International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I7P114

Volume 73 Issue 7, 164-178, July 2025
© 2025 Seventh Sense Research Group®

Original Article

Detecting Fake News in Social Media in Early Publishing Stages: Deep Fake Detect

Vedpriya Dongre[1], Pragya Shukla[2]

[1,2]Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore, India.

[1]Corresponding Author : vdongre@ietdavv.edu.in

Received: 27 March 2025 Revised: 07 June 2025 Accepted: 16 June 2025 Published: 30 July 2025

Abstract - *Social media's fake and manipulated content may disturb social harmony and peace. Therefore, identifying and preventing harmful content posting on social media is an essential but complex task. Social media may circulate hateful and fake information as a result of the publication of harmful content. This paper presents a deep learning model using a multi-model feature fusion technique to deal with the harmful content flooding in social media. The proposed model includes the InceptionV3 for deep image feature extraction, and the GloVe pre-trained model has been used to capture the content and contextual features. Next, the extracted features are fused using the concatenation layer, and a stack of dense layers is used with different filter sizes to learn and classify the harmful content to prevent posting on social media. The experiments have been done, and parameters have been tuned to increase the detection capabilities. The model can provide 65% correct recognition of harmful content.*

Keywords - *Data mining, Deep Learning, Machine Learning, Multi-model feature fusion, Transfer learning, Deep feature learning.*

1. Introduction

In this age of digital media and social networking, a large number of people are influenced by social media news [1]. Every time you check your social media, you can get new and fresh content. Therefore, everyone is kept attracted to social media news [2]. Social media has recently emerged as a powerful force in shaping and dismantling narratives, evident in cases like the Ukraine-Russia War, the Hamas-Israel conflict, and the Bangladesh regime change debate [3]. During these incidents, a significant amount of content on social media was flooded. However, some of the content is real, some of the content is artificially created (AI-generated images), and some of the content is old. Therefore, social media content also contains fake news, fabricated narratives, and hateful content [4]. This kind of content is fake, but it can change public opinion. Therefore, social media can be a tool for modern warfare, running propaganda and distributing fake and hate content [5]. Fake content can raise a geopolitical agenda against the ground truth or disturb peace and harmony in society. In this context, to control misinformation, it is necessary to identify fake and hateful social media posts and control public chaos [6]. However, several research efforts have been made by researchers and engineers to counter fake news on social media. Recent fake news detection methods have contributed by using machine learning techniques [7]. That considers the problem of fake news detection as a text classification problem. However, these methods have become less effective due to images. In this context, some methods are available based on image classification. Next, the multi-model fusion-based approaches are utilized for improving fake and hateful content detection accuracy [8]. But, images with text classification using multi-model fusion classification are different from meme classification. Meme-based content is an indirect method of running false narratives and propaganda [9]. But, the meme content classification techniques are less accurate than other forms of image and text format of data classification. The meme contains indirect semantics and meaning [10]. Therefore, the meme image classification is more complex than normal hate-based image classification. This paper presents a model of deep learning for reducing the spread of false information and offensive content across social media networks. The model is aimed at accepting social media content as input and classifying the data into suspicions. Next, based on the Text and Image, the suspected data is again classified to determine whether the social media post is fake. The model employs pre-trained language representations such as GloVe and BERT (Bidirectional Encoder Representations from Transformers) for text classification tasks, and for dealing with the images, the VGG16 and Inception-3 have been used. Further, the results and the conclusion of the work are discussed. Finally, some policies are discussed to implement for a healthier social media environment.

2. Literature Survey

Social media enables quicker news consumption with minimal filtration. This, in turn, results in faster distribution to a wider audience. This can lead to societal drawbacks such as user deception and opinion manipulation [11]. Creative and sophisticated methods using provocative texts and inviting imagery make fake news detection an extremely arduous [12]. Current fake news exploration methods utilize screening original news articles and user responses. Fake news, when compared with conventional news, often contains thought-provoking images for context. This presents a challenging scenario for detecting fake news in a multi-channel environment [13].

Recent studies have shown that various techniques for fake news filtration make use of sequential neural networks to incorporate societal context and news topics [14]. The analysis of text sequences was initially conducted in a unidirectional manner. Therefore, adopting a bidirectional training approach becomes crucial for effectively capturing the contextual information in fake news, ultimately enhancing classification performance and improving the understanding of sentence-level dependencies. Since fake news is presented in an approach matching the original, AI authentication often becomes challenging without proper context and information. E. Amour et al [15] have shed light on fake news research and current methods of detection and prevention of fake news. They showcase the problems in research and active challenges, discuss current pathways, and highlight directions of future research.

B. Hu et al [16] survey on fake news detection suggests summarizing the three essential qualities by studying its distribution process: purpose creation, irregular transmission, and unconventional response. The other focuses on why there is a need to publish and propagate fake news. Lastly, what are the different user points of view regarding fake news? This study discusses detection approaches, trends, and future research directions. J. Alghamdi et al [17] review to recognize and battle fake news focuses on fake news definition and its related terms. It also discusses upcoming and current ML and DL techniques, which focus on three subsections: content, context, and hybrid features. Moreover, it highlights the qualities of fake news, datasets, and methodologies. It also recognizes future investigation requirements and challenges.

R. K. Kaliyar et al [14] showcase the FakeBERT approach based on a BERT-powered deep learning model, which integrates multiple parallel branches of a single-layer deep CNN using unconventional filter sizes and granularities. This can aid in resolving obscurity, which remains the primary challenge in language learning. Outcomes highlight FakeBERT's efficiency of 98.90% compared to current approaches. S. Ni et al [18] focus on solving fake news detection issues in a real scenario. The model relies solely on source tweets and retweet user data, utilizing a neural network combined with Multi-View Attention Networks (MVAN) for effective fake news detection. To ensure clue capture, the system utilizes attention to propagation structure and text context. These enable clue capture in texts and suspicious users. Finding highlights that this system outperforms others by 2.5% in accuracy and offers logical reasoning. R. K. Kaliyar et al [19] use a two-pronged approach focusing on news and echo chambers. News fusion with tensor utilizes a couple of matrix-tensor factorization methods for inactive depiction of content and context. Multiple filters with opt-outs are used to categorize information separately and compound. This model has acquired a validation accuracy of 92.30%. A multi-channel method proposal by B. Singh et al [20] detects fake images. The sentence transformer is used for text analysis, and for images, the CNN model EfficientNetBO is used. Dense layer passing and fusion of features rooted in visual imagery and text are used to predict fake images. To verify the efficiency, testing occurs on a real dataset, i.e., MediaEval and Weibo. A predicted accuracy of 85.3% and 81.2%, respectively, is noticed. Validation is also performed against the newer Twitter dataset.

A. Giachanou et al [21] suggest that a multi-channel, multi-image system is employed for fake news detection, integrating textual, visual, and semantic information. This approach leverages BERT for textual representation and VGG-16 for visual feature extraction, enabling the fusion of diverse data sources for improved accuracy. Text-image similarities based on cosine similarities between the title and image tags are referred to. Results suggest feature combination is an efficient method for fake news detection. T. Zhang et al [22] suggest BERT-Based Adaptation of Neural Network Domain (BDANN) for multi-channel fake news detection. It consists of three branches: multimodal feature extraction, domain classification, and fake news detection. Like above, BERT and VGG-16 models are used for text and image feature extraction, respectively. The features are integrated and supplied to the detector. Twitter and Weibo datasets were used for performance verification.

B. Hu et al [23] inquire into the LLMs in fake news detection. Studies have been performed to sort out an LLM to unearth fake news and offer reasoning for multiple points of view. The proposal also suggests that existing LLMs may not replace fine-tuned SLMs in fake news detection but rather can be a supplement by providing reasoning. An Adaptive Rationale Guidance (ARG) network has also been designed. ARG-D, a rationale-free version of ARG, has also been derived, and it works without questioning LLMs. Research on two separate datasets exhibits better performance of ARG and ARG-D as compared to baseline processes, namely, SLM-based, LLM-based, and integrated. To bridge the gap between news semantic features and the decision-making space, L. Peng et al. [24] propose Contextual Semantic Representation Learning for Multimodal Fake News Detection (CSFND). Their approach incorporates unattended context learning to capture local contextual features, which are then integrated with semantic features to enhance the understanding of contextual semantic characterization.

