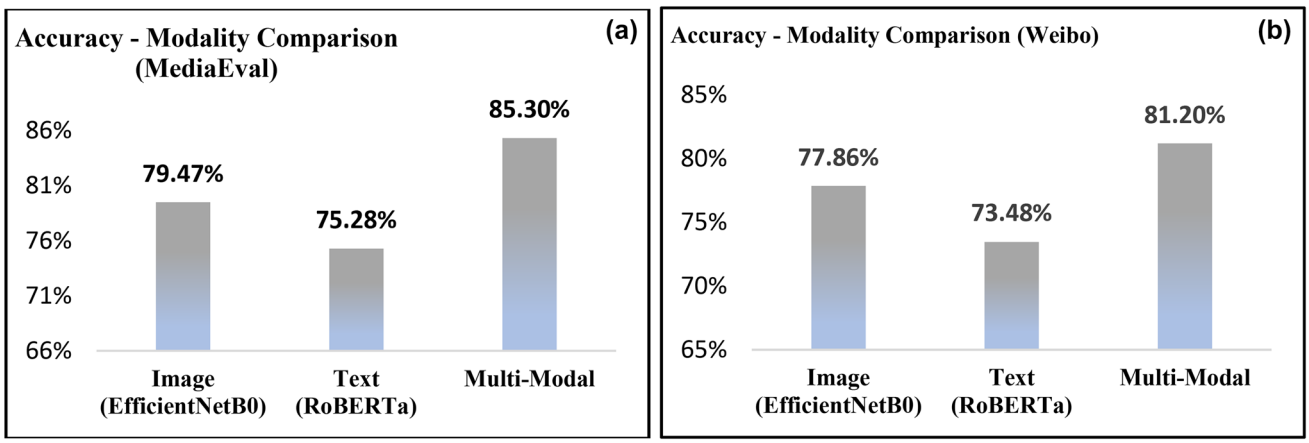

Fig. 6 Accuracies comparison of modalities **a** MediaEval **b** Weibo

Table 4 Performance comparison-proposed model to other models

Dataset	Model	Accuracy	Fake			Real		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
MediaEval (Twitter)	VQA [49]	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	att-RNN [37]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN- [38]	0.648	0.81	0.498	0.617	0.584	0.759	0.66
	SpotFake [42]	0.7777	0.751	0.9	0.82	0.832	0.606	0.701
	MVAE [39]	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	Proposed	0.853	0.821	0.943	0.877	0.913	0.745	0.820
Weibo (Chinese)	VQA [49]	0.736	0.797	0.634	0.706	0.695	0.838	0.76
	att-RNN [37]	0.772	0.797	0.713	0.692	0.684	0.84	0.754
	EANN- [38]	0.795	0.827	0.697	0.756	0.752	0.863	0.804
	SpotFake [42]	0.8923	0.902	0.964	0.932	0.847	0.656	0.739
	MVAE [39]	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	Proposed	0.812	0.851	0.784	0.816	0.744	0.826	0.782

A little lower accuracy is due to two reasons. The images in MediaEval are more related to natural disasters and natural images, while in Weibo dataset contains more image of people and human faces. Second, the translation is not very accurate due to its complexity as it is the Chinese dataset. Also, the posts were long in Weibo dataset, as compared to concise tweets on Twitter.

4.4 Error analysis

There were some observations from the wrongly detected fake images. It was observed that high-resolution images wherein only a small region was manipulated were not detected correctly. Images which were not compressed had few instances of the failed cases. Also, the images which were authentic but had more irrelevant text posts were not correctly predicted.

4.5 Experiment classification graphs analysis

Figures 7 and 8 illustrate the metrics graphs captured during the experiment's learning and validation phases. The AUROC graph shows that the area under the curve, which is 0.87, is good and supports the model's accuracy. Another important graph to observe is the Precision–Recall graph. The Precision–Recall graph provides better information than the ROC graph while evaluating the imbalanced datasets' binary classification problem. The proposed model has a higher recall value which is a positive sign in fake image classification; it is due to additional text analysis that supports the images data.

