Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features

Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features

Eftekhar Hossain [a], Omar Sharif [b], Mohammed Moshiul Hoque [b,*], M. Ali Akber Dewan [c], Nazmul Siddique [d], Md. Azad Hossain [a]

[a] Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering & Technology, Chittagong 4349, Bangladesh
[b] Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Chittagong 4349, Bangladesh
[c] School of Computing and Information Systems, Faculty of Science and Technology, Athabasca University, Athabasca, AB T9S 3A3, Canada
[d] School of Computing, Engineering and Intelligent Systems, Ulster University, Londonderry BT47 7JL, UK

ARTICLE INFO

Article history:
Received 3 March 2022
Revised 16 June 2022
Accepted 17 June 2022
Available online xxxx

Keywords:
Multimodal learning
Multimodal data
Multilingual offense detection
Ensemble
Multimodal fusion

ABSTRACT

In recent years, memes have become a common medium of promulgating offensive views by the content polluters in social media. Due to their multimodal nature, memes can easily evade the content regulators' eyes. The proliferation of these undesired or harmful memes can cause a detrimental impact on social harmony. Therefore, restraining offensive memes on social media is of utmost importance. However, analyzing memes is very complicated as they implicitly express human emotions. Previous studies have not explored the joint modelling of multimodal features and their counteractive unimodal features (i.e., image, text) to classify undesired memes. This paper presents a framework that utilizes the weighted ensemble technique to assign weights to the participating visual, textual and multimodal models. The state-of-the-art visual (i.e., VGG19, VGG16, ResNet50) and textual (i.e., multilingual-BERT, multilingual-DistilBERT, XLM-R) models are employed to make the constituent modules of the framework. Moreover, two fusion approaches (i.e., early fusion and late fusion) are used to combine the visual and textual features for developing the multimodal models. The evaluations have demonstrated that the proposed weighted ensemble technique improves the performance over the investigated unimodal, multimodal, and ensemble models. The result shows that the proposed approach achieves superior outcomes on two multilingual benchmark datasets (MultiOFF and TamilMemes), with 66.73% and 58.59% weighted f_1-scores, respectively. Furthermore, the comparative analysis reveals that the proposed approach outdoes other existing works by improving approximately 13% and 2% weighted f_1-score gain.

1. Introduction

With the phenomenal rise of social media platforms, the world is witnessing a growing epidemic of online offensive and abusive behaviour. A significant portion of social media users has either experienced or witnessed some form of online offense (Duggan, 2017). In these platforms, the users have the freedom to post, comment or share content without modification or intervention of any legal authority (Jørgensen and Zuleta, 2020). This freedom allows some malign users to dispense offensive content, spread rumor/fake news, harass communities or individuals and damage communal harmony. This proliferation of objectionable content in public spaces has detrimental impacts on society (Bannink et al., 2014). Therefore, to maintain social harmony and ensure the quality of the social network ecosystem it is important to expel such content. To date, many works have been conducted to detect and mitigate the spread of objectionable content on online platforms. The majority of the works (Aroyehun et al., 2018; Pavlopoulos et al., 2019; Sharif et al., 2021) focused on only textual modality to identify troll and offensive contents. The SemEval offensive language identification task provides a multilingual dataset to detect the type and target of offensive texts (Zampieri et al., 2020). Kumar et al. (2020) summarize the system's outcome developed on the multilingual troll and aggression dataset. Developing a system that can automatically flag offensive contents is still an arduous problem due to the implicit nature, multi-modality and complicated structure of the contents. The inherent ambiguity of language, computational complexity to audit a large amount of

* Corresponding author.
E-mail addresses: eftekhar.hossain@cuet.ac.bd (E. Hossain), omar.sharif@cuet.ac.bd (O. Sharif), moshiul_240@cuet.ac.bd, moshiul240@cuet.ac.bd (M.M. Hoque), adewan@athabascau.ca (M.A. Akber Dewan), nh.siddique@ulster.ac.uk (N. Siddique), azad@cuet.ac.bd (M.A. Hossain).

https://doi.org/10.1016/j.jksuci.2022.06.010
1319-1578/

Please cite this article as: E. Hossain, O. Sharif, Mohammed Moshiul Hoque et al., Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2022.06.010

content, the issue of low-resource language, and the contextual understanding of natural language are the major obstacles (Zhou et al., 2021; Davidson et al., 2017). Moreover, the mode of communication in social media platforms is dramatically transforming day by day. To deceive the existing NLP system for offensive language detection, content polluters adapt new strategy to the changing system.

Posting and sharing *memes* has recently become a popular form of modality to disseminate information on social media since *memes* can propagate information humorously or sarcastically. A *meme* is an image or screenshot with some text embedded into it. Offensive content creators combine image and text in such a way that can attract and mislead the viewers. They often misstate or fabricate the fact with highly sentimental content to facilitate rapid dissemination. Consider the example of Fig. 1 (c), the image is benign, which shows the photographs of two south Indian actors. However, together with the caption, it insults their marriage by indicating their age gap. It is cumbersome to correctly infer the meaning of a *meme* considering only visual or textual modality. This multimodal nature of the memes makes it very challenging to differentiate between benign and malign contents. It also aids the propagation of abusive content. Such memes are increasingly used as a way to abuse individuals or attack communities based on their race, gender, religion, sexual orientation, or physical appearance (Williams et al., 2016; Drakett et al., 2018). The pervasiveness of these contents poses a direct threat to social peace and communal harmony.

It is a challenging task to develop an automated system that can detect offensive memes. Developing a multimodal offense detection system is intrinsically tricky and complex because it requires a holistic understanding of visual and textual information. The implicit meaning of the memes, presence of ambiguous, humorous, sarcastic terms and usage of attractive, comical, theatrical images have made meme classification even more complicated. Moreover, the absence of baseline methods to capture features from multiple modalities and the prevalence of multilingual texts have further increased the complexity. Despite the growing body of research in meme analysis, these issues are not addressed to date. Suryawanshi et al. (2020a) applied the late fusion technique to combine multimodal features. Their work employed stacked LSTM and VGG16 to extract textual and visual features. In another work, authors classify troll memes using image features without considering the textual features (Suryawanshi et al., 2020b). Sharma et al. (2020) organized a SemEval task to analyze the sentiment and humour of the memes. Their study revealed that the multimodal fusion techniques are effective in combining visual and textual features. Few works train textual and visual models independently and combine the models outcome rather than training a joint multimodal network (Morishita et al., 2020; Bonheme et al., 2020). Most past studies considered only a single modality (image or text) for offense or troll detection, but they did not exploit the advanced techniques to extract the multimodal features. To utilize the multimodal features concerning modalities should be simultaneously processed. Therefore, the key research question explored in this work is how to develop a framework leveraging features from visual and textual modality to identify offense and troll from memes.

This work proposes a multimodal architecture to learn joint representation simultaneously from visual and textual modality to address the research question mentioned above. The proposed architecture comprises four constituent modules: (i) Visual feature extraction module, (ii) Textual feature extraction module, (iii) Multimodal decision fusion module and (iv) Multimodal feature fusion module. Each of the modules is trained independently. To extract image features, pre-trained visual (i.e., VGG16, ResNet50, Inception, Xecption) models are used. Extensive investigation is carried out with deep neural networks (i.e., CNN, BiLSTM, Attention) and transformers (i.e., m-BERT, Distil-BERT, XLM-R) to extract the textual features. Decision and feature fusion modules are responsible for performing aggregation of the extracted features. We perform extensive experimentation on English offensive meme (Suryawanshi et al., 2020a), and Tamil troll meme (Suryawanshi et al., 2020b) dataset using the modules mentioned above. After investigating models' predictions, this work proposes a weighted ensemble technique that exploits the strength of individual visual, textual and multimodal models. The proposed method (Section 4.5) can readdress the softmax probabilities of the partaking models depending on their prior results. Moreover, the effectiveness of the proposed model is empirically validated on multilingual datasets. The key contributions of this work illustrate in the following:

- Present the detailed statistics of the dataset that facilitate the preparation of the models providing useful insights.
- Propose a model that exploits visual, textual and multimodal features of the memes. Moreover, we investigate the multimodal decision fusion, and feature fusion approaches with contemporary visual and textual models. Finally, we employ an ensemble technique that automatically assigns appropriate weight to the participating modules based on their prior performance on the dataset.
- Empirically evaluates the proposed model on multilingual (English & Tamil) datasets and demonstrates how ensemble technique can enhance the classifier's performance.
- Perform extensive experimentation and compare the performance with a set of visual, textual, and multimodal models. The proposed model outperforms all other techniques with a significant margin, thus setting up a benchmark to compare with in the future.

(a) Offensive (b) Troll (c) Troll

Fig. 1. Few examples where textual content does not convey any exaggerated views, however, when combine with the visual information, it eventually becomes an offensive/troll meme.

The research outcomes presented in the paper is one of the pioneering works that leverage multimodal features to classify multilingual offense and troll from memes to the best of our knowledge. It expects that the resources and system presented in this paper will facilitate further research in this domain. The remaining of the paper is structured as follows. Section 2 provides a summary of few existing works on undesired language detection concerning unimodal and multimodal approaches. Various terminologies of offense and troll classes are presented with detailed statistics of the dataset in Section 3. Section 4 discusses the techniques, hyperparameters and architectures of the constituent modules of the proposed system. Section 5 reports the experimental findings and extensive error analysis of the models. Section 6 points out the prospects of future development with concluding remarks.

2. Related work

Although a considerable body of works have been conducted to identify troll (Mojica, 2018; Mut Altin et al., 2020), aggression (Safi Samghabadi et al., 2020; Aroyehun et al., 2018), hate speech (Basile et al., 2019; Fortuna and Nunes, 2018) and abusive (Pamungkas et al., 2019; Vidgen et al., 2019) contents from a single modality (i.e. image, text), it is often cumbersome to understand and categorize the contents of a meme considering only one modality. Therefore, it is important to investigate both visual and textual modalities to detect offensive memes. However, researches focus on detecting such contents from multiple modalities is still in infancy. This section briefly summarizes previous works on undesired contents (i.e., offense, abuse, hate, aggression, troll) detection considering unimodal and multiple modalities.

2.1. Unimodal based undesired contents detection

In the past few years, a series of tasks have been organized to identify offense (Zampieri et al., 2020; Chakravarthi et al., 2021), abuse (Roberts et al., 2019; Akiwowo et al., 2020), hate speech (Mandl et al., 2020; Bosco et al., 2018) and troll (Kumar et al., 2020; Kumar et al., 2018) from social media. These tasks aimed to detect and categorize abusiveness from multilingual (*English, Arabic, Greek, Tamil, Hindi, and Bengali*) texts. Zampieri et al. (2019) develop an English offensive language text dataset. Baseline experimentation is performed with CNN, BiLSTM and SVM techniques where CNN obtained the maximum macro-f_1 score of 0.80 for the detection task. Wang et al. (2020) applied a knowledge distillation method on soft labels to categorize multilingual offensive texts. Tulkens et al. (2016) trained multiple SVMs with handcrafted dictionary-based features to identify racist texts. Their system achieved a f_1-score of 0.46, although it does not care about the context of the texts. Zhou et al. (2020) employed the deep learning-based fusion approach to identify hate in SemEval-2019 dataset (Basile et al., 2019). Their work applied CNN, BERT, and ELMo to extract the textual features. Fusion of BERT and CNN achieved the highest weighted f_1-score of 0.947. Sharif and Hoque (2021) built an aggressive text identification corpus in Bengali using hierarchical annotation schema. They applied a wide range of machine and deep learning techniques. The combined CNN and BiLSTM acquired the best f_1-score of 0.87 and 0.80 in coarse and fine-grained classification. Saha et al. (2021) employed a genetic algorithm-based ensemble strategy to identify offense from multilingual texts. Transformers (BERT, mBERT, DistilBERT) have been used as the ensemble base and achieved 0.78, 0.74 and 0.97 weighted f_1-score in Tamil, Malayalam and Kannada languages, respectively. A recent work (Sharif et al., 2021) showed that transformer-based models outdo ML and DL based methods to detect multilingual offensive texts. Statistical features (number of comments, replies, positive, negative votes) are utilized to find trolls in news community forums by Mihaylov et al. (2015). SVM technique with RBF kernel obtained 82–95% accuracy for various feature combinations. Andrew (2021) performed experimentation with SVM, LR, RF, KNN to detect offensive code-mixed YouTube comments. Their work did not consider any semantics and contextual features for the classification. Davidson et al. (2017) offered a multiclass hate speech dataset of 25K English tweets. Logistic regression with l2 regularizer and term frequency inverse document frequency (tf-idf) feature achieves 0.90 macro f_1-score. Bhardwaj et al. (2020) applied SVM, LR, RF and MLP techniques with m-BERT embedding to detect multi-label hostile Hindi posts where SVM achieved the highest f_1-score of 0.84 in coarse-grained classification. Their work did not adopt any deep learning methods to extract the sequential features. Gambäck et al. (2017) tried CNN to classify tweets into four (*racism, sexism, racism* & *sexism, non-hate*) classes. Experimentation is carried out with random vectors, Word2Vec and character n-grams where the model acquired 0.78 f_1-score with Word2Vec features. Sadiq et al. (2021) developed a combined CNN-BiLSTM based method over a cyber-troll dataset of 20 k tweets. This system can identify cyber-aggressive texts with 92% accuracy, but its performance is inferior for short texts.