Table 1. Literature summary

Ref.	Overview	Type	Used Features	Method	Results
[14]	BERT-based deep learning approach (FakeBERT).	Research	NA	multiple parallel bits of single- layer deep CNN	FakeBERT efficiency of 98.90%.
[15]	Shed light on fake news research and current methods of detection and prevention of fake news.	Review	NA	Showcase the problems in research and challenges.	Discuss pathways, and highlight directions of future research.
[18]	Fake news detection. It enables clue capture in texts and suspicious users.	Research	Uses source tweets and retweet users with a for fake news detection.	neural network and Multi-View Attention Networks (MVAN)	This system outperforms others by 2.5% in accuracy
[19]	Two approaches for the news and echo chambers.	Research	Utilizes a couple of matrix-tensor factorization methods.	Multiple filters with opt-outs are used.	Model acquired a accuracy of 92.30%.
[20]	A multi-channel method to detect fake images.	Research	Dataset, i.e., MediaEval and Weibo. Validation on Twitter dataset.	For text, the sentence transformer is used, and for images, CNN model EfficientNetBO is used.	Accuracy of 85.3% and 81.2%.
[21]	Suggest a multi-channel, multiple image system that merges information from different channels for fake news detection.	Research	Textual, visual, and semantic information are integrated.	BERT and VGG-16 are used for textual and visual features. Similarities based on cosine are used.	Results suggest feature combination is an efficient method for fake news detection.
[22]	BERT-based Adaptation of Neural Network (BDANN) for fake news detection.	Research	Twitter and Weibo datasets were used for performance verification.	BERT and VGG-16 are used for text and image feature extraction.	
[23]	Inquire into the possibilities of LLMs in fake news detection.	Research		An Adaptive Rationale Guidance (ARG) and ARG-D to works without questioning LLMs.	Better performance of ARG and ARG-D as compared to baseline processes.
[16]	Survey on fake news detection	Survey	Why is a need to publish and propagate fake news, and what are the different user points of view?	Summarize using the sharing process, purpose, transmission, and response.	Discusses detection approaches, trends, and future research directions.
[17]	A review to recognize and battle fake news focuses on the fake news definition and its related terms.	Survey	Discusses upcoming and current ML and DL methods, focusing on: content, context, and hybrid features.	Highlights qualities of fake news, datasets, and methodologies.	Recognizes future investigation requirements and challenges.
[24]	For bridging the gap between news semantic features and decision space.	Research	Unattended context learning and integration with semantic features to understand context.	Contextual testing strategy.	CSFND surpasses ten state-of-the-art challengers on two multimodal datasets.
[25]	Utilization of style-	Research	SheepDog has been	Acquires resistance	

	related features is also stressed for style-based attacks.		introduced, which is a content-based style fake news detector.	using: 1) news rephrasing to blend with various styles; 2) Consistent accuracy through style-critical method; 3) Content-focused.	
[26]	Fake news detection by News Semantic Environment Perception (NSEP).	Research	1) Time-bound semantic environment intervals; 2) Semantic inconsistency perception through graph CNN; 3) Veracity confirmation.		Highlighted accuracy of 86.8% for the Chinese dataset, which was 14.1% higher.
[27]	ML and DL methods were combined with FastText word embeddings.	Research	WELFake, FakeNewsNet, and FakeNewsPrediction.	CNNs and LSTM, used with FastText embeddings, were used to create a hybrid model	with accuracy and F1-scores of 0.99, 0.97, and 0.99.
[28]	Multilingual Fake News Detection (MFND) is proposed.	Research	It capitalizes on an encapsulation strategy to evoke meaningful content from the news.	Critical information is protected with length reduction by feeding the data into mBERT.	This approach has proved to be superior in assessments.

Semantically differing fake news is singled out and differentiated. In addition, a contextual testing strategy for differentiation between real and fake news, which have similar semantics, has also been conceived. CSFND surpasses ten state-of-the-art challengers when studies conducted on two multimodal datasets are analysed. Utilization of style-related features is also stressed for style-based attacks, as per J. Wu et al [25]. Ill-natured players have been enabled by LLMs to imitate reliable news source styles. A study shows that veiled LLM fake news content sabotages the potency of text-dependent detectors. To overcome this issue, SheepDog has been introduced, which is a content-based style fake news detector. It acquires resistance through the following methods: 1) Customized LLM-enabled news rephrasing to blend with various styles; 2) Consistent accuracy forecasting through a style-critical training method; 3) Content-focused instructions for discrediting fake news through content-centric reliability assignments. A fake news detection framework by News Semantic Environment Perception (NSEP) was suggested by Fang et al. [26]. This involves three major processes: 1) Time-bound semantic environment intervals; 2) Semantic inconsistency perception through graph convolutional networks; 3) Veracity confirmation by presenting evidence of semantic paradoxes between news content and posts. Observations on Chinese and English datasets highlighted that an accuracy value of 86.8% was achieved for the Chinese dataset, which was 14.1% higher. E. Hashmi et al [27] have presented a method that employs three datasets: WELFake, FakeNewsNet, and FakeNewsPrediction. Multiple ML and DL methods were combined with FastText word embeddings, which improved and rectified the algorithms. CNNs and LSTM, supplemented with FastText embeddings, were used to create a hybrid model that exceeded other methods in classification performance, showcasing accuracy and F1-scores of 0.99, 0.97, and 0.99. Transformer-based models, which exceeded traditional RNN-based frameworks, were also employed. Lastly, justifiable AI was implemented utilizing Local Interpretable Model- Agnostic Explanations, and Latent Dirichlet Allocation. Multilingual Fake News Detection (MFND) is proposed by J. Alghamdi et al [28]. It capitalizes on an encapsulation strategy to evoke meaningful content from the news. Critical information is protected while achieving length reduction. This is followed by feeding the data into mBERT for categorization. This approach has proved to be superior in assessments.

2.1. Existing Work and Research Gap

The existing works involve deep learning techniques for classifying fake news on social media platforms. More recent approaches are also concentrated on a multi-model fusion approach for more detection. But most of the articles are classifying Twitter, Facebook and Weibo datasets. This dataset contains the problem of direct fake news identification by comparing it to older versions of the image and by using AI-based generated similar versions. However, these methods are not performing well when using indirect methods of fake news distribution, such as meme images. Therefore, in this paper, a method is proposed to deal with fake news using meme image classification. The proposed technique can be directly applied to a social media news feed.

3. Proposed System

Fake news detection in social media is a complex task due to the inclusion of visual and textual data. Additionally, the meme content includes indirect semantics. Therefore, the traditional models are becoming less effective. However, many different models are also available for fake news detection. Most of the models are based on multi-model feature fusion-based deep learning approaches. The main issue with these models is learning from both, i.e., image context and associated text with the image. Therefore, both visual and textual features of the images are used for learning.

In order to learn with both types of features, it is necessary to incorporate both types of features before passing them to a model for training. This technique can improve the classification results. This paper introduces a system to detect fake and harmful content from social media. This model detects and eliminates harmful content before it goes viral on social media. The flow of the proposed system is demonstrated in Figure 1.

This diagram includes the key components of the proposed system. The system accepts the content in image format. These images contain the visuals and text. Because on social media, the images with text are the popular post type. The input image passes through the Optical Character Recognition (OCR) system. The OCR extract the text from the input image. The visual and text are obtained by using an OCR system. This content is used for further processing. Next, the text and image feature extraction techniques can be used to identify the image and text features. The extracted features (both the information text and visuals) are fused to combine the features into one. This process is called multi-model feature fusion. The system is operated in two modes.

- Mode 1: The Model uses the fused features and classes to perform the training of the deep neural network. After training, the model is preserved. This preserved model is termed the trained model.
- Mode 2: This Model utilizes the pre-trained models to extract and produce fused features. The model predicts the class labels (i.e. Offensive and non-Offensive).

If the predicted class indicates the harmful or Offensive content, then the system prevents the post from being published. Additionally, the system notifies the user to review the social media post. Otherwise, if it is not harmful, the post is published on social media.

