TrollMeta@DravidianLangTech-EACL2021: Meme classification using deep learning

Manoj Balaji J
BITS, Pilani
manojbalaji1@gmail.com

Chinmaya HS
Cambridge Institute of Technology, Bengaluru
chinmayasbhat4@gmail.com

Abstract

Memes act as a medium to carry one's feelings, cultural ideas, or practices by means of symbols, imitations, or simply images. Whenever social media is involved, hurting the feelings of others and abusing others are always a problem. Here we are proposing a system, that classifies the memes into Troll or Non-Troll memes . The system implements resnet-50, a deep residual neural network architecture.

1 Introduction

Social media are the interactive platforms that are in and around the daily life of most of the people (Thavareesan and Mahesan, 2019, 2020a,b). Memes has become integral part of daily life and they play a crucial role in sociopolitical, cultural and behaviour of the people (a P K et al., 2020). Memes are not only media for conveying information or disrupting the socio-political situation but, they serve as the main source of sharing humour and laughter. Internet humour can create a logical bonding between people and can act as a common platform for like minded people (Laineste and Voolaid, 2017). Memes with colorful/even much dull background powered by splashy texts can create amazing sense of humour (a P K et al., 2020; Laineste and Voolaid, 2017).

On the other hand, meme trolling is considered as online activism (a P K et al., 2020), creating awareness and new kind of marketing too. Troll is a person who starts the flame of insulting/ upsetting the feelings of another person on internet. This can be done by posting messages in the online community such as social media or even in private chats. Trolling has caused in one's personal and professional life even (Bishop, 2013). The content in any mainstream media are closely monitored but, for social media there are no such monitoring systems. Internet has opened the road to post anything and everything on social platforms. Offensive content memes are spreading the emotion of fear, misguided phobia and they are misleading the population as they spend most of their time on the internet(Suryawanshi et al., 2020a).

Considering the psychological, socio-political and social impacts of memes, especially memes we are proposing a deep learning based meme classifier that can differentiate the trolling one's from non-trolling counterparts.

2 Related Works

Suryawanshi et al. (2020a) in their work stated the importance of meme classification/filtering. Their work included the dataset of memes that contains 2000+ annotated images with captions. Their work showed a path towards importance of monitoring social media about the postings from regional languages.

Bishop (2013) discussed the application of criminal procedure over the internet bully's and insulting one's feelings over the internet. Memes not only contain text data, but the visual appearance should also be considered, Smitha et al. (2018) proposed a technique to extract text data from images and analysing the emotions on the image and later classifying them into one of the 6 predefined classes, and they concluded that the work helps to understand the behaviour of the community and information for any further data analysis (Smitha et al., 2018).

Suryawanshi et al. (2020b) observed that the offensive contents are not only in the form of text but, they can also be offensive in visual aspect. Therefore they proposed a technique to classify this multimodel memes, which uses a stacked LSTM + VGG16 model.

3 Dataset

The dataset used here, 'Tamil Troll Memes' dataset, which was made available by Suryawanshi et al. (2020a). The dataset contained total of 2,969 im-

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 277–280
April 20, 2021 ©2021 Association for Computational Linguistics

ages of memes, most which contained text embedded images.

NSK
உலகின் தலை சிறந்த சொல்
EPPAYAACHUM OLARNAA PARAVAALA
EPPAYUMAE OLARNAA EPDI
'கூட்டணி'
IN VETTAIYADU VILAYADU MOVIE (2006)
THAN SAAVAI SATTAI PAIYIL VAITHU ENGAYUM SELGINREN
KAMAL
GOD
When 90s kids watching This serial
~90s kids
Namaku veelaikkaari amaiyula...
veelaikkaari vaikura alavuku panavasithyum amaiyula...

Figure 1: Sample memes from TamilMemes dataset with Latin transcripted or Tamil text

4 Methodology

Residual networks are playing a crucial role in image classification. They find a way to overcome the gradient loss, which are a major problem in deep neural networks. Residual networks, which uses residue from the previous layers to prevent loss of gradients due to depth of the network (Suryawanshi and Chakravarthi, 2021).

