TRO(F)LL or ROFL ? : Exploring Troll Detection in Tamil Memes

Aditya Krishna P[1], Swetha J[1], Rajalakshmi Sivanaiah[1]
Department of Computer Science and Engineering[1]
Sri Sivasubramaniya Nadar College of Engineering,
Chennai 603 110, Tamil Nadu, India
adityakrishna18010@cse.ssn.edu.in
swetha2110037@ssn.edu.in
rajalakshmis@ssn.edu.in

Abstract

The advent of social networks has deeply improved and enhanced the ways in which people communicate. However, along with the positives, there are negatives as well. The rapid dissemination of information via various means, be it tweets, Whatsapp forwards or memes has led to widespread misinformation and online abuse. The increasing prevalence of misinformation and "trolling", whether it is intended for amusement or with malicious intent has necessitated the development of methods to distinguish between the two. This is hard enough for languages that are spoken by most of the population, and even harder for languages that are not. Low-resource Languages struggle to combat the spread of hate and misinformation due to the dearth of computational and semantic resources. This study demonstrates the use of machine learning models, specifically BERT-based architectures like MuRIL to tackle these challenges. Troll memes were identified with an accuracy of 97% using MuRIL, demonstrating its capabilities in identifying trolls. The focus is on leveraging advanced natural language processing (NLP) techniques to build classifiers that can accurately parse the text from images using OCR and then translate the tamil text and finally predict if the meme is a troll or not.

1 Introduction

Social media has revolutionized the way that people interact with and share content. It offers a platform for people to express their ideas and opinions freely without strict surveillance as is the case in media sources like television, radios and newspapers. While freedom of expression is vital, leaving it unchecked can lead to people taking advantage of it. Although most posts on the internet are lighthearted and harmless, with most being targeted at current events to get more eyes on them, sometimes people cross the fine line between humour and insensitivity and hurt others. As social media continues to evolve, fostering a respectful environment while preserving open dialogue remains a crucial challenge for both platforms and users alike.

Trolling is a form of online bullying that involves harassing, criticizing, or antagonizing someone through provocatively disparaging or mocking public statements, postings, or acts. The one performing these acts is known as a troll (March and Marrington, 2019). Identifying whether or not a meme is a troll or not is still not as accurate as it can be (Suryawanshi et al., 2020) despite recent advancements in the field of natural language processing. (Wang and Wen, 2015) studied the variation of memes, and according to them a meme combines images with witty phrases and/or sarcastic or humorous text, making it clear that this is an image classification problem.

Memes have come to be the most passive-aggressive way to threaten or harm people nowadays. In a country as diverse as India, with multiple languages being spoken in each state, sharing memes in low resource languages has come to be a growing concern. Exploiting the lack of familiarity with certain languages, these memes, disguised as humorous content, can spread harmful or offensive messages without detection. This not only complicates moderation efforts but could also deepen social divides, as harmful content flies under the radar in regional dialects, evading scrutiny while still inflicting damage on specific communities. Memes are not always written in regional languages. They could also be transliterated or feature a mix of both english as well as another language.

Figure 1: A scene from "Manadhai Thirudivittai"

Figure 1 is an image of the tamil comedian Vadivelu with a smug expressian and the text in tamil translates to "The one who doesn't study is putting his fingerprint on the paper (because he does not know to write)... The one who studies is putting his fingerprint on the phone (modern security mechanisms)... That's the whole point." This suggests a comparison between two types of people: those who make an effort (the one who studies) and those who do not (the one who doesn't study). It highlights a humorous or sarcastic observation about how some people might claim credit or success without actually doing the work, while others who put in the effort are recognized in a different way.

Figure 2: A scene from "Friends"

Figure 2 is a scene from the hugely popular tamil movie "Friends". The image is from a scene featuring tamil actors Surya(In white) and comedian Vadivelu(in blue). In the image, Surya who is scrubbing the wall is working slowly. Vadivelu asks him *"Why are you rushing, slow down you have one and a half years until diwali. Keep it slow and steady. Such a mess, such a mess"*. The meme is a sarcastic jab at him for working so slowly that there is no need to worry about deadlines, that has been exaggerated for comedic effect. This is an example of a meme that is not harmful, and is meant to be seen and laughed at without offending or mocking anyone.

