Noname manuscript No.
(will be inserted by the editor)

Do Images really do the Talking?

Analysing the significance of Images in Tamil Troll meme classification

Siddhanth U Hegde · Adeep Hande · Ruba Priyadharshini · Sajeetha Thavareesan · Ratnasingam Sakuntharaj · Sathiyaraj Thangasamy · B Bharathi · Bharathi Raja Chakravarthi

Received: date / Accepted: date

Abstract A meme is a part of media created to share an opinion or emotion across the internet. Due to their popularity, memes have become the new form of communication on social media. However, they are used in harmful ways such as trolling and cyberbullying progressively due to their nature. Various data modelling methods create different possibilities in feature extraction and turn them into beneficial information. The variety of modalities included in data plays a significant part in predicting the results. We try to explore the significance of visual features of images in classifying memes. Memes are a blend of both image and text, where the text is embedded into the picture. We consider a meme to be trolling if the meme in any way tries to troll a particular individual, group, or organisation. We try to incorporate the memes as a troll and non-trolling memes based on their images and text. We evaluate if there is any major significance of the visual features for identifying whether a meme is trolling or not. Our work illustrates different textual analysis methods and contrasting multimodal approaches ranging from simple merging to cross attention to utilising both worlds' - visual and textual features. The fine-tuned cross-lingual language model, XLM, performed the best in textual analysis, and the multimodal transformer performs the best in multimodal analysis.

Keywords Feature Extraction · Memes · Troll and non-troll · Transformer

Siddhanth U Hegde
University Visvesvaraya College of Engineering, Bangalore University
siddhanthhegde227@gmail.com

Adeep Hande
Indian Institute of Information Technology Tiruchirappalli, Tamil Nadu, India
adeeph18c@iiitt.ac.in

Ruba Priyadharshini
ULTRA Arts and Science College, Madurai, Tamil Nadu, India
rubapriyadharshini.a@gmail.com

Sajeetha Thavareesan, Ratnasingam Sakuntharaj
Eastern University, Sri Lanka
{*sajeethas,sakuntharaj*}*@esn.ac.lk*

Sathiyaraj Thangasamy
Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamil Nadu, India.
sathiyarajt@skacas.ac.in

B Bharathi
SSN College of Engineering, Tamil Nadu, India.
bharathib@ssn.edu.in

Bharathi Raja Chakravarthi*
Insight SFI Research Centre for Data Analytics, National University of Ireland Galway, Galway, Ireland
bharathi.raja@insight-centre.org

1 Introduction

Easier internet access has aided social media platform users in communicating and expressing their opinions about anything without censorship [1]. Memes have been used to communicate over the last decade, representing users' intentions about specific topics. Memes come in a variety of forms, including image, text, and video. They are frequently used to disseminate knowledge, emotions, ideas, and talents. Because of their widespread popularity, government agencies and industry professionals use memes on their social media accounts to promote awareness programmes, advertise their products and ideas, and so on [2]. A meme that is funny to one person may be offensive to another. The main feature of a meme is that it can be changed, recreated, and frequently taken out of context for sarcastic purposes [3,4]. Several memes, however, are created to denigrate people based on their gender, sexual orientation, religious beliefs, or any other opinions, which is often regarded as trolling and may cause distress in the online community [5].

Image with text (IWT) memes are the most popular memes available on social media apps. IWT memes [6] make decoding its intention or any other characteristics [7,8]. The comprehensive analysis of IWT could elucidate the socio-political and societal factors, their implications on cultures, and the values promoted by them. One alternative to manually moderating memes on social media platforms is to create automated systems that can determine whether a meme is trolling or not.

IWT memes can be found in everyday conversations and on all social media platforms. In multilingual countries like India, where languages represent cultures, a meme could represent a culture and then be used to mock specific cultures and lifestyles. We develop a multimodal approach to classifying images and texts in IWT memes by employing pretrained language models for texts and pretrained vision models for images.

In a meme, attributes such as sarcasm, satire, and irony are usually expressed in the captions, with the images being referenced on occasion. As a result, we test a number of multimodal approaches. The performance of these models is then compared to the performance of transformer-based models for code-mixed Tamil texts. In theory, when combined with textual features during multimodal analysis, visual features should improve the model's overall performance. Despite using transfer learning and smaller datasets (2500 images), it was discovered that visual features had no direct influence on overall performance in identifying memes that troll people. This drop in performance could be attributed to the dataset's nature, as the dataset was obtained by scraping memes from the Internet, which had texts written over the images, and automated image transformation did not remove all of the texts. We believe that the main issue that the models face is a lack of data.

Our contribution are:

- We have created meme classification model for under-resourced Tamil language.
- In this paper, we create architecture by avoiding the use of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in favour of a completely attention-based architecture for both textual and visual analysis (RNN). It was a pure transformer-transformer architecture in which both image and text encoders extracted features using transformers.
- The model performs better than other previously published research.

The rest of the paper is organised as follows. Section 2 shows previous work on analysing memes using unimodal and multimodal approaches. Section 3 consists of a detailed description of the datasets for our purposes. Section 5 talks about the proposed model with Sections 5.1 and 5.2using unimodal and multimodal techniques respectively. We describe the experimental results and error analysis in Sections 6 and Section 7 respectively, and conclude our work and discuss potential directions for our future work in Section 8.

2 Related Work

The analysis of the meme's image falls under the purview of computer vision. However, several methods and techniques for simulating the human visual system have now been developed. The introduction of convolutional neural networks [9] resulted in a breakthrough in this field. Object detection [10], image segmentation [11], biomedical applications [12] and many more. Recent trends include generating high quality images [13] and image translations [14]. Transfer learning has simplified deep learning model training by transferring weights from larger tasks to smaller downstream tasks. Moreover, studies show that transfer learning results and fine-tuning models outperform training models from scratch [15]. Transformers for images significantly improved computer vision through their attention mechanism on images [16]. These consider images as a series of patches and try to attain the conventional attention mechanism on them. Many more researchers followed this in their work of computer vision for numerous other tasks.

As we try to analyse text in our task, we look into Natural Language Processing(NLP). The final goal of analysing texts is to perform repetitive tasks like summarization [17,18], language translation[19,20], spam classification[21] and many more [22,23]. NLP is performed by preprocessing texts and converting them into meaningful numbers/vectors [24,25]. Bag of words [26] methodology was introduced to simplify and retrieve information from the text. The point was to give numbers and convert them into n-grams. For example, a bi-gram would be a phrase with two words. Different words will be added to the next n-gram, and the shortage of words is handled by padding the n-grams. This extra data caused input data to have more features, thus decreasing performance due to the curse of dimensionality. It also assumed words are independent of each other. So every word was represented in a vector space [27], providing a meaning to the words and thus helped in sequential models. Nevertheless, the breakthrough in NLP was with the introduction of attention mechanism [28] on the sentences. The model performed immensely well on almost all NLP tasks. Additionally, such models could be fine-tuned for downstream tasks. The models had millions of parameters and required higher computational power. In our work, we try to accommodate transformers for text analysis.

