The Visual Computer
https://doi.org/10.1007/s00371-023-03224-y

ORIGINAL ARTICLE

A smart video analytical framework for sarcasm detection using novel adaptive fusion network and SarcasNet-99 model

Jamuna S. Murthy[1] · **G. M. Siddesh**[2]

Accepted: 9 December 2023

Abstract
Sarcasm is often related to something that has created a mass confusion among the general uninformed public. It is always associated with a mockery tone or trenchancy facial expression or weird language. Existing literatures that are profound in the field of sarcasm detection mainly focused on text-based input with sarcastic comments or facial expression-based analysis, i.e., image input. But both text and image input are not sufficient to analyze the underlying sarcasm behind the scene. This kind of analysis can also be misleading sometimes as the emotional expression can change with social circumstances (i.e., audio tone) over time. Hence to address these challenges, "A Smart Video Analytical framework for Sarcasm Detection using Deep Learning" is introduced where sarcasm detection is done by considering video modality. Proposed model extracts three important features from the video, i.e., text using proposed Enhanced-BERT, image using ImageNet and audio using Librosa. After extraction, each modality is addressed individually and is finally fused using proposed adaptive early fusion approach. The final task prediction of classification is done using novel deep neural network called "SarcasNet-99" to detect sarcasm in video over distributed framework called Apache Storm. TedX and GIF Reply datasets are used for model training and testing with around 10,000 + video clips. When compared against existing state-of-the-art techniques such as AlexNet, DenseNet, SqueezeNet and ResNet, the proposed model predicted accuracy 99.005% with LeakyReLU activation function.

Keywords Sarcasm · TedX · Twitter · LeakyReLU · BERT

1 Introduction

Social media and other online blogs are hugely adopted by the public in recent years; it has become easier than ever to spread sarcastic news. Such news tends to manipulate the general public and influence their decisions to an extent that it might possibly have lasting repercussions. With the advent of social media, people can now share, reshare as well as download sarcastic information instantaneously without even having a chance to validate it. This problem has worsened to an extent that false information has become indistinguishable from real [1, 2].

The main reason why people fall victim to false information is due to Confirmation Bias and Naive Realism [3]. Confirmation bias refers to the human tendency to favor data that confirms their existing views. In such cases, people tend to ignore the authenticity of the news with the sole purpose of reinforcing their thoughts. Even when presented with authentic facts, people tend to stick to their views and ironically label those who disagree with them as "uninformed" or "biased," illustrating the problem of Naive Realism [4, 5].

Along with the susceptibility to believing sarcastic news, a portion of the society still finds it difficult to identify sarcasm due to its sheer variety and subtleties. The figurative nature of sarcasm also makes it difficult when performing sentiment analysis. Identifying both the metaphorical as well as literal meanings is crucial to interpreting the true meaning behind any source of information.

Sarcasm is one of the most common entities on the web today and hence the progress of the proposed research work

✉ Jamuna S. Murthy
jamunamurthy.s@gmail.com

G. M. Siddesh
siddeshgm14@gmail.com

[1] Department of Computer Science and Engineering, Ramaiah Institute of Technology, India Affiliated to Visvesvaraya Technological University, Bengaluru, India

[2] Department of Artificial Intelligence and Data Science, Ramaiah Institute of Technology, India Affiliated to Visvesvaraya Technological University, Bengaluru, India

Published online: 22 January 2024

can help discover the human's honest thoughts in more realistic way so that it can be applied for Opinion Mining, Review Analysis and Harassment Detection on web. The existing literatures profound in the field of sarcasm detection are categorized as text, image and image-text combination, i.e., memes.

Most of the times sarcasm is a part of natural language, i.e., text data and it goes undetected during the conversations on social media platforms or online sites. The level of truth is completely difficult to detect by the humans as everyone have different thinking capabilities. This weakness is completely used as targeted marking by the news paper editors and fake bloggers who put sarcasm in their headline to grab the attention.

The study proposes an intelligent machine learning-based model for detecting and classifying sarcasm on social media platforms. The research identifies the limitations of existing sarcasm detection methods and addresses the challenges of analyzing complex and varied text data on social networks. The proposed model utilizes a deep learning algorithm that incorporates various features to enhance the accuracy of sarcasm detection. The study concludes that the proposed model outperforms existing methods, achieving high accuracy in detecting and classifying sarcasm in social media data. The findings of this research contribute to the development of advanced natural language processing models for social media data analysis [6].

Godara et al. [7] propose an ensemble classification approach for sarcasm detection, which utilizes multiple machine learning algorithms to achieve higher accuracy. The study presents promising results, indicating that the proposed approach outperforms traditional machine learning algorithms in sarcasm detection. The authors also discuss the importance of sarcasm detection in natural language processing and its potential applications in various fields. Overall, the study provides valuable insights into the development of effective methods for detecting sarcasm in text data.

The research article by Farha and Magdy [8] evaluates the performance of various transformer-based language models for Arabic sentiment and sarcasm detection. The authors utilized six different pre-trained models and benchmarked them on two standard datasets. The results indicate that the transformer-based models outperform traditional machine learning algorithms in both tasks. Moreover, the study highlights the importance of selecting an appropriate pre-trained model for the specific task at hand. Overall, the article provides valuable insights into the application of transformer-based models for Arabic sentiment and sarcasm detection, which can have practical implications in various fields such as social media analysis and customer feedback analysis.

The research article presents an Artificial Intelligence (AI)-based approach for detecting misogyny and sarcasm from Arabic texts. The authors conducted experiments using three different datasets and various machine learning techniques to evaluate the effectiveness of the proposed approach. The results show that the approach achieved high accuracy in detecting misogyny and sarcasm from Arabic texts. The study contributes to the development of effective AI-based tools for detecting hate speech and offensive language in Arabic, which could be useful for social media platforms and online communities in the Middle East [9].

This article [10] proposes a novel approach for detecting sarcasm and irony in text using a combination of transformer-based word embeddings and Convolutional neural networks (CNNs). The authors provide a comprehensive literature review on related works in the field of sarcasm and irony detection, including traditional machine learning techniques and deep learning methods. The experimental results demonstrate the effectiveness of the proposed approach in detecting sarcasm and irony with high accuracy, outperforming the state-of-the-art methods. The study contributes to the advancement of natural language processing techniques in the field of sentiment analysis.

The next category of data used for sarcasm detection is images. Yao et al. [11] presented a literature review on sarcasm detection in social media and proposed a novel approach that imitates the human brain's cognitive processes. Their method combines natural language processing techniques with a cognitive model that mimics how the brain processes sarcasm. The proposed approach achieves high accuracy in detecting sarcasm on Twitter, surpassing state-of-the-art models. The authors suggest that incorporating cognitive models into natural language processing may lead to more effective and human-like language understanding.

