SSN_MLRG1@DravidianLangTech-ACL2022: Troll Meme Classification in Tamil using Transformer Models

Shruthi Hariprasad, Sarika Esackimuthu, Saritha Madhavan
Rajalakshmi Sivanaiah, Angel Deborah Suseelan
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai 603 110, Tamil Nadu, India
{shruthi2010101, sarika2010128, sarithamadhesh}@ssn.edu.in
{rajalakshmis, angeldeborahs}@ssn.edu.in

Abstract

The ACL shared task of DravidianLangTech-2022 for Troll Meme classification is a binary classification task that involves identifying Tamil memes as troll or not-troll. Classification of memes is a challenging task since memes express humour and sarcasm in an implicit way. Team SSN_MLRG1 tested and compared results obtained by using three models namely BERT, ALBERT and XLNet. The XLNet model outperformed the other two models in terms of various performance metrics. The proposed XLNet model obtained the 3rd rank in the shared task with a weighted F1-score of 0.558.

1 Introduction

Memes are interesting ideas that spread the emotions of the people or culture across the internet. Social media plays a pivotal role in facilitating the spread of memes. Memes have become a powerful tool of everyday communication that uses humour to catch the attention of the people and also help in sharing the views (Ghanghor et al., 2021a,b; Yasaswini et al., 2021). And though a good number of them are harmless, there are many memes that are not. Troll memes are such memes whose main goal is to provoke the audience with intent to offend or demean them (Priyadharshini et al., 2021; Kumaresan et al., 2021). Since the internet aids in the widespread propagation of memes it can be utilised in trolling (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). Considering the adverse mental effects of troll memes on individuals, the task is to identify and label memes as troll or not in an effort to monitor what is being posted on the internet.

Tamil is one of the world's longest-surviving classical languages. According to A. K. Ramanujan, it is "the only language of modern India that is recognizably continuous with a classical history." Because of the range and quality of ancient Tamil literature, it has been referred to as "one of the world's major classical traditions and literatures." For almost 2000 years, there has been a recorded Tamil literature. The earliest period of Tamil literature, known as Sangam literature, is said to have lasted from from 600 BC to AD 300 (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018). Among Dravidian languages, it possesses the oldest existing literature. The earliest epigraphic documents discovered on rock edicts and 'hero stones' date from the third century BC (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). In Tamil Nadu, the Archaeological Survey of India discovered over 60,000 of the 100,000 odd inscriptions discovered in India. The majority of them are in Tamil, with just around 5% in languages other than Tamil. Inscriptions in Tamil inscribed in Brahmi script have been unearthed in Sri Lanka, as well as on trade products in Thailand and Egypt (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021).

The task of classifying troll memes is challenging as it needs to discover the intention of the meme. Moreover, memes often use offensive words to express feelings. We used XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019) and BERT (Devlin et al., 2018) models that classify memes as troll and not a troll. The training data set provided contains 2300 memes that have been annotated, out of which 1610 memes were used for training and the rest were used for development of the model. The troll memes are very subjective and the usage of colloquial language, emojis, references, symbols and images without text add more challenges in predicting the trolls.

2 Related Work

Many researchers in the field of Artificial Intelligence and Natural Language Processing have

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 132 - 137
May 26, 2022 ©2022 Association for Computational Linguistics

been working to detect hateful memes (B and A, 2021b,a). In the past couple of years, social media usage has increased drastically and so the data available has also increased (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2021).

Shaheen et al. (2020) studied the performance of NLP transformer models BERT, RoBERTa , DistilBERT, XLNet and M-BERT in Large Multi-Label Text Classification (LMTC) and found that RoBERTa and BERT yield best results.

Afridi et al. (2020) presented an inclusive study on meme classification and proposed a generalised framework for visual-linguistic problems. Suryawanshi et al. (2020) created a dataset that has been used for classifying the memes. They have used image classification to address the difficulties in the state-of-the-art methods and concluded that such an image classifier is not feasible for classifying memes.

The findings of previous shared tasks (Suryawanshi and Chakravarthi, 2021), (Suryawanshi et al., 2022) have been shared where submissions show multiple ways to approach the problem. Smitha et al. (2018) stress the importance of visual memes and recommend a framework that could be utilized to categorize the internet memes by certain visual and textual features.

Hegde et al. (2021b) illustrates different textual analysis methods and contrasting multi-modal methods from simple merging to cross attention to using both visual and textual features. Cross-lingual language model XLM was found to perform the best in textual analysis, and the multi-modal transformer performed the best in multi-modal analysis. It was noted that the distribution of the test set does matter and the type of images was different in the test set which could mainly affect the performance of the ImageNet models while fine-tuning.