Multimodal analysis is the interpretation and interpretation of qualitative data in projects that combine verbal and nonverbal forms of information [1]. It is the extraction of infor-

[1] https://methods.sagepub.com/foundations/multimodal-analysis

mation retrieval process that accepts the joint representation of all modalities used in the system. Because it combines the properties of various aspects, it may appear that adding more valuable data will improve results; however, this is not always the case. It can also add constant noise and other losses [29]. Our work focuses on the fundamental ideas underlying how image and text data can be combined, as well as the importance of multimodal analysis in the classification of Tamil troll memes. There has been signninficant research involved in multimodal analysis for hate speech detection using models such as BERT and InceptionV3[30] and generating captions for images in order to improve context, and hence relying on textual analysis [**?**]. Researchers developed a multimodal sentiment analysis system by devising a deep neural network that combines both visual and text analysis to predict the emotional state of the user using Tumblr posts [31]. This analysis highlights the significance of analysing emotions hidden behind the user's uploads based on their day-to-day life cycle. Such sentiments can be discovered and thus discarded or retained by the administrator based on the application's specific criteria. Several systems were submitted as part of a shared task on classifying Tamil Troll memes on Dravidian languages [5]. A total of ten systems were submitted to the shared task, with the majority of the researchers treating it as a multimodal task, attempting to address the textual and visual features simultaneously by developing deep neural networks that jointly learn from the features. Several researchers have collaborated on the creation of multimodal datasets for video, audio, and text. [32,33].

Our contribution is an extension of our previous work [34]. This paper highlighted its architecture by avoiding the use of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in favour of a completely attention-based architecture for both textual and visual analysis (RNN). It was a pure transformer-transformer architecture in which both image and text encoders extracted features using transformers. The model received a perfect F1 score of 1.0 on the train and validation sets, but received an F1 score of 0.47 on the test dataset due to the same issues discussed in Section 1.

3 Dataset

We use the troll classification dataset of Tamil Memes [35]. It consists of 2,699 memes, of which most of the images have text embedded within them. We were also provided with captions for all images. Fig.1(a) and Fig.1(b) have the code-mixed Tamil-English captions embedded into the image. We have added these images for easier understanding. Fig.1(a) tries to troll the pack of chips, stating that *If you buy a packet of air, 5 chips are completely free*, while Fig.1(b) does not intend to be trolling, only intending to sarcastically be apologetic to *girlfriend*. The images are certain still frame from a

(a) Example of an image belonging to a troll class

(b) Example of an image belonging to non-troll class

Fig. 1: Examples of the dataset

movie or TV shows in Tamil languages. The distribution is shown in Table 1. The dataset consists of two classes:

- **Troll:** A troll meme is an implicit image that intends to demean or offend an individual on the internet.
- **Non-Troll:** A meme that does not intend to demean or offend anyone on the internet is non-troll.

Class	Train	Validation	Test
Troll	1,154	128	395
Non-Troll	917	101	272
total	2,071	229	667

Table 1: Dataset Distribution

4 Motivation

Internet trolls are gaining increasing power in society as a result of the rapid rise of social media. A troll farm is a group

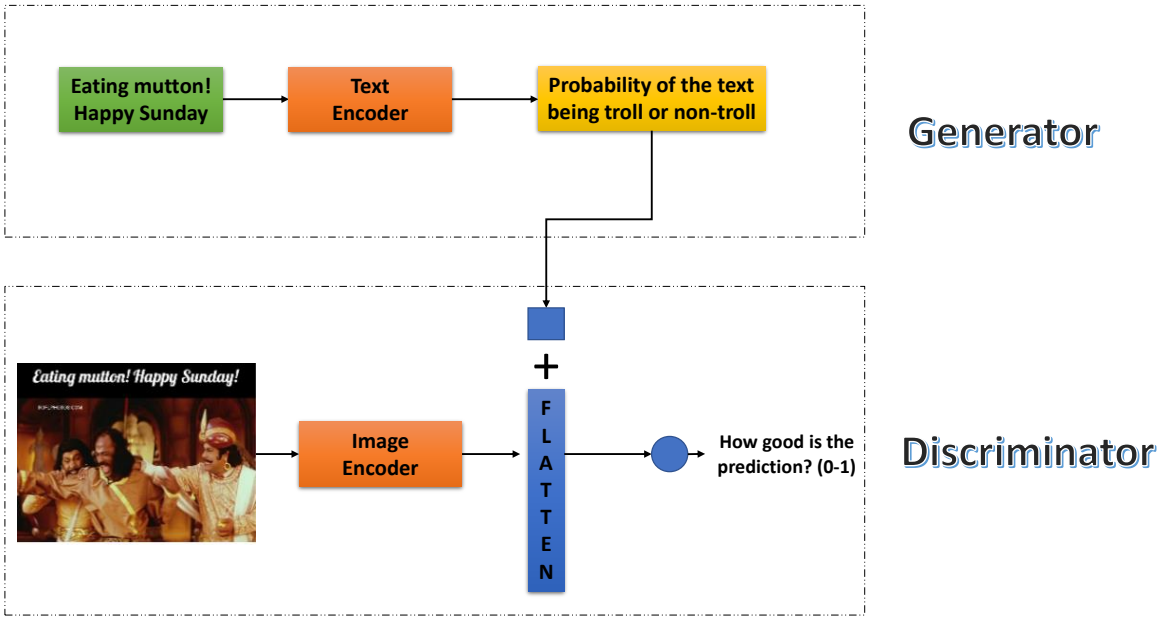

Fig. 2: Architecture of Adversarial learning

of internet trolls who are paid to spread particular information or ideas. Troll farms are difficult to notice since the trolls try to fit in with their surroundings. This study analyses if troll farms can be identified using sentiment analysis on Tamil memes and modelling it as a multi-modal analysis. The project entailed gathering and categorising features from photographs and words, as well as attempting to combine their retrieved features. The findings of various combinations of qualities show that there is no obvious link between the photographs and the phrases. The models used are state of the art and have been performing well in the benchmarking datasets[36].

Trolling can be particularly painful, distressing, and inexplicable for inexperienced or vulnerable Internet community members who trust trolls, are emotionally involved, or communicate private information; given the distributed and asynchronous nature of online discussions, this may have long-term consequences. Although clearly harmful, these acts are popular and generally allowed, in part because libertine beliefs prevalent on the Internet regard offensive speech as a form of expression. Malicious users can utilise CMCs to perform crimes like defamation, stealing other people's identities, and cyberbullying. To combat this, several online communities incorporate identity verification methods and restrict features that allow users to communicate with one other at the same time. Even if trolling does not come as a direct attack, it can still be a threat because it might appear in subtler ways, such as as a tool to alter people' ideas. Indeed, the emergence of the Internet has enabled businesses, organisations, and governments to openly distribute false rumours, misinform and speculate, and engage in other unethical practises to polarise public opinion. It has been demonstrated that the remarks of other users can impact a user's attitude on specific products or politics. Companies and political parties who use reputation management services, i.e. those paid to hijack people's opinions on their behalf, might achieve more popularity this way.The complexity of current social media makes identifying and banning trolls a difficult process. Although it is necessary to educate users about trolling, such warnings do not lessen the occurrence.