The article by Liang et al. [12] presents a multi-modal approach for detecting sarcasm in text using interactive in-modal and cross-modal graphs. The study proposes a novel method of integrating textual, visual, and audio cues to improve sarcasm detection accuracy. The authors report a significant improvement in the detection of sarcastic comments using their proposed method. The paper offers valuable insights into the potential of multi-modal analysis in natural language processing and lays the groundwork for future research in the area of sarcasm detection.

Sharma et al. [13] proposed a hybrid auto-encoder-based model to detect sarcasm on social media platforms. The model incorporates both word-level and character-level information to improve the accuracy of sarcasm detection. The authors trained and tested their model on a dataset of sarcastic and non-sarcastic tweets, achieving a high accuracy of 95.14%. The study highlights the importance of considering both linguistic and contextual cues in detecting sarcasm on social media. The proposed model has potential applications in various fields, including sentiment analysis and social media monitoring.

The article by Liang et al. [14] proposes a novel approach for multi-modal sarcasm detection using a cross-modal graph Convolutional network. The study presents an extensive evaluation of their model using several benchmark datasets and compares it with state-of-the-art methods. The results show that the proposed approach outperforms existing models and can effectively capture the complex relationships between different modalities. The study provides a significant contribution to the field of natural language processing and demonstrates the potential of cross-modal approaches for sarcasm detection.

The research article called "Cat-bigru" [15] for detecting self-deprecating sarcasm. The model combines convolutional and attentional neural network layers with bi-directional gated recurrent units (GRUs) to capture both local and global context information from text. The authors evaluate the proposed model on a publicly available dataset and report significant improvements in accuracy compared to existing methods. Overall, the study presents a promising approach to detecting sarcasm in text using deep learning techniques, which can have applications in various domains such as social media analysis and sentiment analysis.

Nevertheless, these approaches have a drawback in terms of their slow training speed and their disregard for crucial information. Specifically, a substantial portion of the data in video modality is extraneous to sarcasm detection, such as contextual details in the background.

This paper presents a method called a smart video analytical framework for Sarcasm Detection using Deep Learning. The approach combines text, speech, and face image features using an adaptive feature fusion strategy to create a single vector for prediction. The process involves three stages: multi-modal feature extraction, adaptive feature fusion, and feature classification. The method extracts text and speech features using BERT and Librosa and uses DLIB's face detection tool to cut out face images, which are then stitched horizontally using SarcasNet-99 to obtain face image features. The adaptive fusion strategy uses a fusion weight parameter to control information inconsistency between different modalities and achieve high performance. Finally, the fused vector is sent to the fully connected layer for prediction. The results indicate that the fusion of image features from face regions is more effective than simply concatenating the three types of components.

The major contribution of the proposed work includes these three major objectives:

1. A new sarcasm detection model using deep learning has been proposed to address the limitation of current models that use **text, speech, and image** as input, but only consider the entire image, resulting in excessive redundant data that affects accuracy. The new model focuses on facial information to capture emotional cues associated with sarcasm. It performs a face recognition operation to obtain image data of the final input model by horizontally stitching the detected face regions.
2. Conventional methods of merging features in different modes link or combine their distinct characteristics. However, speech modality features are often dismissed as noise due to their numerical differences from other modes. Hence, a novel **adaptive feature fusion approach** is suggested, allowing for flexible fusion weights between modalities to account for their inconsistencies.
3. A Novel Neural Network architecture called "SarcasNet-99" is introduced for final classification of sarcastic videos which has 99 fully connected dense layers.
4. To tackle over fitting in deep learning training, a data augmentation technique is applied using the **TedX and GIF Reply datasets**. Proposed approach is shown to be effective through several experiments, as evidenced by a 10% increase in accuracy compared to the previous baseline method for sarcasm detection.

2 Literature survey

The use of single-mode methods to detect sarcasm is no longer adequate for investigating this complex linguistic phenomenon. Recent research on sarcasm detection has focused primarily on multi-modality approaches over the past few years. There are three primary categories of sarcasm detection methods that based on the survey conducted: rule-based methods, machine learning-based methods, and deep learning-based methods.

The paper [16] proposes a novel approach for detecting sarcasm in social media using coupled-attention networks (CANs). The authors demonstrate the effectiveness of their method on three benchmark datasets, achieving state-of-the-art performance. The paper contributes to the growing body of research on sarcasm detection, and the proposed CANs model has the potential to improve the accuracy of sentiment analysis in social media contexts. However, the paper could benefit from a more detailed discussion of the limitations and future directions of the proposed method.

This research article [17] proposed a new approach for multi-modal sarcasm detection using a cross-modal graph convolutional network. The authors conducted experiments on several datasets and achieved state-of-the-art performance compared to existing methods. The article contributes to the field of computational linguistics by demonstrating the effectiveness of using cross-modal information in sarcasm detection, which can be applied in various natural language processing tasks. However, the article could benefit from further analysis and explanation of the model's limitations and potential biases.

The article by Liu, Wang, and Li [18] in 2022 presents a novel approach to detect sarcasm in multi-modal data, including text, image, and audio. The proposed method employs hierarchical congruity modeling to capture the congruity between the sentiment expressed in different modalities and utilizes knowledge enhancement to enhance the model's performance. The authors also introduce a new multi-modal sarcasm dataset to evaluate their approach's effectiveness. Overall, the article presents a promising approach to addressing the challenging problem of multi-modal sarcasm detection.

The authors of [19] describe the UMUTeam's approach to the SemEval-2022 Task 5, which focuses on automatic misogyny identification through a combination of image and textual embeddings. The study proposes a model that uses a pre-trained convolutional neural network to extract features from images, and a BERT-based model to process textual data. The results of the study show that the proposed model outperforms the baseline models in terms of identifying misogyny in both textual and visual domains. The study highlights the potential of using multi-modal approaches in identifying hate speech and misogyny online.

This paper [20] explores the use of machine learning algorithms to detect irony and sarcasm in public figure speeches. The study analyzes the performances of four different machine learning models on a dataset of public figure speeches to identify the most effective algorithm for detecting irony and sarcasm. The results of the study show that the Support Vector Machine (SVM) algorithm outperforms the other models and achieves a high accuracy rate of 78%.

Next the primary aim of the sarcasm detection approach that uses speech data is to recognize the sound characteristics linked to sarcasm. The article [21] explores the effectiveness of machine learning models in detecting irony and sarcasm in public figure speeches. The authors utilized a dataset consisting of speeches from prominent public figures and trained a classifier to detect instances of irony and sarcasm. The study found that the machine learning model was effective in detecting both irony and sarcasm in public figure speeches with high accuracy. The authors suggest that such models can be useful in analyzing and understanding the nuances of public speeches, which may have important implications for education and communication studies.