Memes are ubiquitous on the internet, but there is no satisfactory way to classify the memes as "Troll" or "Not Troll". This work uses the dataset created by (Suryawanshi et al., 2020) on two models, and compares their accuracy. The findings have been summarized using precision, recall, and F-score metrics.

2 Troll Meme

A troll meme is a form of internet content that combines humor, sarcasm, or provocative elements with the intent to offend, provoke, or elicit strong reactions from a target audience. Unlike traditional memes or Non-Troll memes that aim to entertain, troll memes are created to manipulate emotions, often stepping into offensive territory. They might use images and text in various combinations such as offensive text paired with neutral images, or benign text coupled with disturbing visuals to create a jarring contrast. The underlying goal of troll memes is to distract or disrupt conversations, provoke anger, or mock individuals, groups or social issues. These memes often thrive on the anonymity and rapid sharing that social media platforms provide, making them a powerful tool for spreading disruptive content. Figure 3 illustrates an example of a trolling meme targeting the Tamil television channel 'Vijay TV' for repeatedly airing the same movies. The translation of the text mocks the way the channel celebrates Bogi festival. The exact translation says "Discharge the old patients Raja rani, Bahubali, Saatai which refer to the names of the movies that have been aired by the channel multiple times in the past. Admit the new patients Pariyerum Perumal, Vadachennai, Saamy 2 referring to the new movies that are going to be aired. The template is from the movie Vasool Raja MBBS which is used to mock the channel for its repeated airing tendency. This meme is classified as a troll meme as it tries to tamper the reputation of the television channel.

Similarly, Figure 4 presents a trolling meme directed at the character 'Vijay' from the movie Theri. The reference image captures a scene where Vijay confronts the antagonist. The translated text reads: 'He (the antagonist) didn't even know you were the one who killed his son. So why did you reveal it yourself and invite trouble, ultimately leading to Samantha's (the female lead) death?'"

In conclusion, both examples demonstrate how troll memes leverage humor, sarcasm and cultural references to criticize or provoke, often targeting popular media figures or institutions. These memes are a powerful tool for social commentary but can also serve to damage reputations or stir controversy.

Figure 3: Example of Troll Meme 1

Figure 4: Example of Troll Meme 2

3 Literature Survey

Recent research on online trolling and aggression has highlighted various aspects of this complex issue. (Suryawanshi et al., 2020) introduced a dataset for classifying trolls in Tamil memes, which contributes to understanding online behavior across cultures. (Atanasov et al., 2019) examined the role of political trolls in social media discussions, while (Kumar et al., 2018) benchmarked aggression identification methods. (Clarke and Grieve, 2017) focused on the dimensions of abusive language on Twitter, and (Galery et al., 2018) explored aggression identification using multilingual word embeddings. Additionally, (Dinakar et al., 2012) addressed cyberbullying through common sense reasoning, and (Hosseinmardi et al., 2015) analyzed labeled cyberbullying incidents on Instagram. Research on machine translation for under-resourced languages has been advanced by Chakravarthi et al. (2019a, 2019b) (Chakravarthi et al., 2019a,b), who studied different orthographies and WordNet gloss translation. (Chakravarthi et al., 2020) created a corpus for sentiment analysis in code-mixed Tamil-English text, while (Rao and Lalitha Devi, 2013) looked into Tamil-English cross-lingual information retrieval. (Hariprasad et al., 2022) used three different transformer models namely BERT, ALBERT and XLNET on the same dataset to try and classify memes as troll or not troll. Finally, (Dash et al., 2015) highlighted the importance of generating bilingual texts for cross-lingual fertilization. This body of work underscores the necessity for improved tools and datasets to effectively address trolling and aggression in diverse linguistic environments.

