Findings of the Shared Task on Offensive Span Identification from Code-Mixed Tamil-English Comments

Manikandan Ravikiran[†*], **Bharathi Raja Chakravarthi**[‡], **Anand Kumar Madasamy**[⋆]
Sangeetha Sivanesan [°], **Ratnavel Rajalakshmi** [⊕], **Sajeetha Thavareesan**[◇]
Rahul Ponnusamy [⊖], **Shankar Mahadevan**[⋈]
[†]Georgia Institute of Technology, Atlanta, Georgia
[‡]Data Science Institute, National University of Ireland Galway
[⋆]National Institute of Technology Karnataka Surathkal, India
[°]National Institute of Technology, Trichy, India
[⊕]Vellore Institute of Technology, Chennai, India
[◇]Eastern University, Sri Lanka
[⊖]Indian Institute of Information Technology and Management, Kerala, India
[⋈]Thiagarajar College of Engineering, Madurai, India
mravikiran3@gatech.edu, bharathi.raja@insight-centre.org

Abstract

Offensive content moderation is vital in social media platforms to support healthy online discussions. However, their prevalence in code-mixed Dravidian languages is limited to classifying whole comments without identifying part of it contributing to offensiveness. Such limitation is primarily due to the lack of annotated data for offensive spans. Accordingly, in this shared task, we provide Tamil-English code-mixed social comments with offensive spans. This paper outlines the dataset so released, methods, and results of the submitted systems.

1 Introduction

Combating offensive content is crucial for different entities involved in content moderation, which includes social media companies as well as individuals (Kumaresan et al., 2021; Chakravarthi and Muralidaran, 2021). To this end, moderation is often restrictive with either usage of human content moderators, who are expected to read through the content and flag the offensive mentions (Arsht and Etcovitch, 2018). Alternatively, there are semi-automated and automated tools that employ trivial algorithms and block lists (Jhaver et al., 2018). Though content moderation looks like a one-way street, where either it should be allowed or removed, such decision-making is fairly hard. This is more significant, especially on social media platforms, where the sheer volume of content is overwhelming for human moderators especially. With ever increasing offensive social media contents focusing "racism", "sexism", "hate speech", "aggressiveness" etc. semi-automated and fully automated content moderation is favored (Priyadharshini et al., 2021; Chakravarthi et al., 2020b; Sampath et al., 2022). However, most of the existing works (Zampieri et al., 2020; Chakravarthi et al., 2022a; Bharathi et al., 2022; Priyadharshini et al., 2022) are restricted to English only, with few of them permeating into research that focuses on a more granular understanding of offensiveness.

Tamil is a agglutinative language from the Dravidian language family dating back to the 580 BCE (Sivanantham and Seran, 2019). It is widely spoken in the southern state of Tamil Nadu in India, Sri Lanka, Malaysia, and Singapore. Tamil is an official language of Tamil Nadu, Sri Lanka, Singapore, and the Union Territory of Puducherry in India. Significant minority speak Tamil in the four other South Indian states of Kerala, Karnataka, Andhra Pradesh, and Telangana, as well as the Union Territory of the Andaman and Nicobar Islands (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). It is also spoken by the Tamil diaspora, which may be found in Malaysia, Myanmar, South Africa, the United Kingdom, the United States, Canada, Australia, and Mauritius. Tamil is also the native language of Sri Lankan Moors. Tamil, one of the 22 scheduled languages in the Indian Constitution, was the first to be designated as a classical language of India (Subalalitha, 2019; Srinivasan and

[*]Corresponding Author

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 261 - 270
May 26, 2022 ©2022 Association for Computational Linguistics

Subalalitha, 2019; Narasimhan et al., 2018). Tamil is one of the world's longest-surviving classical languages. The earliest epigraphic documents discovered on rock edicts and "hero stones" date from the 6th century BC. Tamil has the oldest ancient non-Sanskritic Indian literature of any Indian language (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018). Despite its own script, with the advent of social media, code-switching has permeated into the Tamil language across informal contexts like forums and messaging outlets (Chakravarthi et al., 2019, 2018; Ghanghor et al., 2021a,b; Yasaswini et al., 2021). As a result, code-switched content is part and parcel of offensive conversations in social media.