The research article [22] proposes a multi-modal fusion method for detecting sarcasm. The study employs late fusion techniques to combine textual, acoustic, and visual features for improved detection accuracy. The proposed approach demonstrates superior performance compared to existing methods, achieving an F1-score of 81.98%. The article provides a comprehensive review of existing literature on sarcasm detection and discusses the challenges of using multiple modalities for detecting sarcasm. The research highlights the potential of multi-modal fusion techniques for improving the accuracy of sarcasm detection in various applications.

The authors of [23] proposes a novel approach to hate speech detection using a combination of Convolutional Neural Networks (CNN), Bi-directional Gated Recurrent Unit (BiGRU) and Capsule Network. The proposed method called HCovBi-caps achieves promising results on two public datasets and outperforms other state-of-the-art methods in terms of accuracy, *F1* score, and AUC-ROC. The study contributes to the growing body of research on hate speech detection by introducing a hybrid approach that combines multiple deep learning architectures, which can help improve the performance of hate speech detection systems.

The article by Zhang et al. [24] presents a novel approach for stance-level sarcasm detection using BERT and stance-centered graph attention networks. The study highlights the importance of identifying the stance of a statement in detecting sarcasm, as sarcasm often involves contradicting or opposing a particular stance. The proposed approach achieved state-of-the-art performance on the SARC 2.0 dataset, demonstrating the effectiveness of incorporating stance information and graph attention mechanisms in sarcasm detection. The study contributes to the field of natural language processing and has practical applications in detecting sarcasm in online communication.

Juyal's [25] research article presents a study on multi-modal sentiment analysis of audio and visual data using machine learning. The paper focuses on the integration of audio and visual features to enhance the accuracy of sentiment analysis. The study proposes a model that combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to classify sentiment in audio and visual data. The results show that the proposed model outperforms traditional models in accuracy, demonstrating the potential of using multi-modal data for sentiment analysis.

3 Sarcasm detection using proposed enhanced-BERT, adaptive fusion network and SarcasNet-99

This new revolution in the field on Natural Language processing and deep Learning paves a way for researchers to deal with challenges related to sarcasm and fake news detection addressed in literature review. To address those challenges proposed framework involves four major modules such as Data Collection, Data Processing, Data Analytics-Prediction as shown in Fig. 1.

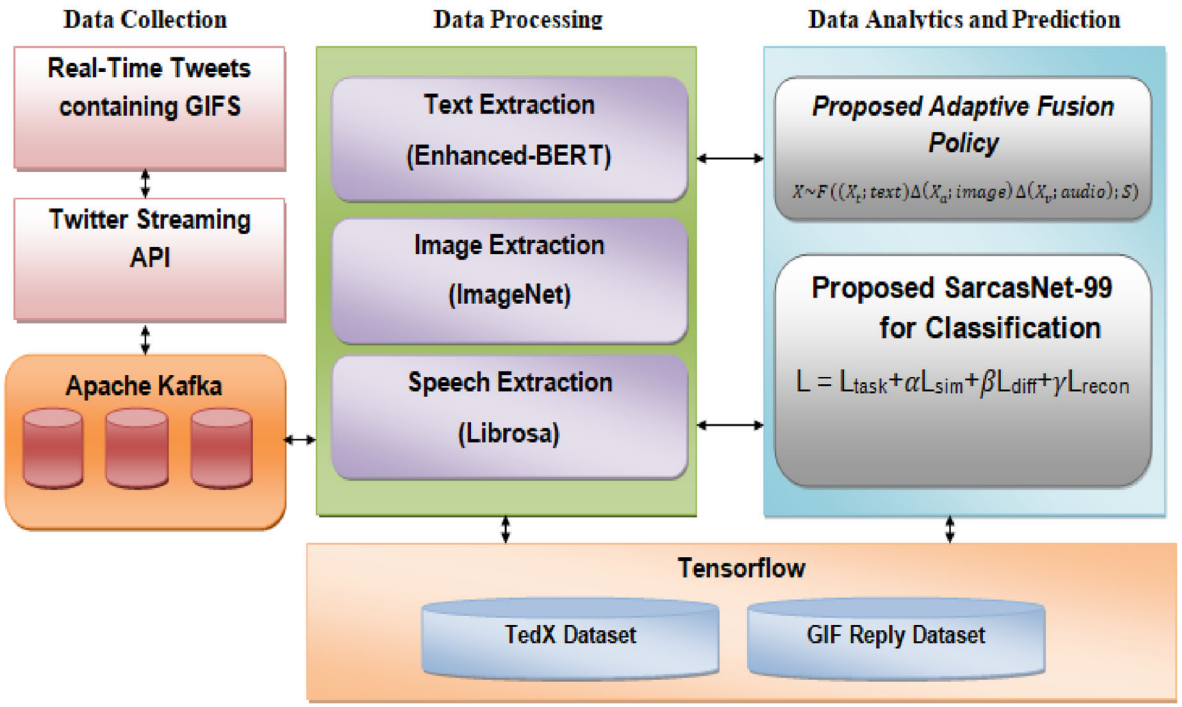

Fig. 1 Proposed architecture for sarcasm detection using deep learning

3.1 Data collection

Twitter is considered as one of the best social media platforms for sharing information related to hot topics around the world. Billions of users are registered to twitter all over the globe. Twitter Users use tweets with trending HashTags for creating a trend on particular hot topic or news. Tweets are reshared in other social media platforms such as Facebook, Instagram to discuss about the hot news. Millions of tweets get generated on web everyday and hence in the proposed work twitter is considered as one of the major data source. Also Twitter provided Twitter Streaming API through which we can analyze the data over real-time.

Apache storm a distributed big data processing engine is used to provide a solution for handling huge amount of data over real-time listed as one of the major challenges of existing systems. Apache storm consist of spouts (i.e., input unit) and bolts (i.e., processing units) which are implemented as map reduce programming model.

Over the real-time using Twitter Streaming API implemented using tweepy python library, related tweets of particular topic is streamed and are continuously collected using Apache Kafka which is an input processing unit of proposed framework. Kafka can stream up to 10 Lakhs messages over real-time on the distributed platform. Parallelly the bolts, i.e., processing units are used for data pre-processing which in turn is used for extraction of important features that can affect the classification of data as sarcastic or real using deep neural network in the later stages.

3.2 Data processing

Raw Tweets are unstructured in nature and hence have to be pre-processed using Natural language processing techniques. Also the feature extraction in the proposed framework plays a very important role since they affect the strategy of building the huge language models for sarcasm and fake news detection. Here, there are four important features extracted from the tweets, i.e., text data, emoji, image data and image-text data.