5 Models

5.1 Textual Analysis

This section discusses the natural language models that are used to determine whether a given meme is a troll or not based on its captions. The Tamil Troll meme dataset [35] is made up of two parts: images and captions, which are both provided separately. Six pretrained language models have been fine-tuned.

5.1.1 Bidirectional Encoder Representation Transformer (BERT):

BERT is a language model that pretrains unlabelled data using deep bidirectional representations [37]. Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)

are the two pretraining strategies used by BERT (NSP). The authors had masked 15% of the words that would be predicted later during pretraining. Next Sentence Prediction predicts whether a given sentence will follow the previous sentence. BERT has been pre-trained on eleven downstream NLP tasks. The tokeniser inserts unique tokens, [CLS] and [SEP], at the beginning and end of each sentence during tokenisation. During fine-tuning, we extract the output of the pooled layer ([CLS] token) to predict the test set. We make adjustments. Multilingual BERT, a multilingual language model pretrained for the top 104 languages available on Wikipedia, dumps.. In mBERT [38], the high-resource languages are downsampled to address the data imbalance during pretraining.

5.1.2 ALBERT

In today's State of the Art (SoTA) LMs, there are hundreds of millions, if not billions, of parameters. The memory constraints of computational hardware such as GPUs or TPUs would restrict our ability to scale the models. It was also discovered in the BERT-large model (340M parameters) that increasing the number of hidden layers can lead to poor performance. Several approaches to parameter reduction have been used to reduce the size of models without affecting their performance. As a result, ALBERT: A lite BERT[39] was proposed to use self-supervised language representation learning for language representation learning. ALBERT solves the problem of high memory usage by implementing multiple memory reduction strategies, such as Factorized Embedding Parameterization, cross-layer parameter sharing, and Sentence Ordering Objectives.

5.1.3 XLM-RoBERTa:

XLM-RoBERTa is a multilingual language model that improves its overall performance in cross-lingual understanding through the use of self-supervised techniques. XLM-R has been pretrained on over 2.5 TB of unlabeled data, with a focus on low-resource languages. XLM-R outperforms its siblings, RoBERTa and XLM, by employing BPE (Byte-Pair Encoding) as a preprocessing technique rather than the workpiece tokeniser used in BERT. XLM-Ruses dual-language modelling with Translated Language Modeling (TLM) pretrained over BPE to achieve cutting-edge results in a variety of downstream tasks, outperforming BERT. There are three language modelling strategies in XLM-R.
(i) **Masked Language Modeling (MLM):** This language modelling is similar to the approach used in monolingual 'Vanilla' BERT.
(ii) **Translated Language Modeling (TLM):** To achieve TLM, every training sample consisted of texts in two languages, with the intention that one model uses the context of one language to predict the tokens of the other language while retaining the same strategy of masking the words randomly.
(iii) **Causal Language Modeling (CLM):** In CLM, a given training sample is trained only based on the existence of previous words while not using any masking strategies.

5.1.4 XLM:

The XLM model was proposed in cross-lingual language model pretraining [40]. For different languages, this model employs a shared vocabulary. Byte-Pair Encoding (BPE) was used to tokenize the text corpus. The goal of Causal Language Modelling (CLM) is to maximise the likelihood of a token x_t to appear at the tth position in a given sequence. Both CLM and MLM perform well on monolingual data.

5.1.5 Multilingual Representations for Indian languages (MuRIL):

MuRIL is a language model that focuses on Indian languages, which is not observed in other multilingual models, as the latter are pretrained over hundreds of languages, inherently resulting in the smaller representations of Indian languages [41].MuRIL, which supports 17 Indian languages including English, was introduced to address the low representations in other multilingual language models. It is based on the architecture of the BERT base model, with the only difference being the pretraining strategies and data used. MuRIL, like XLM-R, employs both supervised and unsupervised language modelling approaches, with conventional MLM employing monolingual data for pretraining and TLM employing both translated and transliterated document pairs during pretraining. To smooth the data and address data imbalance during pretraining, low-resource data are upsampled while high-resource data are downsampled. The model is pre-trained from the ground up using Wikipedia[2], Common Crawl[3], PMINDIA[4] and Dakshina corpora [42].

5.1.6 Roberta

RoBERTa [43] is a robustly optimised BERT-based model. The key difference is in the masking technique. Because BERT only performs masking once during the input processing phase, which is essentially a static mask, the final model tends to see the same type of masks across multiple training rounds. RoBERTa was designed to create a dynamic mask within itself, changing the masking pattern with each input

[2] https://www.tensorflow.org/datasets/catalog/wikipedia
[3] http://commoncrawl.org/the-data/
[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html

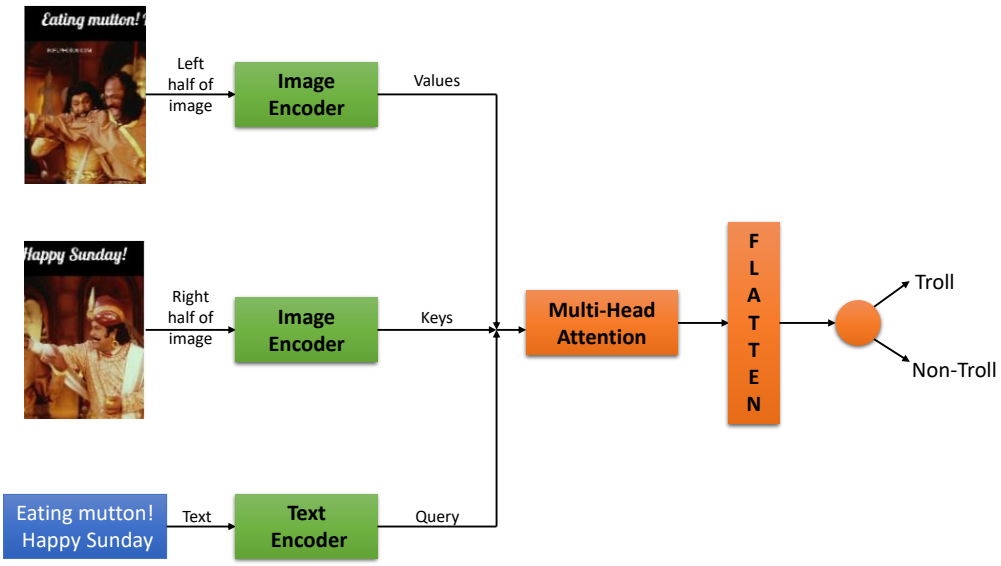

Fig. 3: Architecture of EmbracenNet

sequence, which was critical for pretraining. Byte-Pair Encoding (BPE)[44] was a hybrid encoding between character and word level encoding that allows for easier management of large text corpora by relying on subwords rather than whole words. The model was built to predict the words using an auxiliary NSP loss. Even BERT was trained on this loss, and it was discovered that without it, pretraining had a negative impact on performance, with significant reductions in QNLI and MNLI scores.