Despite many recent NLP advancements, handling code-mixed offensive content is still a challenge in Dravidian Languages (Sitaram et al., 2019) including Tamil owing to limitations in data and tools. However, recently the research of offensive code-mixed texts in Dravidian languages has seen traction (Chakravarthi et al., 2021, 2020a; Priyadharshini et al., 2020; Chakravarthi, 2020). Yet, very few of these focus on identifying the spans that make a comment offensive (Ravikiran and Annamalai, 2021). But accentuating such spans can help content moderators and semi-automated tools which prefer attribution instead of just a system-generated unexplained score per comment. Accordingly, in this shared task, we provided code-mixed social media text for the Tamil language with offensive spans inviting participants to develop and submit systems under two different settings. Our CodaLab website[1] will remain open to foster further research in this area.

2 Related Work

2.1 Offensive Span Identification

Much of the literature related to offensive span identification find their roots in SemEval Offensive Span identification shared task focusing on English Language (Pavlopoulos et al., 2021), with development of more than 36 different systems using a variety of approaches. Notable among these include work by Zhu et al. (2021) that uses token labeling using one or more language models with a combination of Conditional Random Fields (CRF). These approaches often rely on BIO encoding of the text corresponding to offensive spans. Alternatively, some systems employ post-processing on these token level labels, including re-ranking and stacked ensembling for predictions (Nguyen et al., 2021). Then, there are exciting works of Rusert (2021); Pluciński and Klimczak (2021) that exploit rationale extraction mechanism with pre-trained classifiers on external offensive classification datasets to produce toxic spans as explanations of the decisions of the classifiers. Lexicon-based baseline models, which uses look-up operations for offensive words (Burtenshaw and Kestemont, 2021) and run statistical analysis (Palomino et al., 2021) are also widely explored. Finally, there are a few approaches that employ custom loss functions tailored explicitly for false spans. For code-mixed Tamil-English to date, there is only preliminary work by Ravikiran and Annamalai (2021) that uses token level labeling.

3 Task Description

Our task of offensive span identification required participants to identify offensive spans i.e, character offsets that were responsible for the offensive of the comments, when identifying such spans was possible. To this end, we created two subtasks each of which are as described. Example of offensive span is shown in Figure 1

3.1 Subtask 1: Supervised Offensive Span Identification

Given comments and annotated offensive spans for training, here the systems were asked to identify the offensive spans in each of the comments in test data. This task could be approached as supervised sequence labeling, training on the provided posts with gold offensive spans. It could also be treated as rationale extraction using classifiers trained on other datasets of posts manually annotated for offensiveness classification, without any span annotations.

3.2 Subtask 2: Semi-supervised Offensive Span Identification

All the participants of subtask 1 were also encouraged to submit a system to subtask 2 using semi-supervised approaches. Here in addition to training data of subtask 1, more unannotated data was provided. Participants were asked to develop systems using both of these datasets together. To this end, the unannotated data was allowed to be used in anyway as necessary to aid in overall model

[1] https://competitions.codalab.org/competitions/36395

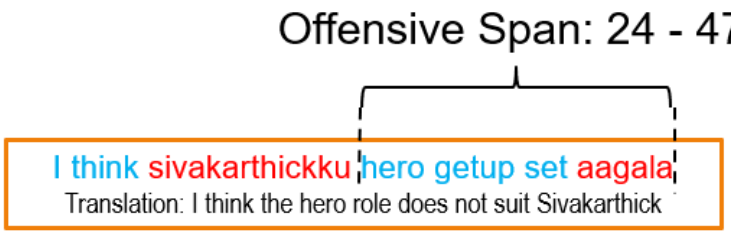

Figure 1: Example Offensive Span Identification from Code-Mixed Tamil-English Text