3.2.1 BERT for text feature extraction

A pre-training language model called BERT (Bi-directional Encoder Representations from Transformers) has considerably enhanced the study of natural language processing (NLP).However, the BERT algorithm suffers a number of difficulties, including:

1. *Limited comprehension of context:* Despite being a strong NLP tool, BERT still has issues comprehending the context of language. For instance, BERT might not be able to comprehend irony or sarcasm, which could result in inaccurate forecasts .

2. *Training data bias:* BERT, like any machine learning model, is susceptible to bias from the training data that was used to develop it. This can result in incorrect predictions or confirm preexisting prejudices.

Hence in the proposed work an "*Enhanced-BERT*" model is built with more precise training data which can overcome contextual difficulties such as understanding sarcasm, irony that can in turn improvise the accuracy. The proposed methodology for text feature extraction is given by: Firstly, input the words into $\text{BERT}_{\text{base}}$ model with Transformer layers, $L = 4$ to average the output. Finally, each piece of word is represented as $W_t = 768$ dimensional feature vector X_t.

$$\left\{X_j^t\right\}_{j=1}^{M} = BERT_{base}(T) \tag{1}$$

$$X_t = \frac{1}{L}\left(\sum_{j=1}^{M} X_j^t\right) \in R^{W_t} \tag{2}$$

here, X_j^t represents the out of the last jth transformation layer in $\text{BERT}_{\text{base}}$ model for each word T.

3.2.2 LPCC for speech extraction

Librosa is a library used for speech extraction in the proposed model. The speech data with time series are inputted to Librosa library with sampling rate of 22,000 Hz. Heuristic-based audio extraction technique is used for noise reduction from the sample audios. Next, the local features such as MFCC, Spectral Centroid, Mel-spectrogram are extracted from the audios as non over palling windows, i.e., $W_{a..}$ A joint vector, i.e., $\left\{X_j^a\right\}_{j=1}^{W_a}$ is created by combining all the local features with $W_a = 285$ dimensions. The average value of the join vector is given by:

$$X_j^a = X_i^{MFCC} \oplus X_i^{MFCCdelta} \oplus X_i^{Mel} \oplus X_i^{Meldelta} \oplus X_i^{Spec} \tag{3}$$

$$X_a = \frac{1}{W_t}\left(\sum_{j=1}^{M} X_j^a\right) \in R^{W_a} \tag{4}$$

Here, $\oplus$ concatenates each of the features, i.e., X_i^{MFCC}, $X_i^{MFCCdelta}$, X_i^{Mel}, $X_i^{Meldelta}$, X_i^{Spec} in Eqs. (3) and (4).

3.2.3 ImageNet for image feature extraction

Out of all the modalities facial features clearly explain the emotion of the person for sarcasm detection. For each video input, the emotional change in the facial expression of the person with time series is determined frame by frame and is processed without any background information. Popular library called OpenCV is used for video to frame extraction. Here, usually background information of the image does not play an important role as the more focus in given on the person's topic of interest or news he is talking about. Next face detection is done by one of the popular libraries Histogram of Oriented Gradients or HOG.

For each video frame $\boldsymbol{V_i}$, the number of faces detected is given by $\boldsymbol{Face_i}$. Let $\boldsymbol{H_i}$ be the height of each face detected from $\boldsymbol{V_i}$. Each face in the frame is of different heights and hence to fill the gap, black block $\boldsymbol{Block_i}$ is used as splicing. This confirms the uniformity of heights of all the faces and also horizontal stitching is applied as final version to the faces in image to input to the neural network. The formulization of the above extraction technique is given in Eqs. (5) and (6).

$$A_{Block_{(Face_i)}} = (3,\ Length(Face_i),\ H_i - Height(Face_i)) \tag{5}$$

$$padding(Face_i) = \begin{cases} Face_i, & Height(Face_i) = H_i \\ Face_i \oplus Block_i \in R^{X_{Block_{(Face_i)}}}, & Height(Face_i) < H_i \end{cases} \tag{6}$$

$$Face_i = stithing(padding(Face_i)) \tag{7}$$

Here, $Length(Face_i)$ is Length of the image, $Height(Face_i)$ I height of the image. Also, $A_{Block_{(Face_i)}}$ is the dimension of the block box and $\oplus$ is the vertical join operator to fuse the different features. Finally, $stithing(padding(Face_i))$ represents the padding and horizontal stitching for the incorrect image.

$Face_i$ is the final stitched image after correction. Next, the each image frame is pre-processed to normalize and then fed into ImageNet neural network algorithm with. This neural network is pre-trained with 2048 dimensions for feature extraction from $Face_i$. An average value of feature vector $X_I^{Face_i}$ for visual feature extraction is calculated with $W_{\text{v}} = 2048$ dimensions for each "*frame*" of video. The formularization of the feature extraction using Sarcasnet-99 is given by Eqs. (8) and (9), respectively.