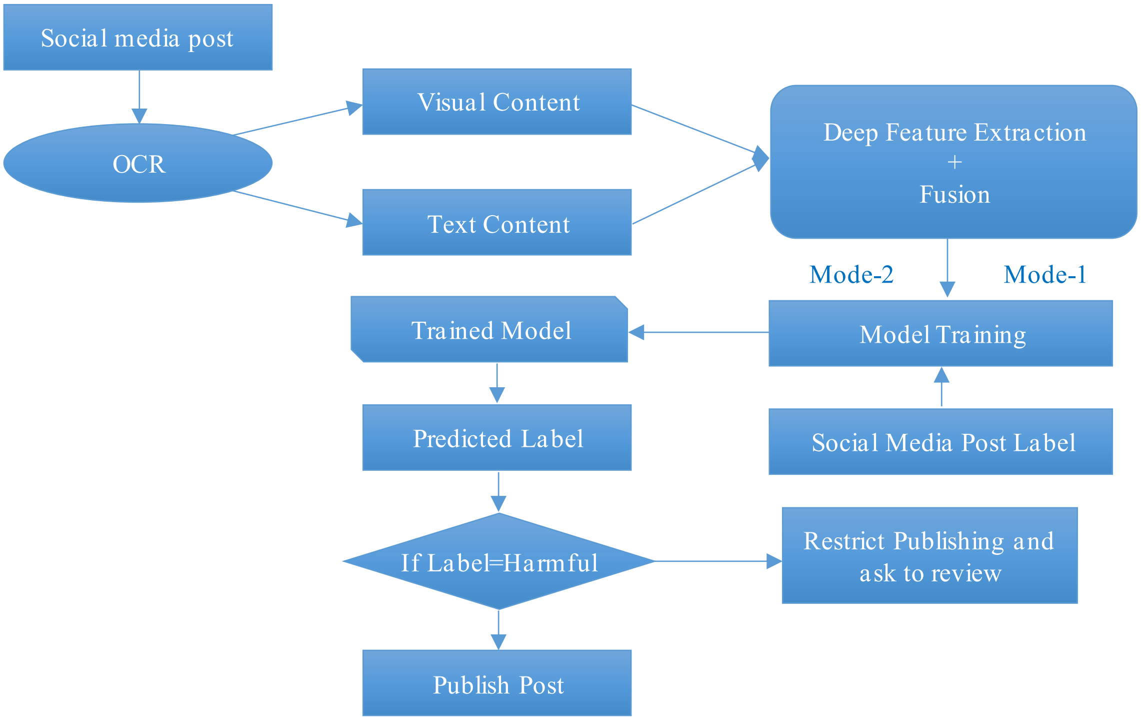


Fig. 1 Proposed system for reducing the risk of social media harmful content

4. Implementation and Experiments

In order to implement the above-discussed model for offensive content detection, a multi-model dataset was obtained from Kaggle [29]. The dataset contains 445 images in the training set and 149 images in the validation set. Thus, a total of 594 images are available. This dataset is composed of 349 images that belong to non-offensive images, and 245 images belong to offensive images. Based on this fact, the data is imbalanced. The class imbalance problem in a dataset can negatively impact classification performance. Therefore, the dataset balancing is essential before utilizing it in the training of the model. There are two types of data balancing techniques available: oversampling and undersampling. In this work, the data is limited; thus, down-sampling can negatively impact the classifier's performance. Thus, oversampling techniques can be beneficial. The Synthetic Minority Oversampling Technique (SMOTE) algorithm has been used to balance the classes. The technique returns a dataset that contains the original samples. It also returns a synthetic minority sample, depending on the percentage. After balancing the dataset, the training and validation samples have been prepared. The training and validation sample ratio is considered 80-20%. 80% of the samples are considered for training, and the remaining 20% samples are used for validation of the model. After the data sampling, a deep neural network implements a fake news detection technique. This model uses a multi-model fusion technique. The model is trained and tuned to find optimal hyperparameters. There are different techniques of parameter tuning available, but most of them are expensive in terms of running cost and resources. Therefore, to manually tune the Model, a series of experiments have been performed. Additionally, the different variants of the multi-model fusion architecture have been prepared. These models utilise pre-trained neural network architectures for computing features from both types of data. This type of model is also known as transfer learning. In this type of learning, pre-trained models are used to extract features, and then a deep model is used for learning with the extracted features. VGG16 and InceptionV3 were used for image data. Additionally, GloVe and BERT models are used for text feature extraction. Two sets of experimental models have been prepared.

1. Set 1: This experiment includes the models based on GloVe for text feature extraction. Additionally, VGG16 and InceptionV3 have been used for image feature extraction. A total of 8 models are implemented.
2. Set 2: This experiment includes the BERT for text embedding, and for image features, VGG16 and InceptionV3 have been used. By using this configuration, a total of 8 models have been developed.

4.1. Experimental Model Set 1

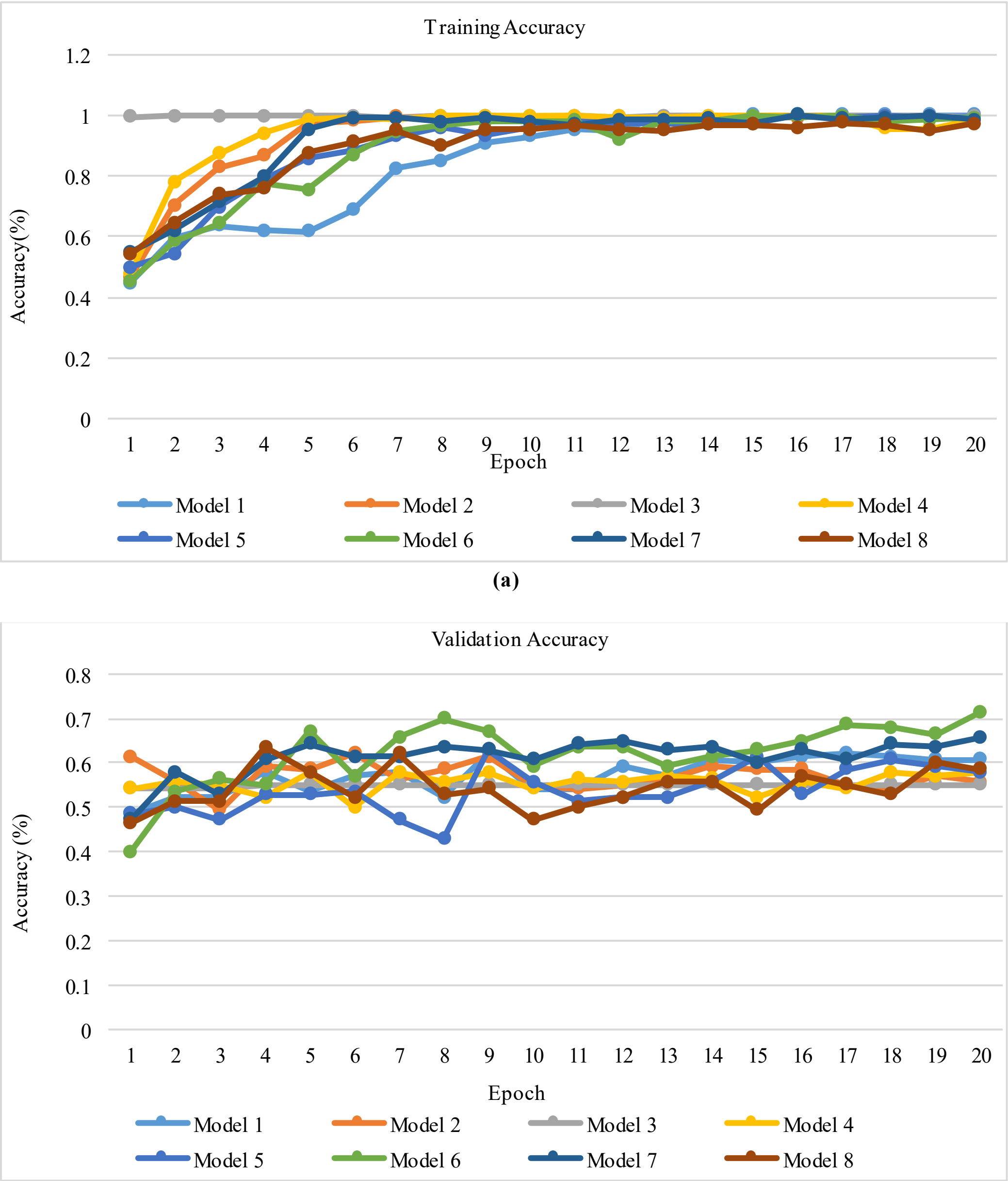
In this section, the first set of models has been discussed. The details of the configured models are given below:

- Model 1: Deep image features are extracted using the pre-trained VGG16 model within this framework. Additionally, the GloVe embedding has been used to encode the text data in this network. The model is compiled with a categorical_crossentropy loss function, and the SGD optimizer has been considered to adjust the weights.
- Model 2: This Model has a similar configuration to Model 1; thus, it uses VGG16 and GloVe embedding for feature extraction. Additionally, the categorical_crossentropy loss function was used with the Adam optimizer.
- Model 3: This Model also uses VGG16 and GloVe for feature extraction. Additionally, the binary_crossentropy loss function and the SGD optimizer are used for compiling the model.
- Model 4: This Model has the same configuration as Model 3; the only difference is in the optimizer function. In this experiment, Adam optimizer has been used in place of SGD.
- Model 5: In this experimental model for extracting the image features, InceptionV3 has been used. This model was developed by Google and is used in different real-world applications to classify images accurately. This model uses the GloVe to deal with the text data. The binary_crossentropy loss function and the Adam optimizer are also used to compile the model.
- Model 6: This Model has a similar configuration to Model 5, which uses InceptionV3, GloVe, and binary_crossentropy—additionally, only a change in the optimizer function and using the SGD optimizer.
- Model 7: This Model is the same as models 5 and 6. Therefore, its usages are InceptionV3, GloVe, and SGD. Only the loss function is changed to categorical_crossentropy.
- Model 8: This Model has the same configuration as Model 7. Therefore, its usage is InceptionV3, GloVe, and categorical_crossentropy. Additionally, the Adam optimizer has been used.

Table 2. Training accuracy and validation accuracy for the first set of models

Epoch	Training accuracy								Validation accuracy							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	0.4468	0.4627	0.9966	0.4758	0.5001	0.4531	0.5483	0.5413	0.4857	0.6143	0.5429	0.5429	0.4857	0.4	0.4714	0.4643
2	0.5977	0.7026	0.9993	0.782	0.5424	0.5859	0.6189	0.6459	0.5214	0.5571	0.5429	0.5571	0.5	0.5357	0.5786	0.5143
3	0.6369	0.8281	0.9967	0.8728	0.6964	0.6429	0.7135	0.739	0.5214	0.4929	0.5429	0.55	0.4714	0.5643	0.5286	0.5143
4	0.6205	0.8682	0.9972	0.9402	0.7867	0.7789	0.7959	0.7566	0.5786	0.5929	0.55	0.5214	0.5286	0.55	0.6071	0.6357
5	0.6165	0.972	0.9972	0.9856	0.8579	0.7554	0.9532	0.8756	0.5357	0.5857	0.55	0.5786	0.5286	0.6714	0.6429	0.5786
6	0.6865	0.9818	0.9994	0.9932	0.8822	0.8688	0.9904	0.9107	0.5714	0.6214	0.55	0.5	0.5357	0.5714	0.6143	0.5214
7	0.826	0.9939	0.9915	0.9892	0.9314	0.9465	0.9929	0.9494	0.5786	0.5643	0.55	0.5786	0.4714	0.6571	0.6143	0.6214

8	0.8511	0.9993	0.9991	0.997	0.9586	0.9674	0.978	0.8986	0.5214	0.5857	0.55	0.5571	0.4286	0.7	0.6357	0.5286
9	0.9077	0.9947	0.9986	0.9979	0.9354	0.9779	0.9919	0.9524	0.6214	0.6143	0.55	0.5786	0.6286	0.6714	0.6286	0.5429
10	0.9298	0.9961	0.9926	0.9987	0.9602	0.9794	0.9781	0.9531	0.5429	0.55	0.55	0.5429	0.5571	0.5929	0.6071	0.4714
11	0.9544	0.9964	0.9978	0.9977	0.9779	0.9837	0.9706	0.9654	0.5429	0.5429	0.55	0.5643	0.5143	0.6357	0.6429	0.5
12	0.9509	0.9959	0.9942	0.9935	0.9729	0.9209	0.9867	0.9521	0.5929	0.55	0.55	0.5571	0.5214	0.6357	0.65	0.5214
13	0.9993	0.9827	0.9988	0.9921	0.972	0.9811	0.9853	0.9489	0.5714	0.5643	0.55	0.5714	0.5214	0.5929	0.6286	0.5571
14	0.9928	0.9931	0.9986	0.9976	0.9692	0.982	0.9893	0.9684	0.6071	0.5929	0.55	0.5643	0.5571	0.6143	0.6357	0.5571
15	1	0.9978	0.9964	0.9972	0.987	0.9995	0.9723	0.9683	0.6	0.5857	0.55	0.5214	0.6143	0.6286	0.6	0.4929
16	1	0.9906	0.9949	0.9976	0.9968	0.9937	1	0.9612	0.6143	0.5857	0.55	0.5571	0.5286	0.65	0.6286	0.5714
17	1	0.9992	0.9982	0.9987	0.9929	0.9977	0.9879	0.9743	0.6214	0.55	0.55	0.5429	0.5857	0.6857	0.6071	0.55
18	1	0.9959	0.9982	0.9553	0.9994	0.9807	0.993	0.9681	0.6143	0.5429	0.55	0.5786	0.6071	0.6786	0.6429	0.5286
19	1	0.9943	0.9967	0.9566	0.9945	0.9857	0.9984	0.9504	0.6071	0.5714	0.55	0.5714	0.5929	0.6643	0.6357	0.6
20	1	0.9941	0.9935	0.9863	0.9925	0.9917	0.9862	0.9729	0.6071	0.5571	0.55	0.5786	0.5786	0.7143	0.6571	0.5857



(a)

(b)

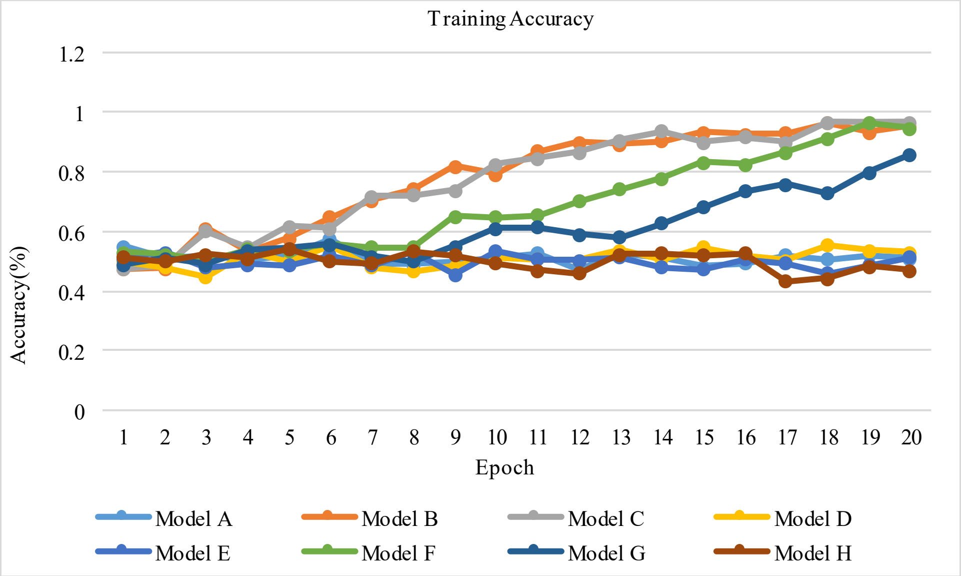
Fig. 2 Comparing deep learning models for (a) Training, and (b) Validation accuracy

In all these models, the Fully Connected Layer (FCL) remains the same. The FCL contains a total of six layers. The first layer combines the output of text and image features. The features are fused using a concatenation operation. Next, a stack of dense layers has been developed with 128, 64, 32, 16, and 2 neurons. The layer with two neurons is working as the output layer. Additionally, experiments were conducted to measure training and validation accuracy. The measured accuracy is visualized in Figure 2, and the values are reported in Table 2. In Figure 2(a), training accuracy is given, and Figure 2(b) shows the validation accuracy. The Y axis represents the accuracy in percentage (%), while the X axis represents the epochs. The models are trained for a total of 20 epochs. According to the results, both kinds of accuracy are increasing with the number of epochs. In this experiment, a total of eight models were compared, which shows the increasing training accuracy of the models. Additionally, most of the models reach 100% training accuracy. On the other hand, the validation accuracy has also been increasing for most of the models, but only model 6 is providing an acceptable level of validation accuracy, of 71.43%. Two deep learning models of image feature extraction have been used in these models, namely VGG-16 and Inception-3. However, the text embedding model is similar in both types of modelling. In this experiment, the GloVe is used for this task. The next set of experiments keeps the BERT model fixed, and the image models are changed.