5.1.7 TaMillion BERT:

TaMillion BERT is a monolingual language model that follows the architecture of Efficiently Learning an Encoder that Classifies Token Replacements Accurately(ELECTRA)[45] pretrained on 11GB of IndicCorp Tamil[5] and the Wikipedia dumps[6] (482 MB) as of October 1, 2020. We use the second version of TaMillion BERT, which has been pretrained on TPU with 224,000 steps. On classification tasks, this model significantly outperforms mBERT [38]. ELECTRA, like GANs, is an architecture that has been trained to distinguish between fake and real tokens.

5.1.8 XLNet

BERT performed admirably on virtually every assignment involving language modelling. Because it could be fine-tuned for any subsequent challenge, it was a game-changing model. However, even this had some flaws of its own. BERT was designed in such a way that it replaces random words in phrases with a specific [MASK] token and attempts to guess what the original word was. During the process, XLNet[46] raised several serious concerns. During fine-tuning the model and other downstream operations, the [MASK] token that was used during training would not display. However, forgetting to replace [MASK] tokens at the end of pretraining could exacerbate the problem. Furthermore, when the input sentence has no [MASK] tokens, the model finds it difficult to train. BERT also makes forecasts separately, which means it is unconcerned about the dependencies between them.

XLNet uses Auto Regressive language modelling to estimate the probability distribution of a text corpus without using the [MASK] token and making parallel independent predictions. It is accomplished through AR modelling, which provides an acceptable method for defining the product rule of factoring the expected tokens' joint probability. XLNet employs a type of language modelling known as "permutation language modelling," in which tokens for a given sentence are predicted in random order rather than sequential order. For all input combinations, the model is forced to learn bidirectional model relationships. It should be noted that it only permutes the factorization order, not the sequence order, and that positional embedding is used to reorganise and restore the original form.

The model is fine-tuned for sentence classifier for sequence classification, and it predicts sentiment rather than

[5] https://indicnlp.ai4bharat.org/corpora/
[6] https://ta.wikipedia.org

tokens based on the embedding. The architecture of the XL-Net is based on the transformer XL[47]. The transformer adds recurrence to the segment level rather than the word level. As a result, fine-tuning is achieved by caching the hidden states of previous states and using them as keys or values in the current sequence. The transformer employs relative embedding rather than positional embedding by recording the relative distance between the words.

5.1.9 Language-agnostic BERT Sentence Embeddings (LABSE):

LABSE are created by modifying a multilingual BERT. Unlike previous multilingual language models that were used to generate English sentence embeddings by fine-tuning pretrained BERT, these models were not used to generate multilingual sentence embeddings. The LABSE model will now combine MLM and TLM pretraining with a translation ranking task utilising bi-directional dual encoders [48]. To train the cross-lingual embedding space productively, LABSE supports 109 languages that use the approach of adopting a pre-trained BERT encoder model to dual encoder model.

Because of their extensive pre-training on massive datasets, all of the models used here are extremely powerful in handling Tamil texts. These transformers would improve the results of categorising memes based on their texts. Millions of trained parameters in these transformers can contribute significantly to our downstream task. Cross-Lingual models such as XLM and XLM-R were used because the representations of one language influence the predictions of another. As a result, these models perform admirably with code-mixed and roman script.

5.2 Multimodal Analysis

We hope to improve the test results by extracting features and including them with the images. When the features of text and images are encoded, they can be combined in a variety of ways. We try to focus on the most popular architectures used in multimodal analysis and experiment with the hyperparameters to get the best model out of them.

5.2.1 Concatenation

Concatenation is a simple method for combining features. Images are fed into a model fine-tuned for this task after being trained on the ImageNet dataset [49], and embedded texts are fed into a multilingual text model that is fine-tuned to obtain the features. The outputs of both sections are then merged by stacking one on top of the other to obtain binary classification predictions. Advanced models compete in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [50] competition, with the goal of classifying millions of images into thousands of classes. Users fine-tune those pretrained models on other downstream tasks, reducing computation power usage and training time. The vision transformer [16] is one of these models. Continuing with the sentence analogy, instead of 1D token embeddings as input, ViT receives a sequence of flattened 2D patches. If H, W are the image's height and width, and (P, P) is the resolution of each patch, then $N = HW/P^2$ is the transformer's effective sequence length. The patches are then projected linearly and multiplied by an embedding matrix to form patched embeddings. The patches, as well as position embeddings, are routed through the transformer. In addition, a [CLS] token is appended to determine the class. The number of classes would be the hidden shape used to encode the features in this case. The same token is used to determine the encoded shape as well as to extract features from the text using multilingual BERT. The encoded parts are combined to form a single layer with the shape of the encoded image + encoded text shape. This encoded text is then carried over to a single output determining the probability of the text being Troll or Non-Troll. The architecture is shown in Fig.4.

5.2.2 EmbraceNet

For multimodal classification, citechoi2019embracenet, a robust deep learning architecture, was used. In addition, an extra layer of multi-head attention was used to blend the features. Instead of feeding a whole image, the architecture suggested feeding half of the images, either vertically or horizontally cropped, into two distinct models. Because memes can contain multiple images, dividing the photos into halves increases the likelihood of dividing several images. To extract features, the photos were split vertically down the middle and loaded into various imagenet models. The text was run through a multilingual transformer, as is customary. The model generates three outputs: two from images and one from text. Using multi-head attention, these three were combined using encoded text as a query and encoded halves of the images as keys and values. The output of this layer was fed into linear layers, which generated classification probabilities to determine whether memes were trolls or not. Multi-Head Attention is a concept in which multiple workers, referred to as heads, perform self-attention tasks at the same time. It is also known as scaled dot-product attention, and it is calculated mathematically using three vectors from two image encoders and one text encoder, *Query, Key* and *Value* vectors. *Key* and *Value* assume dimensions d_k and d_v respectively. A softmax function is applied on the dot product of queries and keys to compute the weights of the values. In reality, the attention function is computed continuously on a set of queries and then stacked into a matrix Q, being packed into a matrix Q. The *Keys* and *Values* are packed into matrices K and V. The matrix of outputs is computed as follows:

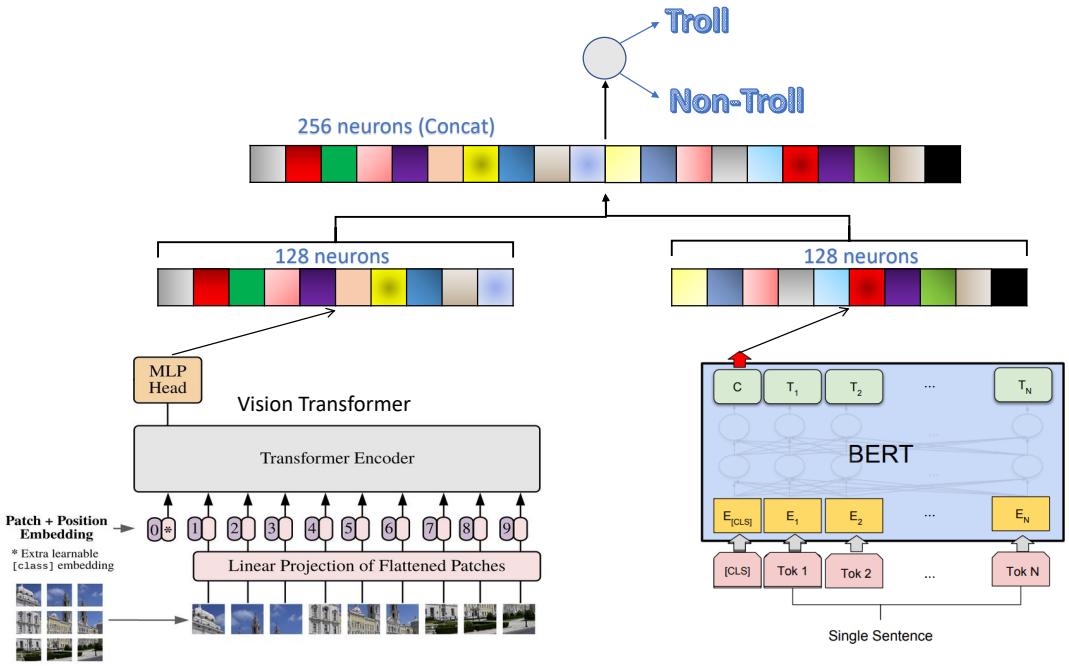

Fig. 4: Concatenating textual and visual features [34]

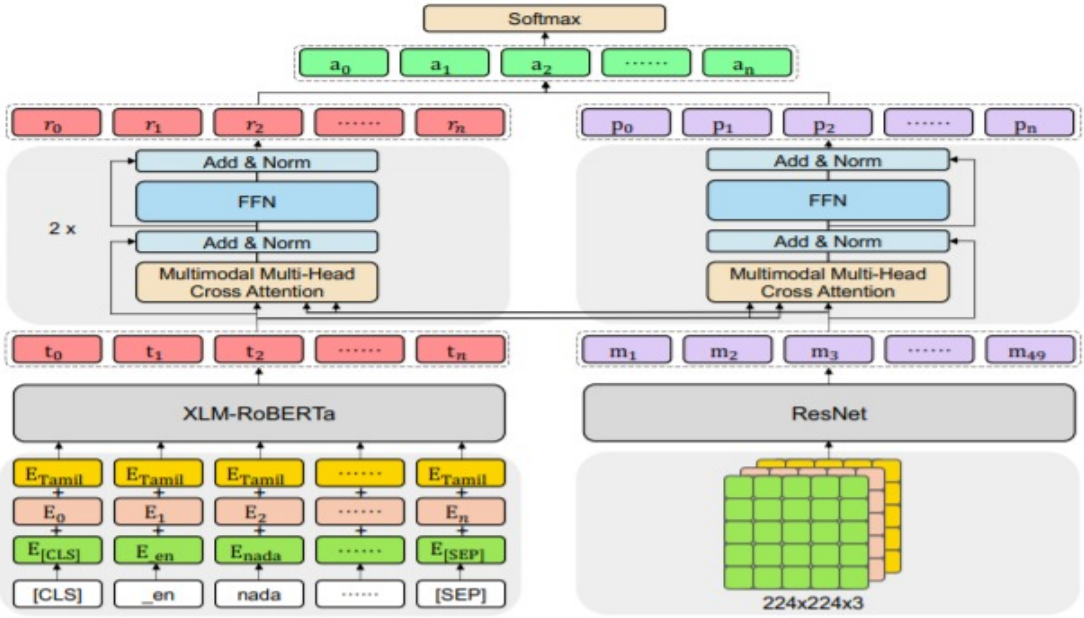

Fig. 5: Architecture of Multimodal transformers [51]

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

This equation represents a single head of self-attention. Similarly, multiple heads are parallel, and each head computes a different set of attention weights and outputs. The architecture of EmbraceNet is shown in Fig.3. We use different image encoders for EmbraceNet. There are many models which attained state of the art. We mainly focus on VGG, ResNet and InceptionV3.

- VGG: TThe impact of convolutional neural network depth on accuracy is the primary focus of this [52] design. Following preprocessing, the input photos are routed through these weight layers. The training images are sent through a stack of convolution layers. There are 13 convolutional layers and three fully linked layers in the VGG16 architecture. Instead of large filters, VGG has smaller (3*3) filters with greater depth. It now has the same effective receptive field as if it had only one 7 7 convolutional layer. Another VGGNet variant has 19 weight layers, 16 convolutional layers, 3 fully connected layers, and the same 5 pooling layers. In both variations, VGGNet has two completely connected layers with 4096 channels each, followed by another fully connected layer with 1000 channels to predict 1000 labels. Softmax layer is used for categorization in the last fully linked layer. We use the VGG-19 to evaluate the scores.
- ResNet: This design [53] introduces the concept of the Residual Network to solve the vanishing/exploding gradient problem. In this network, we use a technique known as skip connections. The skip connection skips a few stages of training and connects directly to the output. We let the network fit the residual mapping instead of allowing layers to learn the underlying mapping. The creation of a network is the result of stacking these networks. ResNet-50 is formed by stacking 50 such networks. Such networks have proven to be effective for a variety of backbones, including object recognition [54] and image classification [55].
- Inception: Inception Networks (GoogLeNet/Inception v1 [56]) have been shown to be more computationally efficient than VGGNet, both in terms of the number of parameters generated and the cost incurred (memory and other resources). When changing an Inception Network, special care must be taken to ensure that the computational advantages are not lost. As a result of the unpredictable efficiency of the new network, adapting an Inception network for multiple use cases becomes a challenge. In an Inception v3 [57] model, several strategies for improving the network have been proposed to loosen the restrictions for easier model adaptation. Among the approaches used are factorised convolutions, regularisation, dimension reduction, and parallelized calculations.