$$X_I^{Face_i} = ImageNet(Face_i) \tag{8}$$

$$X_v = \frac{1}{frame}\left(\sum_i X_I^{Face_i}\right) \in R^{W_v} \tag{9}$$

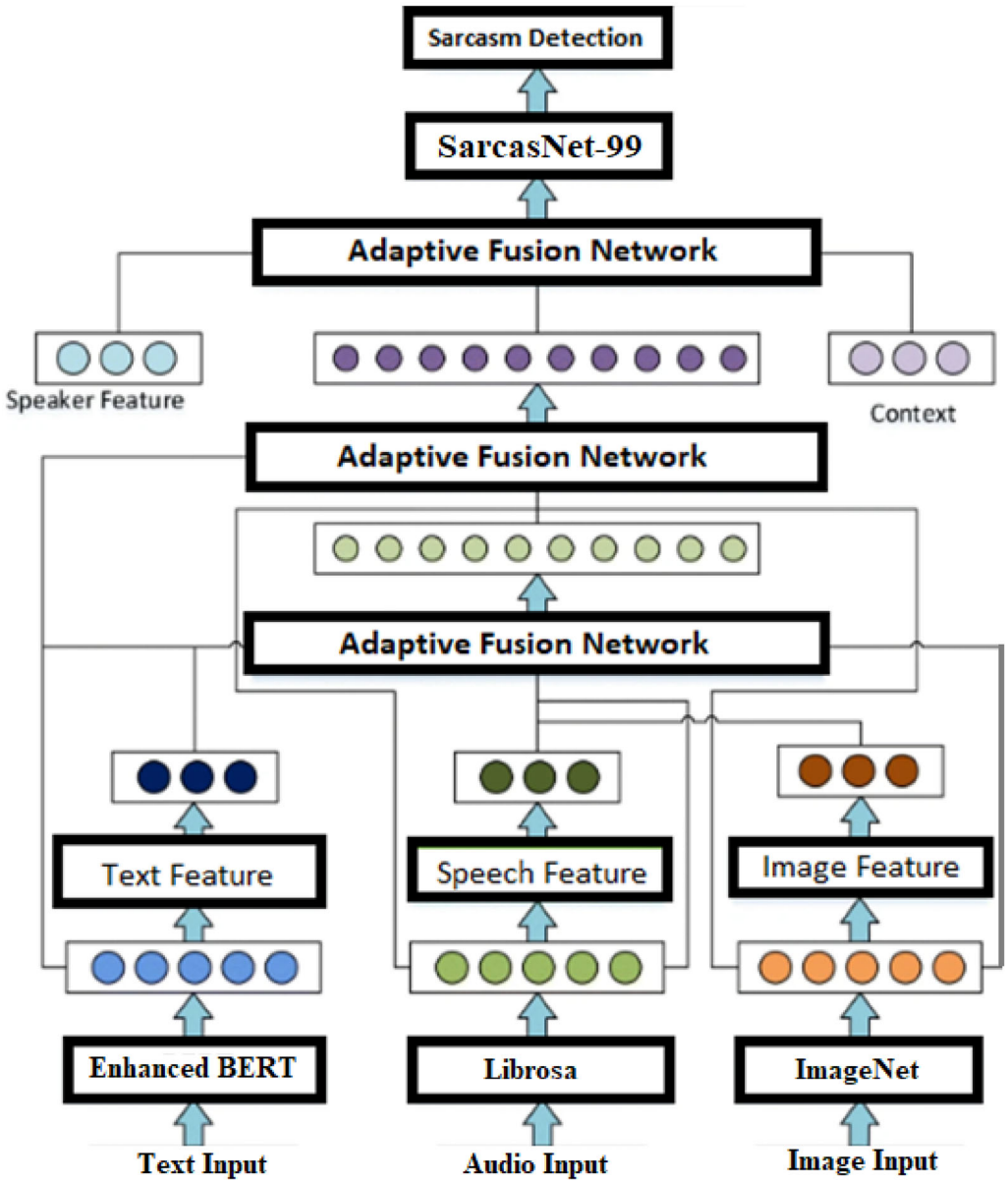

Fig. 2 Proposed adaptive fusion network

3.3 Data analytics and prediction

3.3.1 Proposed adaptive deep fusion policy

After extracting the each modality, they have to be fused to input to the deep neural network for final task prediction. Here, the text feature is represented by X_{t}, image feature is represented by X_{v} and speech feature is represented by X_{a}. Adaptive indicates the flexible fusion learning strategy of the neural network based on the training. Among most of the learning strategy, deep fusion policy learns the best. Hence, in the proposed work, adaptive deep fusion policy is inculcated to fuse the multi-modal data. Let $F(,\,;\,S)$ be the adaptive deep fusion policy, now the fusion of different features is given in Eq. (10):

$$X \sim F((X_t;\, text)\Delta(X_a;\, image)\Delta(X_v;\, audio);\, S) \quad (10)$$

Here, *"text," "image," "audio," and "S"* are the neural network parameters that are updated based on the gradients. Δ is a fusion operator with F as final fusion vector mapped to X. The final representation of adaptive feature fusion network is given in Fig. 2.

Table 1 Hyperparameters for sarcasm classification using novel SarcasNet-99

Sl. No	Parameters	Choice Made
1	Batch size	64
2	Epochs	50
3	L2 Penalty	0.0001
4	Exponential decay	5000 steps with decay factor 0.95
5	Activation function	LeakyReLU

3.3.2 Novel SarcasNet-99 algorithm for classification

Deep neural networks have shown remarkable success in various natural language processing tasks, including sentiment analysis, text classification, and sarcasm detection. These models leverage their ability to automatically learn complex patterns and representations from data. In the case of sarcasm detection in videos, proposed deep neural network is called "SarcasNet-99" with 99 fully connected layers that can be trained to effectively capture the intricate relationships between linguistic, visual, and acoustic features that were extracted earlier.

The fused features of proposed Adaptive Fusion Network are fed into a classification layer, typically implemented as a fully connected network or a LeakyReLU layer, which predicts the presence or absence of sarcasm in the video. The hyperparameter used for turning are mentioned in Table 1 below.

Sarcasm detection in videos is a complex and challenging task, requiring the integration of linguistic, visual, and acoustic cues. Our deep neural network model, designed specifically for sarcasm detection in videos, leverages the power of deep learning to effectively capture the multi-modal features present in video data. By training the model on large annotated datasets, we can enhance its ability to predict sarcasm accurately, thereby contributing to the advancement of sarcasm detection in video content. The proposed model opens up new possibilities for applications in social media analysis, sentiment analysis, and content moderation, enabling a deeper understanding of the complex nature of sarcasm in the digital era.

The overall learning of the model is performed by minimizing the loss function as shown in Eq. (11):

$$L = L_{\text{task}} + \alpha L_{\text{sim}} + \beta L_{\text{diff}} + \gamma L_{\text{recon}} \tag{11}$$

Here, L_{task}, L_{sim},, L_{diff}, L_{recon} denote loss functions. Regularization of loss function $L2$ is carried is determined by interaction weights α, β. To achieve the desired result, each of the loss function is responsible. Now let's see the different loss function listed above:

- L_{task}**:** Task Loss The task-specific loss estimates the quality of prediction during training.
- L_{sim}**:** Similarity Loss is calculated using Cross Modality Discrepancy for adaptive deep fusion strategy
- L_{diff}**:** Difference Loss This loss is to ensure that the loss aspects of different modalities like text, image and speech after modality representations.
- L_{recon}**:** Reconstruction Loss ensures the hidden representations to capture details of their respective modality.

4 Evaluation results

The study aimed to investigate the specific contribution of different types of information in detecting sarcasm. A number of experiments were conducted to assess the effectiveness of each type of information, as well as different combinations of these types. To overcome the problem of over fitting during training, a information expansion method was proposed. Fivefold cross-validation on the dataset was conducted and used the average of the results to evaluate the classifier. The information expansion method was only applied to the training data during each fold.

4.1 Dataset details

TedX Dataset is used for model training and testing and contains 10,000 + video clips extracted from YouTube using the search term "TED talks". These videos are nothing but the speaker's upper body with a maximum of 384 pixel height. The static videos are eliminated where the speaker was not delivering any presentation.

The GIF Reply dataset (https://github.com/xingyaoww/gif-reply/) that has been made available includes a total of 1,562,701 instances of real conversations on Twitter that involve both text and GIFs. Throughout these conversations, a total of 115,586 distinct GIFs were used. Additionally, certain metadata is included with some of the GIFs, such as OCR-derived text, annotated tags, and object names.

4.2 Information expansion for image quality

Proposed Information Expansion methods include:

- Use super pixels to improve the picture quality.
- Apply a blur effect using Gaussian, mean, or median filter.
- Sharpen the image to make it clearer.
- Add an emboss effect to create a 3D illusion on the image.

