

IITK@DravidianLangTech-EACL2021: Offensive Language Identification and Meme Classification in Tamil, Malayalam and Kannada

Nikhil Kumar Ghanghor¹, Prameshwari Krishnamurthy,² Sajeetha Thavareesan³,
Ruba Priyadarshini,⁴ Bharathi Raja Chakravarthi⁵

¹Indian Institute of Information Technology Kottayam

²University of Hyderabad, India, ³ULTRA Arts and Science College, Tamil Nadu, India

⁴ Eastern University, Sri Lanka, ⁵ National University of Ireland Galway

nkghanghor18bcs@iiitkottayam.ac.in

Abstract

This paper describes the IITK team’s submissions to the offensive language identification and troll memes classification shared tasks for Dravidian languages at DravidianLangTech 2021 workshop@EACL 2021. We have used the transformer-based pretrained models along with their customized versions with custom loss functions. State of the art pretrained CNN models were also used for image-related tasks. Our best configuration for Tamil troll meme classification achieved a 0.55 weighted average F1 score, and for offensive language identification, our system achieved weighted F1 scores of 0.75 for Tamil, 0.95 for Malayalam, and 0.71 for Kannada. Our rank on Tamil troll meme classification is 2, and offensive language identification in Tamil, Malayalam, and Kannada is 3, 3 and 4. We have open-sourced our code implementations for all the models across both the tasks on GitHub¹.

1 Introduction

With the rise of deep learning in the past decade, we have seen some groundbreaking milestones achieved by various machine learning (ML) models being built over the era. The AlexNet (Krizhevsky et al., 2012) architecture achieved the first significant milestone on ILVRSC challenge, which showed the state of the art results at their time of recording on the image classification tasks. Although transfer learning was proposed much before in the community, it was not used much due to the lack of computational resources. The success of AlexNet reignited the usage of transfer learning (Bozinovski, 2020) which leverages the knowledge gained by the models having parameters of the size of millions to billions and which are being trained over an enormously large dataset can be further use

for fine-tuning it over other computer vision tasks to achieve promising results.

Another breakthrough in the community was the release of BERT (Devlin et al., 2019) which was being built upon the transformer architecture (Vaswani et al., 2017) with several innovative training strategies at the time of its release topped the GLUE (Wang et al., 2019) benchmarks. The introduction of BERT paved the ways for the transfer learning on the Natural Language Processing (NLP) tasks. However, one particular issue with these huge pretrained models was that they were most of the times not available for under-resourced languages.

The Offensive Language Identification task aimed to detect the presence of offensive content in texts of such under-resourced languages (Mandl et al., 2020; Chakravarthi et al., 2020b; Ghanghor et al., 2021; Yaraswini et al., 2021; Puranik et al., 2021). The texts were in the native language scripts as well as of code-mixed nature (Chakravarthi, 2020a). Code-mixing means the mixing of two or more languages or language varieties in speech (Priyadarshini et al., 2020; Jose et al., 2020). The Meme Classification task focused on classifying memes into a troll and not a troll. Along with the memes, their extracted captions were also being provided in Tamil native scripts and code-mixed nature (Hegde et al., 2021). All the three languages Kannada, Malayalam and Tamil for which the datasets were being provided across the two tasks are under resourced in nature and belong to Dravidian languages’ family (Chakravarthi, 2020b). The Dravidian language family is dated to approximately 4500 years old (Kolipakam et al., 2018). In Jain and Buddhist texts written in Pali and Prakrit in ancient times, they referred to Tamil as Damil, and then Dramil became Dravida and Dravidian. Tamil language family includes Kannada, Telugu, Tulu and Malayalam. The Tamil languages scripts

¹<https://github.com/nikhil6041/OLI-and-Meme-Classification>

are first attested in the 580 BCE as Tamil² script inscribed on the pottery of Keezhadi, Sivagangai and Madurai district of Tamil Nadu, India (Sivanantham and Seran, 2019). The Dravidian languages have its scripts evolved from Tamil scripts.

In this paper, we showcase our efforts for the two shared tasks on Dravidian languages. We participated in the shared task Offensive Language Identification (Chakravarthi et al., 2021) for Tamil, Malayalam and Kannada. We also participated in the shared task on Tamil Troll Meme Classification (Suryawanshi and Chakravarthi, 2021) task. We have used the Transformer and BERT based models in our work. This paper summarizes our works on the two shared tasks. Each of the following sections is being subdivided further into two sections to detail every aspect of the procedures being worked out on them.