4 Dataset

The dataset consists of a variety of images containing memes in Tamil, transliterated Tamil, as well as English, collected by (Suryawanshi et al., 2020). These memes include sentences in monolingual Tamil, transliterated Tamil in Roman script, code-mixed (combining Tamil and English within a single sentence), and code-switched (alternating between Tamil and English at phrase or sentence boundaries) as shown in in the Figure 3.

There are 2,289 memes in the training set, with 1,251 classified as troll images and 1,038 classified as not troll images. The test dataset comprises 659 images, with 389 being troll and 270 being not troll.

This dataset is slightly imbalanced, with troll images being more frequent (about 55% vs. 45%) than non-troll images. The presence of code-mixed and code-switched text further adds to the complexity of the dataset, posing unique challenges for preprocessing and classification tasks.f

5 Model Architecture

This paper makes use of the MuRIL model that is developed specifically for processing code-mixed and colloquial language like Tamil with code-switching. It builds upon the BERT architecture but adds modifications that can better capture the linguistic nuances that are seen in Indian languages, particularly in informal, mixed-language contexts. This makes MuRIL especially well-suited for detecting complex phenomena such as code-switching (i.e., the alternation between languages

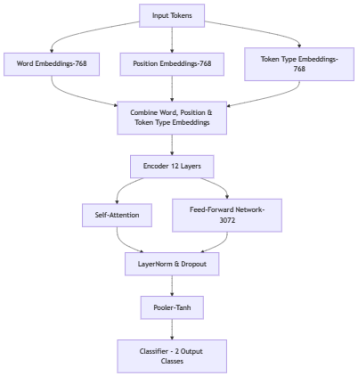

Figure 5: Flow diagram depicting MuRIL Model Architecture

within a sentence) and code-mixing (i.e., the intermingling of words from multiple languages).

The MuRIL model is configured with key hyperparameters that enhance its ability to handle the nuances of colloquial language. This means that an embedding size of 768, accompanied by a vocab size of 197,285, would allow the model to efficiently capture words and contextual meanings in the source as well as target languages. With an intermediate size of 3072 and 12 attention heads in self-attention mechanism, this would allow the model to process relationships between words efficiently and capture context and meaning for mixed-language phrases. The use of 12 layers in the encoder as well as max position embeddings up to 512 ensures a longer sequence would not reduce important sequential information in its processing.

Furthermore, various regularization techniques in the forms of hidden dropout probability, attention dropout that are 0.1 help keep the model free from the problem of overfitting, thus providing generalizations even when considering noisy or informal language usual in data in real scenarios. These hyperparameters are therefore critical to MuRIL's ability to be able to identify linguistic patterns and anomalies in code-mixed Tamil sentences, particularly for the task at hand such as abusive language or hate speech, since these often use a blend of languages. This sets up MuRIL appropriately to deal with the complexity involved in colloquialism in code-switching or code-mixing cases, hence it is aptly suited for this kind of model.

6 Methodology

This work used (Das et al., 2022) to train the datasets. MuRIL (Multilingual Representations for Indian Languages) is a pre-trained language model specifically designed for Indian languages and code-mixed languages based on the BERT architecture. One of the biggest advantages of it is the fact that it has been trained on a large corpus of indian languages and hence captures cultural nuances better than most other models. The referenced work performed a large-scale analysis of multilingual abusive speech in Indic languages and examined different interlingual transfer mechanisms and observed the performance of various multilingual models for abusive speech detection for eight different Indic languages.

Hyperparameter	Value
`_name_or_path`	Hate-speech-CNERG/tamil-codemixed-abusive-MuRIL
`embedding_size`	768
`hidden_size`	768
`num_attention_heads`	12
`num_hidden_layers`	12
`intermediate_size`	3072
`max_position_embeddings`	512
`vocab_size`	197285
`attention_probs_dropout_prob`	0.1
`hidden_dropout_prob`	0.1
`initializer_range`	0.02
`torch_dtype`	float32
`transformers_version`	4.46.3

Table 1: Important Model Configuration Hyperparameters

Both proposed approaches use OCRTamil (Prasath, 2024) (Rajendran, 2023) to extract tamil words from an image, which is then translated into english, before training the models.