5.2.3 Multimodal Transformers

This is a concept that was created to achieve cross attention between two different modalities [51]. The text encoder was XLM-Roberta [58], which was trained on the 100 languages Common Crawl dataset, and the image feature extraction was ResNet [59] with 152 layers, followed by linear transformation to match the word embedding size of the text encoder. The next type of attention is multihead cross attention, which is similar to the conventional multi-head attention described in the previous section. The first step was to create word representations for each image that were attentive. If Q is the query and K is the key, then R is given by,

$$A = LN(Q + Attention(Q, K, K)) \quad (2)$$

$$R = LN(A + FFN(A)) \quad (3)$$

where, LN is a layer normalisation and FFN is feed forward network. Similarly attentive image representations are generated for each word which now takes in K as query and Q as key, and is represented by I,

$$Z = LN(K + Attention(K, Q, Q)) \quad (4)$$

$$I = LN(Z + FFN(Z)) \quad (5)$$

Finally, both the representations were concatenated and subjected to average pooling to then get the probability of the class as shown in Fig.5.

5.2.4 Adversarial Learning

This type of concept was initially developed by simultaneously training two models: a generative model G that captures the data distribution and a discriminative model D that estimates the probability that a sample came from the training data rather than G [13]. Adversarial learning is a two-player minimax game in which one player tries to minimise its loss while increasing the loss of the other player and vice versa. The idea originated with the goal of image generation, and this process can be used in multimodal analysis to test how well texts are classified. Using a transformer, we create a generative model G for Tamil text classification. As shown in Fig 2, this generator returns a value indicating the likelihood of the text being troll or non-troll.

G uses multilingual BERT for encoding, and the model is followed by linear layers and ReLU activation functions, yielding a single probability with sigmoid activation indicating the text class. D begins with an ImageNet model, such as ResNet [59], for image encoding, and is followed by linear layers with a ReLU activation function. The two models are linked by the fact that the probability of G is concatenated with the encoded image features and then subjected to a single probability that indicates how accurate the prediction of G was. The two models compete with each other to learn the parameters. To predict the classes, the test set was fed into the only G.

Team Name	Precision	Recall	F1-Score
Our Approach	**0.60**	0.58	**0.57**
Codewithzichao [51]	0.57	**0.60**	0.55
IIITK [1]	0.56	0.59	0.54
NLP@CUET [60]	0.55	0.58	0.52
SSNCSE NLP [61]	0.58	0.60	0.50
Simon [62]	0.53	0.58	0.49
TrollMeta [63]	0.45	0.41	0.48
UVCE-IIITT [34]	0.60	0.60	0.46
HUB [64]	0.50	0.54	0.40
IIITDWD [2]	0.52	0.59	0.30

Table 2: Comparisons of the existing models developed for the Tamil Troll meme dataset, as a part of the shared task [5]

Model	Troll			Not-troll			Overall			
	P	R	F1	P	R	F1	Acc	W_{avg}(P)	W_{avg}(R)	W_{avg}(F1)
mBERT	0.60	**0.95**	0.73	0.55	0.08	0.14	0.59	0.58	0.59	0.49
DistilmBERT	0.60	0.88	0.72	0.51	0.17	0.26	0.59	0.56	0.59	0.53
XLM-R_$base$	0.60	0.90	0.72	0.52	0.15	0.23	0.59	0.57	0.59	0.52
XLM	**0.62**	0.81	0.70	0.52	0.29	0.37	0.60	0.58	0.60	**0.57**
MuRIL	0.60	0.85	0.71	0.49	0.19	0.28	0.58	0.56	0.58	0.53
TamillionBERT	0.42	0.36	0.39	0.59	0.65	0.62	0.53	0.52	0.53	0.52
LABSE	0.50	0.20	0.29	0.61	**0.85**	**0.71**	0.59	0.56	0.59	0.54
ALBERT	0.59	0.74	0.70	0.48	0.20	0.27	0.51	0.50	0.53	0.52
RoBERT	0.59	0.87	0.69	0.50	0.13	0.20	0.57	0.55	0.54	0.51
XLNeT	0.59	0.88	0.70	0.52	0.15	0.23	0.58	0.58	0.59	0.52
Concatenation	0.60	0.98	**0.74**	**0.60**	0.03	0.06	0.60	0.60	0.60	0.47
Multimodal Transformers	0.61	0.75	0.67	0.46	0.31	0.37	**0.61**	0.57	0.60	**0.55**
EmbraceNet	0.60	0.95	0.74	0.56	0.09	0.15	0.60	0.58	0.60	0.50
Adversarial Model	0.59	0.96	0.73	0.39	0.03	0.06	0.58	0.51	0.58	0.46

Table 3: Results of Textual architectures and Multimodal architectures considering both images and texts

6 Results and Analysis

All the experiments were conducted on Google Colaboratory[7] accelerated by GPU. For the textual analysis, the transformers were fine-tuned with an optimal learning rate of 2e-5 with Adam optimiser and warmup with the scheduler. The batch size used was chosen among 16,32, and 64 sets according to the computational power. The results of the test set can be observed in Table 5.2.3. Among all the transformers, XLM scored the highest with a weighted F1 score of 0.57 and next to it is the Language agnostic BERT with a weighted F1 score of 0.54. We have listed the heatmap of the confusion matrix of XLM in Fig.6. In other textual models, we observe that the performance of the models is quite similar to each other.

Surprisingly, the Adversarial Model and Concatenation have the lowest scores among the multimodal techniques. Because transformers are massive models with millions of parameters, it's difficult to train them perfectly for small datasets. It used thousands of phrases and overfits both the training and validation sets by more than 0.9. The test set findings, on the other hand, are lower due to the models' substantial variation.

[7] https://colab.research.google.com/

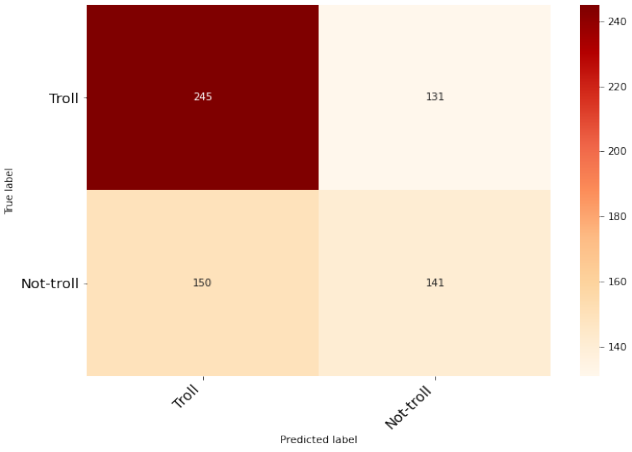

Fig. 6: Heatmap of Confusion matrix for the best performing model

In comparison to the submitted systems for the Tamil Troll meme detection shared task, and we observe that the text-based unimodal models perform better than multimodal models, as we achieve the best-weighted F1-Score among all the teams. The scores of all the teams are listed in Ta-

Fig. 7: Losses during training

Fig. 8: Example image in the test set belongs to troll class

Fig. 9: Example image in the test set belongs to non-troll class

ble 2. Most of the teams submitted unimodal systems that relied on pretrained natural language models to capture its textual features [1], while some teams resorted to multimodal approaches that performed worse than the textual models [51,64,60]. The use of images could have improved the outcome if the images contained valuable features that could predict the probability. The results in Table 3 show that there is no significant increase in the metrics. The Adam optimiser was used to determine the exact optimal learning rate. The Multimodal Transformers, on the other hand, received the highest score, with a weighted F1 score of 0.55. The transformers overfit the training and validation sets, and the cross attention added more variance to the training set as well. Despite the fact that the images were cropped in the centre to remove the standard text at the top and bottom, the test set had multiple images merged in them and text all over the meme. The merged text was a significant disadvantage because ImageNet models could not extract features due to the difference in kernel size and the text present. The vision models would dismiss them as mere noise. Even a sequential feature extractor, such as a vision transformer, was unable to extract features from the image's embedded text. Thus, images are an extra computation that contributes nothing to the model in this task. It is preferable to analyse only the text rather than combining both modalities. Textual analysis is a preferred solution for Tamil Meme Classification, with both computational power and performance as deciding factors.