Du et al. (2020) provided the first large-scale analysis of who shares the image with text (IWT) memes, relative to other forms of expression and provided an analysis of the relationship between the demographics of users and their meme sharing patterns. They also developed an accurate, publicly available classifier to identify IWT memes in other data sets.

Hegde et al. (2021a) have put forth a model using vision transformer for images and Bidirectional Encoder Representations from Transformers (BERT) for captions of memes to achieve an overall F1-score of 0.59 on the test set. They believed that the preprocessing of the images was a huge factor for achieving a great F1-score on the validation set.

In (Sivanaiah et al., 2020), we worked to identify the presence of offensive language in social media posts using BERT. Deep network model with BERT embeddings was found to achieve better F1 score when compared to 1D-CNN model trained with GloVe pretrained embeddings, 2D-CNN and BiLSTM models with Word2Vec embeddings.

ColBERT (Contextualized Late Interaction over BERT), a modification of BERT was used to detect offense and humor in text in (Sivanaiah et al., 2021) which outperformed the machine learning models tested by a large margin.

3 Methodology

The architecture diagram for the offensive text classification is shown in Figure 1.

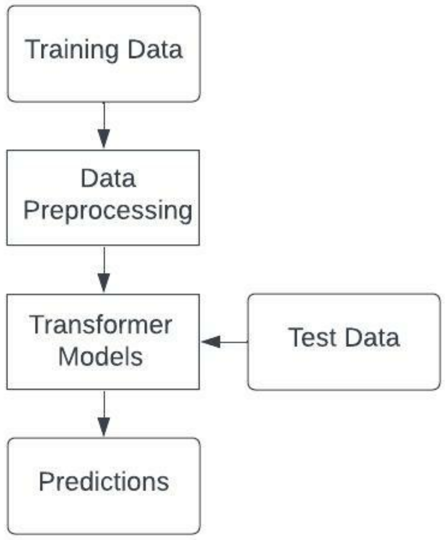

Figure 1: Architecture of the Proposed System

3.1 Dataset

The dataset consists of troll and non-troll images with their captions as text. The training data set provided contained 1018 troll and 1282 not troll memes. And for the test data, 395 of the total memes were annotated with troll and the remaining 272 were not troll. The data distribution of the train and test set is in Table 1.

The data set (Suryawanshi et al., 2020) provided by the organisers was used for the task. It contains image files of the memes as well as transcriptions

Class Label	Train Set	Test Set
Troll	1018	395
Not Troll	1282	272

Table 1: Distribution of Data Set

of the text embedded in the images. Every single one of these memes have been annotated as either troll or not troll and this label is embedded in the file name.

3.2 Data Preprocessing

Data preprocessing is an essential step where the given data set has to be regulated as much as possible in order to have some consistency. The data set was cleaned and processed using methods from NLTK (Loper and Bird, 2002) and spacy (Honnibal and Johnson, 2015) toolkit. In pre-processing, the following changes were made to the data: annotation of emojis and emoticons, conversion of uppercase letters to lowercase, expanding contractions, removal of URLs, reduction of lengthened words, removal of accented characters, removal of stopwords, removal of extra whitespaces, and lemmatization of text.

3.3 Model Description

We classified memes using three models namely BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019) and XLNet (Yang et al., 2019). The first model we used is BERT - Bidirectional Encoder Representations from Transformers. It is a deep learning model based on transformers. The directional models usually read the text input sequentially, but BERT reads the entire sequence of words at once. This characteristic allows the model to learn the context of a word based on all of its surroundings. The BERT model was trained for 4 epochs. The second model that we used is ALBERT, a Lite BERT, shrinks BERT in size while maintaining the performance. This model was trained for 5 epochs.

The next model that we used is XLNet, which is a BERT like pretrained model that has outperformed BERT in some NLP tasks including text classification. It captures bi-directional context using a mechanism called "permutation language modeling". XLNet does not suffer from pre-train fine-tune discrepancy since it does not depend on data corruption. We trained the XLNet model for 4 epochs.

4 Results

The models were tested on the test data with labels provided. The accuracy obtained by the proposed XLNet model is 0.59. BERT and ALBERT showed an accuracy of 0.58 and 0.57 respectively. XLNet was found to have the best performance out of all three models. Weighted F1-score, precision and recall are other performance metrics that have been used to measure the effectiveness of the model. Table 2 shows the results obtained with all three models.

Performance Metrics	BERT	ALBERT	XLNET
Accuracy	0.58	0.57	**0.59**
Weighted Avg. F1-score	0.54	0.54	**0.558**
Weighted Avg. Recall	**0.58**	0.57	0.565
Weighted Avg. Precision	0.55	0.54	**0.555**

Table 2: Performance Metrics of Transformer models

We secured a rank of 3 with the XLNet model in the shared task. The first rank had obtained a weighted average F1 score 0.596 while we obtained 0.558.