The difference between both approaches is that the first model used train_test_split where the dataset is split into 80% training and 20% validation. The model is trained and evaluated only once on this specific split. It does not use early stopping, meaning the model always trains for the full number of epochs.

The second model used K-Fold Cross Validation with Early Stopping where the dataset is split into 5 subsets (folds), and the model is trained 5 times, each time using 4 folds for training and 1 fold for validation. This approach evaluates the model across multiple splits of the dataset, leading to a more reliable estimate of model performance.

Training stops early if validation loss does not improve for 1 epoch, potentially saving time and preventing overfitting. The model is evaluated on each fold's validation set after training, and the results are recorded for each fold.

7 Results and Discussion

The train-test split model, as shown in Table 2, demonstrated a relatively high recall (0.85) for the "Not Troll" class, indicating that it correctly identified 85% of the non-troll memes. However, its precision was low at 0.40, suggesting a high rate of false positives, as many troll memes were incorrectly classified as non-troll.

For the "Troll" class, the model exhibited poor performance, with a recall of only 0.11, meaning it correctly identified only 11% of actual troll memes. Although the precision was 0.50, the low recall and F1-score of 0.17 highlight the model's struggle in accurately detecting troll memes. This imbalance indicates a significant limitation in its ability to effectively classify troll content, which is crucial for the task at hand.

On the other hand, the K-Fold cross-validation model with MuRIL, shown in Table 3, achieved an accuracy of 59% and an F1-score of 0.74. With a precision of 0.59 and a recall of 0.97, the model correctly identified most troll memes but was less effective at predicting "Not Troll" content.

For the "Not Troll" class, the model's performance was weak, with a precision of 0.52 and an extremely low recall of 0.04, meaning it identified only 4% of actual non-troll memes. The F1-score of 0.08 underscores this poor balance between precision and recall, as 258 of the 270 "Not Troll" instances were misclassified as "Troll."

In contrast, the "Troll" class showed much stronger performance, with a recall of 0.97 indicating that 97% of troll memes were correctly classified. The precision of 0.59 suggests that while the majority of troll predictions were accurate, some false positives still occurred. Overall, the F1-score of 0.74 reflects good performance for this class.

From Table 4 it is clear that our work, which makes use of MuRIL as well as cross validation with early stopping, has siginificantly higher F1-Scores as well as Recall and Precision than that of (Hariprasad et al., 2022) that does not use MuRIL or cross validation.

The model's performance clearly exhibits an imbalance, effectively identifying troll memes but struggling significantly to recognize non-troll content. This discrepancy between the two classes suggests the need for further improvements, particularly in enhancing the model's ability to distinguish non-troll memes.

	Precision	Recall	F1-Score	Support
Not Troll	0.40	0.85	0.54	270
Troll	0.50	0.11	0.17	389
Macro avg	0.45	0.48	0.36	659
Weighted avg	0.46	0.41	0.32	659

Table 2: Classification Report with Train-Test Split

	Precision	Recall	F1-Score	Support
Not Troll	0.52	0.04	0.08	270
Troll	0.59	0.97	0.74	389
Macro avg	0.56	0.51	0.41	659
Weighted avg	0.56	0.59	0.47	659

Table 3: Classification Report with Cross Validated MuRIL (CVM Model)

Model	Accuracy	F1-score	Recall	Precision
BERT	0.58	0.54	0.58	0.55
ALBERT	0.57	0.55	0.57	0.54
XLNET	0.59	0.565	0.555	0.558
CVM Model	0.59	0.74	0.97	0.59
Train-Test Split	0.41	0.17	0.11	0.50

Table 4: Performance Metrics for BERT, ALBERT, XLNET, CVM Model, and Train-Test Split

7.1 Error Analysis

The confusion matrix in Figure 6 shows the classification results for the CV Model. A true positive was obtained for 378 memes, and a true negative was recorded for 12 memes. On the other hand, false positives were found in 258 cases, and false negatives in 11 cases. This indicates that while the model effectively identifies troll memes, it struggles with non-troll meme classification.