2 Related Work

Offensive content detection from texts has been part of some conferences as challenging tasks. Some of the most popular tasks have been part of the SemEval³ organized by International Workshop on Semantic Evaluation. The tasks covered in SemEval varied from offensive content detection in monolingual corpus to code-mixed corpus collected from social media comments. However, most of these competitions had their tasks on datasets based on English and other high resourced languages from the west. HASOC-Dravidian-CodeMix-FIRE2020⁴ was one such competition being organized for under-resourced languages such as Tamil and Malayalam.

Ranasinghe et al. (2020) at HASOC-Dravidian-CodeMix-FIRE2020 used traditional Machine Learning (ML) methods like Naive Bayes Classifier (NBC), Support Vector Machines (SVMs) and Random Forest along with the pretrained transformers models like XLM-Roberta (XLMR) (Conneau et al., 2020) and BERT for the offensive content identification in code-mixed datasets (Tamil-English and Malayalam-English). Arora (2020) at HASOC-Dravidian-CodeMix-FIRE2020 used ULMFit (Howard and Ruder, 2018) to pre-train on a synthetically generated code-mixed dataset and then fine-tuned it to the downstream tasks of text classification.

²This is also called Damili or Dramili or Tamil-Brahmi

³<https://semeval.github.io/>

⁴<https://competitions.codalab.org/competitions/25295>

Meme Classification has been one of the major tasks that various social media platforms aim to solve in the past decade. Various methods have been proposed in order to classify memes. However, it is still an unsolved problem because of its multimodal nature, making it extremely hard to have a system correctly classify the memes (Suryawanshi et al., 2020a). Classification of memes needs a visual and linguistic approach to be followed. Hu and Flaxman (2018) did work similar to memes analysis. They used the Tumblr dataset using memes, which had images with several tags associated with them. They dealt with texts and visual features independently and built a model to predict sentiments behind those pictures with corresponding tags and obtained good results. Using this idea of dealing with different modalities independently, different researchers came forward with different implementations for this problem of memes analysis. This particular problem became a flagship the task of SemEval-2020. The team having the best performance for sentiment analysis was obtained by the (Keswani et al., 2020) where they used several different models for addressing the problem by considering both the modalities and coming up with several combinations using both the modalities.

However, the best performance was given by considering the textual features only using BERT as the architecture for doing so for sentiment analysis of the memes. The overview paper (Sharma et al., 2020) of memotion analysis has summarized the works of the participants and gives a clearer picture of the status of the task of memotion analysis its reporting. Recently Facebook AI conducted hateful memes challenge⁵ in 2020 whose winner has achieved the best results using several visual, linguistic transformer models. Zhu (2020) used an ensemble of four different Visual Linguistic models VL-BERT (Su et al., 2020), UNITER (Chen et al., 2020), VILLA (Gan et al., 2020), and ERNIE-Vil (Yu et al., 2020).

3 Dataset Description

The competition organizers have released datasets for three different languages namely the Kannada dataset (Hande et al., 2020), Malayalam dataset (Chakravarthi et al., 2020a) and Tamil dataset (Chakravarthi et al., 2020c) languages for the Offen-

⁵<https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/>

sive Language Identification shared task. However, for the Meme Classification shared task the dataset (Suryawanshi et al., 2020b) was only released for the Tamil language. The train, dev and test set distributions for both of them are as follows:

| Distribution | Kannada | Malayalam | Tamil |
|--------------|---------|-----------|--------|
| Train | 6,217 | 16,010 | 35,139 |
| Dev | 777 | 1,999 | 4,388 |
| Test | 778 | 2,001 | 4,392 |

Table 1: Offensive Language Dataset Distribution

For the Meme Classification task, the competition organizers have provided us with images and the extracted texts present in them.

| Distribution | Data |
|--------------|-------|
| Train | 2,300 |
| Test | 667 |

Table 2: Meme Dataset Distribution

4 Methodology

4.1 Offensive Language Identification Task

Analysing the sentiment behind the texts have been one of the central tasks of the NLP community. Over time, many advancements have been made in order to analyse the sentiments behind the texts (Chakravarthi et al., 2020d). However, it is still an unsolved problem because of the linguistic diversity existing across the globe and the requirement of the difference in ways the texts have to be addressed depending upon its nature. Here, by nature, we intend to highlight that in languages like English, the casing of words matters; however, it is not even a thing to consider in languages like Malayalam. There could be many more differences across the languages on similar lines across the languages many details on which can be found on the internet.