4.6 Performance over Indian dataset

Both MediaEval and Weibo datasets are old datasets about specific events occurring in the 2012–2016 timeframe. There have been changes in the usage of social media

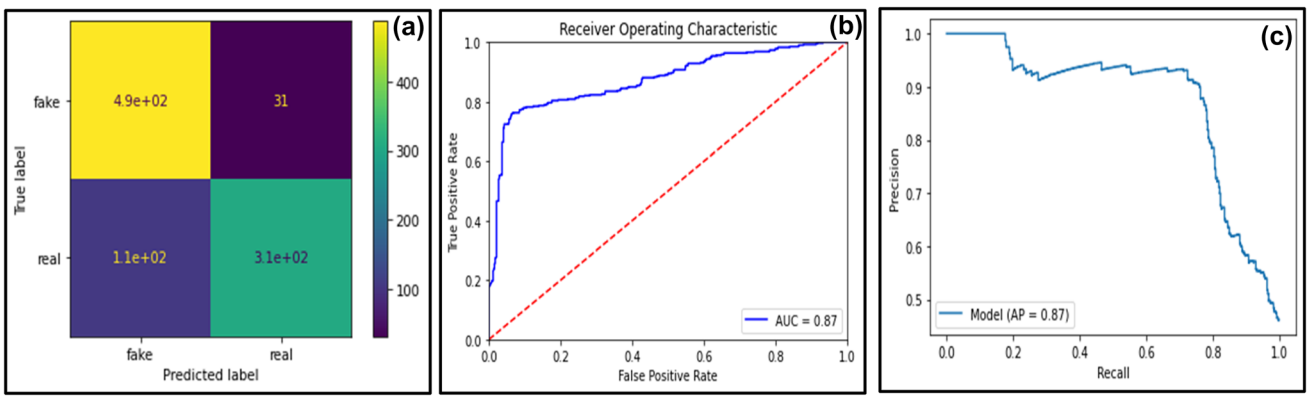

Fig. 7 MediaEval dataset—**a** confusion matrix **b** ROC curve **c** Precision–Recall graph

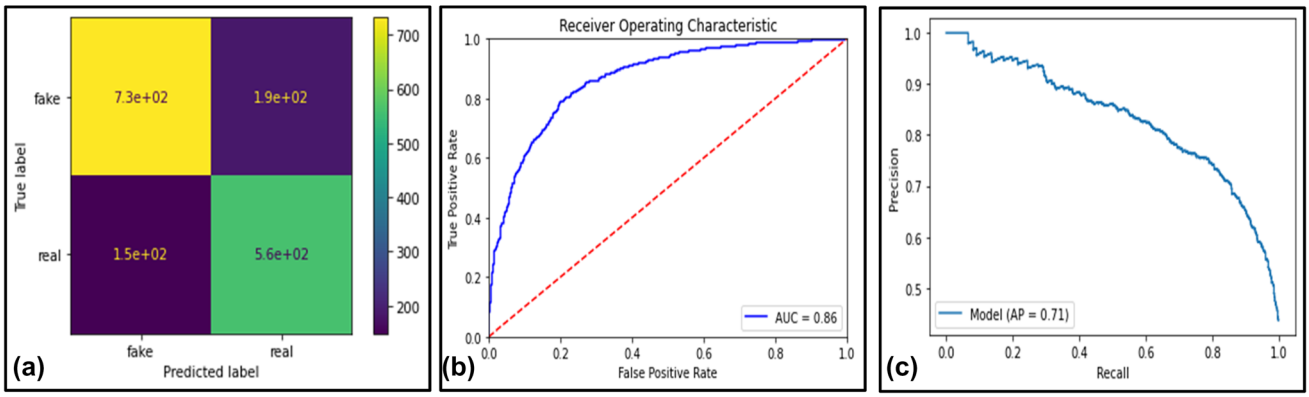

Fig. 8 Weibo dataset—**a** confusion matrix **b** ROC curve **c** Precision–Recall graph