Very few researches have been conducted focusing on image-based features to detect offense and troll since existing models largely depend on textual features. Gandhi et al. (2019) developed a system to detect and remove offensive contents form e-commerce catalog. Pre-trained visual models are employed that achieved f_1-score of 0.62. Suryawanshi et al. (2020b) released a dataset containing troll and not-troll memes in Tamil. They used pre-trained (ResNet, MobileNet) image classification methods to differentiate between meme classes. Although the system achieved a 0.52 macro f_1-score, it performed poorly in the troll class with the recall value of 0.37. This system is failed when the same image with different texts has a heterogeneous interpretation. Manoj and Chinmaya (2021) developed a visual feature-based meme classification model. They directly employed the ResNet50 model without any modifications in the layers, resulting in a very poor weighted f_1-score of 0.48. A CNN based system is proposed to identify aggression from symbolic images (Kumari et al., 2019) which achieved a weighted f_1-score of 0.89 on a holdout validation set. Connie et al. (2017) developed a CNN based adult content recognition system. Their system used a weighted sum of multiple CNNs, which outperformed a single and average weighted CNN.

2.2. Multimodal based undesired contents detection

Recently, multimodal learning has gained much attention due to its ability to efficiently combine information from multiple modalities into a single learning framework (Morency and Baltrušaitis, 2017). This method already showed good performance on tasks that involve both visual and linguistic understanding such as Visual Question Answering (Hudson and Manning, 2019) and Visual Reasoning (Suhr et al., 2018). Therefore, researchers are adopting the multimodal technique to detect offensive content from memes since such contents have detrimental impact on society (Mishra et al., 2019). To advance research in this domain, Facebook launched a challenge to detect hate speech from multimodal memes (Kiela et al., 2020). To address this challenge, Lippe et al. (2020) developed a multimodal framework using an ensemble of UNITER (UNiversal Image-TExt Representation) (Chen et al., 2020) which received 0.8053 AUROC scores. Velioglu and Rose

(2020) proposed a solution with VisualBERT which is a "BERT variant of vision and language" (Li et al., 2019). They adopted an ensemble strategy that helps to achieve an accuracy of 0.765. Few other works have also aggregated linguistic and visual information to detect hateful memes and gained promising performance (Zhang et al., 2020; Das et al., 2020; Sandulescu, 2020). Gomez et al. (2020) offered a multimodal hate speech dataset containing images and corresponding tweets. Exploration was carried out with unimodal and multimodal architectures, but results revealed that multimodal methods could not outdo the unimodal counterparts. Perifanos and Goutsos (2021) developed a multimodal dataset considering *hateful, xenophobic and racist* tweets. They applied pre-trained Resnet and BERT models for extracting visual and textual features that achieved a weighted f_1-score of 0.947. Rather than BERT, authors did not employ other variants such as mBERT, XLM-R which might improve the performance. Nakamura et al. (2020) introduced a benchmark dataset for multimodal fake news detection. The authors developed a hybrid (text + image) model to perform fine-grained classification. Maximum accuracy on different classes is achieved with pre-trained BERT (text) and ResNet50 (image) models. Xue et al. (2021) proposed a novel multimodal consistency network leveraging the multimodal fusion technique. This method is validated in four widely used multimodal datasets. In another similar work, cross-modal attention residual and multichannel convolutional neural networks were adopted by Song et al. (2021). Kumari et al. (2021) proposed a hybrid model where pre-trained VGG-16 is employed to pick out the image features while layered CNN extracted the textual features. These features are optimized by binary particle swarm optimization technique that helps to achieve 0.74 weighted f_1-score. The authors do not experiment with any transformer-based models to comprehend the textual features. Hosseinmardi et al. (2016) showed that user metadata and visual features are useful to predict cyberbullying incidents. A variety of textual, visual and multimodal features are analyzed to detect cyberbullying events by Singh et al. (2017). Their results showed that aggregation of both features helps to improve the model's performance. In a similar work, the authors presented a CNN based unified representation of text and image to detect cyberbullying (Kumari et al., 2020). In the extended work, they optimized the features using Genetic Algorithm (Kumari and Singh, 2021). Results indicate that model's performance has been improved about 4% with the updated set of features.

Suryawanshi et al. (2020a) built a multimodal dataset of 743 offensive and non-offensive memes related to the 2016 U.S. presidential election. They adopted the early fusion approach to combine the multimodal features. Although the combined model obtained a 0.50 f_1-score, the text-based CNN model outperformed this by achieving a f_1-score of 0.54. A shared task is organized in EACL-2021 to classify multimodal troll memes (Suryawanshi and Chakravarthi, 2021). The dataset contains images and associated transcripted texts of the memes. Li (2021) developed a multimodal model leveraging the pre-trained BERT and ResNet152 architectures. The multimodal attention layer is applied to map text and image features in the same semantic space in this work. The developed model won the shared task by achieving the weighted f_1-score of 0.55. Hossain et al. (2021) put together image and text features using the late fusion approach. In the multimodal approach, BiLSTM is employed to extract the textual features while it can be done with transformers. Results revealed that the textual model with XLNet outdoes others by obtaining the f_1-score of 0.52. Hegde et al. (2021) experimented with a state-of-the-art vision transformer to extract the image features. However, the system does not perform well and achieved only 0.46 f_1-score. Mishra and Saumya (2021) combined features from image and text modalities using a hybrid approach. They used CNN and BiLSTM to obtain the image and text features. The system performed very poorly and attaining only a f_1-score of 0.30. Table 1 presents a summary of few works concerning the modality of the dataset, methods, results and their limitations.

The majority of the studies discussed earlier focused on meme classification considering either text or image. Existing works employing multimodal techniques for memes classification mostly used the late fusion approach. Very few works have been carried out that explored multimodal fusion approaches to identify offense and troll memes. The proposed work performs extensive experimentation with state-of-the-art visual and textual models. Besides, features from both modalities are combined with early (feature) fusion and late (decision) fusion techniques. Moreover, model architectures (i.e., No. of neurons, No. of layers) and hyperparameters (i.e., epochs, batch size, dropout rate, learning rate) are fine-tuned to get the optimal model. Finally, this work proposes a weighted ensemble model leveraging textual, visual and multimodal features. The proposed model evaluates empirically in English offensive meme (Suryawanshi et al., 2020a) and Tamil troll meme (Suryawanshi et al., 2020b) dataset. Evaluation results exhibit that the proposed model outdoes all the existing techniques and facilitates multilingual offense classification from memes.

3. Description of the task and dataset

The research objective of this work is to develop a framework **(F)** to identify offense and troll from memes. The **F** analyzes a set of memes $M = \{m_1, m_2, \ldots, m_n\}$ and categorize them as offense/troll ($\mathbf{c} = \mathbf{1}$) or not ($\mathbf{c} = \mathbf{0}$). Each meme $m_i \epsilon M$ consists of visual **(v)** and textual **(t)** information and the **F** utilize these information to classify m_i. The task is represented as a mapping, $\mathbf{F} : \mathbf{M}(\mathbf{v}, \mathbf{t}) \rightarrow \mathbf{c}\ \epsilon\ (\mathbf{0}, \mathbf{1})$. Following subsections provides the definition of various meme classes and a brief analysis of datasets.

3.1. Task definition

Two benchmark datasets have been utilized to attain the goal: (i) English offensive meme or MultiOFF (Suryawanshi et al., 2020a), and (ii) Tamil troll meme or TamilMemes (Suryawanshi et al., 2020b). For ease of understanding, MultiOFF and TamilMemes datasets are denoted as dataset-1 (D1) and dataset-2 (D2), respectively. The first dataset contains offensive and non-offensive memes related to U.S. presidential election. The second consists of troll and not-troll memes where captions are written in Tamil-English code mixed language. Previous studies (Suryawanshi et al., 2020a; Suryawanshi et al., 2020b) have manually accumulated these memes from various social media platforms such as Facebook, Whatsapp, Instagram, Twitter and Pinterest. It is crucial to have a clear understanding of the class labels to develop a successful computational model. The authors (Suryawanshi et al., 2020a; Suryawanshi et al., 2020b) defined offense and troll as the following:

- **Offense:** memes that spread an idea/emotion with the intention to demean social identity, harass targeted individuals, community or a minority group.
- **Not-offense:** memes without any offensive content.
- **Troll:** memes which contain offensive texts or images and intend to provoke, offend, abuse or insult individuals, group, or a race.
- **Not-troll:** memes not having any trolling content.

Table 1
Brief literature summary concerning undesired text classification using unimodal and multimodal methods. Here A, F, MF, WF denote accuracy, f_1-score, macro and weighted f_1-score respectively.

Article	Modality of Dataset	Approach	Results	Limitation/Gap
Zampieri et al. (2019)	Text [English tweets]	CNN, SVM, BiLSTM	0.80 (MF)	Model biased towards not offensive class. Performance degrades as the number of class increases
Gambäck et al. (2017)	Text [English tweets]	CNN with word embedding and character ngrams	0.78 (F)	Incapable of capturing sequential features as recurrent networks are not used
Tulkens et al. (2016)	Text [Dutch posts]	SVM with dictionary based features	0.46 (F)	Failed to capture the context
Mihaylov et al. (2015)	Text [English trolls]	Statistical features applied on SVM with RBF kernel	0.82–0.96 (A)	Content features (keywords, named entities, topics) and other ML methods are not considered
Andrew (2021)	Text [Tamil, Malayalam and Kannada posts]	Tf-idf features employed on set of baseline machine learning classifiers	[0.61, 0.63, 0.93] (WF)	Used only tf-idf features and no counter measures is taken to handle code-mixing of texts
Bhardwaj et al. (2020)	Text [Hindi comments]	mBERT embedding employed on set of ML classifiers	0.84 (F)	Ignored the sequential information and limited number of training texts in fine-grained classes
Suryawanshi et al. (2020a)	Multimodal meme	Late fusion of stacked LSTM and VGG-16	0.50 (F)	Performance can be improved by pretrained language models
Gandhi et al. (2019)	Image	Pre-trained object detector model (ResNet50, Inception-V3)	0.62 (F)	Do not incorporate the textual signals
Manoj and Chinmaya (2021)	Image	ResNet50	0.48 (WF)	Model is overfitted, much higher deviation between validation and test results
Suryawanshi et al. (2020b)	Image	Variations of ResNet and MobileNet	0.52 (MF)	Embedded texts in the images are ignored
Perifanos and Goutsos (2021)	Multimodal Greek tweets	Combine pretrained BERT and ResNet models	0.94 (F)	Other variants of transformers are not considered rather than BERT
Kumari et al. (2021)	Multimodal posts	VGG-16 and layered CNN with binary particle swarm optimization	0.74 (WF)	Unable to capture the semantic information of the textual modality
Gomez et al. (2020)	Multimodal tweets	Employ feature concatenation, spatial concatenation and text kernel models with CNN + RNN	0.68 (A)	Unimodal models achieve better results than the multimodal ones
Hossain et al. (2021)	Multimodal meme	Late fusion of textual (BiLSTM) and visual (ResNet50, CNN) features	0.52 (WF)	Textual features can be extracted with transformers
Mishra and Saumya (2021)	Multimodal	Combines image and text features using CNN and BiLSTM	0.30 (WF)	Do not employ state of the art models

3.2. Dataset analysis

Each dataset consists of two parts: an image with embedded text and an associated caption. In dataset-1, all the captions are written in English. Most of the captions of dataset-2 are written in Tamil, and a few in Tamil-English code mixed language. Dataset-1 has 743 memes, of which 303 are offensive, and the remaining are not offensive. Dataset-2 is four-time as much more extensive than dataset-1. Out of 2967 instances, 1677 memes are labelled as trolls, while the remaining 1290 memes belong to the not-troll class. For model building and evaluation, datasets are partitioned into three mutually exclusive sets: train, validation and test. A summary of both datasets is presented in Table 2.

The training set is analyzed to get more insights about the data. Table 3 shows the training set statistics, which exhibits both datasets are imbalanced. Not-offensive and troll classes have a higher number of total words and unique words compared to their counterparts. On average, each category on the offensive dataset has 21 words per caption. On the other hand, the captions of the troll dataset are much shorter. The troll class has approximately 12 words per caption, while the not-troll type has only 9 words long. It may be a challenging task to classify trolls due to their shorter text length accurately. Fig. 2 depicts the number of captions that fall into various length ranges for each of the classes. It is observed that approximately 55% of the captions have less than 20 words. Only a fraction of instances have higher than 40 words. This distribution gives an idea of selecting the input text (based on caption length) during the training phase. Finally, Fig. 3 presents few sample memes in each class.

Table 2
Number of instances in train, validation and test set for each dataset.

	Dataset-1		Dataset-2	
	Offensive	Not-Offensive	Troll	Not-Troll
Train	187	258	1026	814
Validation	58	91	256	204
Test	58	91	395	272
Total	303	440	1677	1290

4. Methodology

The primary concern of this work is to classify offense and troll from memes on social media. Usually, memes contained multimodal content such as visual and textual. In order to accomplish the task, we investigate several computational models considering only visual, only textual, and combination of both modalities. State of the art pre-trained convolutional neural networks (i.e., VGG19, VGG16, Xception, InceptionV3, and ResNet50) architectures are employed for visual feature extraction. On the other hand, to obtain textual features, deep recurrent neural networks (i.e., BiLSTM, Attention) and pre-trained transformers (i.e., m-BERT, XLM-R) are applied. This section briefly describes the methods and strategies employed to classify offensive and troll memes. Furthermore, to acquire more robust inferences about the content, both visual and textual features are exploited, and several models are developed by employing multimodal fusion approaches. Fig. 4 shows the abstract view of the overall system. Architectures and parameters of the different modules are discussed in the subsequent subsections.

Table 3
Training set statistics for textual content.