Several different techniques which are in existence to address the sentiment analysis problem have been developed over the past years. They encompass from the primitive ML models built upon Bag Of Words (BOW), TF-IDF, and N-grams representations to the neural network models like RNNs having LSTMs (Hochreiter and Schmidhuber, 1997), GRUs (Chung et al., 2014), and BiLSTMs (Schuster and Paliwal, 1997) as their key components to the current models having state of the results on various GLUE benchmarks based

upon the popular transformer architecture. We also have word embeddings (Bengio et al., 2003) available for the representations of words in a high dimensional vector space which are being used as a starting component of different neural network models.

We have used several pretrained multilingual models in our works and used them to fine-tune the offensive language identification task. Most of the work has been done by using huggingface transformers library⁶. One key issue with the dataset was the skewness of it. We have used the Negative Log Likelihood (NLL) Loss with class weights, Self Adjusting Dice (Li et al., 2020) (Sadice) loss and applied it to address the class imbalance issues with the customised version of models provided by the transformers library. We have also used the original versions of the models available apart from the customised implementations. We have trained each of the models possible combinations as mentioned above and loss functions with each of the language available in the dataset separately. The multilingual models which we have used are multilingual BERT-cased (mBERT-cased), XLM-Roberta and IndicBERT (Kakwani et al., 2020).

Since the IndicBERT model is comparatively much smaller than mBERT-cased and the XLM-Roberta models we have not tried to develop a customised version for it and went with the original version. There are two different multilingual BERT models available one spans 100 languages, and the uncased version, the other one spans 104 languages and is the cased version. BERT’s author’s recommendations are to use the cased versions, so we have used it in our implementations. For the XLM-Roberta model, we have gone with the base model, which covers 100 languages. We have added a fully connected layer of 512 neurons on top of the final transformer layer for the customised versions of these models, which gives us 768-dimensional feature vectors as outputs. This fully connected layer is finally followed by a LogSoftmax layer having the number of neurons the same as our output classes. The base layers of the original models in the customised versions of the models were being frozen. The custom models were trained with two different loss functions: one was the Negative Log-Likelihood Loss (NLL Loss) with class weights being applied. The other loss function we have tried is the Sadice Loss. We have reported

⁶<https://huggingface.co/transformers/>

our results on the development set in Table 3 and the results on the test set are reported at Table 4. The results are being reported in terms of weighted average F1 scores.

4.2 Meme Classification Task

With the rise of social media in the past decade, we have observed the raise of memes. However, the meme is not a new word and dates back to 1970s. To better understand memes, we have two different definitions, each of them unique in their way and helps us understand it from two different perspectives.

The first definition⁷ says that meme can be an element of a culture or system of behaviour passed from one individual to another by imitation or other non-genetic means. Moreover, the other definition⁸ states that meme can be an image, video, and text piece that is typically humorous, copied and spread rapidly by internet users, often with slight variations.

According to researchers, both definitions are equally acceptable and set the required knowledge base to fully understand a meme. As we discuss the type of modality of memes, evident from the definition, the memes are a mixture of modalities with textual and visual contents and audio in videos. Several methods for approaching the multimodal problems spanning from simple unimodal approaches to different ways to concatenating the multiple modalities like fusion and concat.

We have used the unimodal approaches in our work, which has a different model for each different modality. The competition organizers have provided us with images along with the extracted captions from those memes. We list our approach to each of the modalities in the following subsections:

4.3 Text Modality

As discussed in Section 4.1, we know there exist several methods for approaching any sentence classification problem. We have used several pretrained multilingual models, namely mBERT-cased, XLM-Roberta and IndicBERT for this problem and fine-tuned them on our dataset. Since the dataset was not suffering from skewness, we went with the models original pretrained versions.

⁷<https://www.britannica.com/topic/meme>

⁸<https://www.lexico.com/definition/meme>

4.4 Image Modality

Image classification was the task which can be attributed to recoiling the interest in neural networks and bringing the deep learning era because of its exceptional performance in ILVRSC challenges. Image classification techniques have evolved from the usage of Support Vector Machines(SVM) and other primitive machine learning models for classification with handcrafted image features to the LeNet (Lecun et al., 1998). LeNet brought CNNs into the picture which proceeded further with AlexNet, ResNet (He et al., 2016), Inception (Szegedy et al., 2014) and other popular architectures giving the state of the art results at their time of introductions and bringing some revolutionary ideas into the picture resulting in pushing the performance each time. These architectures were trained on massive datasets having 100s or even 1000s of classes in them. We have used the ResNet50 and inception architectures with their pretrained weights and fine-tuned them over our classification task.