platforms in the last few years. So, as part of our research, we created India perspective dataset "*Indian dataset v.01*" over the Twitter platform. All news events in the dataset represent events in the period from October 2019 to November 2020. This dataset comprises 110 photographs and their corresponding associated tweets. India has multiple languages and has tweeted in local languages. We have considered tweets in the English language only. The news is majorly from politics, religion, and the Bollywood arena. Figure 9 shows some of the examples from the India dataset. All three images shows the morphed images and tampered with face or poster or placard.

There are differences from previous Twitter datasets. These differences are due to three primary reasons. First, Twitter platform rules were updated for tweets. Twitter extended its 140-character limit to 280 in 2017. So, people started writing long tweets. These long textual comments impact the learning from short texts. This confirms the concern raised in the paper [42]. Second, the latest technological software available for manipulations. Due to advanced software, people can edit a minimal area of the image. Identifying small, manipulated regions has resulted in low accuracy. Third, the evolution of people's mindset in using social platforms in India. Owing to its broader reach, Twitter is currently used as a complaint forum for elected people. Therefore, several tweets were irrelevant to the image as individuals shared their grievances and complaints in tweets rather than tweeting on the associated image.

When we ran our India dataset over the same model accuracy of only 58.3% was observed. Figure 10 shows the evaluation graphs from the Indian dataset. The low accuracy is due to long textual posts and the majority of being irrelevant to the images. Secondly, as a small region of the image was tampered with, CNN models learnt main image features. This shows that with changing times, textual and image cues have changed, and models need to be continuously trained on new data to improve accuracy.

This calls for substantial new datasets of fake images that need to be created to keep up with the changing technological advances in the digital platform industry. Older databases are not going to work well with the current social media network trends. However, a recent dataset created like Fakeddit [10] its source is Reddit, a web aggregator and not a social media platform. Another new dataset "New Politifact" [50] is also not from the

Fig. 9 Examples of fake images from the Indian dataset—**a** Kamala Harris morphed photo **b** an altered Kamal Nath poster **c** a morphed placard held by CPM party people

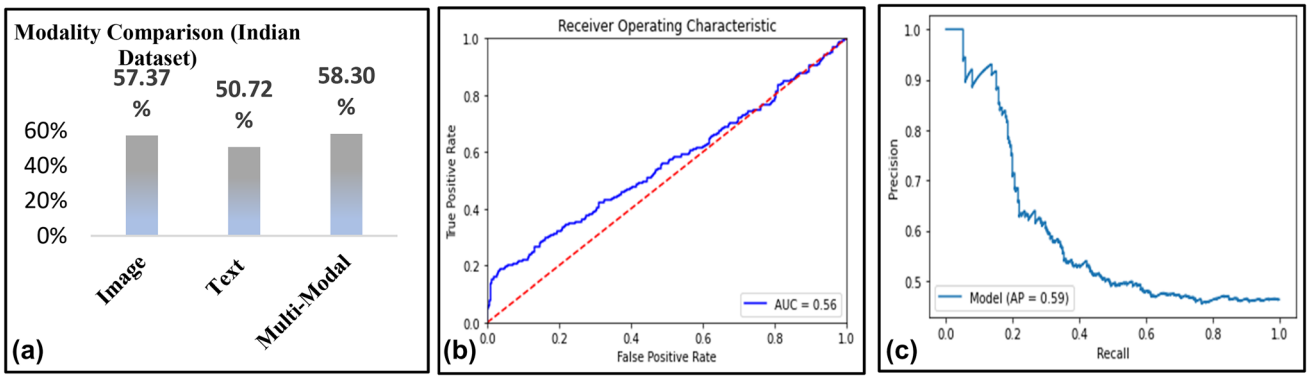

Fig. 10 Indian Dataset v.01 **a** accuracy comparison across modalities **b** ROC curve **c** Precision–Recall

microblogging platform. It was observed in the CIGI-IPSOS survey that microblogging platforms spread more fake news content than the rest of the websites [3]. Authors are working on to create new dataset solely based on Twitter images considering the latest events of 2020 and 2021.