	Class	Total words	Unique words	Max text length (words)	Avg. No. of words per texts
Dataset-1	Offensive	4064	2065	148	21.73
	Not-Offensive	5428	2569	139	21.03
Dataset-2	Troll	12652	6200	61	12.33
	Not-Troll	4402	2487	29	9.39

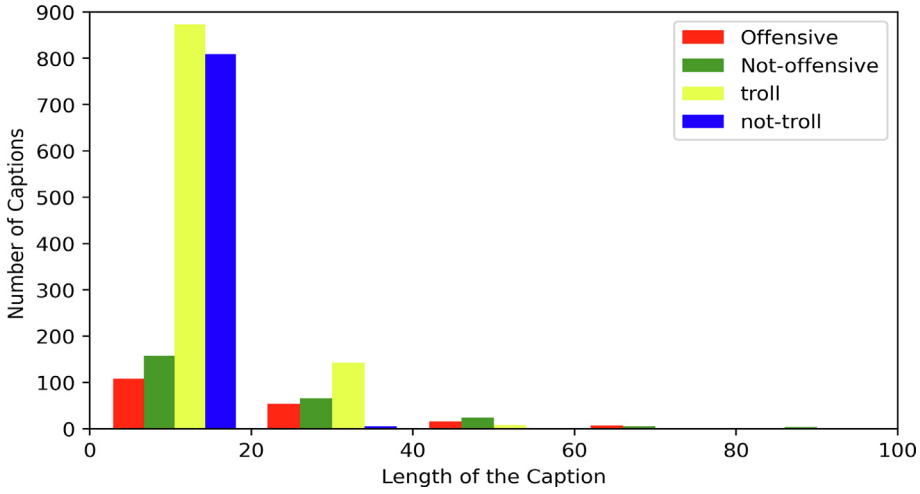
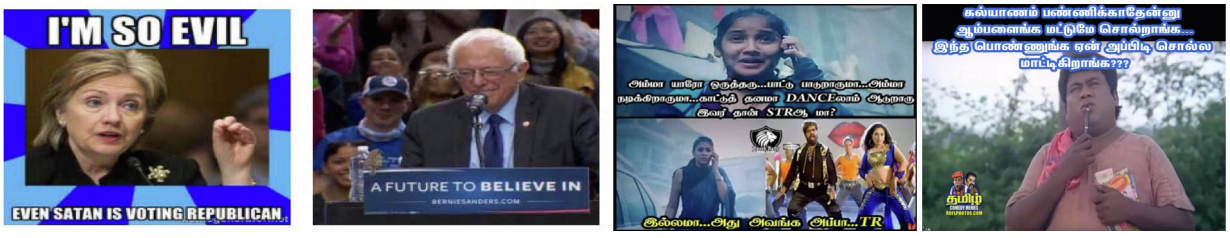


Fig. 2. Distribution of captions with various length in each classes.



(a) Offensive (b) Not-Offensive (c) Troll (d) Not-Troll

Fig. 3. Sample memes of each class: dataset-1 (a,b) and dataset-2 (c,d).

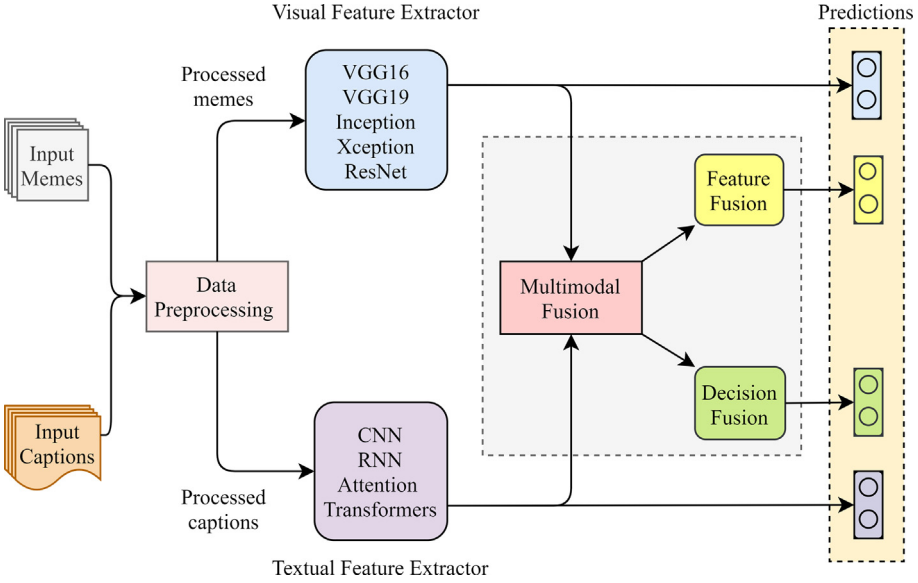


Fig. 4. Abstract view of multimodal offense and troll detection system.

4.1. Data preprocessing

Deep learning techniques are not effective at learning from unprocessed images and texts. Thus, preprocessing is required before feeding them into the networks. For the visual modality, images are transformed into equal size of $150 \times 150 \times 3$. The normalization is performed over the pixel matrix of the images to map the pixel intensity values between 0 and 1. Moreover, the Keras[1] image preprocessing function is used to make the input images suitable before driving them into the CNN models.

For textual modality, deep neural networks (DNN) and transformer-based models are utilized. Both architectures take input in a specific format. For DNN, the input texts are converted into a vector of unique numbers. The mapping of this word to the index is obtained using the Keras tokenizer function. Post padding technique is adopted to get equal length vectors. The maximum text length is determined by analyzing the text length-frequency distribution for each dataset. We choose 50 and 30 as the maximum length for dataset-1 (D1) and dataset-2 (D2), respectively. Similarly, for transformers, we follow the transformer tokenization method for the respective models. After instantiating the tokenizer[2] object, 'encode_plus' method is used to encode the inputs texts. This method adds a special [CLS] and [SEP] token at the start and end of an input text. It also converts the texts into a vector of unique ids and pad 0's to the shorter vectors than the maximum length. Besides, an attention mask is enabled so that the model emphasizes the tokens having unique ids. These 'ids' and 'attention masks' are given as input to the transformer models.

4.2. Visual feature extraction modules

Visual features are extracted by using convolutional neural networks. Rather than developing a custom network, the transfer learning approach is employed in this work. In this approach, parameters of a neural network are trained with a large dataset to solve the problem with a smaller dataset for a different task. Several pre-trained CNN architectures such as VGG16, VGG19, ResNet50, InceptionV3, and Xception are considered here. VGG16 and VGG19 are the variants of the VGG (Simonyan and Zisserman, 2015) model consist of 16 and 19 convolution layers, respectively. Both architectures use a fixed kernel size (3×3) in every convolution layer. However, VGG16 and VGG19 models are expensive to evaluate as they use much memory and parameters. InceptionV3 (Szegedy et al., 2015) is an extended version of GoogLeNet (Szegedy et al., 2015), having several inception modules. The modules consist of series of stacked convolutional filters $(1 \times 1, 3 \times 3$, and $5 \times 5)$ that make the Inception more powerful in learning higher representations with fewer parameters. The standard Inception modules are replaced by 'depthwise separable convolutions' in Xception (Chollet, 2017) architecture. It slightly outperforms the Inception model in several large image classification tasks. ResNet50 (He et al., 2016) is another deep CNN network consists of 50 weight layers. It utilizes the skip connection between layers to resolve overfitting problems largely present in the existing deep neural networks.

To accomplish the purpose, the upper layers of all the models keep non-trainable and only use the weights already pre-trained on the ImageNet (Russakovsky et al., 2015) dataset for 1000 classes. The top two layers of the models are excluded; instead, a fully connected (FC) layer of 50 neurons is added, accompanied by a softmax layer for prediction. Finally, the models are fine-tuned on dataset-1 and dataset-2. Hyperband (Li et al., 2016) optimization technique is adopted to maximize the performance and find the appropriate hyperparameters (i.e., optimizer, learning rate, and so on). The Keras tuner (O'Malley et al., 2019) is utilized to implement the optimization process. Several values have been experimented with for each hyperparameter, where the optimum value is selected based on the maximum validation accuracy. Table 4 shows the list of hyperparameters chosen for each dataset. All the visual models have been trained with the 'adam' optimizer. Learning rate settled to $1e^{-3}$ for D1 and $1e^{-4}$ for D2. Furthermore, the models are compiled using the *categorical cross-entropy* loss function and trained for 30 epochs with a batch size of 16 (for D1) and 32 (for D2). Keras checkpoint is utilized to stop further training when validation accuracy remains unchanged up to five consecutive epochs.

4.3. Textual feature extraction modules

Various deep learning architectures are implemented to obtain features from the textual content. The primary investigation is carried out using RNN and CNN architectures, namely BiLSTM, BiLSTM with CNN, and BiLSTM with attention. Word embedding (Mikolov et al., 2013) features are used to train these models. Embeddings are generated through the Keras embedding layer that transformed each word into a 64-element vector. These vectors convey the semantic meaning of the words, which makes learning more accessible, especially for the deep neural networks. Pre-trained transformers are also exploited to develop cutting-edge models. The implementation of various textual models are described in the following:

- **BiLSTM**: BiLSTM architecture is considered due to its ability to capture long-term dependencies by utilizing both past and future information of a text (Hossain et al., 2020). The constructed network consists of two BiLSTM layers with 64 and 32 units, respectively. The outputs of the second BiLSTM layer are passed to a fully connected layer of 20 neurons. Afterwards, a softmax layer is used for performing the classification. Before the softmax operation, a dropout layer is added with a 10% dropout rate.
- **CNN**: Embedding features are propagated into a two-layer CNN architecture. Convolutional layers are equipped with 64 and 16 filters of kernel size (1×2). The extracted features are downsampled by a pooling window of (1×2). An FC layer having 20 neurons takes the pooling features and creates the final hidden representations. Finally, the softmax layer uses this representation for classification.
- **BiLSTM + CNN**: This combined network is constructed by slightly modifying the BiLSTM described earlier and the CNN architecture. The embedding features are passed to the BiLSTM layer of 32 units. This layer's last time step output vectors are propagated to a convolutional layer having 16 filters of kernel size (1×2). CNN features further downsampled by a window of size (1×2). The last three layers (i.e., FC, dropout, softmax layer) and their parameters remain unaltered.

Table 4
Optimum hyperparameters value used for visual models. Here, D1, D2 denote dataset-1 and dataset-2.

Hyperparameters	Optimum Value
Number of neurons	50
Optimizer	'adam'
Learning Rate	$1e^{-3}$ (D1), $1e^{-4}$ (D2)
Batch Size	16 (D1), 32 (D2)
Epochs	30

[1] https://keras.io/.
[2] https://huggingface.co/transformers/main_classes/tokenizer.html.

- **BiLSTM + Attention**: Though BiLSTM effectively captures long-range dependencies, it cannot emphasize the words that are significant for classification. Architecture is defined by adopting the attention mechanism (Bahdanau et al., 2015) with a BiLSTM network consisting of 32 units to reconcile the weakness of BiLSTM. The forward and backward hidden representations of each word are concatenated and then passed into an attention layer with 20 neurons. Attention weights are assigned to the words through this layer. The higher the significance of a word, the more the weights. Finally, the obtained attention vector of weights is propagated to the softmax layer for the prediction.
- **Transformers**: In recent years, models like transformers (Vaswani et al., 2017) trained on multilingual and cross-lingual settings achieved the state of the art performance in solving several NLP problems (Sun et al., 2019; Liu et al., 2020; Lukovnikov et al., 2019). As we deal with datasets of two different languages, only multilingual and cross-lingual pretrained transformer models are considered for the investigation to avoid ambiguity in experiments. This work employs three transformer models, namely Multilingual Bidirectional Encoder Representations for transformers (m-BERT), a lighter version of BERT (m-DistilBERT), and a cross-lingual version of robustly optimized BERT (XLM-R). The models culled from the huggingface[3] transformers library and fine-tuned on our datasets with varying hyperparameters. Multilingual-BERT (Devlin et al., 2019) is a large model trained on over 104 languages. We use the 'bert-base-multilingual-cased' model with 12 transformer blocks and 110 million parameters. The distilled version of m-BERT (i.e., m-DistilBERT Sanh et al., 2019) with 6 transformer blocks is also considered. This model alleviated the computational cost and maintained the overall system performance up to 95%. The 'distilbert-base-multilingual-cased' version is procured for the implementation. XLM-Roberta (Conneau et al., 2020) is a transformer model trained in cross-lingual fashion over 100 languages having 125 million parameters. It outperformed BERT in several multilingual benchmark problems (Hossain et al., 2021; Ou and Li, 2020). To accomplish our purpose 'xlm-roberta-base' version is utilized. Transformer models take 'token ids' and 'attention masks' as input and provide a contextualized embedding vector as output. The obtained vector dimension is 768, and it is taken from the first output of the last hidden state of the transformer models. The embedding vector is then passed to a fully connected layer with 32 neurons, followed by a softmax layer for prediction. The dropout technique is used with a 10% rate before the softmax classification. Similar construction and parameters are used in the last three layers (FC, dropout, and softmax layer) for all the models. All the textual models are trained with different hyperparameter combinations. The value of the hyperparameters is listed in Table 5. A Hyperband tuner is used to find the optimum hyperparameter values. In this implementation, the BiLSTM, CNN, and BiLSTM + CNN models are compiled using 'adam' optimizer with learning rate of $1e^{-5}$ and $1e^{-4}$ respectively for dataset-1 (D1) and dataset-2 (D2). Similarly, in the case of D1, a learning rate of $1e^{-5}, 2e^{-5}$, and $3e^{-5}$ are chosen for m-BERT, XLM-R, and m-distilBERT models, respectively. On the other hand, $1e^{-4}$ (m-BERT), $1e^{-5}$ (XLM-R), and $3e^{-4}$ (m-distilBERT) are selected as the learning rate for D2. A batch size of 16 and 32 is chosen for D1 and D2. All the models trained for 30 epochs with Keras checkpoint to stop the overtraining.

