UVCE-IIITT@DravidianLangTech-EACL2021: Tamil Troll Meme Classification: You need to Pay more Attention

Siddhanth U Hegde[1], **Adeep Hande**[2], **Ruba Priyadarshini**[3], **Sajeetha Thavareesan**[4], **Bharathi Raja Chakravarthi**[5]
[1] University Visvesvaraya College of Engineering, Bangalore University,
[2]Indian Institute of Information Technology Tiruchirappalli, Tamil Nadu,
[3]ULTRA Arts and Science College, India,[4]Eastern University, Sri Lanka
[5]National University of Ireland Galway
siddhanthhegde227@gmail.com

Abstract

Tamil is a Dravidian language that is commonly used and spoken in the southern part of Asia. In the era of social media, memes have been a fun moment in the day-to-day life of people. Here, we try to analyze the true meaning of Tamil memes by categorizing them as troll and non-troll. We propose an ingenious model comprising of a transformer-transformer architecture that tries to attain state-of-the-art by using attention as its main component. The dataset consists of troll and non-troll images with their captions as text. The task is a binary classification task. The objective of the model is to pay more attention to the extracted features and to ignore the noise in both images and text.

1 Introduction

Over the past decade, memes have become a ubiquitous phenomenon over the internet. Memes can come in several formats such as images, video, etc. Memes can take a combined form of both text and images too. Due to its vast popularity, different people perceive memes distinctively. Recent studies have prompted the usage of memes as a mode of communication across social media platforms. The presence of text in images makes it harder to decode the sentiment or any other characteristic (Avvaru and Vobilisetty, 2020). Regardless of the type of the meme, they may be changed, recreated over social media networks, and tend to be used in contexts involving sensitive topics such as politics, casteism, etc, to add a sarcastic perspective (French, 2017; Nave et al., 2018). Due to its multimodality, conscientious analysis of memes can shed light on the societal factors, their implications on culture, and the values promoted by them (Milner, 2013). In addition to that, analyzing the intended emotion of a meme could help us acknowledge fake news, offensive content that is being propagated using the internet memes as a medium, thus helping in eradicating the spread of misinformation and hatred to the large user base in social media (Chakravarthi et al., 2020b,a). It is plausible that memes might become an integral part of most of the people, as it is used to understand racial and gender discourse on social media platforms such as Reddit (Milner, 2013; Ghanghor et al., 2021b,a). One of the approaches to overcome this is manually monitoring and moderating user-generated content. But due to the amount of data being generated on the internet every day, it would be ideal to develop automated systems to moderate them (Kumar et al., 2018; Yasaswini et al., 2021; Puranik et al., 2021; Chakravarthi et al., 2020c; Mandl et al., 2020).

Consider countries with huge populations such as India, several memes are directed towards targeted communities. To address these issues of identifying if a given meme is trolling a person's sentiments, a dataset for memes that were suspected to troll a particular community. We participate in the shared task on meme classification based on the troll classification of Tamil Memes (Suryawanshi et al., 2020). Tamil (ISO 639-3: tam) language is spoken in South Asia (Chakravarthi, 2020). The earliest inscription in India dated from 580 BCE was the Tamil inscription in pottery and then the Asoka inscription in Prakrit, Greek, and Aramaic dating from 260 BCE. The earliest known inscriptions in Sanskrit are from the inscriptions of the 1st century BCE. Tamil is the official language of Tamil Nadu, India, as well as of Singapore and Sri Lanka (Chakravarthi et al., 2018, 2019). The task primarily consists of identifying whether a meme is a *troll* or a *non-troll* (Suryawanshi and Chakravarthi, 2021). We use the images and captions that are provided to achieve the most efficient model to classify the memes. We use a combination of Vision Transform (ViT) (Dosovitskiy et al., 2021) and mBERT (Pires et al., 2019) over other pretrained models used for image classification as described in (Venkatesh et al., 2020, 2021).