5 Conclusion

This paper has proposed an explicit deep learning-based multi-modal approach to detect fake images shared over social media platforms. Forensic methods have their limitations. In the proposed model, the visual and textual modalities are learned on respective channels and later fused to get the feature sets from both modalities. There are no additional components required for understanding the correlation between modalities. The model uses EfficientNet-B0 and sentence transformer RoBERTa for extracting the features of images and text, respectively. The ELA-generated images are used as input to the CNN model. ELA images further support better learning of image manipulations. The EfficientNetB0 has been verified against CASIA2.0 dataset as well, which is dataset comprising of tampered images. An efficiency of 87.13% is recorded over CASIA 2.0. The experiment has been tested against Twitter and Weibo datasets. Accuracy of 85.3% and 81.2% is achieved. These results surpass other previous state-of-art models. The research proves that a multi-modal deep learning model can detect fake images over social media platforms. We further created a new Twitter dataset using the latest 2020 events from an Indian perspective. The observation was that currently, there are many changes in image and textual cues from the previous dataset, which lowers the accuracy of models trained over old data. This indicates dire need to create social media images dataset based on the latest trends to keep up with the microblogging industry's changing trends.

The detection of satire images is not covered. Also, the proposed solution is not verified against fake images generated through generative adversarial networks. The text written over the images is also not considered. This will be taken up as a further part of the research.

Declarations

Conflict of interest The authors declare that there is no actual or potential conflict of interest in relation to this paper. 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.

References

1. Sharma, S., Sharma, D.K. (2020). Fake News Detection: A long way to go. In: 4th International Conference on Information Systems and Computer Networks (ISCON)2019. pp 816–821, https://doi.org/10.1109/ISCON47742.2019.9036221.
2. Stoll, J. (2020) Reading fake news about the Coronavirus in Norway 2020. https://www.statista.com/statistics/1108710/reading-fake-news-about-the-coronavirus-in-norway-by-source/ (2020). Accessed on May 2, 2020
3. CIGI-Ipsos Global Survey on Internet Security and Trust (2019). https://www.cigionline.org/internet-survey-2019. Accessed on 15 January 2021
4. Fazio, L. (Feb 2020). Curbing fake news: Here's why visuals are the most potent form of misinformation from https://scroll.in/article/953395/curbing-fake-news-heres-why-visuals-are-the-most-potent-form-of-misinformation last accessed on 2021/1/2
5. Eveleth R (2014) Hurricane Sandy: Five ways to spot a fake photograph. https://www.bbc.com/future/article/20121031-how-to-spot-a-fake-sandy-photo. Accessed on August 9, 2020
6. Adobe blog https://blogs.adobe.com/creative/files/2015/12/Adobe-State-of-Content-Report.pdf
7. McCarthy N (2020) Report: Facebook Poses A Major Threat To Public Health. https://www.statista.com/chart/22660/health-misinformation-on-facebook/ Accessed on September 3, 2020
8. Sharma S, Sharma DK (2020) Comment filtering based explainable fake news detection. In: 2nd International conference on computing, communication and cyber-security (IC4S) (2020)