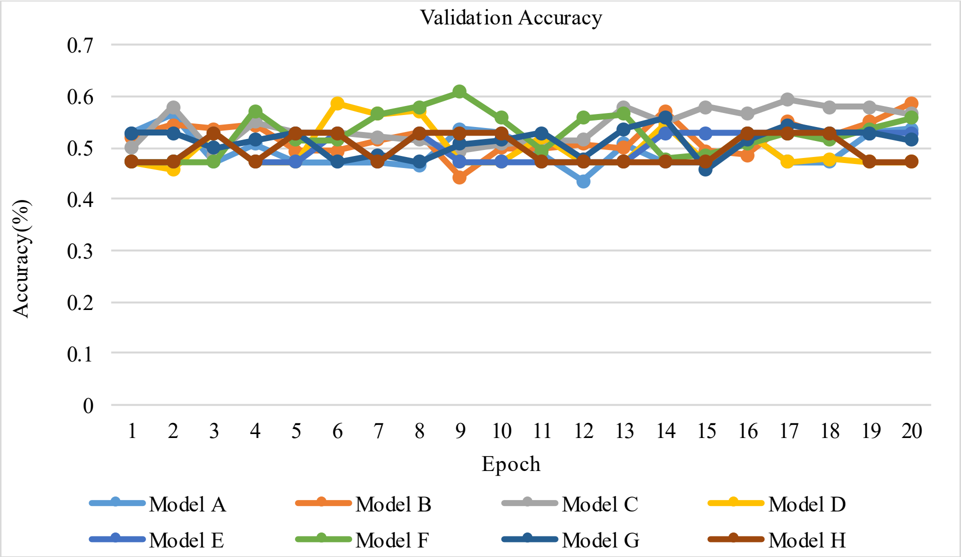
4.2. Experimental Model Set 2

In this experiment, the combination of InceptionV3 and GloVe Model provides the highest accuracy. Therefore, in the next set of experiments, only the embedding is changed to the BERT model. By making a simple change in the above implemented models with the help of a BERT embedding layer, the following eight models are configured:

- Model A: In this Model, for image features, InceptionV3 has been used, and for text features, BERT is used. Additionally, to compile the models, categorical_crossentropy as a loss function, and Adam as optimizer are used.
- Model B: In this Model, InceptionV3 was used for images, and BERT was used for text features. Additionally, categorical_crossentropy is used as a loss function, and the SGD optimizer is used for compiling the model.
- Model C: This Model is similar to the above two models and uses InceptionV3 and BERT for feature extraction. But the loss function is changed to Binary_crossentropy, and optimizer SGD is used.
- Model D: This Model is also similar to the above models and uses InceptionV3 and BERT for feature extraction. Additionally, the loss function Binary_crossentropy and optimizer Adam are used.
- Model E: This Model uses VGG16 for image feature extraction, and BERT is fixed for text embedding. Moreover, Binary_crossentropy is used as a loss function, and Adam is used as an optimizer.
- Model F: This Model also uses VGG16 for image features and BERT for text embedding. Additionally, Binary_crossentropy is used for loss calculation, and SGD is used as an optimizer.
- Model G: This Model is also similar to the above model, where VGG16 is used for image and BERT is used for text features. Additionally, the loss function is changed to categorical _cross-entropy, and optimizer SGD is used.
- Model H: In this Model, VGG16 is used for image features, and BERT is used for text features. Additionally, categorical _cross-entropy loss function is used with Adam optimizer.



(a)



(b)

Fig. 3 Comparing deep learning models for (a) Training, and (b) Validation accuracy

Table 3. Training and validation accuracy for the second set of experimental models

	Training accuracy								**Validation accuracy**							
Epoch	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
1	0.5513	0.4735	0.4743	0.5012	0.5112	0.5291	0.4887	0.5143	0.5286	0.5214	0.5	0.4714	0.4714	0.4714	0.5286	0.4714
2	0.5097	0.4769	0.4816	0.4811	0.5318	0.5242	0.511	0.4997	0.5643	0.5429	0.5786	0.4571	0.4714	0.4714	0.5286	0.4714
3	0.5152	0.6101	0.6019	0.447	0.4813	0.5011	0.491	0.523	0.4714	0.5357	0.4857	0.5286	0.5286	0.4714	0.5	0.5286
4	0.5061	0.5337	0.5479	0.5286	0.4888	0.5392	0.5385	0.5097	0.5071	0.5429	0.55	0.4714	0.4714	0.5714	0.5143	0.4714
5	0.5177	0.5758	0.6185	0.4999	0.4856	0.5374	0.5449	0.5407	0.4714	0.4929	0.5286	0.4714	0.4714	0.5143	0.5286	0.5286
6	0.5741	0.6488	0.6109	0.5505	0.5205	0.5554	0.5563	0.4994	0.4714	0.4929	0.5286	0.5857	0.5286	0.5143	0.4714	0.5286
7	0.4923	0.7049	0.717	0.4787	0.4861	0.5464	0.5181	0.4933	0.4714	0.5143	0.5214	0.5643	0.4714	0.5643	0.4857	0.4714
8	0.4912	0.7423	0.7219	0.4669	0.5294	0.5454	0.5005	0.5345	0.4643	0.5286	0.5143	0.5714	0.5286	0.5786	0.4714	0.5286
9	0.4951	0.8205	0.7398	0.4839	0.4559	0.6501	0.5512	0.5194	0.5357	0.4429	0.4929	0.4714	0.4714	0.6071	0.5071	0.5286
10	0.5121	0.7925	0.8252	0.5106	0.5338	0.6467	0.6092	0.4925	0.5286	0.5	0.5071	0.4714	0.4714	0.5571	0.5143	0.5286
11	0.5268	0.8685	0.8452	0.5054	0.5064	0.6529	0.6125	0.4696	0.4857	0.5	0.5143	0.5214	0.4714	0.4929	0.5286	0.4714
12	0.4722	0.8993	0.8669	0.5038	0.5044	0.7016	0.5904	0.459	0.4357	0.5071	0.5143	0.4714	0.4714	0.5571	0.4786	0.4714
13	0.5257	0.8941	0.9059	0.5383	0.5132	0.7428	0.579	0.5237	0.5071	0.5	0.5786	0.4714	0.4714	0.5643	0.5357	0.4714
14	0.5145	0.9028	0.9355	0.5082	0.4785	0.777	0.626	0.5256	0.4714	0.5714	0.55	0.5429	0.5286	0.4786	0.5571	0.4714
15	0.4844	0.9322	0.8993	0.5482	0.4724	0.8326	0.6796	0.5194	0.4714	0.4929	0.5786	0.4714	0.5286	0.4857	0.4571	0.4714
16	0.4942	0.928	0.9175	0.5211	0.5059	0.8244	0.7355	0.5276	0.5286	0.4857	0.5643	0.5286	0.5286	0.5071	0.5143	0.5286
17	0.5196	0.9284	0.8997	0.5016	0.4929	0.8672	0.76	0.4315	0.4714	0.55	0.5929	0.4714	0.5286	0.5286	0.5429	0.5286
18	0.5083	0.9626	0.966	0.5535	0.4582	0.9135	0.7286	0.4439	0.4714	0.5214	0.5786	0.4786	0.5286	0.5143	0.5286	0.5286
19	0.52	0.9323	0.9674	0.5381	0.4861	0.9631	0.7998	0.483	0.5286	0.55	0.5786	0.4714	0.5286	0.5357	0.5286	0.4714
20	0.509	0.9524	0.9651	0.5313	0.5138	0.9476	0.8582	0.4711	0.5357	0.5857	0.5643	0.4714	0.5286	0.5571	0.5143	0.4714