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 180–186

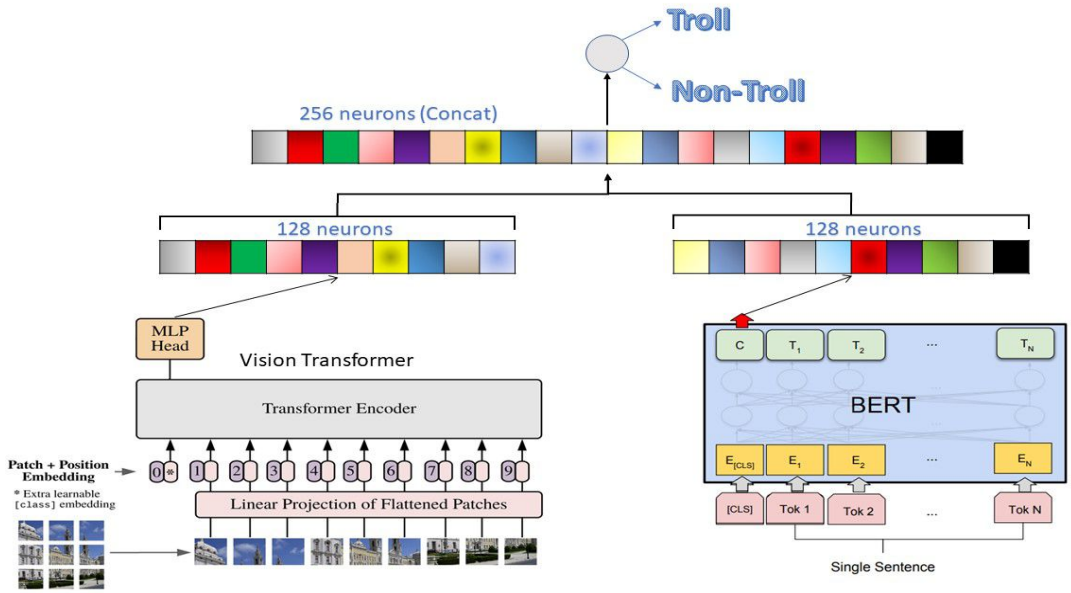


Figure 1: System Architecture (Dosovitskiy et al., 2021; Devlin et al., 2019)

2 Related Work

Internet memes have been a subject of interest for both Computer Vision and Natural Language Processing researchers. The type of memes that are being used illustrates the context of discussions on social media platforms. People are using memes to express themselves, and in the making, showcase their stance on a certain social issue, be it in acknowledgment or rejection of the issue (French, 2017; Boinepelli et al., 2020; Gal et al., 2016). There exist several reasons that suggest the spread of memes. Some of the reasons include novelty, simplicity, coherence. It also includes an emotional attachment, its ability to have different meanings, depending on how a person perceives it (Nave et al., 2018; Stephens, 2018; Chielens and Heylighen, 2002). Hu and Flaxman developed a multimodal sentiment analysis by developing a deep neural network that combines both visual analysis and text analysis to predict the emotional state of the user by using Tumblr posts.

3 Data

We use Troll Classification dataset of Tamil Memes (Suryawanshi et al., 2020). It consists of 2,699 memes, of which most of the images have text embedded within them. We are also provided with captions for all images. The distribution is shown is Table 1.

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

Table 1: Dataset Distribution

4 System Description

Multimodal deep learning is a robust and efficient way of addressing the main goals of artificial intelligence by integrating and combining multiple communicative modalities to obtain crucial results which usually improves the outcome of the single models trained. As deep learning models tend to extract features on their own, the objective can easily be achieved with the help of neural networks.

Given the images of Tamil Memes, along with the embedded text on the images, scrutiny of images and texts independently and then picking out relevant information for further process plays a climacteric role in our system. At the end of the training, the model has to output a single value stating the given meme is Troll or Non-Troll. The specialty of our model was to neither use the Convolutional Neural Networks (CNN) nor Recurrent Neural Networks (RNN). As the title of the paper points out, the model tries to gain more attention towards the salient portions of text and images. The proposed solution makes an effort to convey the importance of attention gain and its relation

	Precision	Recall	F1-Score	Support
Non-Troll	0.96	0.95	0.96	101
Troll	0.96	0.97	0.96	128
Accuracy			0.96	229
Macro Avg	0.96	0.96	0.96	229
Weighted Avg	0.96	0.96	0.96	229

Table 2: Classification report of ViT to images of validation set

	Precision	Recall	F1-Score	Support
Non-Troll	0.87	0.99	0.93	101
Troll	0.99	0.88	0.93	128
Accuracy			0.93	229
Macro Avg	0.93	0.94	0.93	229
Weighted Avg	0.94	0.93	0.93	229

Table 3: Classification report when memes are classified based on captions on validation set

to the performance of the model. The model is put forward to compute the classification is **Vision transformer** (Dosovitskiy et al., 2021) for images and **Bidirectional Encoder Representations from Transformers (BERT)** (Devlin et al., 2019) for captions of memes. This corresponds to a *transformer-transformer* architecture as shown in Fig 1.