[3] https://huggingface.co/.

4.4. Multimodal fusion module

Learning from multiple modalities (i.e., image, text, speech, etc.) has become a prominent research issue in recent years. Multimodal learning is widely used for various NLP problems, including image captioning (Huang et al., 2019) and visual question answering (Agrawal et al., 2015). The joint feature representation of more than one modality is utilized in multimodal tasks (Illendula and Sheth, 2019; Solieman and Pustozerov, 2021). However, classification problems can also be tackled using the same idea (Zou and Yang, 2018; Mouzannar et al., 2018). Two approaches used mainly in multimodal problems are decision fusion (Zhou, 2009) and feature fusion (Nojavanasghari et al., 2016). In the decision fusion, the softmax outputs of the visual and textual models are combined while an arbitrary hidden layer from multiple modalities is aggregated in the feature fusion technique. After the fusion operation, a single layer neural network or FC layer is trained in both approaches by feeding the combined decision outcomes or hidden feature representations as input. In this approach, the neural network works as a meta learner. For final classification, the softmax operation is performed over the learned features obtained from the meta learner.

Algorithm 1: Process of selecting best 3 visual and textual models

Input: Weighted f_1-scores
Output: Best visual and textual models
$V_f \leftarrow [vf_1, vf_2, ..., vf_N]$ (Weighted f_1 scores of visual models);
$T_f \leftarrow [tf_1, tf_2, ..., tf_M]$ (Weighted f_1 scores of textual models);
$V_m \leftarrow []$;
$T_m \leftarrow []$;
sort$(V_f, V_f + N)$;
sort$(T_f, T_f + M)$;
//choosing best 3 visual and textual models
for $i \epsilon (1, 3)$ **do**
    $V_m.append(V_f[i])$;
    $T_m.append(T_f[i])$;
    $i = i + 1$;
end

This work applies both fusion approaches to develop computational models by utilizing multimodal features. A set of visual $VN = \{v_1, v_2, \ldots., v_N\}$ and textual $TM = \{t_1, t_2, \ldots., t_M\}$ models have already been developed (in Sections 4.2 and 4.3) to classify offense and troll memes. Here, N = 5 and M = 7, which denotes the total number of visual and textual models, respectively. The splicing of each visual model with each textual model for decision and feature fusion approach results in a total of $((N \times M) \times 2) = ((5 \times 7) \times 2) = 70$ different multimodal models. However, the training of these abundant amounts of models is computationally expensive. It also requires a lot of memory and time. Therefore, this work considers only the best three models from each modality for ease of analysis to develop the multimodal models. The best models are chosen based on their weighted f_1-score on the validation set. The selection procedure of these models is illustrated in algorithm 1. Empirical observations revealed that VGG16, VGG19, and ResNet50 are the best visual models, whereas m-BERT, m-DistilBERT, and XLM-R are the best textual

Table 5
Optimum hyperparameters value utilized for training the textual models. Here, D1, and D2 represents the dataset-1 and dataset-2.

Hyperparameters	CNN (C)	BiLSTM (B)	B + C	B + Attention	m-BERT	m-DBERT	XLM-R
Input Length				50 (D1), 30 (D2)			
Embedding Dimension		64			–	–	–
Filters (layer-1)	64	–	16	–	–	–	–
Filters (layer-2)	16	–	–	–	–	–	–
Pooling type	'max'	–	'max'	–	–	–	–
Kernel Size	2	–	2	–	–	–	–
LSTM units (layer-1)	–	64	32	32	–	–	–
LSTM units (layer-2)	–	32	–	–	–	–	–
Neurons (last FC layer)		20		–		32	
Dropout		0.1		–		0.1	
Optimizer		'adam'		'RMSprop'		'adam'	
Learning rate (D1)		$1e^{-5}$		$4e^{-7}$	$1e^{-5}$	$3e^{-5}$	$2e^{-5}$
Learning rate (D2)		$1e^{-4}$		$1e^{-5}$	$1e^{-4}$	$3e^{-4}$	$1e^{-5}$
Epochs				30			
Batch Size				16(D1), 32(D2)			

models. Thus considering these six models, we obtain a total of $((3 \times 3) \times 2) = 18$ multimodal models where each fusion approach (i.e., decision, feature) contributed 9 different models.

4.4.1. Decision fusion based models

The architectures of the visual (VGG16, VGG19, and ResNet50) and textual (m-BERT, DistilBERT, and XLM-R) models have remained the same as described in Sections 4.2 and 4.3. Instead of acquiring decisions from the softmax layer of visual and textual models, the softmax outputs of individual models are combined in this approach. Consider, d_{ip}^{V} and d_{jp}^{T} are the softmax outputs for p^{th} sample provided by the visual model $v_i \in VN$ and textual model $t_j \in TM$. Then the decision fusion output can be obtained by Eq. (1).

$$DF_{ij} = d_{ip}^{V} \oplus d_{jp}^{T} \tag{1}$$

where $\oplus$ denotes the concatenation operation, and $DF_{ij} \in \mathbb{R}^{1\times 2C}$ represents the decision fusion vector containing softmax probabilities of visual, and text modalities. C indicates the number of classes in the dataset.

The vector DF_{ij} is passed to a fully connected layer with 10 neurons. Eventually, the predictions are obtained from a softmax layer. By utilizing this construction, nine multimodal decision fusion based models namely VGG19 + m-BERT, VGG16 + m-BERT, ResNet50 + m-BERT, VGG19 + DistilBERT, VGG16 + DistilBERT, ResNet50 + DistilBERT, VGG19 + XLM-R, VGG16 + XLM-R, and ResNet50 + XLM-R are developed. The models take pre-processed image, token ids, and attention masks as input. Due to the language and parametric diversity, we did not find any common hyperparameters for all the models. In case of D1, 'RMSprop' optimizer with learning rate of $2e^{-3}$, and $2e^{-4}$ is used for VGG19 + m-BERT and ResNet50 + m-BERT. Contrarily, VGG16 + m-BERT models are utilized 'adam' with a learning rate of $1e^{-5}$. 'Adam' and 'RMSprop' are chosen respectively for VGG16 + DistilBERT, and ResNet50 + DistilBERT where the learning rate is settled at $7e^{-4}$. Meanwhile, VGG16 + XLM-R, VGG19 + XLM-R, and ResNet50 + XLM-R are compiled using 'RMSprop' optimizer with a learning rate of $1e^{-5}$, $1e^{-4}$, and $5e^{-5}$ respectively. On the other hand, all the models with D2 were compiled using 'RMSprop'. Moreover, the learning rate is settled at $3e^{-5}$ for all of them except the ones having XLM-R ($2e^{-5}$).

4.4.2. Feature fusion based models

The feature fusion technique takes advantage of the hidden features extracted by visual and textual models. At first, the softmax layers are excluded from the single modality models. Following this, an FC layer with 20 neurons is added at each modality side. Let, for p^{th} sample, h_{ip}^{V} and h_{jp}^{T} are the hidden or FC layers output provided by the visual model $v_i \in VN$ and textual model $t_j \in TM$. A combined representation of visual and textual features are attained through Eq. (2).

$$FF_{ij} = h_{ip}^{V} \oplus h_{jp}^{T} \tag{2}$$

where, $FF_{ij} \in \mathbb{R}^{1\times 2h_n}$ represents the feature fusion vector containing features of both modalities and h_n denotes the number of hidden neurons. Subsequently, this unified feature vector (FF_{ij}) is fed into a fully connected layer (with 10 neurons) which is followed by a softmax layer. The number of neurons in the last FC layer is kept unaltered for all the constructed feature fusion models. The model names are similar as described in the earlier paragraph. However, different values of hyperparameters are utilized here. For D1, the visual models (VGG16, VGG19, and ResNet50) with DistilBERT combination are compiled using 'RMSProp' where the learning rate is settled at $2e^{-4}$. Likewise, VGG16 + m-BERT used a learning rate of $1e-4$, while other two models (VGG19 + m-BERT, ResNet50 + m-BERT) used a rate of $2e-4$. However, in the case of visual models with XLM-R, 'adam' is utilized with a learning rate of $5e-4$ (for ResNet50 + XLM-R) and $2e^{-5}$ (for VGG16 + XLM-R, and VGG19 + XLM-R). On the other hand, for D2, all the models used 'RMSprop' (lr = $2e-5$) except ResNet50 + m-BERT ('adam', lr = $1e-4$) model. All the models are trained for 30 epochs with a batch size of 8 (for D1) and 16 (for D2). Other hyperparameter values have remained the same as described earlier.

4.5. Proposed ensemble method

The aforementioned developed models can provide acceptable performance in classifying offense and troll memes. Nevertheless, language variation and dataset size largely influence the models' outcomes. Owing to these, distinct models achieved the highest performance for the two datasets. Therefore, to develop a standard method that can acquire superior outcomes on both datasets, this work proposes a weighted ensemble technique. This approach exploits the strength of multiple models and tries to increase the overall system predictive accuracy. Fig. 5 shows the overall architecture of the proposed method. It comprises four different models, namely, VGG19, DistilBERT, VGG19 + DistilBERT with decision fusion, and VGG19 + DistilBERT with feature fusion approach. Models are chosen based on their performance (i.e., highest weighted f_1 score) on the validation set.

Model-1 (VGG19) accepts preprocessed memes (m) as input and provides the semantic expression of the visual part by

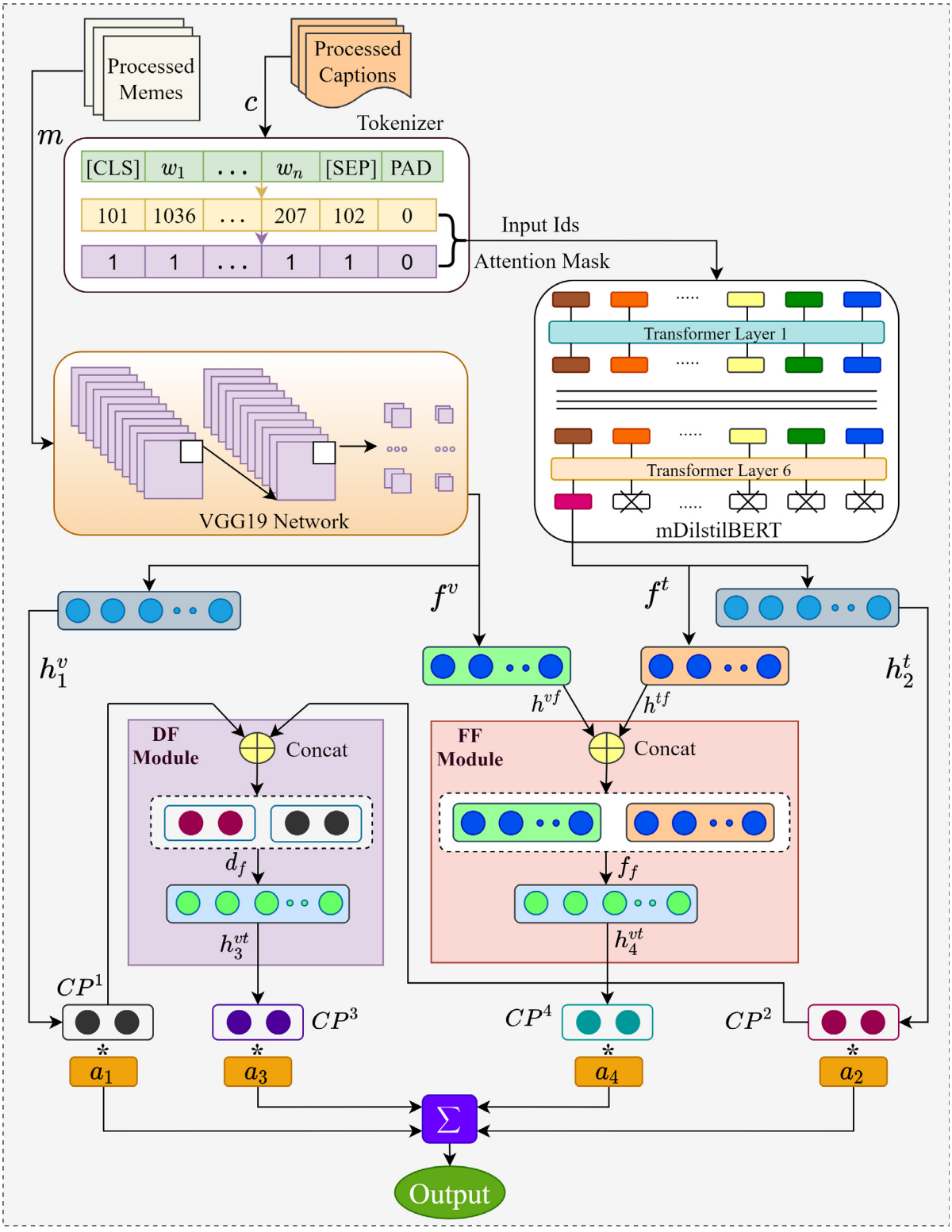


Fig. 5. Overall architecture of the proposed framework for offensive/troll meme identification.

extracting suitable features f^v. The features are then encoded by a 50-dimensional FC layer and passed to a softmax function. The process can be expressed by Eqs. (3) and (5).

$$f^v = VGG19\ (m) \tag{3}$$

$$h_1^v = \left[NN\ (f^v)\right]^{50\times 1} \tag{4}$$

$$CP^1 = \left[Softmax\ (h_1^v)\right]^{2\times 1} \tag{5}$$

here, h_1^v, and CP^1 represent the visual features obtained from the neural network (NN) layer, and the class probabilities predicted by model-1.