- Detect edges in the original image, assign them a value of 0 or 255, and superimpose them on the original image.
- Add Gaussian noise to the image to introduce randomness.
- Set a certain percentage of pixels to black or replace them with black squares.
- Invert the intensity of some pixels with a probability of 5%.
- Randomly add or subtract a number between -10 and 10 to each pixel in the image.
- Multiply each pixel in the image by a random number between 0.5 and 1.5.
- Adjust the contrast of the entire image by halving or doubling it.
- Distort the local area of the image to create interesting effects.
- Move the pixels around to create a sense of motion or fluidity in the image.

4.3 Experimental setup

The proposed SarcasNet-99 model is tested against the existing state-of-the-art techniques such as AlexNet, DenseNet, SqueezeNet and ResNet. The taxonomy of each of the algorithms is discussed as follows:

1. **AlexNet:**

AlexNet is a seminal deep convolutional neural network (CNN) architecture. It played a crucial role in popularizing deep learning for computer vision tasks. The network consists of five convolutional layers followed by three fully connected layers. It introduced novel features such as rectified linear units (ReLU) for activation and local response normalization (LRN) for normalization. AlexNet's architecture is defined by the following formula:

- Convolutional Layer: Conv(filter size, number of filters, stride, padding)
- ReLU Activation: ReLU()
- Max Pooling Layer: MaxPool(pool size, stride)
- Fully Connected Layer: Dense(number of units)
- Softmax Activation: Softmax()

2. **DenseNet:**

DenseNet is a densely connected convolutional network architecture that addresses the vanishing gradient problem. It introduces skip connections between all layers, enabling each layer to directly access the feature maps of preceding layers. This dense connectivity enhances information flow and encourages feature reuse. DenseNet's formula is as follows:

- Dense Block: [Conv(filter size, number of filters), ReLU()]*N
- Transition Layer: [Conv(filter size, number of filters), ReLU(), AvgPool(pool size, stride)]

3. **SqueezeNet:**

SqueezeNet is a compact CNN architecture designed to reduce model size while maintaining accuracy. It achieves this by employing 1×1 pointwise convolutions to reduce the number of input channels and expand them back to capture complex patterns. SqueezeNet also incorporates fire modules consisting of squeeze and expand layers. The formula for SqueezeNet is as follows:

- Fire Module: [Conv(1×1, squeeze filters), ReLU(), Conv(1×1, expand filters), ReLU()]
- Skip Connection: Concatenate()
- Convolution Layer: Conv(filter size, number of filters)
- ReLU Activation: ReLU()

4. **ResNet:**

ResNet (short for Residual Network) is a groundbreaking CNN architecture that introduces residual connections to alleviate the vanishing gradient problem. Residual connections enable the network to learn residual mappings by directly propagating the original input to subsequent layers. This architecture facilitates the training of extremely deep networks. The formula for ResNet is as follows:

- Residual Block: [Conv(filter size, number of filters), BatchNorm(), ReLU(), Conv(filter size, number of filters), BatchNorm()] + Skip Connection
- Shortcut Connection: Addition or Concatenate()
- ReLU Activation: ReLU()

The metrics used for comparison are the activation functions that play a very important role in the performance of any neural network. The activation function used here are Sigmoid, Tanh, ReLU and LeakyReLU. The taxonomy for the sane is given below.

Activation functions are essential components of neural networks that introduce nonlinearity, allowing models to learn complex patterns and make accurate predictions. Here are brief explanations of four popular activation functions along with their formulas:

Table 2 Performance of proposed SarcasNet algorithm with *TedX* dataset

Sl. No	Method	Sigmoid	Tanh	ReLU	LeakyReLU
1	AlexNet	78.12	76.2	80.15	83.20
2	DenseNet	54.11	59.23	83.45	78.50
3	SqueezeNet	83.12	74.48	83.50	79.25
4	ResNet	87.45	92.30	94.84	87.66
5	Proposed SarcasNet-99	91.60	91.23	93.45	98.45

Table 3 Performance of proposed SarcasNet algorithm with GIF reply dataset

Sl. No	Method	Sigmoid	Tanh	ReLU	LeakyReLU
1	AlexNet	74.12	74.54	90.15	78.11
2	DenseNet	65.12	56.89	88.76	54.13
3	SqueezeNet	77.34	69.11	79.50	68.54
4	ResNet	82.13	84.54	89.56	85.39
5	Proposed SarcasNet-99	89.38	92.58	96.25	99.56

1. *Sigmoid:* The sigmoid function is a smooth, S-shaped curve that squashes the input into the range (0, 1). It is commonly used in binary classification tasks where the output represents the probability of belonging to a particular class. The formula for the sigmoid activation function is shown in Eq. (11):

$$\sigma(x) = 1/(1 + exp(-x)) \tag{11}$$

2. *Tanh:* The hyperbolic tangent (tan$h$) function is similar to the sigmoid function but maps the input to the range $(-1, 1)$. It is symmetric around the origin and introduces negative values. The tanh function is effective in capturing both positive and negative relationships in the data. The formula for the tanh activation function is shown in Eq. (12):

$$tanh(x) = (exp(x) - exp(-x))/(exp(x) + exp(-x)) \tag{12}$$

3. *ReLU (Rectified Linear Unit):* The rectified linear unit (ReLU) is a popular activation function that has gained prominence in deep learning. It replaces negative values with zero, effectively introducing nonlinearity. The ReLU function is defined as shown in Eq. (11):

$$ReLU(x) = max(0, x) \tag{13}$$

4. *LeakyReLU:* The Leaky ReLU is a variant of the ReLU function that addresses the issue of "dying" neurons by allowing small negative values. It introduces a small slope for negative inputs, which helps alleviate the vanishing gradient problem. The formula for the LeakyReLU activation function is as follows:

$$LeakyReLU(x) = max(\alpha x, x) \tag{14}$$

where α is a small positive constant (e.g., 0.01).

Tables 2 and 3 represent the results of different methods or models on the task sarcasm detection over two different datasets, where each row corresponds to a specific method/model, and each column represents a different activation function used in the model. The activation functions compared in this table are Sigmoid, Tanh, ReLU (Rectified Linear Unit), and LeakyReLU (Leaky Rectified Linear Unit).

These results indicate the performance of each model with different activation functions. The higher the accuracy percentage, the better the model's performance on the given task. In this case, Proposed SarcasNet-99 achieved the highest accuracy overall, particularly when using the LeakyReLU activation function. The graphical analysis of the same is given in Figs. 3 and 4.