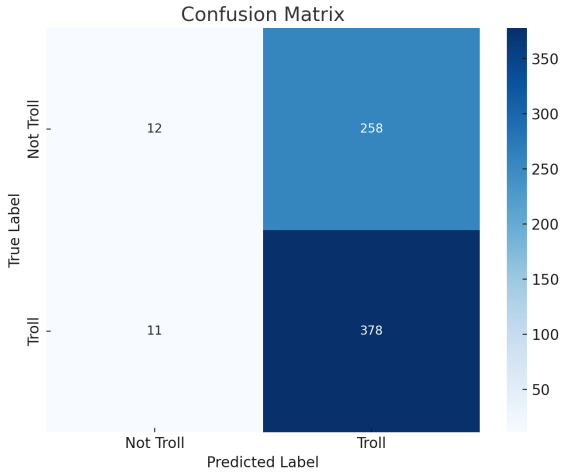

Figure 6: Confusion Matrix for CVM Model

8 Conclusion and Future Work

As seen from the results in Table 3, The model is proficient at predicting "Troll" memes with great accuracy, improving on the great work done by (Suryawanshi et al., 2020) and (Hariprasad et al., 2022). But improving the performance on "Not Troll" memes is critical for achieving balanced and effective classification. Future work could involve exploring multimodal approaches that better capture the intricate relationship between the image and text components of troll memes, and using a more generalized OCR Model to predict troll memes in other low-resource languages.

To summarize, while the model does demonstrate high accuracy in identifying troll memes, its bias toward misclassifying 'Non-Troll' memes as 'Troll' limits its practical application and can be worked upon.

References

Alexander Atanasov, Gianmarco DeFrancisci Morales, and Preslav Nakov. 2019. Predicting the role of political trolls in social media. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1023–1034, Hong Kong, China. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019a. Comparison of different orthographies for machine translation of under-resourced dravidian languages. In *Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 6:1–6:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019b. 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, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced Languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020)*, Marseille, France. European Language Resources Association (ELRA).

Ian Clarke and Jack Grieve. 2017. Dimensions of abusive language on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages.

N. S. Dash, A. Selvraj, and M. Hussain. 2015. Generating translation corpora in indic languages: Cultivating bilingual texts for cross-lingual fertilization. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 333–342, Trivandrum, India. NLP Association of India.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (ACM Trans. Interact. Intell. Syst.)*, 2(3).

Thiago Galery, Emmanouil Charitos, and Yanyan Tian. 2018. Aggression identification and multilingual word embeddings. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 74–79, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shruthi Hariprasad, Sarika Esackimuthu, Saritha Madhavan, Rajalakshmi Sivanaiah, and Angel S. 2022. SSN_MLRG1@DravidianLangTech-ACL2022: Troll meme classification in Tamil using transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 132–137, Dublin, Ireland. Association for Computational Linguistics.

Homa Hosseinmardi, Sara A. Mattson, Rahat I. Rafiq, Ruoyun Han, Qian Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International Conference on Social Informatics*, pages 49–66. Springer.

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.

Evita March and Jessica Marrington. 2019. A qualitative analysis of internet trolling. In *Cyberpsychology, Behavior, and Social Networking*, volume 22. Mary Ann Liebert.

D Gnana Prasath. 2024. Tamil ocr.

Prof. R. Rajendran. 2023. Optical character recognition (ocr) of textbook material in tamil. In *Language in India*.

T. P. R. K. Rao and Sobha Lalitha Devi. 2013. Tamil english cross-lingual information retrieval. In Prasenjit Majumder et al., editors, *Multilingual Information Access in South Asian Languages*, pages 269–279. Springer Berlin Heidelberg, Berlin, Heidelberg.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Varma, Mihael Arcan, John P. McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of tamil memes. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille. Language Resources and Evaluation Conference (LREC 2020), European Language Resources Association (ELRA). Licensed under CC-BY-NC.

William Yang Wang and Miaomiao Wen. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 355–365. Association for Computational Linguistics.