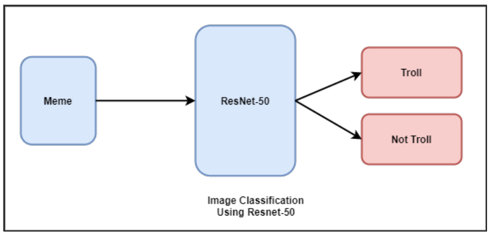

Figure 2: Proposed system architecture (ResNet 50) that distinguishes troll from not troll Tamil memes

Here we are using a deep residual neural network based on resnet architecture. The architecture has 50 layers, with which the network can grasp the features from different parts and levels of the image. The images are resized to 64x64, with all three color channels, are fed into the algorithm, which then passed to sub-sequent layers of the neural network to finally give the output. Softmax activation in the last layer gives the probability value for each class and class with max probability is selected as output. We used fastai library, with which we are able to train our model much faster(He et al., 2015). The framework provides easy to access layered API architecture, which has the APIs to access most of the deep learning algorithms. Pause and train architecture(Howard and Gugger, 2020) helps to improve the control over model training as one can have a look at loss, accuracy and other parameters. Manual early stopping mechanism helps to monitor the model improvement easily(Howard and Gugger, 2020) which is integrated as a part of fastai library, which employs cycle training that allows to monitor each cycle and freeze the weights if required.

5 Results and Evaluation

The work was performed to classify the memes in to trolling and non-trolling. The resnet-50 model which was trained on the dataset containing 2349 images belonging to two different classes, for 50 epochs.

The variation of loss and accuracy for image classification are shown in the figure below.

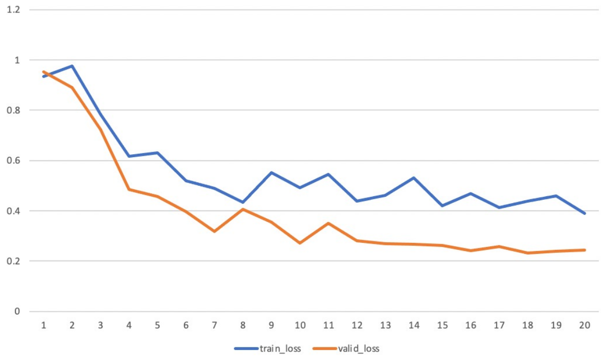

Figure 3: Graph showing variation of loss during the course of training. (Blue line shows decrease in training loss and orange line shows the variation of validation loss)

To analyze the results of the model, a confusion matrix was constructed, and the weighted f1 is calculated. The model trained with Binary Cross Entropy loss function for 10 cycles with imagenet weights and performed with the following results,

The values for validation are more than that of training, even though it looks like overfitting the values remained same for the K-fold validation also, therefore the model is not overfitting over a set of data but it suffers lack of data, which we think can be improved with accumulation of data from different sources.

	precision	recall	F1-score	support
Train	0.87	0.87	0.87	1840
Val	0.92	0.92	0.92	460
Test	0.45	0.41	0.48	

Table 1: Evaluation values for training, validation and test (weighted avg. values)

However the values in the test are much lesser compared to that of training and validation, which can be result of model overfitting.

6 Limitations

During the course of work we faced many challenges and observed some limitations. It is hard to identify sarcasm. Most of the times text and the expression in the image contradict each other and create a wave of sarcasm. An example of such meme is where a character is associated with ”Ahaan”. This is highly contextual as this was part of a Tamil movie comedy scene.

The text/caption in the image is in Tamil but written in Italic, and mixture of Tamil and English will also pose a huge problem while building the model.

7 Conclusion

Memes poses as a medium which are powerful enough to disrupt politics, induce and spread humour. In the mean time they are capable of causing considerable harm to socio-political and personal health of individuals via trolls. Our work, meme classification using deep learning sheds light towards filtering the memes to trolling and non-trolling ones with training weighted f1 of 0.87, validation weighted f1 of 0.92 and overall testing accuracy of 0.48. There’s work needed to be done as a part of building model with better generalization.

The work helped us to understand the impacts and effects of the memes in day today life of the people and how they evolved to be an integral part of the internet and social media.

8 Future Work

The work carried out here proposes the method to classify memes based only on images, but the future work is intended to consider both text and image from the memes. In simple words, it includes, extracting text from images, emotion detection using some NLP technique, combining the results with image classification model to form a hybrid model that can classify the emotions more efficiently.

References

Jonathan Bishop. 2013. The effect of de-individuation of the internet troller on criminal procedure implementation: An interview with a hater. *International Journal of Cyber Criminology*, 7:28.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Jeremy Howard and Sylvain Gugger. 2020. Fastai: A layered api for deep learning. *Information*, 11(2):108.

L. Laineste and Piret Voolaid. 2017. Laughing across borders: Intertextuality of internet memes. *The European Journal of Humour Research*, 4:26–49.

Abdul Rasheed a P K, Carmel Maria, and Anju Michael. 2020. Social media and meme culture: A study on the impact of internet memes in reference with ’kudathai murder case’.

E. S. Smitha, S. Sendhilkumar, and G. S. Mahalaksmi. 2018. Meme classification using textual and visual features. In *Computational Vision and Bio Inspired Computing*, pages 1015–1031, Cham. Springer International Publishing.

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, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. 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).

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.