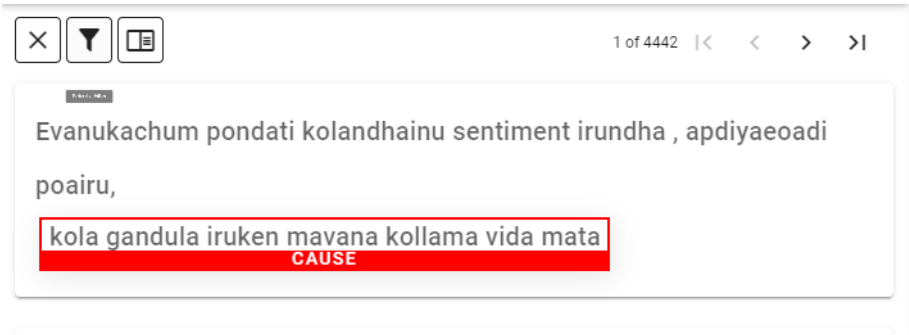

Figure 2: Annotation of offensive spans using Doccano.

development including creating semi-supervised annotations, ranking based on similarity etc.

4 Dataset

For this shared task, we build upon dataset from earlier work of Ravikiran and Annamalai (2021), which originally released 4786 code-mixed Tamil-English comments with 6202 offensive spans. We released this dataset to the participants during training phase for model development. Meanwhile for testing we extended this dataset with new additional annotated comments. To this end, we use dataset of Chakravarthi et al. (2022b) that consist of 10K+ offensive comments. From this we filter out comments that were already part of train set resulting 4442 comments suitable for annotation. Out of this we created (a) 3742 comments were used for creating the test data and (b) 700 comments were used for training phase of subtask 2.

Split	Train	Test
Number of Sentences	4786	876
Number of unique tokens	22096	5362
Number of annotated spans	6202	1025
Average size of spans (# of characters)	21	21
Min size of spans (# of characters)	4	3
Max size of spans (# of characters)	82	85
Number of unique tokens in spans	10737	1006

Table 1: Dataset Statistics used in this shared task

In line with earlier works (Ravikiran and Annamalai, 2021) for the 3742 comments we create span level annotations where at least two annotators annotated every comment. Additionally, we also employ similar guidelines for annotation, anonymity maintenance etc. Besides, no annotator data was collected other than their educational background and their expertise in the Tamil language.

Additionally, all the annotators were informed in prior about the inherent profanity of the content along with an option to withdraw from the annotation process if necessary. For annotation, we use doccano (Nakayama et al., 2018)which was locally hosted by each annotator. Within doccano, all the annotators were explicitly asked to create a single label called **CAUSE** with label id of 1, thus maintaining consistency of annotation labels. (See Figure 2).

To ensure quality each annotation was verified by one or more annotation verifier, prior to merging and creating gold standard test set. The overall dataset statistics is given in the Table 1. Compared to train set, we can see that the test set consists of significantly lesser number of samples, this is because many of the comments were either small or were hard to clearly identify the offensive spans. Overall for the 876 comments we obtained Cohen's Kappa inter-annotator agreement of 0.61 inline with Ravikiran and Annamalai (2021).

5 Competition Phases

5.1 Training Phase

In the training phase, the train split with 4786 comments, and their annotated spans were released for model development. Participants were given training data and offensive spans. No validation set was released; rather, participants were emphasized on cross-validation by creating their splits for preliminary evaluations or hyperparameter tuning. In total, 30 participants registered for the task and downloaded the dataset.

5.2 Testing Phase

Test set comments without any span annotation were released in the testing phase. Each participating team was asked to submit their generated span predictions for evaluation. Predictions are submitted via Google form, which was used to evaluate the systems. Though CodaLab supports evaluation inherently, we used google form due to its simplicity. Finally, we assessed the submitted spans of the test set and were scored using character-based F1 (See section 7.2).

6 System Descriptions

Overall we received only a total of 4 submissions (2 main + 2 additional) from two teams out of 30 registered participants. All these were only for subtask 1. No submissions were made for subtask 2. Each of their respective systems are as described.