4.1 Error Analysis

The confusion matrix for the results obtained with the XLNet model is shown in figure 2. True positive was obtained for 311 memes and true negative was obtained for 69 memes. False positive and false negative was obtained for 203 memes and 84 memes respectively.

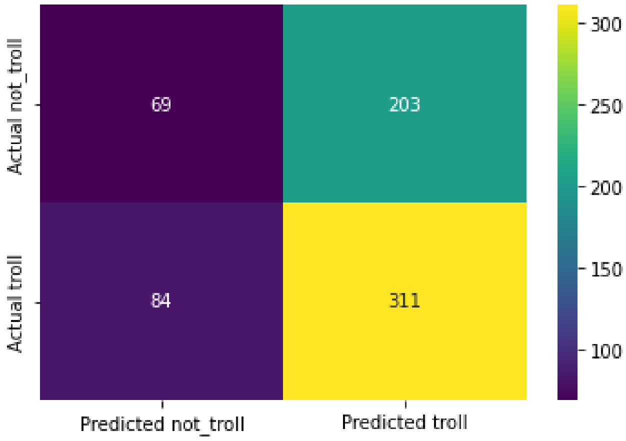

Figure 2: Confusion matrix for results with XLNet

5 Conclusion

We built an XLNet based model for the task Troll Meme Classification in Tamil. The model uses the preprocessed data done using NLTK, which we believe is a key factor for improved accuracy. Classifying a meme based only on the text is a reason for the reduced accuracy. How a meme is perceived is based upon a multitude of factors and cannot be judged using simple conventional models. This is another reason for the reduced precision of classification. The words present in a meme are too intuitive for the models to detect accurately. We intend to further proceed by adding multiple hidden layers and building a complex network structure.

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2020. A multimodal memes classification: A survey and open research issues. In *The Proceedings of the Third International Conference on Smart City Applications*, pages 1451–1466. Springer.

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

R Anita and CN Subalalitha. 2019b. Building discourse parser for Thirukkural. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 18–25.

Bharathi B and Agnusimmaculate Silvia A. 2021a. 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*, pages 336–339, Kyiv. Association for Computational Linguistics.

Bharathi B and Agnusimmaculate Silvia A. 2021b. SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318, Kyiv. Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. 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 and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual YouTube comments. *arXiv preprint arXiv:2109.00227*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164.

Nikhil 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*, pages 222–229, Kyiv. Association for Computational Linguistics.

Nikhil 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*, pages 197–203, Kyiv. Association for Computational Linguistics.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021a. Uvce-iiitt@ dravidianlangtech-eacl2021: Tamil troll meme classification: You need to pay more attention. *arXiv preprint arXiv:2104.09081*.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, Ratnasingam Sakuntharaj, Sathiyaraj Thangasamy, B Bharathi, and Bharathi Raja Chakravarthi. 2021b. Do images really do the talking? analysing the significance of images in tamil troll meme classification. *arXiv preprint arXiv:2108.03886*.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1373–1378.

Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P McCrae. 2021. Findings of shared task on offensive language identification in Tamil and Malayalam. In *Forum for Information Retrieval Evaluation*, pages 16–18.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada. In *Forum for Information Retrieval Evaluation*, pages 4–6.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishor Kumar Ponnusamy, and Santhiya Pandiyan. 2022. Findings of the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. 2020. Large scale legal text classification using transformer models. *arXiv preprint arXiv:2010.12871*.

Rajalakshmi Sivanaiah, S Milton Rajendram, Mirnalinee Tt, Abrit Pal Singh, Aviansh Gupta, Ayush Nanda, et al. 2021. Techssn at semeval-2021 task 7: Humor and offense detection and classification using colbert embeddings. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1185–1189.

Rajalakshmi Sivanaiah, Angel Suseelan, S Milton Rajendram, and Mirnalinee Tt. 2020. Techssn at semeval-2020 task 12: Offensive language detection using bert embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2190–2196.

ES Smitha, Selvaraju Sendhilkumar, and GS Mahalaksmi. 2018. Meme classification using textual and visual features. In *Computational Vision and Bio Inspired Computing*, pages 1015–1031. Springer.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

C. N. Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

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, Susan Levy, Paul Buitaleer, Prasanna Kumar Kumaresan, Rahul Ponnusamy, and Adeep Hande. 2022. Findings of the second shared task on Troll Meme Classification in Tamil. In *Proceedings of the Second 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).

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.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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*, pages 187–194, Kyiv. Association for Computational Linguistics.