This set of experiments has also been conducted to evaluate the models' classification accuracy. The accuracy of the configured deep learning model's training and validation is shown in Figure 3 and Table 3. In this case, Figure 3(a) displays the training accuracy while Figure 3(b) displays the validation accuracy. The Y axis in these pictures contains the accuracy, whereas the X axis displays the epochs. During experiments, it was found that the majority of the models could not even complete the training appropriately; as a result, the accuracy of four models decreased. Additionally, only four Model configurations are able to complete training appropriately and provide acceptable training accuracy. On the other hand, when considering the validation accuracy, it is found that the models configured with the BERT model cannot provide an accuracy higher than 60.71%. Therefore, it is concluded that the configuration of the multi-model fusion technique works better with two pre-trained models, InceptionV3 and GloVe. Next, a comparison of all the models was performed based on the accuracy of the validation. The aim is to find the most appropriate configurations of the deep learning models for offensive image classification. Figure 4 and Table 4 demonstrate the comparative validation accuracy of all the models. The accuracy reported is based on the highest accuracy obtained in experiments.

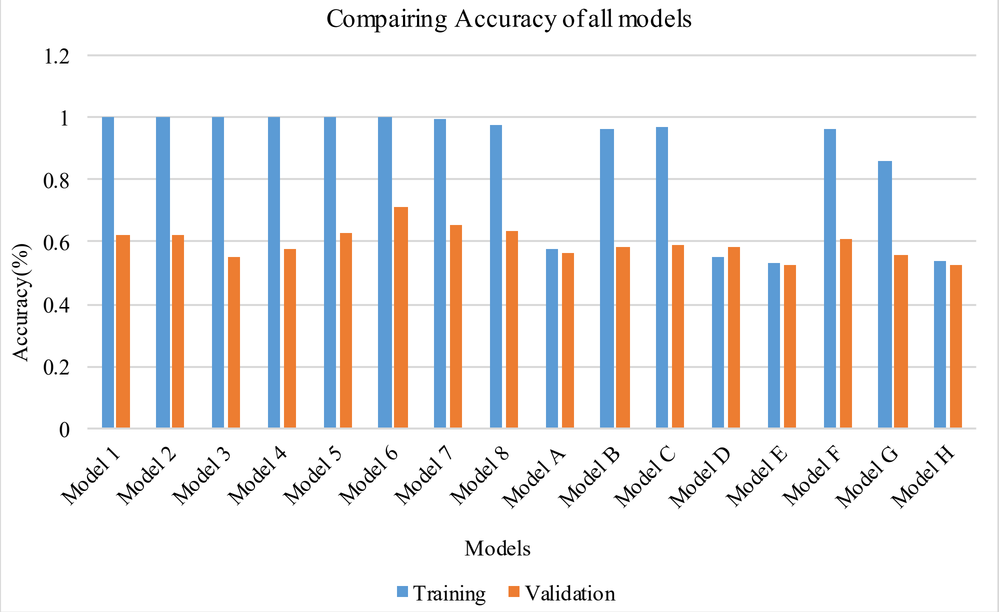


Fig. 4 Comparative analysis of the employed models' training and validation accuracy

In this diagram, the X axis shows the models used, and the Y axis shows the highest training and validation accuracy. Blue colour bars represent training accuracy, and orange colour bars indicate the accuracy of validation. According to the results obtained, Model 6 provides higher accuracy than all the other implemented models. Thus, based on this optimal performing model, a final image classification model has been prepared and demonstrated in the next section for fake news identification.

Table 4. Comparative training and validation accuracy of all the implemented models

Models	Training	Validation
Model 1	1	0.6214
Model 2	0.9993	0.6214
Model 3	0.9994	0.55
Model 4	0.9987	0.5786
Model 5	0.9994	0.6286
Model 6	0.9995	0.7143
Model 7	0.9984	0.6571
Model 8	0.9729	0.6357
Model A	0.5741	0.5643
Model B	0.9626	0.5857
Model C	0.9674	0.5929
Model D	0.5535	0.5857
Model E	0.5318	0.5286
Model F	0.9631	0.6071
Model G	0.8582	0.5571
Model H	0.5407	0.5286

5. Deep Learning Model for Multi-Model Hate Speech Detection

After experimentation with different combinations of deep feature extractors and hyperparameters, the most promising model has been identified, which offers an acceptable level of accuracy for training and validation. In this model, the image is utilized with the InceptionV3 layer, which is a pre-trained deep learning model. It is a type of CNN model initially introduced by researchers at Google and is known as GoogLeNet. InceptionV3 is the third improved version of this CNN architecture. This model is trained on a large number of images. Additionally, the model's weights are used to apply the knowledge gained from previous training. It is frequently used as a deep feature extractor in computer vision applications. On the other hand, GloVe embedding is used to extract contextual text and content features. It is also a pre-trained model and utilizes to perform learning of the context of the text content. The previously performed training is incorporated by embedding these specialized layers. Further, the extracted features from InceptionV3 and GloVe need to be combined. Therefore, a concatenation operator is used as a layer to combine both features. That layer is used to fuse the multi-model features into a common feature of the same length. This process is known as a multi-model fusion method. These combined features are now fed into a stack of dense layers. Therefore, five dense layers have been implemented with filter sizes 128, 64, 32, and 16. These layers are configured with the 'ReLu' activation function. Additionally, the last layer is also a dense layer, which is considered the output layer. This layer is configured with 2 neurons and a 'softmax' activation function. The discussed model is visualized using Figure 5. The given model accepts two inputs, the first is an image of 120*120*3 size. Additionally, the second input is the text obtained from the input image. The text data is used with the GloVe model. This model encodes the data and creates a uniform length embedding of size 300, after extracting features from both the techniques, i.e. InceptionV3 and GloVe. The vectors are combined to create a common feature. However, the dataset has suffered from a class imbalance problem. Therefore, the model also utilizes the SMOTE model to deal with the class imbalance problem of the training dataset.

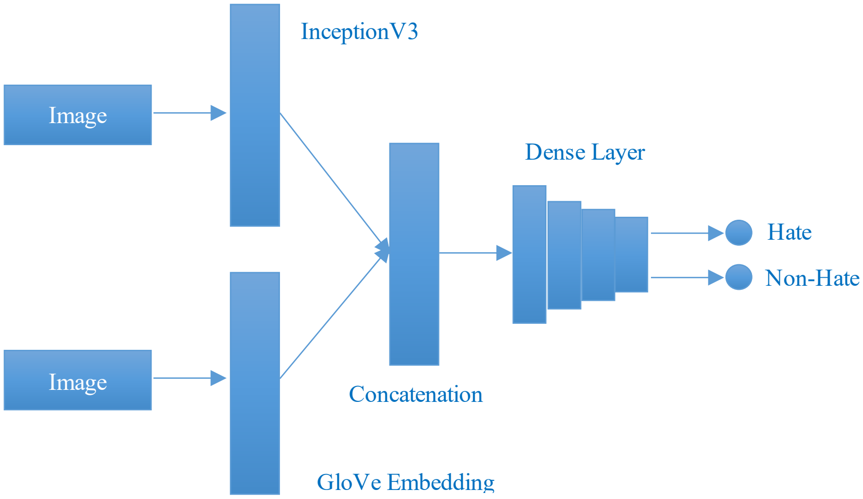


Fig. 5 Proposed multi-model fusion architecture for classifying hate speech images

Therefore, during the data preparation, input images and encoded data are merged according to the class labels. Additionally, the SMOTE is applied to balance the data and class labels. After dataset balancing, the images and associated text are separated and utilized with the above-described model for performing the training and validation of the implemented model.