6.1 The NITK-IT_NLP Submission

The best performing system from NITK-IT_NLP (Hariharan RamakrishnaIyer LekshmiAmmal, 2022) experimented with rationale extraction by training offensive language classifiers and employing model-agnostic rationale extraction mechanisms to produce toxic spans as explanations of the decisions of the classifier. Specifically NITK-IT_NLP used MuRIL (Khanuja et al., 2021) classifier and coupled with LIME (Ribeiro et al., 2016) and used the explanation scores to select words suitable for offensive spans.

6.2 The DLRG submission

The DLRG team (Mohit et al., 2022) formulated the problem as a combination of token labeling and span extraction. Specifically, the team created word-level BIO tags i.e., words were labelled as B (beginning word of a offensive span), I (inside word of a offensive span), or O (outside of any offensive span). Following which word level embeddings are created using GloVe (Pennington et al., 2014) and BiLSTM-CRF (Panchendrarajan and Amaresan, 2018) model is trained.

6.3 Additional Submission

After testing phase, we also requested each team to submit additional runs if they have variants of approaches. Accordingly we received two additional submissions from NITK-IT_NLP where they replaced MuRIL from their initial submission with (i) Multilingual-BERT (Devlin et al., 2019) and (ii) ELECTRA (Clark et al., 2020) respectively without any other changes. More details in section 7.2.

7 Evaluation

This section focuses on the evaluation framework of the task. First, the official measure that was used to evaluate the participating systems is described. Then, we discuss baseline models that were selected as benchmarks for comparison reasons. Finally, the results are presented.

7.1 Evaluation Measure

In line with work of Pavlopoulos et al. (2021) each system was evaluated F1 score computed on character offset. For each system, we computed the F1 score per comments, between the predicted and the ground truth character offsets. Following this we calculated macro-average score over all the 876 test comments. If in case both ground truth and predicted character offsets were empty we assigned a F1 of 1 other wise 0 and vice versa.

7.2 Benchmark

To establish fair comparison we first created two baseline benchmark systems which are as described.

- `BENCHMARK 1` is a random baseline model which randomly labels 50% of characters in comments to belong to be offensive. To this end, we run this benchmark 10 times and average results are presented in Table 2.
- `BENCHMARK 2` is a lexicon based system, which first extracted all the offensive words from the train set and during inference these words were searched in comments from testset and corresponding spans were extracted.
- `BENCHMARK 3` is RoBERTA (Liu et al., 2019; Ravikiran and Annamalai, 2021) model

trained using token labeling approach with BIO encoded texts corresponding to annotated spans.

Table 2: Official rank and F1 score (%) of the 2 participating teams that submitted systems. The baselines benchmarks are shown in red.

RANK	TEAM	F1 (%)
1	NITK-IT_NLP	44.89
BASELINE	BENCHMARK 1	39.75
BASELINE	BENCHMARK 2	37.84
BASELINE	BENCHMARK 3	38.61
2	DLRG	17.28

Table 2 shows the scores and ranks of two teams that made their submission. NITK-IT_NLP (Section 6.1) was ranked first, followed by DLRG (Section 6.2) that scored 27% lower was ranked second. The median score was 31.08%, which is far below the top ranked team and the benchmark baseline models. Meanwhile the additional submission post testing phase are excluded from ranked table. Instead they are presented separately in Table 3.

BENCHMARK 1 achieves a considerably high score and, hence, is very highly ranked with character F1 of 39.83%. Combination of MuRIL with LIME interpretability by model NITK-IT_NLP is ahead of BENCHMARK 1 by 11%, indicating the language models ability to effectively rationalize and identify the spans. This is inline the results of Rusert (2021) which show higher results than random baseline. Meanwhile BENCHMARK 2 and BENCHMARK 3, also shows F1 of 37.84% and 38.61% which again NITK-IT_NLP model tend to beat significantly. On contrary we could see that DLRG model to show least results of 17.28% lesser than akk the baselines as well as the top performing system. The lexicon-based BENCHMARK 2 and RoBERTA based BENCHMARK 3 too score very high. Especially as it overcomes, the submission of DLRG. This may be attributed to dataset domain itself. Especially, since much of the dataset was collected from Youtube comments section of Movie Trailers, often we see usages of same word or similar words. Such behavior is well established across social media forums including Youtube (Duricic et al., 2021), which begs to ask if indeed the dataset construction needs to be revisited, which forms one potential exploration for immediate future.