In the case of model-2 (m-DistilBERT), we utilized the textual features extracted by the pre-trained multilingual DistilBERT

model. The initial features are transformed into a 32 dimensional vector. Then class probabilities are calculated by a softmax function (Eqs. (6)–(8)).

$$f^t = [mDistilBERT\ (c)]^{768\times 1} \tag{6}$$

$$h_2^t = \left[NN\ (f^t)\right]^{32\times 1} \tag{7}$$

$$CP^2 = \left[Softmax\ (h_2^t)\right]^{2\times 1} \tag{8}$$

where c denotes the processed caption, f^t represents the embedding vector provided by DistilBERT, h_2^t indicates the text feature representation done by the neural network, and CP^2 denotes the predicted class probabilities.

Afterwards, using decision fusion approach, model-3 is constructed simply by aggregating the class probabilities CP^1 and CP^2 respectively obtained from model-1 and model-2. These combined probabilities are then propagated to a NN resulted in a 10-dimensional feature vector. Eqs. (9)–(11) described the process of computation.

$$d_f = \left[Concat\ (CP^1, CP^2)\right]^{4\times 1} \tag{9}$$

$$h_3^{vt} = \left[NN\ (d_f)\right]^{10\times 1} \tag{10}$$

$$CP^3 = \left[Softmax\ (h_3^{vt})\right]^{2\times 1} \tag{11}$$

where d_f denotes the concatenated class probabilities, h_3^{vt} resembles the feature vector containing both visual and textual part, and CP^3 indicates the class probabilities predicted by model-3.

For developing model-4, each visual and textual feature are represented by a 20-dimensional vector. By employing the feature fusion approach, these two vectors are combined and passed to a neural network with 10 neurons, as conferred in Eqs. (12)–(16).

$$h^{vf} = \left[NN\ (f^v)\right]^{20\times 1} \tag{12}$$

$$h^{tf} = \left[NN\ (f^t)\right]^{20\times 1} \tag{13}$$

$$f_f = \left[Concat\ (h^{vf}, h^{tf})\right]^{40\times 1} \tag{14}$$

$$h_4^{vt} = \left[NN\ (f_f)\right]^{10\times 1} \tag{15}$$

$$CP^4 = \left[Softmax\ (h_4^{vt})\right]^{2\times 1} \tag{16}$$

where f_f denotes the feature fusion vector, h_4^{vt} resembles the feature vector containing both visual and textual information, and CP^4 indicates the class probabilities.

To sum up, a set of models $U = \{M_1, M_2, .., M_l\}$ is obtained (where $l = 4$) from the all aforementioned models. From 'm' validation set of samples, a model classifies each instances m_i into one of n predefined classes. For each m_i, model U_j provides a class probability vector of size 'n', $CP_i^j[n]$. Thus, the output of the models become: $\left\langle CP_1^1[], CP_2^1[], \ldots.., CP_m^1[]\right\rangle$, $\left\langle CP_1^2[], CP_2^2[], \ldots.., CP_m^2[]\right\rangle$, and $\left\langle CP_1^l[], CP_2^l[], \ldots.., CP_m^l[]\right\rangle$. Prior that, the accuracy of the individual models on validation set also measured which can be represented as $a_1, a_2, \ldots., a_l$. Utilizing these values as weights, the proposed technique compute the final output as described in Eq. (17).

$$E_p = argmax\left(\frac{\forall_{i\epsilon(1,m)}\sum_{j=1}^{l} CP_i^j * a_j}{\sum_{j=1}^{l} a_j}\right) \tag{17}$$

here, E_p denotes the vector of $m \times 1$, which contains the ensemble method predictions. The process of calculating ensemble prediction is described in Algorithm 2. Class probabilities of the models are summed after multiplying with the accuracy. Probability values are normalized by dividing with the sum of accuracy. Finally, the output predictions are computed by taking the maximum from the probabilities.

Algorithm 2: Process of the proposed weighted ensemble technique

```
Input: Class probabilities and Accuracy
Output: Predictions of the W-ensemble
cp ← [] (class probabilities);
a ← [] (accuracy);
sum = [] (weighted sum);
for iϵ(1, m) do
    for jϵ(1, l) do
        sum[i] = sum[i] + (cp_i^j[] * a_j);
        j = j + 1;
    end
    i = i + 1;
end
n_sum = 0;
for jϵ(1, l) do
    n_sum = n_sum + a_j;
    j = j + 1;
end
P = (sum/n_sum) //normalized probabilities;
E_p = arg max(P) // set of predictions;
```

5. Experiments

This section provides a comprehensive performance analysis of the methods employed to identify the offense and troll from social media memes. A GPU facilitated platform, Google colaboratory, is used for conducting the experiments. Data processing and preparation are performed using pandas (1.1.4) and numpy (1.18.5) libraries. Transformers are accumulated from the Huggingface library, and all the models are implemented with Keras (2.4.0) and TensorFlow (2.3.0). For model evaluation, scikit-learn (0.22.2) packages are utilized. The models are developed using train, validation and test set instances. Train set instances are utilized for model learning, while hyperparameter tweaking and selection are performed based on the validation set. Finally, the trained models are evaluated using the test set instances.

5.1. Evaluation measures

Various statistical measures are considered for evaluating and comparing the performance of the systems, such as accuracy (A),

precision (P), recall (R), misclassification rate (MR) and weighted f_1 score (WE).

- Accuracy (A): is the proportion of correctly predicted observations to the total number of observations (m).

$$A = \frac{True\ Positive + True\ Negative}{Number\ of\ Samples} \tag{18}$$

- Precision (P): calculates the proportion of correctly identified positive observations (c) among the total number of predicted observations as class (c).

$$P = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{19}$$

- Recall (R): calculates the proportion of correctly identified positive observations (c) among the total number of actual observations of class (c).

$$R = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{20}$$

- Misclassification Rate (MR): calculates how many samples are wrongly classified among the total number of test samples of class (c).

$$MR = \frac{No.\ of\ Incorrect\ Classification\ in\ Class\ (c)}{Number\ of\ Samples\ in\ Class\ (c)} \tag{21}$$

- f_1-score: calculated by averaging precision and recall ($F = \frac{2PR}{P+R}$). However, considering the data imbalance problem, we calculate the weighted f_1-score (WF) which is defined as,

$$WF = \frac{1}{m}\sum_{j=1}^{c} F_j N_j,\ m = \sum_{j=1}^{c} n_i \tag{22}$$

here, m, F_j and n_j denotes total samples in test set, f_1-score and number of samples in class (j) respectively.
The weighted f_1-score metric considered to determine the superiority of the models. On the other hand, accuracy metric is utilized as weights in weighted ensemble method. Other scores such as P, R, and MR are also reported to get more insights about the model's performance on the individual classes.

5.2. Results

Table 6 presents the performance comparison of the various models developed considering only image and text modality. Concerning visual models, the results exhibited that VGG19 achieved the highest f_1-score of 0.614 and 0.514 respectively for D1 and D2. However, ResNet50 also shows good outcomes of 0.606 (D1) and 0.503 (D2), which is slightly less than the VGG19 f_1-score. Other visual models such as InceptionV3 and Xception perform poorly on both datasets. On the other hand, in the case of textual approach, transformer models obtained outstanding performance whereas other model's (CNN, BiLSTM, BiLSTM + CNN, BiLSTM + Attention) performance vacillating between $50-56\%$ (D1) and $50-53\%$ (D2). Among the transformer models, XLM-R achieved f_1-score of 0.576 (D1) and 0.556 (D2) while m-BERT score increased $\approx 5\%$ (f_1 = 0.626) for D1 and $\approx 1\%$ (f_1 = 0.561) for D2. However, m-distilBERT outdoes all the models by achieving the highest f_1-score of 0.654 (for D1) and 0.573 (for D2), respectively. The obtained result is approximately $4-6\%$ higher (in both datasets) than the best visual model (i.e., VGG19) outcomes.

The investigation is further continued where we utilized both visual and textual information and developed several unified models using two different approaches (i.e., decision fusion, feature fusion). The three best visual and textual models are chosen for developing the multimodal models. The outcome of different multimodal models is reported in Table 7. It is observed that, in the case of decision fusion based models, ResNet50 + m-BERT obtained an f_1-score of 0.562 (D1) and 0.517 (D2) while other visual models (VGG16, VGG19) with m-BERT do not perform well. Similarly, XLM-R with visual models got the lowest f_1-score ranging between $\approx 50-53\%$ (D1) whereas only VGG16 and VGG19 with XLM-R obtained acceptable outcome (f_1-score $\approx$ 57%) for D2. However, VGG19 + m-distilBERT model achieved the highest f_1-score of 0.595 and 0.583 for D1 and D2, respectively. Meanwhile, among feature fusion based models, VGG19 + m-distilBERT also got highest performance with both D1 (f_1-score = 0.660) and D2 (f_1-score = 0.557). Other models performance vacillating between $\approx 50-60\%$ (D1) and $\approx 48-54$ (D2) and thus lags almost $6-16\%$ (for D1) and $1-7\%$ (for D2) compared to the best feature fusion model. Thus, the results confirmed that the best feature fusion and decision fusion model outperformed all the unimodal and multimodal models on both datasets. It is not surprising that multimodal approaches have proven superior in identifying offense and troll memes, as the aggregation of both modals' information surely provides significant insights about a meme's overall expression. The best multimodal model obtained an f_1-score of 0.660 (D1) and 0.583 (D2), which is slightly higher than the best unimodal model (i.e., m-DistilBERT) f_1-score 0.654 (D1), and 0.573 (D2), respectively.

The results, as mentioned earlier, confirmed that VGG19, m-distilBERT, VGG19 + m-distilBERT (DF), and VGG19 + m-distilBERT (FF) is the best performing model in visual, textual, and multimodal contents. Finally, average and weighted ensemble techniques are applied to the various combination of these four models. Table 8 presents the outcomes of both ensemble approaches. Results indicate that averaging visual, textual, and feature fusion models improves the performance with f_1-score of

Table 6
Performance comparison of visual and textual models on test set where A, P, R, f_1-score denotes accuracy, precision, recall and weighted f_1-score.

Approach	Models	Dataset-1 (D1)				Dataset-2 (D2)			
		A	P	R	f_1-score	A	P	R	f_1-score
Visual	VGG16	0.577	0.581	0.577	0.579	0.596	0.572	0.596	0.502
	VGG19	0.610	0.621	0.610	**0.614**	0.575	0.536	0.575	**0.516**
	ResNet50	0.624	0.607	0.624	0.606	0.592	0.560	0.592	0.503
	InceptionV3	0.604	0.562	0.604	0.532	0.509	0.456	0.509	0.464
	Xception	0.503	0.493	0.503	0.497	0.572	0.506	0.572	0.478
Textual	CNN	0.510	0.502	0.510	0.506	0.559	0.523	0.559	0.518
	BiLSTM	0.530	0.487	0.530	0.496	0.595	0.568	0.595	0.530
	BiLSTM + CNN	0.590	0.556	0.590	0.550	0.595	0.569	0.595	0.536
	BiLSTM + Attention	0.597	0.568	0.597	0.564	0.548	0.509	0.548	0.507
	m-BERT	0.638	0.625	0.638	0.626	0.608	0.591	0.608	0.561
	m-DistilBERT	0.671	0.662	0.671	**0.654**	0.601	0.583	0.601	**0.573**
	XLM-R	0.591	0.573	0.591	0.576	0.601	0.578	0.601	0.556

Table 7
Performance comparison of multimodal models on test set. Here, (+) sign denoted the aggregation of visual and textual models. m-DBERT represents the multilingual DistilBERT model.

Approach	Models		Dataset-1 (D1)				Dataset-2 (D2)			
			A	P	R	f_1-score	A	P	R	f_1-score
Decision Fusion	m-BERT +	VGG16	0.483	0.488	0.483	0.485	0.583	0.539	0.583	0.499
		VGG19	0.544	0.541	0.544	0.542	0.589	0.555	0.589	0.513
		ResNet50	0.577	0.558	0.577	0.562	0.513	0.532	0.513	0.517
	m-DBERT +	VGG16	0.537	0.523	0.537	0.528	0.601	0.579	0.601	0.547
		VGG19	0.591	0.628	0.591	**0.595**	0.582	0.583	0.582	**0.583**
		ResNet50	0.570	0.576	0.570	0.573	0.574	0.556	0.574	0.556
	XLM-R +	VGG16	0.497	0.523	0.497	0.503	0.592	0.579	0.592	0.579
		VGG19	0.497	0.528	0.497	0.502	0.567	0.559	0.567	0.567
		ResNet50	0.604	0.563	0.604	0.532	0.574	0.551	0.574	0.548
Feature Fusion	m-BERT +	VGG16	0.584	0.564	0.584	0.567	0.580	0.556	0.580	0.549
		VGG19	0.577	0.547	0.577	0.549	0.604	0.588	0.604	0.529
		ResNet50	0.584	0.567	0.584	0.570	0.568	0.511	0.568	0.489
	m-DBERT +	VGG16	0.604	0.592	0.604	0.595	0.589	0.563	0.589	0.546
		VGG19	0.685	0.681	0.685	**0.660**	0.591	0.568	0.591	**0.557**
		ResNet50	0.611	0.598	0.611	0.600	0.597	0.571	0.597	0.528
	XLM-R +	VGG16	0.570	0.582	0.570	0.574	0.586	0.539	0.586	0.487
		VGG19	0.530	0.524	0.527	0.502	0.568	0.518	0.568	0.499
		ResNet50	0.577	0.589	0.577	0.581	0.608	0.618	0.609	0.508

Table 8
Performance comparison of various models on test set utilizing the *average and weighted ensemble* method. Here, V, T, DF, and FF represents best visual (VGG19), textual (m-distilBERT), decision fusion (VGG19 + m-distilBERT) and feature fusion (VGG19 + m-distilBERT) models respectively.