9. Singh V, Ghosh I, Sonagara D (2020) Detecting fake news stories via multimodal analysis. J Assoc Inf Sci Technol 72(1):3–17. https://doi.org/10.1002/asi.24359
10. Nakamura, K., Levy, S., Wang, W.Y. (2020) Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In: Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020), pages 6149–6157. https://www.aclweb.org/anthology/2020.lrec-1.755
11. Tan M, Le Q V (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6105–6114. http://proceedings.mlr.press/v97/tan19a.html
12. Liu, Yinhan & Ott, Myle & Goyal, Naman & Du, Jingfei & Joshi, Mandar & Chen, Danqi & Levy, Omer & Lewis, Mike & Zettlemoyer, Luke & Stoyanov, Veselin. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.
13. https://github.com/bhuvaneshsingh80/Indian-Dataset-github link for Indian Dataset
14. Li, G., Wu, Q., Tu, D., & Sun, S. (2007). A sorted neighborhood approach for detecting duplicated regions in image forgeries based on DWT and SVD, In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME '07), IEEE, Beijing, China. pp 1750–1753
15. Mahmood T, Nawaz T, Irtaza A, Ashraf R, Shah M, Mahmood MT (2016) Copy-move forgery detection technique for forensic analysis in digital images. Hindawi Publ Corp Math Probl Eng 8713202:13
16. Jwaid, M.F., & Baraskar, T.N. (2017) Study and analysis of copy-move & splicing image forgery detection techniques. In: International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC) pp 697–702
17. Alamro L, Nooraini Y (2017) Copy-move forgery detection using integrated DWT and SURF. J Telecommun Electro Comput Eng (JTEC) 2017:67–71
18. Hussain M, Qasem S, Bebis G, Muhammad G, Aboalsamh H, Mathkour H (2015) Evaluation of image forgery detection using multiscale weber local descriptors. Int J Artif Intell Tools 24(4):1540016
19. Chen B, Yu M, Su Q, Shim HJ, Shi Y (2018) Fractional quaternion Zernike moments for robust color image copy-move forgery detection. IEEE Access 2018:56637–56646
20. Popescu AC, Farid H (2005) Exposing digital forgeries in color filter array interpolated images. IEEE Trans Signal Process 53(10):3948–3959
21. Ferrara P, Bianchi T, Rosa AD, Piva A (2012) Image forgery localization via fine-grained analysis of CFA artifacts. IEEE Trans Inf Forensics Secur 7(5):1566–1577
22. Sheng H, Shen X, Lyu Y, Shi Z, Ma S (2018) Image splicing detection based on Markov features in discrete octonion cosine transform domain. IET Image Proc 12(10):1815–1823
23. Mazumdar, A., Bora, P K. (2016) Exposing splicing forgeries in digital images through dichromatic plane histogram discrepancies. In: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing 62, pp 1–8
24. Jaiswal AK, Srivastava R (2020) A technique for image splicing detection using a hybrid feature set. Multimed Tools 17:11837–11860
25. Huang Q, Zhou C, Wu J, Liu L (2020) Wang B Deep spatial-temporal structure learning for rumor detection on Twitter. Neural Comput Appl. https://doi.org/10.1007/s00521-020-05236-4
26. Rao, Y., Ni, J. (2016) A deep learning approach to detection of splicing and copy-move forgeries in images. In: IEEE International Workshop on Information Forensics and Security
27. Bayar, B., & Stamminger, M.C. (2016) A deep learning approach to universal image manipulation detection using a new convolutional layer. IH&MMSec, In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp 5–10
28. Rehman YAU, Po LM, Liu M (2018) LiveNet: Improving features generalization for face liveness detection using convolution neural networks. Expert Syst Appl 108:159–169
29. Liu B, Pun C-M (2020) Exposing Splicing Forgery in Realistic Scenes Using Deep Fusion Network. Inf Sci 2020(526):133–150
30. Mangal, D., & Sharma, D.K. (2020) Fake News Detection with Integration of Embedded Text Cues and Image Features. In: 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (2020), pp 68–72
31. Singh, B., Sharma, D.K. (2021) Image Forgery Over Social Media Platforms - A Deep learning Approach for Its Detection and Localization. In: 15th International Conference on computing for Sustainable Global development—Bharti Vidyapeeth, Delhi
32. Johnston P, Elyan E, Jayne C (2019) Video tampering localization using features learned from authentic Content. Neural Comput Appl 32:12243–12257
33. Ghanem B, Ponzetto S P & Rosso P (2020). FacTweet: Profiling Fake News Twitter Accounts. In: International Conference on Statistical Language and Speech Processing, pp 35–45
34. Vishwakarma DK, Varshney D, Yadav A (2019) Detection and veracity analysis of fake news via scrapping and authenticating the web search. Cognit Syst Res 58:217–229
35. Kaliyar RK, Goswami A, Narang P, Sinha S (2020) FNDNet—A deep convolutional neural network for fake news detection. Cogn Syst Res 61:32–44
36. Wang D, Chen Y (2019) A neural computing approach to the construction of information credibility assessments for online social networks. Neural Comput Appl 31(Suppl 1):S259–S275
37. Jin Z, Cao J, Guo H, Zhang Y, Luo J (2018) Multi-modal fusion with recurrent neural networks for rumor detection on microblogs. ACM on Multimed Conf 2017:795
38. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha K., Su, L., & Gao, J. (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 (KDD '18), pp. 849–857. https://doi.org/10.1145/3219819.3219903.
39. Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal Variational Autoencoder for Fake News Detection. The World Wide Web Conference pp 2915–2921. https://doi.org/10.1145/3308558.3313552
40. Cui, L., Wang, S., & Lee, D. (2019). SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 41–48). https://doi.org/10.1145/3341161.3342894
41. Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-Aware Multi-Modal Fake News Detection. (eds) Advances in Knowledge Discovery and Data Mining, 354–367. https://doi.org/10.1007/978-3-030-47436-2_27
42. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A Multi-modal Framework for Fake News Detection. In: IEEE Fifth International Conference on Multimedia Big Data (BigMM), 39-47. https://doi.org/10.1109/BigMM.2019.00-44
43. Moghaddasi, Z., Jalab, H.A., & Noor, R.M. (2017). Image splicing detection using singular value decomposition. In: Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing,140, pp 1–5. https://doi.org/10.1145/3018896.3036383
44. Ramachandran P, Barret Z, Quoc L(2017) Swish: a Self-Gated Activation Function. Google Brain https://arxiv.org/abs/1710.05941