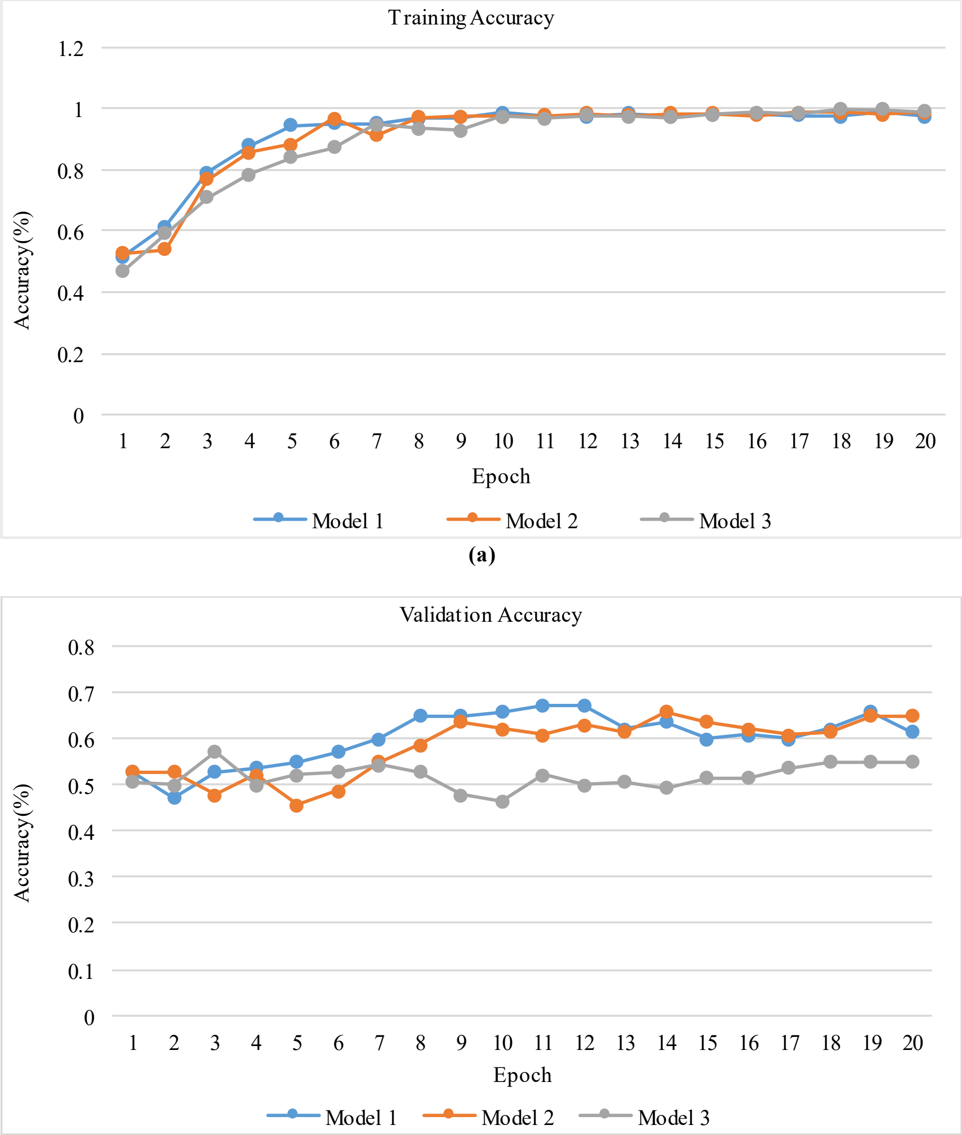
6. Results Analysis

After concluding the most optimal multi-model feature fusion-based classification model, by using the same architecture, two more variants of the model have been created. These models are named as Model 1, Model 2 and Model 3. Additionally, experiments were performed to measure the performance. The performance evaluation of the created models in terms of training and validation accuracy is given in this section. The recorded training and validation accuracy values are given in Table 5 and Figures 6(a) and 6(b). Both diagrams contain the accuracy of the models for training and validation. In Figure 6(a), the Y axis shows the accuracy as a percentage (%), while the X axis shows the experiment's epochs. According to the results, all three models are demonstrating enhanced training and validation accuracy. In addition, all three models provide similar training accuracy. Model 1 provides 97.32% training accuracy, Model 2 provides 98.90%, and Model 3 provides 99.20% accuracy. By using training accuracy, Model 3 provides higher accuracy.

Table 5. Training and validation accuracy for all three better-performing models

Epoch	Training			Validation		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
1	0.5167	0.5265	0.4672	0.5286	0.5286	0.5071
2	0.6142	0.5387	0.5920	0.4714	0.5286	0.5000
3	0.7929	0.7673	0.7096	0.5286	0.4786	0.5714
4	0.8792	0.8584	0.7847	0.5357	0.5214	0.5000

5	0.9436	0.8845	0.8397	0.5500	0.4571	0.5214
6	0.9511	0.9672	0.8718	0.5714	0.4857	0.5286
7	0.9505	0.9136	0.9490	0.6000	0.5500	0.5429
8	0.9676	0.9718	0.9351	0.6500	0.5857	0.5286
9	0.9723	0.9736	0.9283	0.6500	0.6357	0.4786
10	0.9883	0.9747	0.9759	0.6571	0.6214	0.4643
11	0.9759	0.9779	0.9677	0.6714	0.6071	0.5214
12	0.9760	0.9838	0.9787	0.6714	0.6286	0.5000
13	0.9837	0.9764	0.9730	0.6214	0.6143	0.5071
14	0.9734	0.9830	0.9726	0.6357	0.6571	0.4929
15	0.9805	0.9833	0.9800	0.6000	0.6357	0.5143
16	0.9859	0.9774	0.9889	0.6071	0.6214	0.5143
17	0.9772	0.9873	0.9851	0.6000	0.6071	0.5357
18	0.9732	0.9864	0.9981	0.6214	0.6143	0.5500
19	0.9867	0.9810	0.9969	0.6571	0.6500	0.5500
20	0.9732	0.9890	0.9920	0.6143	0.6500	0.5500



(a)

(b)

Fig. 6 shows the accuracy of the obtained models in terms of (a) Training, and (b) Validation

On the other hand, when considering validation accuracy, it is found that Model 1 provides 61.43% validation accuracy, Model 2 offers 65% accuracy, and Model 3 provides a total of 55% accuracy. Thus, in conclusion, Model 2 provides superior accuracy to both similar variants of the multi-model classification model.

Next, to make a comparison among them, three popular performance metrics are considered: precision, recall, and F-score. Class-wise performance is also termed the classification report. The calculation of this performance is also represented in terms of a confusion matrix.

The model's predicted values and the actual target values are compared using a confusion matrix. The instances in a predicted class are represented by each column of the Matrix, whereas the occurrences in an actual class are represented by each row.

A Confusion Matrix has the following elements:

- True Positives (TP): The proportion of positive cases that were anticipated to be positive.
- True Negatives (TN): The number of instances that are expected to be negative.
- False Positives (FP): The number of times negative occurrences are forecasted as positive.
- False Negatives (FN): The frequency with which positive occurrences are forecast as negative.

The confusion matrix for model experiments is given in Figure 7. Figure 7(a) shows the confusion matrix with its components. Additionally, how these components are used to calculate accuracy, precision, recall, and F1-score.

Figures 7(b), 7(c), and 7(d) show the confusion matrix of Model 1, Model 2, and Model 3. By using the confusion matrix, we can understand the effectiveness of the implemented models.

Finally, Table 6 consists of class-wise performance of models for recognizing the classes into "Offensive" and "Non-offensive". Based on the entire performance analysis, Model 2 is the most promising model for utilizing in future research for multi-model social media offensive content detection.