Approach	Models	Dataset-1 (D1)				Dataset-2 (D2)			
		A	P	R	f_1-score	A	P	R	f_1-score
Average Ensemble	V + T	0.617	0.609	0.617	0.612	0.588	0.555	0.588	0.522
	V + DF	0.597	0.614	0.597	0.602	0.574	0.535	0.574	0.516
	V + FF	0.638	0.625	0.638	0.626	0.586	0.548	0.586	0.509
	T + DF	0.678	0.669	0.678	0.663	0.594	0.574	0.594	0.566
	T + FF	0.678	0.678	0.678	0.644	0.603	0.584	0.603	0.571
	DF + FF	0.678	0.673	0.678	0.651	0.594	0.573	0.594	0.563
	V + T + DF	0.570	0.565	0.570	0.567	0.585	0.556	0.585	0.540
	V + T + FF	0.678	0.669	0.678	0.665	0.592	0.566	0.592	0.546
	V + DF + FF	0.604	0.592	0.604	0.594	0.588	0.557	0.588	0.532
	T + DF + FF	0.655	0.656	0.655	0.654	0.601	0.583	0.601	0.573
	V + T + DF + FF	0.671	0.662	0.671	0.659	0.592	0.567	0.592	0.548
Weighted Ensemble	V + T	0.637	0.624	0.637	0.6232	0.583	0.551	0.583	0.5314
	V + DF	0.597	0.614	0.597	0.6019	0.574	0.535	0.574	0.5164
	V + FF	0.644	0.630	0.644	0.6133	0.593	0.564	0.592	0.5292
	T + DF	0.677	0.669	0.677	0.6627	0.594	0.573	0.593	0.5658
	T + FF	0.678	0.678	0.677	0.6444	0.597	0.576	0.596	0.5632
	DF + FF	0.671	0.663	0.671	0.6458	0.594	0.572	0.594	0.5625
	V + T + DF	0.597	0.590	0.597	0.5927	0.587	0.561	0.588	0.5457
	V + T + FF	0.677	0.669	0.677	0.6650	0.592	0.566	0.592	0.5460
	V + DF + FF	0.617	0.602	0.617	0.6041	0.592	0.565	0.592	0.5415
	T + DF + FF	0.685	0.686	0.685	0.6536	0.601	0.583	0.575	0.5734
	V + T + DF + FF	0.677	0.669	0.684	**0.6673**	0.583	0.587	0.585	**0.5859**

0.665 on the test set of D1. Conversely, different behaviour was observed in D2, where the combination of textual, decision fusion, and feature fusion models provides the highest f_1-score (0.573). Unfortunately, the obtained outcome fall behind almost 1% than the best f_1-score (0.5859) on D2. On the contrary, we used the respective best model's validation accuracy as their weights for the weighted ensemble method. The outcomes exhibited that the proposed weighted ensemble method with visual, textual, decision, and feature fusion models acquired the highest f_1-score of 0.6673 (D1) and 0.5859 (D2). These results are the highest attained performance that outperformed all the previous outcomes.

Performance analysis of various models revealed that VGG19 achieved the highest weighted f_1-score among the visual models, whereas m-distilBERT attained maximum performance in textual models. A substantial increase in performance is observed when the visual and textual information is combined. Two distinct fusion approaches with similar models combination (VGG19 + m-distil BERT) outdoes all the unimodal approaches in both datasets. Apart from this, in case of average ensemble, the combination of textual and decision fusion model shows outstanding performance with D1 whereas other models combination did not provide any consistent outcomes. The inferior performance of one or two models might be the reason for deteriorating the overall performance of different average ensemble models. However, the proposed weighted ensemble method outperformed all the unimodal and multimodal models in both datasets (D1 and D2). The proposed method ability to emphasize the model's softmax predictions based on their prior results might be the reason behind the amelioration of performance to a lesser extent.

5.3. Error analysis

The results confirmed that the proposed weighted ensemble is the best performing model in classifying offensive and troll memes

(Table 8). However, to attain more in-depth insights, we performed a thorough analysis of the individual model's error both quantitatively and qualitatively. In order to illustrate the proposed model's preeminence, two other models (i.e., best visual model, best textual model) are considered for the comparison.

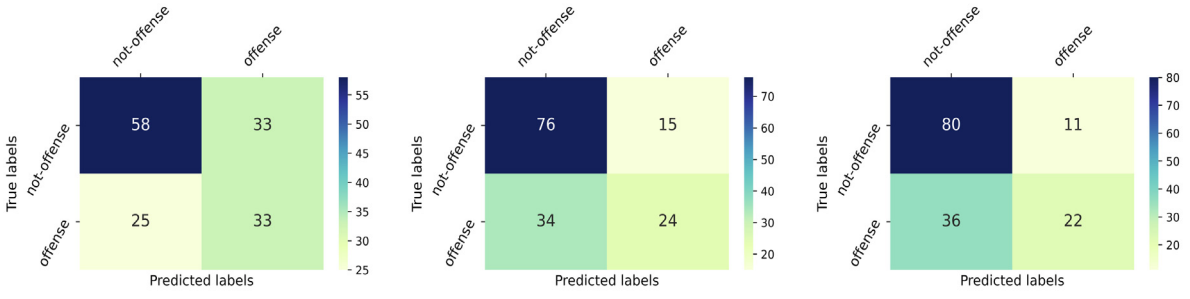
5.3.1. Quantitative analysis

Quantitative analysis of models' performance is performed in D1 and D2 by inspecting their confusion matrices. Fig. 6 shows confusion matrices of three models for the D1 (i.e., offense/not-offense). The confusion matrices (a, b, and c) exhibit that the best visual and best textual model misclassified 33 and 15 samples, respectively, whereas the proposed model incorrectly identified only 11 instances. These are the samples where models infer "Offense"; however, the actual labels say "not-offense" (known as false negatives). The textual model showed a significant boost over the visual model, whereas when multimodal features are incorporated along with unimodal features in the proposed method, the misclassification rate falls significantly from 33 to 11. On the other hand, in the case of the *offense* class, a slight increase is observed in the misclassification of *offense* as *not-offense* (known as false positives) across the models. Fig. 6 shows these mistakes as 25 by the visual model, 34 by the textual model, and 36 instances by the proposed model. Unfortunately, no improvements were observed from the proposed approach as noticed in the "not-offense" class. Meanwhile, an almost similar scenario is observed with D2, which can be visualized from the confusion matrices shown in Fig. 7. It observed that the number of misclassified instances (*not-troll* predicted as *troll*) significantly dropped (227 to 155) from the visual to the proposed model. Though the textual model showed an improvement compared to the visual model, the proposed model reduces the error most for the *not-troll* class. Unfortunately, the error rate dramatically increased in the case of the *troll* class. The mistakes are observed in Fig. 7 where visual and textual models misclassified 56 and 78 instances. On the contrary, the proposed model incorrectly identified 117 instances as the 'not-troll'. In this case, an experienced of undesirable rise in the false-positive rate is observed. Fig. 8 depicts the rate of misclassification (MR) across different classes attained by three models (i.e., best visual model, best textual model, and proposed method) on D1 and D2. From Fig. 8 (a), it is observed that the MR significantly falls from 36.3% to 12.1% for *offense* class, while in *not-offense* MR rose up to 62.1% from 43.1%. Likewise, concerning D2, the MR gradually increases for the *troll* class, whereas it substantially reduced to 57% from 83.5% (*not-troll* class). The error also indicates that the best visual model (i.e., VGG19) is more appropriate in identifying the *offense* and *troll* classes, providing lower predictions. Furthermore, there is a trade-off between individual class performance as when the error of one class decreased, the other's class is increased. Although the proposed method lessens the error in the *not-offense* and *not-troll* classes, it minimized the combined errors for both datasets, which acquired the highest outcomes (found in Section 5.2).

5.3.2. Qualitative analysis

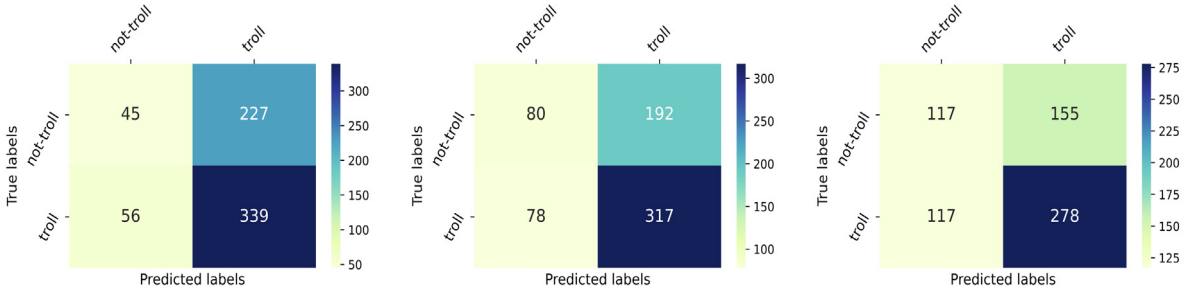
Figs. 9 and 10 show some example memes from D1 and D2 that elucidate how the proposed weighted ensemble model can capture information effectively, and hence, lead to better predictions over the visual and textual models. Besides, to illustrate the mistakes made by the proposed method, some misclassified memes are also presented. Fig. 9 (a) illustrates the correctly detected sample by the visual model as an *offense* meme, whereas Fig. 9 (b) depicts the correctly identified sample as *not-offense* by the textual model. Both samples are also correctly classified by the proposed method, which further signifies the capability of acquiring the information by the model when at least one modality can identify the precise class. However, a more profound advantage of incorporating multimodal features is observed explicitly in Fig. 9 (c), where both visual and textual models reckon the meme as *not-offensive*. On the other hand, the proposed model correctly identified this sample as the *offensive* meme. Concerning D2, the visual model did not find any trolling information from Fig. 10 (b), whereas, textual model labelled it as a troll meme. It is probably due to the presence of words like *expectation*, and *reality* in the textual content. Similarly, in Fig. 10 (c), evaluating visual alone or textual alone yields incorrect predictions; however, when both modalities are jointly evaluated, they provided firm evidence for the proposed model to identify it as a *troll* meme. Furthermore, an interesting case is observed in Figs. 9 (d) and 10 (d), where none of the model detects the actual label of the memes.

To sum up, quantitative analysis revealed that the model's performance becomes biased towards a particular class (i.e., not-offense/not-troll) for both datasets. The possible reason of this incongruity might be due to the extensive appearance of some strong words such as "Trump", "Hilary", "Bernie", "Communist", "Amala", "Sayessha", "boys", "girls", and "Anna" respectively in the textual content of the offense/not-offense and troll/not-troll classes of memes. In addition to that, dataset-1 (i.e., offense/not-offense) is developed using the memes posted during the presidential election period; thus, some world-famous person faces frequently appeared in the memes of both classes. Likewise, dataset-2 also has plenty of memes with common person faces (i.e., south Indian actors) in troll and not-troll classes. The presence of these consistent visual and textual features among the classes of each dataset made it arduous for the models to differentiate the appropriate class. Indeed, these are the major factors that resulted in one modality approach performing well in one class, and another modality approach yielded better outcomes in other classes. Frenda et al. (2022) investigates how the implicit humor of a textual content can reveal the aggressive intention towards a



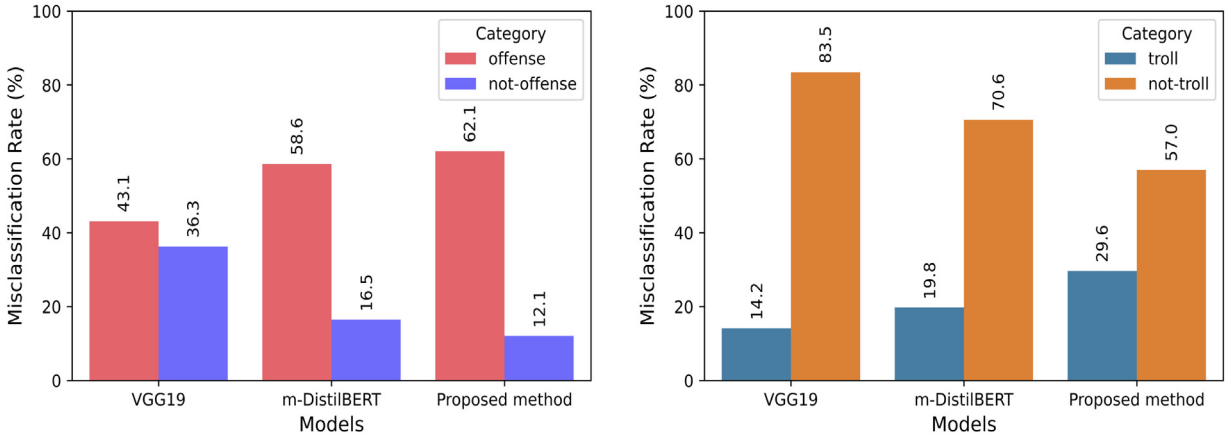
(a) Best Visual (VGG19) (b) Best Textual (m-DistilBERT) (c) Proposed method

Fig. 6. Confusion matrices of different models developed for dataset-1 (D1).