45. Sharma, J. (21 Oct 2017) Experiments with SWISH activation function on MNIST dataset. https://medium.com/@jaiyamsharma/experiments-with-swish-activation-function-on-mnist-dataset-fc89a8c79ff7. Accessed on 10 February 2021
46. Target E (30 July 2019) Facebook Says its New AI Training Recipe Upgrades Google's Natural Language Processing System. https://www.cbronline.com/news/roberta-facebook-nlp#:~:text=Facebook Says its New AI Training Recipe Upgrades Google's Natural Language Processing System&text=Facebook has created and published, benchmark leaderboard Facebook said today. Accessed on 10 February 2021
47. Dong, J., Wang, W., Tan, T. (2013). CASIA Image Tampering Detection Evaluation Database. In: Proceedings IEEE China Summit and International Conference on Signal and Information Processing pp 422-426. https://doi.org/10.1109/ChinaSIP.2013.6625374
48. Boididou C, Papadopoulos S, Dang-Nguyen DT, Boato G, Riegler M, Middleton SE, Petlund A, Kompatsiaris Y (2015) Verifying multimedia use at MediaEval 2015. MediaEval 3(3):7
49. Agrawal A, Lu J, Antol A, Mitchell M, Zitnick CL, Parikh D, Batra D (2017) Vqa: Visual question answering. Int J Comput Vision 123(1):4–31
50. Garg, S., Sharma, D.K. (2020) New Politifact: A Dataset for Counterfeit News. In: 9th International Conference System Modeling and Advancement in Research Trends (SMART) 2020, pp 17–22, https://doi.org/10.1109/SMART50582.2020.9337152.