5 Conclusion and future scope

Sarcasm detection is the ability to identify when someone is using sarcasm in their speech or writing. It is an important skill for natural language processing models, as sarcasm can change the meaning of a sentence entirely. Many approaches have been proposed, including using contextual information and linguistic cues. However, these methods are not always sufficient in analyzing the underlying sarcasm, as emotional expressions can change with social circumstances over time. To address these challenges, a new approach called "A Smart Video Analytical framework for Sarcasm Detection using Deep Learning" was introduced. The proposed methodology had video as its input streamed over real-time using apache storm distributed framework in Data Collection module. Later the video feature extraction was done as text,

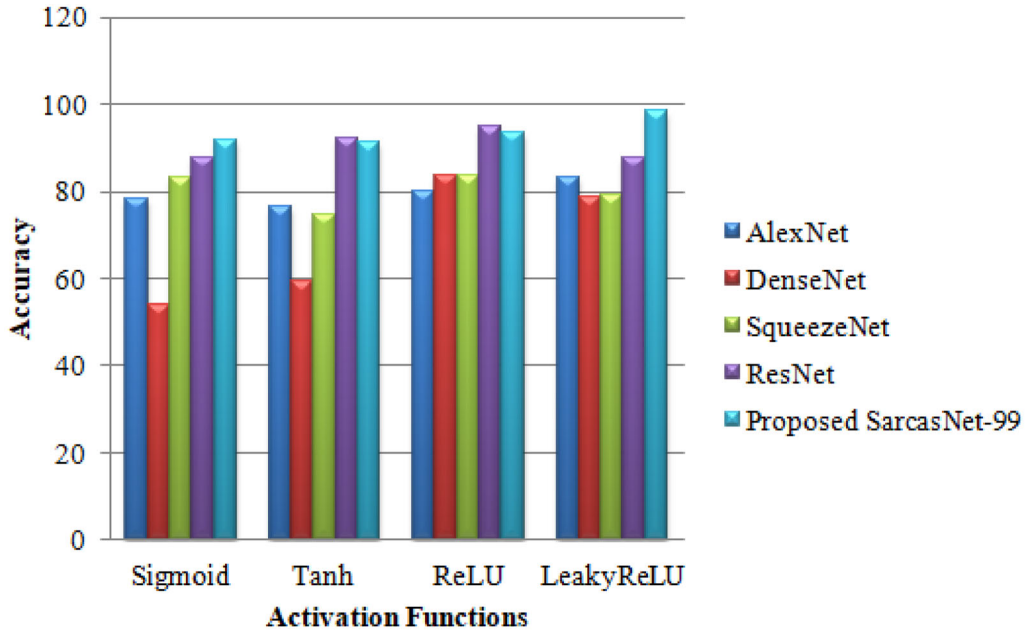

Fig. 3 Performance graphs of proposed SarcasNet algorithm with *TedX* dataset

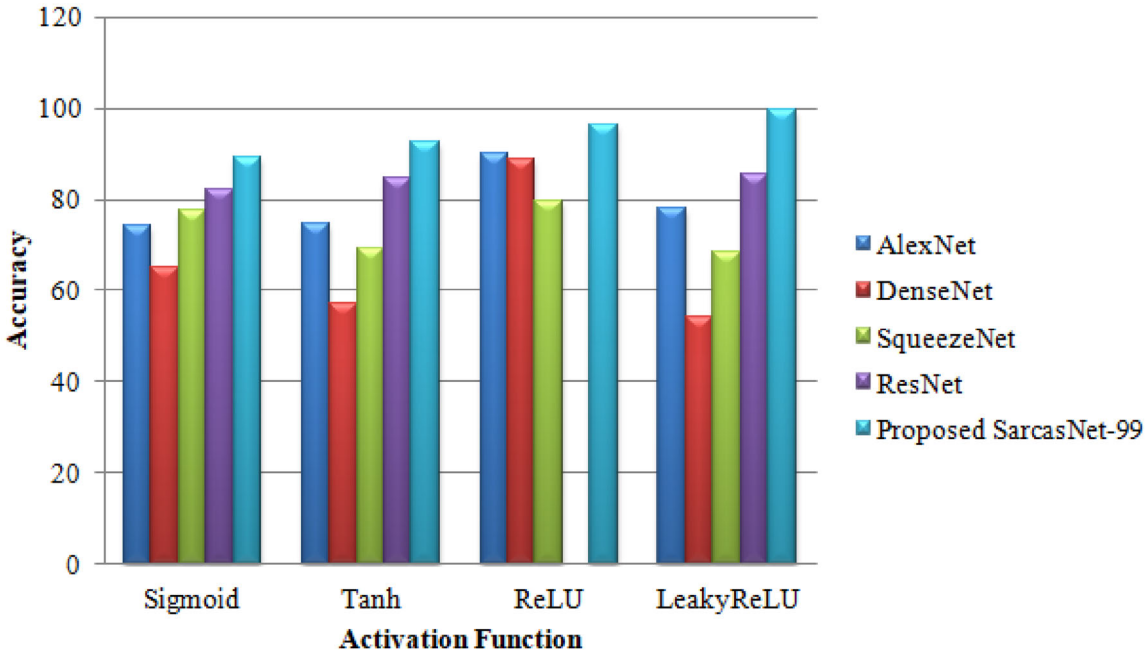

Fig. 4 Performance graphs of proposed SarcasNet algorithm with *GIF Reply* dataset

image, and audio using BERT, SarcasNet-99, and Librosa, respectively. Each modality is addressed individually and then fused using proposed adaptive early fusion approach. The final task prediction is done using proposed deep neural network called "SarcasNet-99" to detect sarcasm in videos. The proposed model was trained and tested on the TedX and GIF Reply Datasets with over 10,000 video clips. Compared to existing state-of-the-art techniques, the proposed model outperformed as one of the best model fit. Hyperparameter tuning with LeakyReLU suppression improved the precision and F1 score by 10%, resulting in a final accuracy of 99.005%.

Acknowledgements This research was supported by Ramaiah Institute of Technology (MSRIT), Bangalore-560054 and Visvesvaraya Technological University, Jnana Sangama, Belagavi-590018.

Data availability The datasets generated during and/or analyzed during the current study are available in the MultiComp Lab repository, http://multicomp.cs.cmu.edu/resources/.

Declarations

Conflict of interest There is no conflict of interest.

Ethics approval We did not use animals and Human participants in the study reported in this work.

Informed consent For this type of study informed consent is not required.

Consent for publication For this type of study consent for publication is not required.