(a) Best Visual (VGG19) (b) Best Textual (m-DistilBERT) (c) Proposed method

Fig. 7. Confusion matrices of different models developed for dataset-2 (D2).



(a) Dataset-1 (D1) (b) Dataset-2 (D2)

Fig. 8. Proportion of misclassification among the classes of dataset-1 (D1) and dataset-2 (D2).

(a) **Visual modality:** Offense (✓) **Text modality:** not-Offense (✗) **Proposed method:** Offense(✓)
(b) **Visual modality:** Offense (✗) **Text modality:** not-Offense (✓) **Proposed method:** not-Offense(✓)
(c) **Unimodal:** Not-offense (✗) **Proposed method:** Offense (✓)
(d) **Unimodal:** not-offense (✗) **Proposed method:** not-offense (✗)

Fig. 9. Few correctly and misclassified examples predicted by the proposed and other approaches on the dataset-1 (D1). The symbol (×) indicates an incorrect classification and the symbol (✓) indicates a correct classification.

(a) **Visual modality:** Not-Troll(✓) **Text modality:** Troll (✗) **Proposed method:** Not-Troll(✓)
(b) **Visual modality:** Not-Troll (✗) **Text modality:** Troll (✓) **Proposed method:** Troll (✓)
(c) **Unimodal:** Not-Troll (✗) **Proposed method:** Troll (✓)
(d) **Unimodal:** Not-Troll (✗) **Proposed method:** Not-Troll (✗)

Fig. 10. Few correctly and misclassified examples predicted by the proposed and other approaches on the dataset-2 (D2).

particular entity. Furthermore, analysis of the incorrect predictions shown in Figs. 9 (d) and 10 (d) rendered some other reasons that lead to the performance degradation across the classes. To shed light on that, we go through the memes of both datasets and found several disparities regarding contextual complexity and annotation. Among them, one reason is that memes contain very short captions (shown in Fig. 11 (a)), specifically having less than two words. Moreover, many memes even do not have any captions at all (shown in Fig. 11 (b)), and their visual content does not provide any meaningful information regarding the class. In particular, out of 743 memes, 65 have a very short caption consisting of less than two words, and 21 memes have no caption (dataset-1). Concerning dataset-2, among 2967 memes, 355 have a short caption (less than two words), whereas 122 memes do not have any caption. Apart from this, it observed that some memes seem offensive and troll; however, the annotated label showed that the memes are from the *not-offensive* and *not-troll* class. For instance, in Figs. 11 (c) and 11 (d), by examining both visual and textual content, it can be unequivocally said that the memes are from offensive and troll classes, respectively. Mistakes in class labelling during annotation are another prime reason for the models failing to yield more improved results. The reasons mentioned above bring forward new challenges in the direction of undesired meme classification that should need to be handled to develop a more efficient model in future.

5.4. Comparison between the proposed and existing methods

We have developed several multimodal models by combining the existing state-of-the-art visual and textual models (such as BERT + VGG19, DistilBERT + ResNet50, XLM-R + VGG16 etc.). The performance of the proposed model compared with the existing state-of-the-art techniques (Suryawanshi et al., 2020a; Mishra et al., 2020; Huang and Bai, 2021; Hegde et al., 2021; Manoj and Chinmaya, 2021; Que, 2021; Bharathi and Agnusimmaculate, 2021; Li, 2021; Suryawanshi et al., 2020b). Table 9 shows the results of the comparison. The results revealed that the proposed method (weighted ensemble) achieved the best weighted f_1 score of 66.73% ($\approx$ 13% $\uparrow$) as compared to the weighted f_1 score of 54% of the baseline model (i.e., Suryawanshi et al. (2020a)) for "MultiOFF" dataset (D1). Similarly, for the "TamilMemes" dataset (D2), the proposed model gained the highest weighted f_1 score of 58.59%(1.59% $\uparrow$) as compared to the outcome of the model developed by Suryawanshi et al. (2020b). Analysis of the comparison confirmed that the proposed technique outperformed other contemporary works on both datasets. In recent years, a few algorithms have been introduced for multimodal learning, such as Visual-BERT (Li et al., 2019), VL-BERT (Su et al., 2019), CLIP (Radford et al., 2021). As far as we know from the most recent literature, these algorithms have not been applied to the offense and troll memes detection problems. However, we aim to investigate these models in the future.

Table 9
Comparative analysis of the proposed method with the existing state-of-the-art techniques. *MultiOFF* and *TamilMemes* indicates the dataset-1 (D1) and dataset-2 (D2).

Techniques	Datasets	WF (%)
Suryawanshi et al. (2020a)	MultiOFF	54
Mishra et al. (2020)	TamilMemes	30
Huang and Bai (2021)	TamilMemes	40
Hegde et al. (2021)	TamilMemes	47
Manoj and Chinmaya (2021)	TamilMemes	48
Que (2021)	TamilMemes	49
Bharathi and Agnusimmaculate (2021)	TamilMemes	50
Li (2021)	TamilMemes	55
Suryawanshi et al. (2020b)	TamilMemes	57
Proposed (weighted ensemble)	MultiOFF	66.73
	TamilMemes	58.59

6. Conclusion

This paper proposes a weighted ensemble-based technique that can effectively learn from all types of features, including visual, textual, and multimodal, for classifying social media memes. Two benchmark multimodal meme datasets viz. *MultiOFF (D1)* and *TamilMemes (D2)* are utilized to evaluate the models. This work investigated various state of the art visual (i.e., VGG19, VGG16, InceptionV3, Xception, ResNet50) and textual (i.e., LSTM, CNN, Attention, m-BERT, m-DistilBERT, XLMR) models. In addition, two different fusion approaches (i.e., decision fusion, feature fusion) are also used to construct several multimodal models utilizing the image and text features. After analyzing all models' performance on the two datasets, this work proposed a weighted ensemble technique for classifying memes. The proposed technique can readdress the softmax probabilities of the participating models based on their previous outcomes on the datasets. The experimented results revealed that the proposed technique outdoes the unimodal (i.e., image, text), multimodal, and average ensemble models by obtaining the highest weighted f_1 score of 66.73% (MultiOFF dataset) and 58.59% (TamilMemes dataset), respectively. Moreover, the comparative analysis indicated that the proposed technique outcomes are approximately 13% (in 'MultiOFF') and 1.69% (in 'TamilMemes') ahead compared to the current state of the art systems. Thus, results ensured the effectiveness of the proposed technique in detecting offensive and troll memes based on multimodal information. Quantitative and qualitative error analysis shows that it is arduous to identify offenses/trolls expressed implicitly or sarcastically. Moreover, the disparity between visual and textual information and the lack of appropriate methods to analyze the multimodal data made the problem more challenging. In the future, we aim to explore visual attention and transformer architectures (i.e., Visual-BERT, VL-BERT, CLIP) to capture strong visual and textual features. Moreover, it will be interesting to investigate how multimodal offense or troll detection can be tackled utilizing the multitask learning approach.

(a) Very short caption (b) No caption (c) Incorrectly annotated as "not-offense" (d) Incorrectly annotated as "not-troll"

Fig. 11. Few ambiguous and complicated memes from D1 and D2 illustrating why models failed to detect the actual label of memes.

Declaration of Competing Interest

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

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D., 2015. Vqa: Visual question answering. Int. J. Comput. Vision 123, 4–31.

Akiwowo, S., Vidgen, B., Prabhakaran, V., Waseem, Z. (Eds.), 2020. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online. URL:/ https://www.aclweb.org/anthology/2020.alw-1.0

Andrew, J.J., 2021. JudithJeyafreedaAndrew@DravidianLangTech-EACL2021: offensive language detection for Dravidian code-mixed YouTube comments. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, pp. 169–174. URL:/ https://www.aclweb.org/anthology/2021.dravidianlangtech-1.22

Aroyehun, S.T., Gelbukh, A., 2018. Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 90–97. URL:/ https://www.aclweb.org/anthology/W18-4411

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015.

Bannink, R., Broeren, S., van de Looij - Jansen, P.M., de Waart, F.G., Raat, H., 2014. Cyber and traditional bullying victimization as a risk factor for mental health problems and suicidal ideation in adolescents. PLOS ONE 9 (4), 1–7. https://doi.org/10.1371/journal.pone.0094026.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M., 2019. SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 54–63. https://doi.org/10.18653/v1/S19-2007. URL:/ https://www.aclweb.org/anthology/S19-2007.

Bharathi, Agnusimmaculate, S., 2021. SSNCSE_NLP@DravidianLangTech-EACL2021: Meme classification for Tamil using machine learning approach. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv. pp. 336–339. URL: https://aclanthology.org/2021.dravidianlangtech-1.49.

Bhardwaj, M., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T., 2020. Hostility detection dataset in hindi. arXiv:2011.03588.

Bonheme, L., Grzes, M., 2020. SESAM at SemEval-2020 task 8: investigating the relationship between image and text in sentiment analysis of memes. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), pp. 804–816. URL:/ https://aclanthology.org/2020.semeval-1.102

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., Maurizio, T., 2018. Overview of the evalita 2018 hate speech detection task. In: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol. 2263, CEUR, pp. 1–9.

Chakravarthi, B.R., Priyadharshini, R., Jose, N., Kumar, A.M., Mandl, T., Kumaresan, P. K., Ponnusamy, R., Hariharan, McCrae, J., Sherly, E., 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, pp. 133–145. URL:/ https://www.aclweb.org/anthology/2021.dravidianlangtech-1.17

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J., 2020. Uniter: universal image-text representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 104–120.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale. ACL.

Connie, T., Al-Shabi, M., Goh, M., 2017. Smart content recognition from images using a mixture of convolutional neural networks. Lect. Notes Electr. Eng., 11–18 https://doi.org/10.1007/978-981-10-6451-7_2.

Das, A., Wahi, J.S., Li, S., 2020. Detecting hate speech in multi-modal memes. CoRR abs/2012.14891. arXiv:2012.14891.

Davidson, T., Warmsley, D., Macy, M., Weber, I., May 2017. Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media 11 (1). URL:/ https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186. doi:10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Drakett, J., Rickett, B., Day, K., Milnes, K., 2018. Old jokes, new media–online sexism and constructions of gender in internet memes. Fem. Psychol. 28 (1), 109–127.

Duggan, M., 2017. Men, women experience and view online harassment differently. pew research center. published july 14.

Fortuna, P., Nunes, S., 2018. A survey on automatic detection of hate speech in text. ACM Comput. Surv. 51 (4). https://doi.org/10.1145/3232676.

Frenda, S., Cignarella, A.T., Basile, V., Bosco, C., Patti, V., Rosso, P., 2022. The unbearable hurtfulness of sarcasm. Expert Syst. Appl. 193,. https://doi.org/10.1016/j.eswa.2021.116398 116398.

Gambäck, B., Sikdar, U.K., 2017. Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, pp. 85–90. doi:10.18653/v1/W17-3013. URL:/ https://www.aclweb.org/anthology/W17-3013

Gandhi, S., Kokkula, S., Chaudhuri, A., Magnani, A., Stanley, T., Ahmadi, B., Kandaswamy, V., Ovenc, O., Mannor, S., 2019. Image matters: detecting offensive and non-compliant content/ logo in product images, CoRR abs/1905.02234. arXiv:1905.02234. URL:/ http://arxiv.org/abs/1905.02234.

Gomez, R., Gibert, J., Gomez, L., Karatzas, D., 2020. Exploring hate speech detection in multimodal publications. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Hegde, S.U., Hande, A., Priyadharshini, R., Thavareesan, S., Chakravarthi, B.R., 2021. UVCE-IIITT@DravidianLangTech-EACL2021: Tamil troll meme classification: You need to pay more attention. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv. pp. 180–186. URL: https://www.aclweb.org/anthology/2021.dravidianlangtech-1.24.

Hossain, E., Sharif, O., Hoque, M., Sarker, I.H., 2020. Sentilstm: A deep learning approach for sentiment analysis of restaurant reviews. In: HIS.

Hossain, E., Sharif, O., Hoque, M.M., 2021. NLP-CUET@DravidianLangTech-EACL2021: investigating visual and textual features to identify trolls from multimodal social media memes. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, pp. 300–306. URL:/ https://www.aclweb.org/anthology/2021.dravidianlangtech-1.43

Hossain, E., Sharif, O., Hoque, M.M., 2021. NLP-CUET@LT-EDI-EACL2021: Multilingual code-mixed hope speech detection using cross-lingual representation learner. In: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Kyiv. pp. 168–174. URL: https://aclanthology.org/2021.ltedi-1.25.

Hosseinmardi, H., Rafiq, R.I., Han, R., Lv, Q., Mishra, S., 2016. Prediction of cyberbullying incidents in a media-based social network. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 186–192. https://doi.org/10.1109/ASONAM.2016.7752233.

Huang, B., Bai, Y., 2021. HUB@DravidianLangTech-EACL2021: Meme classification for Tamil text-image fusion, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv. pp. 210–215. URL: https://aclanthology.org/2021.dravidianlangtech-1.28.

Huang, L., Wang, W., Chen, J., Wei, X.-Y., 2019. Attention on attention for image captioning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4633–4642.

Hudson, D.A., Manning, C.D., 2019. Gqa: a new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Illendula, A., Sheth, A., 2019. Multimodal emotion classification. In: Companion Proceedings of The 2019 World Wide Web Conference.