4.1 Vision Transformer (ViT)

The architecture of the ViT is analogous to the transformer used for Natural Language Processing (NLP) tasks. NLP transformers use self-attention which is a highly cost-inefficient approach in regard to images. Admitting this, the technique applied here was Global Attention. Keeping the analogy of sentences, instead of 1D token embeddings as input, ViT receives a sequence of flattened 2D patches. If H, W is the height and width of the image and (P, P) is the resolution of each patch, $N = HW/P^2$ is the effective sequence length for the transformer (Dosovitskiy et al., 2021). Then the patches are projected linearly and then multiplied with an embedding matrix to eventually form patched embeddings. This along with position embeddings are sent through the transformer. Similar to BERT's [CLS] token, a token is prepended along with the patched embeddings. The transformer consists of an encoder block which consists of alternating layers of multiheaded self-attention blocks to generate attention for specific regions of the images. Layer normalization and residual connections are made comparable to the original NLP transformer.

4.2 BERT

The success of fine-tuning a pretrained model in computer-vision prompted researchers to do the same in Natural Language Processing. Therefore it was the objective of the researchers to develop a model which can be fine-tuned for NLP related works. **B**idirectional **E**ncoder **R**epresentations from **T**ransformers **(BERT)** (Devlin et al., 2019) is a language representation model which was trained on Wikipedia corpus. The training phase had two tasks. First was Masked Language Modelling(MLM), where the sentence had random masks in them and the model has to predict the masked word. The second task Next Sentence Prediction(NSP), where the model has to predict whether the second sentence is the continuation of the first one.

The input to the transformer is the sum of the token segmentation and positional embeddings. As the name suggests, the model is jointly conditioned on both left and right contexts to extract meaning. BERT is comparable to the transformer encoder block of (Vaswani et al., 2017). The NSP task matches the classification task for the objective of the model. During NSP, two sentence separated by [SEP] and [CLS] token are fed in and the output of the [CLS] token is pondered upon to determine the required class. Here, the input is only a single sentence with tokens and the model is fine-tuned as necessary. For this system, *bert-base-multilingual-cased* (L=12, H=768, A=12, Total Parameters=179M) was used. This model is pretrained on largest available Wikipedia dumps

	Precision	Recall	F1-Score	Support
Non-Troll	0.60	0.03	0.06	272
Troll	0.60	0.98	0.74	395
Accuracy			0.60	667
macro Avg	0.60	0.51	0.40	667
Weighted Avg	0.60	0.60	0.47	667

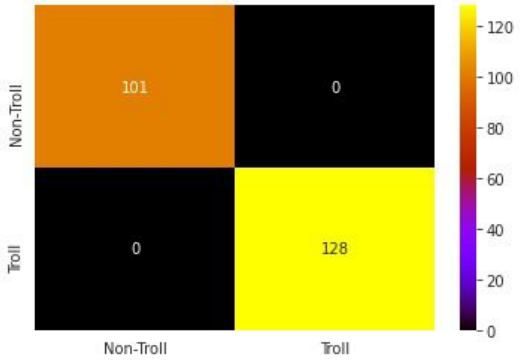
Table 4: Classification report of our system on the test set

of the top 104 different languages, with the largest MLM objective, also making the model case sensitive (Pires et al., 2019).

5 Experiments

All suitable models were implemented using PyTorch version 1.5.0 in a google colaboratory environment. The early stages of this model include preprocessing of images. The dataset had pictures with various resolutions and had to be made equal. The images were resized to 256 X 256 pixels. Most of the images had texts on the top and bottom of the images. Texts in the images were considered as noise for classification, which resulted in performing a center crop for all images. The border of the portions was removed and images of size 224 X 224 were produced. Finally, the images were ready as the input to the transformer by normalizing the RGB channels with mean 0.485, 0.456, 0.406, and standard deviation 0.229, 0.224, 0.225 respectively. No augmentations were made to preserve the meaning of the images. The transformer was originally trained on the ImageNet dataset and had achieved remarkable results. The trained weights are transferred to this downstream task. The base version of ViT is fine-tuned which had default hyperparameters of 16 patches, an embed dimension of 768, 12 layers, 12 attention heads, and a dropout rate of 0.1. The head of the vision transformer, which outputs 1000 classes, is now replaced by a linear layer of 128 neurons. The texts were also preprocessed by removing stopwords, special characters, and punctuation. Texts need to be tokenized before feeding into the BERT configuration. After inserting it into the transformer, the resulting pooled output from the multilingual BERT model is also passed through a linear layer of 128 neurons.