References

1. Chatterjee, S., Bhattacharjee, S., Ghosh, K., Das, A.K., Banerjee, S.: Class-biased sarcasm detection using BiLSTM variational autoencoder-based synthetic oversampling. Soft. Comput. **8**, 1–8 (2023). https://doi.org/10.1007/s00500-023-08045-8
2. Moores, B., Mago, V.: A survey on automated sarcasm detection on Twitter (2022). arXiv preprint https://doi.org/10.48550/arXiv.2202.02516
3. Rahma, A., Azab, S.S., Mohammed, A.: A comprehensive review on arabic sarcasm detection: approaches, challenges and future trends. IEEE Access **8**, 24 (2023)
4. Bhat, A., Jha, G.N.: Sarcasm detection of textual data on online socialmedia: a review. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1981–1985. IEEE (2022). https://10.0.4.85/ICACITE53722.2022.9823869
5. Dutta, P., Bhattacharyya, C.K.: Multi-modal sarcasm detection in social networks: a comparative review. In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 207–214. IEEE (2022). https://10.0.4.85/ICCMC53470.2022.9753981
6. Vinoth, D., Prabhavathy, P.: An intelligent machine learning-based sarcasm detection and classification model on social networks. J. Supercomput. **78**(8), 10575–10594 (2022). https://doi.org/10.1007/s11227-023-05071-z
7. Godara, J., Batra, I., Aron, R., Shabaz, M.: Ensemble classification approach for sarcasm detection. Behav. Neurol. **22**, 2021 (2021). https://doi.org/10.1155/2021/9731519
8. Li, L., Levi, O., Hosseini, P., Broniatowski, D.A.: A multi-modal method for satire detection using textual and visual cues (2020). arXiv preprint https://doi.org/10.48550/arXiv.2010.06671
9. Muaad, A.Y., Jayappa Davanagere, H., Benifa, J.V., Alabrah, A., Naji Saif, M.A., Pushpa, D., Al-Antari, M.A., Alfakih, T.M.: Artificial intelligence-based approach for misogyny and sarcasm detection from Arabic texts. Comput. Intell. Neurosci. **26**, 2022 (2022). https://doi.org/10.1155/2022/7937667
10. Ahuja, R., Sharma, S.C.: Transformer-based word embedding with CNN model to detect sarcasm and irony. Arab. J. Sci. Eng. **47**(8), 9379–9392 (2022). https://doi.org/10.1007/s13369-021-06193-3
11. Yao, F., Sun, X., Yu, H., Zhang, W., Liang, W., Fu, K.: Mimicking the brain's cognition of sarcasm from multidisciplines for Twitter sarcasm detection. IEEE Trans. Neural Netw. Learn. Syst. **24**, 31 (2021)
12. Liang, B., Lou, C., Li, X., Gui, L., Yang, M., Xu, R.: Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4707–4715 (2021). https://doi.org/10.1145/3474085.3475190
13. Bedi, M., Kumar, S., Akhtar, M.S., Chakraborty, T.: Multi-modal sarcasm detection and humor classification in code-mixed conversations. IEEE Trans. Affect. Comput. (2021)
14. Sharma, D.K., Singh, B., Agarwal, S., Kim, H., Sharma, R.: Sarcasm detection over social media platforms using hybrid autoencoder-based model. Electronics **11**(18), 2844 (2022). https://doi.org/10.3390/electronics11182844
15. Kamal, A., Abulaish, M.: Cat-bigru: convolution and attention with bi-directional gated recurrent unit for self-deprecating sarcasm detection. Cogn. Comput. **1**, 1–9 (2022). https://doi.org/10.1007/s12559-021-09821-0
16. Zhao, X., Huang, J., Yang, H.: CANs: coupled-attention networks for sarcasm detection on social media. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2021). https://10.0.4.85/IJCNN52387.2021.9533800
17. Liang, B., Lou, C., Li, X., Yang, M., Gui, L., He, Y., Pei, W., Xu, R.: Multi-modal sarcasm detection via cross-modal graph convolutional network. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, vol 1: Long Papers, pp. 1767–1777 (2022). https://10.0.72.221/v1/2022.acl-long.124
18. Liu, H., Wang, W., Li, H.: Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. arXiv preprint https://doi.org/10.48550/arXiv.2210.03501
19. García-Díaz, J., Caparros-Laiz, C., Valencia-García, R.: UMUTeam at SemEval-2022 Task 5: combining image and textual embeddings for multi-modal automatic misogyny identification. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp. 742–747 (2022). https://10.0.72.221/v1/2022.semeval-1.103
20. Zuhri, A.T., Sagala, R.W.: Irony and sarcasm detection on public figure speech. J. Elem. School Educ. **1**(1), 41–45 (2022)
21. Ray, A., Mishra, S., Nunna, A., Bhattacharyya, P.: A multimodal corpus for emotion recognition in sarcasm (2022). arXiv preprint https://doi.org/10.48550/arXiv.2206.02119
22. Ding, N., Tian, S.W., Yu, L.: A multimodal fusion method for sarcasm detection based on late fusion. Multimed. Tools Appl. **81**(6), 8597–8616 (2022). https://doi.org/10.1007/s11042-022-12122-9
23. Khan, S., Kamal, A., Fazil, M., Alshara, M.A., Sejwal, V.K., Alotaibi, R.M., Baig, A.R., Alqahtani, S.: HCovBi-caps: hate speech detection using convolutional and bi-directional gated recurrent unit with capsule network. IEEE Access **10**, 7881–7894 (2022)
24. Zhang, Y., Ma, D., Tiwari, P., Zhang, C., Masud, M., Shorfuzzaman, M., Song, D.: Stance level sarcasm detection with BERT and stance-centered graph attention networks. ACM Trans. Internet Technol. (2022). https://doi.org/10.1145/3533430
25. Juyal, P.: Multi-modal sentiment analysis of audio and visual context of the data using machine learning. In: 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), pp. 1198–1205. IEEE (2022). https://10.0.4.85/ICOSEC54921.2022.9951988

Jamuna S. Murthy is currently working as an assistant professor in the Department of Computer Science and Engineering at Ramaiah Institute of Technology, Bengaluru. She is pursuing her PhD degree in the area of deep learning. She received Best Project Award for her M.Tech thesis work under year 2017. Her research interests include real-time data analytics, machine learning, natural language processing, bigdata and cloud computing. She also holds a good publication record of peer-reviewed international conferences, journals and book publications. She is a distinguished member of National Society of Professional Engineers (NSPE), ISTE, IEEE, and CSI.

G. M. Siddesh is currently working as Professor and Head in Department of Artificial Intelligence and Data Science, M S Ramaiah Institute of Technology, Bangalore. He is the recipient of Seed Money to Young Scientist for Research (SMYSR) for FY 2014–15, from Government of Karnataka, Vision Group on Science and Technology (VGST). He has published a good number of research papers in reputed international conferences and journals. He is a member of ISTE, IETE, etc. He has authored books on network data analytics, statistical programming in R, Internet of Things with Springer, Oxford University Press and Cengage publishers, respectively. He has edited research monographs in the area of cyber-physical systems, fog computing and energy aware computing, and bioinformatics with CRC Press, IGI Global and Springer publishers, respectively. His research interests include Internet of Things, distributed computing and data analytics.