Predicted classes

Actual classes		Positive	Negative	
	Positive	True Positive	False Negative	Sensitivity $\frac{TP}{TP+FN}$
	Negative	False Positive	True Negative	Specificity $\frac{TN}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	Negative predictive value $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FN+FP}$

Fig. 7(a) Confusion matrix components

Confusion matrix

True label		
Offensive	31	35
Non-offensive	19	55
	Offensive	Non-offensive

Predicted label

Fig. 7(b) Confusion matrix for model 1

Confusion matrix

True label		
Offensive	40	26
Non-offensive	23	51
	Offensive	Non-offensive

Predicted label

Fig. 7(c) Confusion matrix for model 2

Confusion matrix

True label		
Offensive	34	32
Non-offensive	31	53
	Offensive	Non-offensive

Predicted label

Fig. 7(d) Confusion matrix for model 3

Table 6. Precision, recall and f-score matrix for each model

Labels	Precision			Recall			F-Score		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Offensive	0.62	0.63	0.52	0.47	0.61	0.52	0.53	0.62	0.52
Non-Offensive	0.61	0.66	0.57	0.74	0.69	0.58	0.67	0.68	0.58

7. Comparison and Discussion

After the successful implementation and performance evaluation of the proposed techniques for offensive content detection, the proposed model is compared with the two most similar existing techniques. The first technique is discussed in [10], and the second technique is described in [30]. The article [10] discusses a multimodal multi-task framework for meme understanding. Additionally, in article [30], the Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content is discussed. The results of these two papers are compared with the proposed model using Table 7. According to the results obtained, the proposed work is performing better than the models used in the article [30] and shows low performance compared to the results in the article [10].

Table 7. comparison

S. No.	Method	Precision	Recall	F-score
1	Stacked LSTM + VGG16 [30]	0.40	0.66	0.50
2	BiLSTM + VGG16 [30]	0.40	0.44	0.41
3	CNNText + VGG16 [30]	0.38	0.67	0.48
4	Model 1 [10]	74.09	69.59	76.15
5	Model 2	0.66	0.69	0.68

8. Conclusion

Social media is a source of fresh information; therefore, a large segment of individuals globally consumes the NEWS from social media. However, due to fewer restrictions and monitoring, the information on social media has no credibility and results in fake, manipulated, and hateful content flooding social media. In this presented work, first, a machine learning model has been introduced, which evaluates and analyzes the social media content before publishing. Therefore, an architecture of the required model has been prepared, and the different components of the model have been discussed. Further, a series of experiments has been performed to identify the suitable components and hyperparameters of the learning algorithm. These experiments have considered different deep learning models for feature extraction, such as BERT and GloVe, for dealing with the data. Additionally, VGG-16 and InceptionV3 have been considered for dealing with the Image data. Based on the results, the most promising deep learning architecture has been identified, and its detailed architecture has been discussed. This final architecture utilizes GloVe and InceptionV3 for the extraction of deep multi-model features. Additionally, a stack of dense layers has been used to train and classify the extracted deep features. In this experiment, the selection of the loss function and optimizer also influences the performance of the learning architecture. By using the experiment, the binary_crossentropy loss function and the SGD optimizer provide acceptable results. During this analysis, the dataset is found to be imbalanced, which influences the classification accuracy of the models. In this context, to balance the data, the down-sampling technique is not suitable for model development; the SMOTE over-sampling approach has been used to rectify the class imbalance issue. Based on the obtained results, the model provides an acceptable level of validation accuracy with reduced validation loss. The model provides 99.82% training accuracy and 71.43% validation accuracy. Thus, the model is acceptable for utilizing in preventing the social media harmful content.

References

[1] Chin-Wen Chang, and Sheng-Hsiung Chang, “The Impact of Digital Disruption: Influences of Digital Media and Social Networks on Forming Digital Natives’ Attitude,” *SAGE Open*, vol. 13, no. 3, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Gil Appel et al., “The Future of Social Media in Marketing,” *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 79-95, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Niels Frederik Lund, Scott A. Cohen, and Caroline Scarles, “The Power of Social Media Storytelling in Destination Branding,” *Journal of Destination Marketing and Management*, vol. 8, pp. 271-280, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Sarah A. Fisher, Jeffrey W. Howard, and Beatriz Kira, “Moderating Synthetic Content: The Challenge of Generative AI,” *Philosophy & Technology*, vol. 37, no. 4, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Mark A. Flynn, Emery Veilleux, and Alexandru Stana, “A Post from the Woods: Social Media, Well-Being and Our Connection to the Natural World,” *Computers in Human Behavior Reports*, vol. 5, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Rana Ali Adeeb, and Mahdi Mirhoseini, “The Impact of Affect on the Perception of Fake News on Social Media: A Systematic Review,” *Social Sciences*, vol. 12, no. 12, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] M. Sudhakar, and K.P. Kaliyamurthie, “Detection of Fake News from Social Media using Support Vector Machine Learning Algorithms,” *Measurement: Sensors*, vol. 32, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub, “A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion,” *Heliyon*, vol. 9, no. 10, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Tariq Habib Afridi et al., “A Multimodal Memes Classification: A Survey and Open Research Issues,” *The Proceedings of the Third International Conference on Smart City Applications*, Safranbolu, Turkey, pp. 1451-1466, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Bingbing Wang et al., “What Do They ‘‘Meme’’? A Metaphor-Aware Multi-Modal Multi-Task Framework for Fine-Grained Meme Understanding,” *Knowledge-Based Systems*, vol. 294, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Cristian Vaccari, and Andrew Chadwick, “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News,” *Social Media + Society*, vol. 6, no. 1, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Despoina Mouratidis, Andreas Kanavos, and Katia Kermanidis, “From Misinformation to Insight: Machine Learning Strategies for Fake News Detection,” *Information*, vol. 16, no. 3, pp. 1-30, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo, “The Power of Context: A Novel Hybrid Context-Aware Fake News Detection Approach,” *Information*, vol. 15, no. 3, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang, “FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach,” *Multimedia Tools and Applications*, vol. 80, no. 8, p. 11765-11788, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Esma Aïmeur, Sabrine Amri, and Gilles Brassard, “Fake News, Disinformation and Misinformation in Social Media: A Review,” *Social Network Analysis and Mining*, vol. 13, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Bo Hu, Zhendong Mao, and Yongdong Zhang, “An Overview of Fake News Detection: From a New Perspective,” *Fundamental Research*, vol. 5, no. 1, pp. 332-346, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin, “A Comprehensive Survey on Machine Learning Approaches for Fake News Detection, ” *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 51009-51067, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Shiwen Ni, Jiawen Li, Hung-Yu Kao, “MVAN: Multi-View Attention Networks for Fake News Detection on Social Media,” *IEEE Access*, vol. 9, pp. 106907-106917, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, “EchoFakeD: Improving Fake News Detection in Social Media with an Efficient Deep Neural Network,” *Neural Computing and Applications*, vol. 33, no. 14, pp. 8597-8613, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Bhuvanesh Singh, and Dilip Kumar Sharma, “Predicting Image Credibility in Fake News Over Social Media using Multi-Modal Approach,” *Neural Computing and Applications*, vol. 34, no. 24, pp. 21503-21517, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso, “Multimodal Multi-image Fake News Detection,” *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, pp. 647-654, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Tong Zhang et al., “BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection,” *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Beizhe Hu et al., “Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection,” *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-24)*, vol. 38, no. 20, pp. 22105-22113, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Liwen Peng et al., “Not All Fake News is Semantically Similar: Contextual Semantic Representation Learning for Multimodal Fake News Detection,” *Information Processing & Management*, vol. 61, no. 1, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jiaying Wu, Jiafeng Guo, and Bryan Hooi, “Fake News in Sheep’s Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks,” *KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona Spain, pp. 3367-3378, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Xiaochang Fang et al., “NSEP: Early Fake News Detection Via News Semantic Environment Perception,” *Information Processing and Management*, vol. 61, no. 2, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Ehtesham Hashmi et al., “Advancing Fake News Detection: Hybrid Deep Learning with FastText and Explainable AI,” *IEEE Access*, vol. 12, pp. 44462-44480, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo, “Fake News Detection in Low-Resource Languages: A Novel Hybrid Summarization Approach,” *Knowledge-Based Systems*, vol. 296, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Aditya, Hate Speech Detection Dataset, Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/aditya1220/hate-speech-detection-dataset/data.

[30] Shardul Suryawanshi et al., “Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text,” *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, pp. 32-41, 2020. [Google Scholar] [Publisher Link]