The two layers obtained from the transformers are merged together to form a single layer with 256 neurons. This is passed through the ReLu activation function and a dropout to obtain one final neuron which determines the class as Troll or Non-Troll. A learning rate of $2e-5$ was used with a batch size of 16. The maximum length of the captions was truncated to 128 as memes usually do not contain very long sentences. The training was done for 4 epochs and with a linear schedule with warmup. To our surprise, the model learned very rapidly and achieved well progress on the validation set which mimicked the train set. It was also observed that merging the outputs of two different domain models did not harm the training, moreover, it helped in getting better results.



(a) Validation set

(b) Test set

Figure 2: Confusion Matrix

6 Results

We achieve an overall F1-score of 0.96 when we use images for classification using ViT as shown in 2. It is to be noted that using mBERT to clas-

sify memes solely based on the captions achieves 0.93 as F1-score as shown in Table 3. While we achieve such good results in comparison to the baseline scores of 0.59 mentioned in the dataset paper, we feel that if both of representations of ViT and mBERT were concatenated and then fed into a linear layer, the model would learn better. We find that the model achieves a perfect 1.00 weighted F1-score on the validation set. We believe that preprocessing of the images was a major factor for achieving a great F1-score on validation set. This argument is supported by our system's poor performance on the test set, as the test set was not coherent with the training data in terms of the positioning of texts on the images as shown in Table 4. The confusion matrix on validation and test set are as shown in Figures 2(a) and 2(b) respectively.

7 Conclusions

The proposed solution performs at greater heights on the validation and set in the training phase. The validation set mimics the train set as the memes are split looking at the distribution of the classes. The dataset is very small and augmenting it will not help for the optimal results. The algorithm overfits the train set undoubtedly. The reason behind the poor performance is due to the change in the distribution. The memes in the test set had multiple images which were difficult for the ViT to capture features. The model scored a F1 score of 0.46 on the test set and 1.0 on the validation set. Vast difference can be observed due to high bias. Here, in this paper, we have tried to come up with this innovation of transformer-transformer architecture which can achieve extreme results. In the future, we will be performing a wonderful task of having more transformers in parallel computation and syncing them makes an immense difference in this era of deep learning.

.

References

Adithya Avvaru and Sanath Vobilisetty. 2020. BERT at SemEval-2020 task 8: Using BERT to analyse meme emotions. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1094–1099, Barcelona (online). International Committee for Computational Linguistics.

Sravani Boinepelli, Manish Shrivastava, and Vasudeva Varma. 2020. SIS@IIITH at SemEval-2020 task 8: An overview of simple text classification methods for meme analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1190–1194, Barcelona (online). International Committee for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020c. Overview of the Track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 21–24, New York, NY, USA. Association for Computing Machinery.

Klaas Chielens and F. Heylighen. 2002. Operationalization of meme selection criteria : Methodologies to empirically test memetic predictions.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

J. H. French. 2017. Image-based memes as sentiment predictors. In *2017 International Conference on Information Society (i-Society)*, pages 80–85.

Noam Gal, Limor Shifman, and Zohar Kampf. 2016. "it gets better": Internet memes and the construction of collective identity. *New Media & Society*, 18:1698 – 1714.

Nikhil Kumar Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021a. 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*, Online. Association for Computational Linguistics.

Nikhil Kumar Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021b. IIITK@LT-EDI-EACL2021: Hope Speech Detection for Equality, Diversity, and Inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, Online. Association for Computational Linguistics.

Anthony Hu and Seth Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 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, page 29–32, New York, NY, USA. Association for Computing Machinery.

R. M. Milner. 2013. Fcj-156 hacking the social: Internet memes, identity antagonism, and the logic of lulz. *The Fibreculture Journal*, pages 61–91.

Nir Noon Nave, Limor Shifman, and Keren Tenenboim-Weinblatt. 2018. Talking it personally: Features of successful political posts on facebook. *Social Media + Society*, 4.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@LT-EDI-EACL2021-Hope Speech Detection: There is always hope in Transformers. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Niall P Stephens. 2018. Ryan m. milner, the world made meme: Public conversations and participatory media. *International Journal of Communication*, 12:4.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 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.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Venkatesh, Y Nagaraju, S SahanaT, S Swetha, and Siddhanth U Hegde. 2020. Transfer learning based convolutional neural network model for classification of mango leaves infected by anthracnose. *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–7.

Venkatesh, Nagaraju Yallappa, Siddhanth Udayashankar Hegde, and Sangeetha Raj Stalin. 2021. Fine-tuned mobilenet classifier for classification of strawberry and cherry fruit types. *Journal of Computer Science*, 17(1):44–54.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.