We have reported our results on the development and test set in Table 5. The results are being reported in terms of weighted average F1 scores.

5 Results and Discussion

5.1 Offensive Language Identification Task

The task aimed to detect the offensive content in three different languages Tamil, Malayalam and Kannada. We have used the pretrained transformer models checkpoints provided by hugging face transformers library. For the identification task, the original pretrained versions and the customized versions as described in Section 4.1 with NLL Loss and Sadice loss were being used, however, out of all the custom implementations the original versions performed much better when being fine-tuned on our datasets. Across all the models, mBERT-cased and XLMR gave very competitive results across both dev and test sets, which may be attributed to their vast size and training corpus. Although IndicBERT having almost ten times fewer parameters gave almost similar results to mBERT-cased and XLMR. The competitive performance of IndicBERT can be attributed to the fact that it was trained on a corpus having Dravidian languages in a relatively more significant proportion as compared to the other two models where the proportion of the Dravidian languages in the entire training corpus was comparatively lesser. However, the better performance of the XLMR and mBERT-cased

| Model | Tamil | Malayalam | Kannada |
|--|-------|-----------|---------|
| mBERT-cased | 0.76 | 0.96 | 0.68 |
| Custom mBERT-cased with NLL loss and Class weights | 0.58 | 0.82 | 0.35 |
| Custom mBERT-cased with Sadice Loss | 0.61 | 0.87 | 0.55 |
| XLMR | 0.76 | 0.94 | 0.67 |
| Custom XLMR with Sadice Loss | 0.61 | 0.84 | 0.39 |
| Custom XLMR with NLL loss and Class weights | 0.64 | 0.84 | 0.10 |
| IndicBERT | 0.74 | 0.94 | 0.63 |

Table 3: Experiments with Offensive Language Identification development dataset

| Model | Tamil | Malayalam | Kannada |
|--|-------------|-------------|-------------|
| mBERT-cased | 0.75 | 0.95 | 0.71 |
| Custom mBERT-cased with NLL loss and Class weights | 0.59 | 0.82 | 0.34 |
| Custom mBERT-cased with Sadice Loss | 0.61 | 0.87 | 0.53 |
| XLMR | 0.75 | 0.93 | 0.71 |
| Custom XLMR with Sadice Loss | 0.61 | 0.83 | 0.39 |
| Custom XLMR with NLL loss and Class weights | 0.64 | 0.83 | 0.09 |
| IndicBERT | 0.73 | 0.93 | 0.67 |

Table 4: Experiments with Offensive Language Identification test dataset

| Model | Dev | Test |
|---------------|------|-------------|
| mBERT-uncased | 0.90 | 0.50 |
| mBERT-cased | 0.94 | 0.55 |
| XLMR | 0.93 | 0.54 |
| IndicBERT | 0.85 | 0.50 |
| Inception | 0.78 | 0.53 |
| ResNet | 0.89 | 0.46 |

Table 5: Experiments with Meme Classification dataset

on the task datasets can be attributed to the training objective of these models which evaluates their model’s performance on XNLI (Conneau et al., 2018) benchmark thus resulting in their competitive performance when compared to IndicBERT. We have used mBERT-cased for our final submissions across all the three languages as it performed comparatively better on all the three languages dev sets. The one major bottleneck of the model’s lower performance across several languages in the offensive language identification task was the skewness of the dataset. The performance can further be improved by having some techniques to generate synthetic text data in order to handle the class imbalance issues. Since the transformer models expect texts as inputs, and each one has their own tokenizers, the generation of synthetic text data is another challenge to improve these models’ performance.