Jørgensen, R.F., Zuleta, L., 2020. Private governance of freedom of expression on social media platforms: Eu content regulation through the lens of human rights standards. Nord. Rev. 41 (1), 51–67. https://doi.org/10.2478/nor-2020-0003.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D., 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates Inc, pp. 2611–2624. URL: https://proceedings.neurips.cc/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

Kumar, R., Ojha, A.K., Zampieri, M., Malmasi, S. (Eds.), 2018. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA. URL:/ https://www.aclweb.org/anthology/W18-4400

Kumar, R., Ojha, A.K., Lahiri, B., Zampieri, M., Malmasi, S., Murdock, V., Kadar, D. (Eds.), 2020. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France. URL:/ https://www.aclweb.org/anthology/2020.trac-1.0

Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M., 2020. Evaluating aggression identification in social media. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, pp. 1–5. URL:/ https://www.aclweb.org/anthology/2020.trac-1.1

Kumari, K., Singh, J.P., 2021. Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Trans. Emerg. Telecommun. Technol. 32 (2), e3907. doi: 10.1002/ett.3907.

Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P., 2019. Aggressive social media post detection system containing symbolic images. In: Pappas, I.O., Mikalef, P., Dwivedi, Y.K., Jaccheri, L., Krogstie, J., Mäntymäki, M. (Eds.), Digital Transformation for a Sustainable Society in the 21st Century. Springer International Publishing, Cham, pp. 415–424.

Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P., 2020. Towards cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Comput. 24 (15), 11059–11070. doi:10.1007/s00500-019-04550-x.

Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P., 2021. Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. Future Gener. Comput. Syst. 118, 187–197. https://doi.org/10.1016/j.future.2021.01.014.

Li, Z., 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multimodal transformers for meme classification in Tamil language. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv. pp. 352–356. URL: https://www.aclweb.org/anthology/2021.dravidianlangtech-1.52.

Li, L., Jamieson, K.G., DeSalvo, G., Rostamizadeh, A., Talwalkar, A., 2016. Efficient hyperparameter optimization and infinitely many armed bandits. CoRR abs/1603.06560. arXiv:1603.06560.

Li, L.H., Yatskar, M., Yin, D., Hsieh, C.-J., Chang, K.-W., 2019. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Li, L.H., Yatskar, M., Yin, D., Hsieh, C., Chang, K., 2019. Visualbert: a simple and performant baseline for vision and language, CoRR abs/1908.03557. arXiv:1908.03557.

Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., Yannakoudakis, H., 2020. A multimodal framework for the detection of hateful memes, CoRR abs/2012.12871. arXiv:2012.12871. URL:/ https://arxiv.org/abs/2012.12871

Liu, X., Duh, K., Liu, L., Gao, J., 2020. Very deep transformers for neural machine translation. ArXiv abs/2008.07772.

Lukovnikov, D., Fischer, A., Lehmann, J., 2019. Pretrained transformers for simple question answering over knowledge graphs. International Semantic Web Conference, Springer, 470–486.

Mandl, T., Modha, S., Kumar, A.M., Chakravarthi, B.R., 2020. Overview of the hasoc track at fire 2020: hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In: Forum for Information Retrieval Evaluation, FIRE 2020, Association for Computing Machinery, New York, NY, USA, p. 29–32. doi:10.1145/3441501.3441517.

Manoj, B. Chinmaya, 2021. TrollMeta@DravidianLangTech-EACL2021: meme classification using deep learning. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, pp. 277–280. URL:/ https://www.aclweb.org/anthology/2021.dravidianlangtech-1.39.

Mihaylov, T., Georgiev, G., Nakov, P., 2015. Finding opinion manipulation trolls in news community forums. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Beijing, China, pp. 310–314. doi:10.18653/v1/K15-1032. URL:/ https://www.aclweb.org/anthology/K15-1032.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. ICLR.

Mishra, A.K., Saumya, S., 2021. IIIT_DWD@EACL2021: identifying troll meme in Tamil using a hybrid deep learning approach. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, pp. 243–248. URL:/ https://www.aclweb.org/anthology/2021.dravidianlangtech-1.33

Mishra, P., Yannakoudakis, H., Shutova, E., 2019. Tackling online abuse: a survey of automated abuse detection methods, CoRR abs/1908.06024. arXiv:1908.06024. URL:/ http://arxiv.org/abs/1908.06024

Mishra, S., Prasad, S., Mishra, S., 2020. Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at TRAC 2020. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France. pp. 120–125. URL: https://www.aclweb.org/anthology/2020.trac-1.19.

Mojica de la Vega, L.G., Ng, V., 2018. Modeling trolling in social media conversations. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan. URL:/ https://www.aclweb.org/anthology/L18-1585.

Morency, L.-P., Baltrušaitis, T., 2017. Multimodal machine learning: integrating language, vision and speech. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Association for Computational Linguistics, Vancouver, Canada, pp. 3–5. URL:/ https://www.aclweb.org/anthology/P17-5002

Morishita, T., Morio, G., Horiguchi, S., Ozaki, H., Miyoshi, T., 2020. Hitachi at SemEval-2020 task 8: simple but effective modality ensemble for meme emotion recognition. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), pp. 1126–1134. URL:/ https://aclanthology.org/2020.semeval-1.149.

Mouzannar, H., Rizk, Y., Awad, M., 2018. Damage identification in social media posts using multimodal deep learning. ISCRAM.

Mut Altin, L.S., Bravo, A., Saggion, H., 2020. LaSTUS/TALN at TRAC - 2020 trolling, aggression and cyberbullying, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, pp. 83–86. URL:/ https://www.aclweb.org/anthology/2020.trac-1.13.

Nakamura, K., Levy, S., Wang, W.Y., 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 6149–6157. URL: https://www.aclweb.org/anthology/2020.lrec-1.755.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., Morency, L.-P., 2016. Deep multimodal fusion for persuasiveness prediction. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 284–288.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al., 2019. Keras tuner. URL: https://github.com/keras-team/keras-tuner.

Ou, X., Li, H., 2020. Ynu@dravidian-codemix-fire2020: Xlm-roberta for multi-language sentiment analysis. In: FIRE.

Pamungkas, E.W., Patti, V., 2019. Cross-domain and cross-lingual abusive language detection: a hybrid approach with deep learning and a multilingual lexicon. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, pp. 363–370. doi:10.18653/v1/P19-2051. URL:/ https://www.aclweb.org/anthology/P19-2051.

Pavlopoulos, J., Thain, N., Dixon, L., Androutsopoulos, I., 2019. ConvAI at SemEval-2019 task 6: offensive language identification and categorization with perspective and BERT. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 571–576. https://doi.org/10.18653/v1/S19-2102. URL:/ https://www.aclweb.org/anthology/S19-2102.

Perifanos, K., Goutsos, D., 2021. Multimodal hate speech detection in greek social media. In: Preprints. URL:/ https://www.preprints.org/manuscript/202103.0390/v1

Que, Q., 2021. Simon @ DravidianLangTech-EACL2021: Meme classification for Tamil with BERT. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv. pp. 287–290. URL: https://aclanthology.org/2021.dravidianlangtech-1.41.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR. pp. 8748–8763.

Roberts, S.T., Tetreault, J., Prabhakaran, V., Waseem, Z. (Eds.), 2019. In: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy. URL:/ https://www.aclweb.org/anthology/W19-3500

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115, 211–252.

Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G.S., On, B.-W., 2021. Aggression detection through deep neural model on twitter. Future Gener. Comput. Syst. 114, 120–129. https://doi.org/10.1016/j.future.2020.07.050.

Safi Samghabadi, N., Patwa, P., PYKL, S., Mukherjee, P., Das, A., Solorio, T., 2020. Aggression and misogyny detection using BERT: a multi-task approach. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, pp. 126–131. URL:/ https://www.aclweb.org/anthology/2020.trac-1.20.

Saha, D., Paharia, N., Chakraborty, D., Saha, P., 2021. Mukherjee, A. Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, pp. 270–276. URL:/ https://www.aclweb.org/anthology/2021.dravidianlangtech-1.38

Sandulescu, V., 2020. Detecting hateful memes using a multimodal deep ensemble. CoRR abs/2012.13235. arXiv:2012.13235.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv abs/1910.01108.

Sharif, O., Hoque, M.M., 2021. Identification and classification of textual aggression in social media: Resource creation and evaluation. In: Chakraborty, T. et al. (Eds.), Combating Online Hostile Posts in Regional Languages during Emergency Situation. Springer Nature, Switzerland AG, pp. 1–12. https://doi.org/10.1007/978-3-030-73696-5_2.

Sharif, O., Hossain, E., Hoque, M.M., 2021. NLP-CUET@DravidianLangTech-EACL2021: offensive language detection from multilingual code-mixed text using transformers. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, pp. 255–261. URL:/ https://www.aclweb.org/anthology/2021.dravidianlangtech-1.35

Sharif, O., Hossain, E., Hoque, M.M., 2021. Combating hostility: covid-19 fake news and hostile post detection in social media. arXiv:2101.03291.

Sharma, C., Bhageria, D., Scott, W., PYKL, S., Das, A., Chakraborty, T., Pulabaigari, V., Gambäck, B., 2020. SemEval-2020 task 8: memotion analysis- the visuo-lingual metaphor!. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), pp. 759–773. URL:/ https://aclanthology.org/2020.semeval-1.99

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Singh, V.K., Ghosh, S., Jose, C., 2017. Toward multimodal cyberbullying detection. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors

in Computing Systems, CHI EA '17, Association for Computing Machinery. New York, NY, USA. pp. 2090–2099. doi:10.1145/3027063.3053169.

Solieman, H., Pustozerov, E., 2021. The detection of depression using multimodal models based on text and voice quality features. In: 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), pp. 1843–1848.

Song, C., Ning, N., Zhang, Y., Wu, B., 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf. Process. Manage. 58, (1). https://doi.org/10.1016/j.ipm.2020.102437 102437.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J., 2019. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.

Suhr, A., Zhou, S., Zhang, I., Bai, H., Artzi, Y., 2018. A corpus for reasoning about natural language grounded in photographs, CoRR abs/1811.00491. arXiv:1811.00491. URL:/ http://arxiv.org/abs/1811.00491

Sun, C., Qiu, X., Xu, Y., Huang, X., 2019. How to fine-tune BERT for text classification? CoRR abs/1905.05583. arXiv:1905.05583.

Suryawanshi, S., Chakravarthi, B.R., 2021. Findings of the shared task on troll meme classification in Tamil. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv. pp. 126–132. URL: https://www.aclweb.org/anthology/2021.dravidianlangtech-1.16.

Suryawanshi, S., Chakravarthi, B.R., Arcan, M., Buitelaar, P., 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, pp. 32–41. URL:/ https://www.aclweb.org/anthology/2020.trac-1.6.

Suryawanshi, S., Chakravarthi, B.R., Verma, P., Arcan, M., McCrae, J.P., Buitelaar, P., 2020. A dataset for troll classification of TamilMemes. In: Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, European Language Resources Association (ELRA), Marseille, France, pp. 7–13. URL:/ https://www.aclweb.org/anthology/2020.wildre-1.2

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the inception architecture for computer vision. CoRR abs/1512.00567. arXiv:1512.00567.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., Daelemans, W., 2016. A dictionary-based approach to racism detection in dutch social media, CoRR abs/1608.08738. arXiv:1608.08738. URL:/ http://arxiv.org/abs/1608.08738

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010.

Velioglu, R., Rose, J., 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. CoRR abs/2012.12975. arXiv:2012.12975.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., Margetts, H., 2019. Challenges and frontiers in abusive content detection. In: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, pp. 80–93. doi:10.18653/v1/W19-3509. URL:/ https://www.aclweb.org/anthology/W19-3509

Wang, S., Liu, J., Ouyang, X., Sun, Y., 2020. Galileo at SemEval-2020 task 12: Multilingual learning for offensive language identification using pre-trained language models. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), pp. 1448–1455. URL:/ https://www.aclweb.org/anthology/2020.semeval-1.189

Williams, A., Oliver, C., Aumer, K., Meyers, C., 2016. Racial microaggressions and perceptions of internet memes. Comput. Hum. Behav. 63, 424–432. https://doi.org/10.1016/j.chb.2016.05.067.

Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., Wei, L., 2021. Detecting fake news by exploring the consistency of multimodal data. Inf. Process. Manage. 58, (5). https://doi.org/10.1016/j.ipm.2021.102610 102610.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R., 2019. Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 1415–1420. https://doi.org/10.18653/v1/N19-1144. URL:/ https://www.aclweb.org/anthology/N19-1144.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç., 2020. SemEval-2020 task 12: multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), pp. 1425–1447. URL:/https://www.aclweb.org/anthology/2020.semeval-1.188.

Zhang, W., Liu, G., Li, Z., Zhu, F., 2020. Hateful memes detection via complementary visual and linguistic networks. CoRR abs/2012.04977. arXiv:2012.04977.

Zhou, Z.-H., 2009. Ensemble learning. Encyclopedia Biometrics 1, 270–273.

Zhou, Y., Yang, Y., Liu, H., Liu, X., Savage, N., 2020. Deep learning based fusion approach for hate speech detection. IEEE Access 8, 128923–128929. https://doi.org/10.1109/ACCESS.2020.3009244.

Zhou, X., Sap, M., Swayamdipta, S., Smith, N.A., Choi, Y., 2021. Challenges in automated debiasing for toxic language detection. arXiv:2102.00086.

Zou, P., Yang, S., 2018. Multimodal tweet sentiment classification algorithm based on attention mechanism. DMLE/IOTSTREAMING@PKDD/ECML.