8 Analysis and Discussion

Overall we were happy to see the degree of involvement in this shared task with multiple participants registering, requesting access to datasets and potential baseline codes for the shared task. Though only two teams submitted the systems, the resulting diversity of approaches to this problem is fairly encouraging. However we include some of our observations below, from our evaluation and what we have learned from the results.

Table 3: Results of additional runs submitted by NITK-IT_NLP.

Method	F1 (%)
ELECTRA + LIME	37.33
M-BERT + LIME	33.95

8.1 Participation Characteristics

The authors reached out to teams that initially registered but failed to create any systems and the vast majority were undergraduate students who were new into the concept of shared task and were time-limited due to semester exams. The fact that students participated in the task is promising and we plan to consider more ways to introduce Shared tasks on Low-Resource Dravidian Languages in classrooms. To this end, the we used social media and other medium to spread the word around universities.

On the other hand, 60% of the participants did not download dataset after registering and instead chose to participate in other shared tasks, which is problematic and should be addressed. To this end, correspondence with such teams revealed potential favoritism towards classification based problems that are common in undergraduate studies. Moreover we also received multiple queries on the concept of offensive span itself during the training phase, which is a indicates potential need of improving the overall task structure with potential early release of data and task details. Yet, upon extending the number of submissions NITK-IT_NLP submitted additional runs (See Table 3). Additionally both the teams also submitted source codes [2] for their respective models encouraging further development of systems.

[2] https://drive.google.com/drive/folders/1T3kl8mljPt8oXcKVn7OQqaU3d55za2zZ?usp=sharing

Table 4: Results of submitted systems across comments of different lengths.

	F1@30 (%)	F1@50 (%)	F1@>50 (%)
NITK-IT_NLP	42.39	37.05	26.42
DLRG	39.62	23.47	14.05

8.2 General remarks on the approaches

Though neither of teams that made final submissions created any simple baselines, we could see that all the submissions of NITK-IT_NLP use well established approaches in recent NLP focusing on pretrained language models. Meanwhile DLRG used well-grounded Non-Transformer based approach. Yet neither of teams used any ensembles, data augmentation strategies or modifications to loss functions that are seen for the task of span identification in the past across shared tasks.

8.3 Error Analysis

Table 2 shows maximum result of 0.4489 with DLRG failing significantly compared to random baseline. To this end, we wonder if potentially these approaches have any weaknesses or strengths. To understand this, first we study the character F1 results across sentences of different lengths. Specifically we analysis results of (a) comments with less than 30 characters (F1@30) (b) comments with 30-50 characters (F1@50) (c) comments with more than 50 characters (F1@>50). The results so obtained are as shown in Table 4.

Firstly we can see though NITK-IT_NLP shows high results overall for cases of comments with larger lengths the model fails significantly. Specifically, comparing results with ground truth showed that use of LIME often restricts the overall word so selected as the rationale for offensiveness in turn reducing number of character offsets predicted as spans. This is because with larger texts the net score distribution weakens and span extraction is largely off leading to significant drop in results. Meanwhile for DLRG the results are more mixed, especially we can see that for comments with less than 30 characters the model shows improvement in F1. Analysis of results reveal that token labeling is highly accurate, which drops significantly with large size sentences. This may be attributed to non-local interactions between the words that may not be captured by the Bi-LSTM CRF model. Further more much of these sentences often contained only cuss words or clearly abusive words that are easily identifiable and often present in the train set. Also we found few bugs in the training code so used, which was already informed to the authors.

Besides error analysis also showed some implicit challenges in the proposed shared task. First the strong dependency of offensiveness on context makes it particularly difficult to solve as evident from NITK-IT_NLP which used language models. Second, offensiveness often is expressed as sarcasm or even is very subtle. In such cases we often see the offensiveness results to depend only the words bearing the most negative sentiment, meanwhile the ground truth spans annotated are larger thus showing high errors. Finally, many times the nature of offensiveness itself becomes debatable without clear context. Often these are the cases where we find the developed approaches to fail significantly.