5.2 Meme Classification Task

The task aimed to detect the troll memes in the Tamil language. Since the memes are multimodal, we have tried several different pretrained unimodal models to deal with each of the modalities. For the image modality, we have used inception and ResNet50 , and for the text modality, we have used mBERT-uncased , mBERT-cased , XLMR and IndicBERT. The competition organizers have provided us with a training set of 2,300 images along with their captions. We used a validation split of 20% on our training set and evaluated all our models on that validation set. Of all the models, text-based models’ performance was comparatively better than the image-based models except for ResNet performing better than IndicBERT on dev set. Both XLMR and mBERT-cased performed almost similarly on our dev set, but since mBERT-cased had slightly better performance than XLMR on the dev set we submitted our predictions using the mBERT-cased model. For the meme classification task, the reported results show that even after performing well on dev datasets our models have not performed up to the mark on the test datasets which can be attributed to the fact that memes are of multimodal nature. They cannot be solved only by unimodal approaches.

6 Conclusion

We have presented the IITK submission to Tamil troll meme classification, and offensive language identification shared tasks at DravidianLangTech. We experimented with several different pretrained models for the given datasets with their original and customized architecture and different loss functions. Our best model for Tamil troll meme classification was multilingual BERT-cased, and it gave a weighted average F1 score of 0.55. Our best model for offensive language identification was also multilingual BERT-cased, and it gave weighted average F1 scores of 0.75 for Tamil, 0.95 for Malayalam and 0.71 for Kannada.

References

- Gaurav Arora. 2020. [Gauravarora@HASOC-Dravidian-CodeMix-FIRE2020: Pre-training ULMFiT on Synthetically Generated Code-Mixed Data for Hate Speech Detection](#).
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.
- Stevo Bozinovski. 2020. [Reminder of the first paper on transfer learning in neural networks, 1976](#). *Informatica*, 44.
- Bharathi Raja Chakravarthi. 2020a. [HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion](#). In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.
- Bharathi Raja Chakravarthi. 2020b. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.
- 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, M Anand Kumar, John Philip McCrae, Premjith B, Soman KP, and Thomas Mandl. 2020b. Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020)*. *CEUR Workshop Proceedings*. In: *CEUR-WS.org, Hyderabad, India*.
- Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020c. [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, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.
- Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P. McCrae. 2020d. [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.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Uniter: Universal image-text representation learning](#).
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. [Empirical evaluation of gated recurrent neural networks on sequence modeling](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#).
- Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [Xnli: Evaluating cross-lingual sentence representations](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. [Large-scale adversarial training for vision-and-language representation learning](#).

- Nikhil Kumar Ghanghor, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. [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.
- Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. [KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection](#). In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.
- K. He, X. Zhang, S. Ren, and J. Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. [UVCE-IIIT@DravidianLangTech-EACL2021: Tamil Troll Meme Classification: You need to Pay more Attention](#). In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Jeremy Howard and Sebastian Ruder. 2018. [Universal language model fine-tuning for text classification](#).
- 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*.
- Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. [A Survey of Current Datasets for Code-Switching Research](#). In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages](#). In *Findings of EMNLP*.
- Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. 2020. [IITK at SemEval-2020 Task 8: Unimodal and Bimodal Sentiment Analysis of Internet Memes](#).
- Vishnupriya Kolipakam, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. [A Bayesian phylogenetic study of the Dravidian language family](#). *Royal Society open science*, 5(3):171504.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. [Imagenet classification with deep convolutional neural networks](#). In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc.
- Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. [Gradient-based learning applied to document recognition](#). In *Proceedings of the IEEE*, pages 2278–2324.
- Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. [Dice loss for data-imbalanced nlp tasks](#).
- 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.
- Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. [Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding](#). In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.
- Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. [IIIT@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.
- Tharindu Ranasinghe, Sarthak Gupte, Marcos Zampieri, and Ifeoma Nwogu. 2020. [WLV-RIT at HASOC-Dravidian-CodeMix-FIRE2020: Offensive Language Identification in Code-switched YouTube Comments](#).
- Mike Schuster and Kuldip Paliwal. 1997. [Bidirectional recurrent neural networks](#). *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gambäck. 2020. [SemEval-2020 Task 8: Memotion Analysis – The Visuo-Lingual Metaphor!](#)
- R Sivanantham and M Seran. 2019. [Keeladi: An Urban Settlement of Sangam Age on the Banks of River Vaigai](#). *India: Department of Archaeology, Government of Tamil Nadu, Chennai*.

- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. [VL-BERT: Pre-training of Generic Visual-Linguistic Representations](#).
- 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).
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. [Going deeper with convolutions](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#).
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding](#).
- Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavaresan, 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.
- Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. [ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph](#).
- Ron Zhu. 2020. [Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution](#).