7 Error Analysis

Textual models performed better than multimodal approaches. One reason for this is that ImageNet-trained features are inadequate for detecting troll classes. ImageNet and ResNet models are used to classify everyday objects such as people, cars, and food. These items can be found in the meme collection, but they also have text and images embedded in them. Because these models are complex and deep, training them without fine-tuning them results in a high level of variance, which can lead to over-fitting. Higher-level information must be extracted from the photos. The sample size of the dataset, on the other hand, is one of the study's shortcomings. The dataset contains only 1154 training photos. According to the analysis of the photos in the dataset, the emotion-related variables do not contribute to detecting troll or non-troll classes. From Fig.8 & 9, it has been noted that both images and facial expressions are almost the same however the classes are different. Another reason is that few images in the test set had only text embedded in all over the image which is shown in Fig.9

8 Conclusion

This study investigates whether it is necessary to include all modalities in a study. It is not always true that features

Fig. 10: Example image with only the text embedded

from all types of data contribute significantly to outcome prediction. Multimodal analysis topologies range from the most basic vanilla concatenation model to the most complex multimodal transformers with cross attention. None of the multimodal models outperformed the textual model XLM, which had a weighted F1 score of 0.57. It is important to consider the distribution of the test set, and the types of photos in the test set were distinct. This could have a significant impact on the fine-tuning performance of ImageNet models. Taking into account all modalities results in a massive improvement in the outcome in some cases, and the processing power is not squandered. In experiments like this one of meme categorization with limited data, it is always preferable to use unimodal analysis rather than multimodal analysis. We plan to translate these works into other languages in the future. This will be expanded to include other projects.

Acknowledgements The author Bharathi Raja Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2), co-funded by the European Regional Development Fund and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

Funding

This research has not been funded by any company or organisation

Compliance with Ethical Standards

Conflict of interest: The authors declare that they have no conflict of interest.

Availability of data and material: The dataset used in this paper are obtained from `https://zenodo.org/record/4765573/`.

Code availability: The data and approaches discussed in this paper are available at `https://github.com/adeepH/MemeClassification`.

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Ghanghor, N., Krishnamurthy, P., Thavareesan, S., Priyadharshini, R., Chakravarthi, B.R.: IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 222–229. Association for Computational Linguistics, Kyiv (2021). URL `https://www.aclweb.org/anthology/2021.dravidianlangtech-1.30`
2. Mishra, A.K., Saumya, S.: 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, pp. 243–248. Association for Computational Linguistics, Kyiv (2021). URL `https://www.aclweb.org/anthology/2021.dravidianlangtech-1.33`
3. Grundlingh, L.: Memes as speech acts. Social Semiotics **28**(2), 147–168 (2018)
4. French, J.H.: Image-based memes as sentiment predictors. In: 2017 International Conference on Information Society (i-Society), pp. 80–85 (2017). DOI 10.23919/i-Society.2017.8354676
5. Suryawanshi, S., Chakravarthi, B.R.: 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, pp. 126–132. Association for Computational Linguistics, Kyiv (2021). URL `https://www.aclweb.org/anthology/2021.dravidianlangtech-1.16`
6. Du, Y., Masood, M.A., Joseph, K.: Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. Proceedings of the International AAAI Conference on Web and Social Media **14**(1), 153–164 (2020). URL `https://ojs.aaai.org/index.php/ICWSM/article/view/7287`
7. Avvaru, A., Vobilisetty, S.: BERT at SemEval-2020 task 8: Using BERT to analyse meme emotions. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1094–1099. International Committee for Computational Linguistics, Barcelona (online) (2020). URL `https://www.aclweb.org/anthology/2020.semeval-1.144`
8. Nave, N.N., Shifman, L., Tenenboim-Weinblatt, K.: Talking it personally: Features of successful political posts on facebook. Social Media + Society **4** (2018)
9. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. ArXiv e-prints (2015)
10. Park, H., Park, S., Joo, Y.: Detection of abandoned and stolen objects based on dual background model and mask r-cnn. IEEE Access **8**, 80010–80019 (2020). DOI 10.1109/ACCESS.2020.2990618
11. Gurusamy, V., Kannan, S., Nalini, G.: Review on image segmentation techniques. J Pharm Res **20125**, 4548–4553 (2013)