9 Conclusion

Overall this shared task on offensive span identification we introduced a new dataset for code-mixed Tamil-English language with total of 5652 social media comments annotated for offensive spans. The task though has large participants, eventually had only two teams that submitted their systems. In this paper we described their approaches and discussed their results. Surprisingly rationale extraction based approach involving combination MuRIL and LIME performed significantly well. Meanwhile Bi-LSTM CRF model was found showing sensitivity towards shorter sentences, though it performed significantly worse than the random baseline. Also extracting offensive spans for long sentences were found to be difficult especially as they are context dependent. To this end, we release the baseline models and datasets to foster further research. Meanwhile in the future we plan to re-do the task of offensive span identification where we could require the participants to identify offensive spans and simultaneously classify different types of offensiveness.

Acknowledgements

We thank our anonymous reviewers for their valuable feedback. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors only and does not reflect the view of their employing organization or graduate schools. The shared task was result of series projects done during CS7646-ML4T (Fall 2020), CS6460-Edtech Foundations

(Spring 2020) and CS7643-Deep learning (Spring 2022) at Georgia Institute of Technology (OMSCS Program). Bharathi Raja Chakravarthi were supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2), co-funded by the European Regional Development Fund and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

References

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.

Andrew Arsht and Daniel Etcovitch. 2018. The human cost of online content moderation. *Harvard Journal of Law & Technology*.

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.

Ben Burtenshaw and Mike Kestemont. 2021. UAntwerp at SemEval-2021 task 5: Spans are spans, stacking a binary word level approach to toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 898–903, Online. 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, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

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

Bharathi Raja Chakravarthi 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, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020a. 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, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022a. 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, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022b. DravidianCodeMix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

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*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on*

Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tomislav Duricic, Volker Seiser, and Elisabeth Lex. 2021. Cross-platform analysis of user comments in youtube videos linked on reddit conspiracy theory forum. *CoRR*, abs/2109.01127.

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.

Manikandan Ravikiran Hariharan RamakrishnaIyer LekshmiAmmal, Anand Kumar Madasamy. 2022. Nitk-it_nlp@tamilnlp-acl2022: Transformer based model for toxic span identification in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shagun Jhaver, Sucheta Ghoshal, Amy S. Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput. Hum. Interact.*, 25(2):12:1–12:33.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

More Mohit, Naga Shrikriti Bhamatipati, Saharan Gitansh, Hanchate Samyuktha, Nandy Sayantan, and Rajalakshmi Ratnavel. 2022. Dlrg@tamilnlp-acl2022: Offensive span identification in tamil using bilstm-crf approach. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from `https://github.com/doccano/doccano`.

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.

Viet Anh Nguyen, Tam Minh Nguyen, Huy Quang Dao, and Quang Huu Pham. 2021. S-NLP at SemEval-2021 task 5: An analysis of dual networks for sequence tagging. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 888–897, Online. Association for Computational Linguistics.

Marco Palomino, Dawid Grad, and James Bedwell. 2021. GoldenWind at SemEval-2021 task 5: Orthrus - an ensemble approach to identify toxicity. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 860–864, Online. Association for Computational Linguistics.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection (to appear). In *Proceedings of the 15th International Workshop on Semantic Evaluation*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Kamil Pluciński and Hanna Klimczak. 2021. GHOST at SemEval-2021 task 5: Is explanation all you need? In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 852–859, Online. Association for Computational Linguistics.

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.

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. IEEE.

Manikandan Ravikiran and Subbiah Annamalai. 2021. DOSA: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17, Kyiv. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 97–101. The Association for Computational Linguistics.

Jonathan Rusert. 2021. NLP_UIOWA at Semeval-2021 task 5: Transferring toxic sets to tag toxic spans. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 881–887, Online. 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, and Santhiya Ponnusamy, Kishor Kumar 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.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and A. Black. 2019. A survey of code-switched speech and language processing. *ArXiv*, abs/1904.00784.

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*.

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.

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.

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.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çagri Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1425–1447. International Committee for Computational Linguistics.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.