12. Yin, P., Yuan, R., Cheng, Y., Wu, Q.: Deep guidance network for biomedical image segmentation. IEEE Access **8**, 116106–116116 (2020). DOI 10.1109/ACCESS.2020.3002835
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134 (2017)
15. Venkatesh, Yallappa, N., Hegde, S.U., Stalin, S.R.: Fine-tuned mobilenet classifier for classification of strawberry and cherry fruit types. Journal of Computer Science **17**(1), 44–54 (2021). DOI 10.3844/jcssp.2021.44.54. URL https://thescipub.com/abstract/jcssp.2021.44.54
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Moratanch, N., Gopalan, C.: A survey on extractive text summarization. pp. 1–6 (2017). DOI 10.1109/ICCCSP.2017.7944061
18. Obaidullah, S.M., Santosh, K., Halder, C., Das, N., Roy, K.: Automatic indic script identification from handwritten documents: page, block, line and word-level approach. International Journal of Machine Learning and Cybernetics **10**(1), 87–106 (2019)
19. Eria, K., Jayabalan, M.: Neural machine translation: A review of the approaches. Journal of Computational and Theoretical Nanoscience **16**, 3596–3602 (2019). DOI 10.1166/jctn.2019.8331
20. Ali, M.N.Y., Rahman, M.L., Chaki, J., Dey, N., Santosh, K.: Machine translation using deep learning for universal networking language based on their structure. International Journal of Machine Learning and Cybernetics pp. 1–12 (2021)
21. Dhanaraj, K.R., Thiag, H., Chakkaravarthi, M., Surya, P.: Spam classification based on supervised learning using machine learning techniques. ICTACT Journal on Communication Technology **2** (2011). DOI 10.1109/PACC.2011.5979035
22. Ghosh, M., Roy, S.S., Mukherjee, H., Obaidullah, S.M., Santosh, K., Roy, K.: Understanding movie poster: transfer-deep learning approach for graphic-rich text recognition. The Visual Computer pp. 1–20 (2021)
23. Guha, R., Das, N., Kundu, M., Nasipuri, M., Santosh, K.: Devnet: an efficient cnn architecture for handwritten devanagari character recognition. International Journal of Pattern Recognition and Artificial Intelligence **34**(12), 2052009 (2020)
24. Soumya, U., Swarnendu, G., Md, O.S., Santosh, K., Kaushik, R., Nibaran, D.: Improved word-level handwritten indic script identification by integrating small convolutional neural networks. Neural Computing & Applications **32**(7), 2829–2844 (2020)
25. Obaidullah, S.M., Santosh, K., Das, N., Roy, K.: Handwritten indic script identification–a multi-level approach. In: Computational Intelligence, Communications, and Business Analytics: Second International Conference, CICBA 2018, Kalyani, India, July 27–28, 2018, Revised Selected Papers, Part II, vol. 1031, p. 109. Springer (2019)
26. Yan, D., Li, K., Gu, S., Yang, L.: Network-based bag-of-words model for text classification. IEEE Access **8**, 82641–82652 (2020). DOI 10.1109/ACCESS.2020.2991074
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
29. Guo, W., Wang, J., Wang, S.: Deep multimodal representation learning: A survey. IEEE Access **7**, 63373–63394 (2019). DOI 10.1109/ACCESS.2019.2916887
30. Singh, B., Upadhyay, N., Verma, S., Bhandari, S.: Classification of hateful memes using multimodal models. In: Data Intelligence and Cognitive Informatics, pp. 181–192. Springer (2022)
31. Hu, A., Flaxman, S.: Multimodal sentiment analysis to explore the structure of emotions. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018). DOI 10.1145/3219819.3219853. URL http://dx.doi.org/10.1145/3219819.3219853
32. Jain, N., Gupta, V., Shubham, S., Madan, A., Chaudhary, A., Santosh, K.: Understanding cartoon emotion using integrated deep neural network on large dataset. Neural Computing and Applications pp. 1–21 (2021)
33. Ghosh, M., Mukherjee, H., Obaidullah, S.M., Santosh, K., Das, N., Roy, K.: Lwsinet: A deep learning-based approach towards video script identification. Multimedia Tools and Applications pp. 1–34 (2021)
34. U Hegde, S., Hande, A., Priyadharshini, R., Thavareesan, S., Chakravarthi, B.R.: 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, pp. 180–186. Association for Computational Linguistics, Kyiv (2021). URL https://www.aclweb.org/anthology/2021.dravidianlangtech-1.24
35. Suryawanshi, S., Chakravarthi, B.R., Verma, P., Arcan, M., McCrae, J.P., Buitelaar, P.: A dataset for troll classification of TamilMemes. In: Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pp. 7–13. European Language Resources Association (ELRA), Marseille, France (2020). URL https://www.aclweb.org/anthology/2020.wildre-1.2
36. Hande, A., Hegde, S.U., Priyadharshini, R., Ponnusamy, R., Kumaresan, P.K., Thavareesan, S., Chakravarthi, B.R.: Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. arXiv preprint arXiv:2108.03867 (2021)
37. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). DOI 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423
38. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996–5001. Association for Computational Linguistics, Florence, Italy (2019). DOI 10.18653/v1/P19-1493. URL https://www.aclweb.org/anthology/P19-1493
39. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. CoRR **abs/1909.11942** (2019). URL http://arxiv.org/abs/1909.11942
40. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2978–2988. Association for Computational Linguistics, Florence, Italy (2019). DOI 10.18653/v1/P19-1285. URL https://www.aclweb.org/anthology/P19-1285
41. Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D.K., Aggarwal, P., Nagipogu, R.T., Dave, S., Gupta, S., Gali, S.C.B., Subramanian, V., Talukdar, P.: Muril: Multilingual representations for indian languages. CoRR **abs/2103.10730** (2021). URL https://arxiv.org/abs/2103.10730

42. Roark, B., Wolf-Sonkin, L., Kirov, C., Mielke, S.J., Johny, C., Demirşahin, I., Hall, K.: Processing South Asian languages written in the Latin script: the Dakshina dataset. In: Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 2413–2423 (2020). URL https://www.aclweb.org/anthology/2020.lrec-1.294
43. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019). URL http://arxiv.org/abs/1907.11692
44. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany (2016). DOI 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162
45. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. In: ICLR (2020). URL https://openreview.net/pdf?id=r1xMH1BtvB
46. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. CoRR **abs/1906.08237** (2019). URL http://arxiv.org/abs/1906.08237
47. Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. CoRR **abs/1901.02860** (2019). URL http://arxiv.org/abs/1901.02860
48. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. CoRR **abs/2007.01852** (2020). URL https://arxiv.org/abs/2007.01852
49. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). DOI 10.1109/CVPR.2009.5206848
50. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). DOI 10.1007/s11263-015-0816-y
51. Li, Z.: 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, pp. 352–356. Association for Computational Linguistics, Kyiv (2021). URL https://www.aclweb.org/anthology/2021.dravidianlangtech-1.52
52. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
53. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
54. Mo, N., Yan, L., Zhu, R., Xie, H.: Class-specific anchor based and context-guided multi-class object detection in high resolution remote sensing imagery with a convolutional neural network. Remote Sensing **11**, 272 (2019). DOI 10.3390/rs11030272
55. Ma, Y., Zhang, P., Tang, Y.: Research on fish image classification based on transfer learning and convolutional neural network model. In: 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 850–855 (2018). DOI 10.1109/FSKD.2018.8686892
56. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)
57. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)
58. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
59. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
60. Hossain, E., Sharif, O., Hoque, M.M.: 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, pp. 300–306. Association for Computational Linguistics, Kyiv (2021). URL https://aclanthology.org/2021.dravidianlangtech-1.43
61. B, B., A, A.S.: 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, pp. 336–339. Association for Computational Linguistics, Kyiv (2021). URL https://aclanthology.org/2021.dravidianlangtech-1.49
62. Que, Q.: Simon @ DravidianLangTech-EACL2021: Meme classification for Tamil with BERT. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 287–290. Association for Computational Linguistics, Kyiv (2021). URL https://aclanthology.org/2021.dravidianlangtech-1.41
63. J, M.B., Hs, C.: TrollMeta@DravidianLangTech-EACL2021: Meme classification using deep learning. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 277–280. Association for Computational Linguistics, Kyiv (2021). URL https://aclanthology.org/2021.dravidianlangtech-1.39
64. Huang, B., Bai, Y.: HUB@DravidianLangTech-EACL2021: Meme classification for Tamil text-image fusion. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 210–215. Association for Computational Linguistics, Kyiv (2021). URL https://aclanthology.org/2021.dravidianlangtech-1.28