MUCIC@TamilNLP-ACL2022: Abusive Comment Detection in Tamil Language using 1D Conv-LSTM

F. Balouchzahi [1, a], **M. D. Anusha** [2, b], **H. L. Shashirekha** [2, c], **G. Sidorov** [1, d]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, Mexico
[2]Department of Computer Science, Mangalore University, Mangalore, India
{[b]anugowda251, [c]hlsrekha}@gmail.com,
{[a]fbalouchzahi2021, [d]sidorov}@cic.ipn.mx

Abstract

Abusive language content such as hate speech, profanity, and cyberbullying etc., which is common in online platforms is creating lot of problems to the users as well as policy makers. Hence, detection of such abusive language in user-generated online content has become increasingly important over the past few years. Online platforms strive hard to moderate the abusive content to reduce societal harm, comply with laws, and create a more inclusive environment for their users. In spite of various methods to automatically detect abusive languages in online platforms, the problem still persists. To address the automatic detection of abusive languages in online platforms, this paper describes the models submitted by our team - MUCIC to the shared task on "Abusive Comment Detection in Tamil-ACL 2022". This shared task addresses the abusive comment detection in native Tamil script texts and code-mixed Tamil texts. To address this challenge, two models: i) n-gram-Multilayer Perceptron (n-gram-MLP) model utilizing MLP classifier fed with char-n gram features and ii) 1D Convolutional Long Short-Term Memory (1D Conv-LSTM) model, were submitted. The n-gram-MLP model fared well among these two models with weighted F1-scores of 0.560 and 0.430 for code-mixed Tamil and native Tamil script texts, respectively. This work may be reproduced using the code available in Gthub[1].

1 Introduction

Abusive language refers to the usage of words for any type of insult, vulgarity, profanity, sexism, or misogyny (Butt et al., 2021) that debases the target, as well as anything that causes aggravation (Spertus, 1997). The term abusive language is often reframed as offensive language (Razavi et al., 2010) and hate speech (Djuric et al., 2015; Chakravarthi et al., 2021b). In recent years, an increasing number of users have witnessed the offensive behavior on social media (Duggan, 2017) targeting individuals, group or community. In spite of many social media companies using a variety of tools such as human reviewers, user reporting procedures, etc., to censor the offensive language, the problem is growing day by day mainly because the offensive/abusive language detection algorithms fail to capture the subject and context-dependent characteristics of the text (Chatzakou et al., 2017; Priyadharshini et al., 2021; Kumaresan et al., 2021). For example, an individual message may appear harmless, but when viewed in the context of previous threads, it may appear abusive, and vice versa. It is challenging even for human beings to detect such abusive language.

Social media texts are usually written mixing regional languages such as Tamil, Kannada, Malayalam, etc., with English at sub-word, word or sentence level (Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). Further, the usage of internet slangs, words in short forms, words of other languages, emojis etc., adds to the problem of tackling abusive language (Balouchzahi and Shashirekha, 2021; Anusha and Shashirekha, 2020). The focus of abusive comment detection algorithms on low-resources like Tamil is rarely explored due to scarcity and unavailability of annotated dataset Amjad et al. (2021b).

"Abusive Comment Detection in Tamil-ACL 2022"[2] shared task (Priyadharshini et al., 2022) encourages researchers to develop models for detecting comments in native Tamil script texts as well as code-mixed Tamil texts. The objective of the shared task is to identify the abusive content in Tamil and categorize it into predefined abusive language categories. To address the challenges of the shared task, we - team MUCIC, submitted two models: i) n-gram-MLP model utilizing MLP classifier fed with char-n gram features and ii) 1D

[1]https://github.com/anushamdgowda/abusive-detection

[2]https://competitions.codalab.org/competitions/36403

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 64 - 69

Conv-LSTM model, to detect abusive comments in Tamil. This paper describes the methodology of the proposed models and the results obtained.

The rest of the paper is arranged as follows: A review of related work is included in Section 2, and the methodology is discussed in Section 3. Experiments, and results are described in Section 4 followed by concluding the paper with future work in Section 5.

2 Related Work

Most of the abusive comment detection works focus on high-resource languages like English, leaving the low-resource languages such as Dravidian languages, Arabic, Persian, Urdu, etc., unexplored for the task (Amjad et al., 2021a).

A brief description of some of the recent abusive language detection works are given below:

The main problem with low-resource languages are the annotated datasets for abusive language detection. Even human annotators find it difficult to annotate some of the comments as abusive because of which building a large and reliable dataset becomes challenging. Chatzakou et al. (2017) found that datasets openly available for abusive language detection on Twitter ranged from 10K to 35K in size and are insufficient to train Deep Learning (DL) models.

Ashraf et al. (2021) explored abusive comment detection in YouTube comments using several Machine Learning (ML) and DL models as baselines and used n-grams features and pre-trained Glove embeddings to train ML and DL models respectively. Ada-boost (ML model) and 1-Dimensional Convolutional Neural Network (1D-CNN) (DL model) models obtained 87.29 and 89.24 F1-scores on comments without replies. Adding replies as conversational context enhanced the results to 91.96 and 91.68 F1-scores for Ada-boost and 1D-CNN respectively.

Lee et al. (2018) compared various learning models using Hate and Abusive Speech Twitter dataset (Founta et al., 2018). In addition to traditional ML approaches (NB, LR, SVM, and RF), they also investigated Neural Network (NN) models (CNN, Recurrent Neural Networks (RNN) and Bidirectional Gated Recurrent Unit (BiGRU)). Term Frequency-Inverse Document Frequency (TF-IDF) of word vectors and pre-trained GloVe vectors were used to train ML and NN models. Further, Latent Topic Clustering (LTC) which extracts latent topic information from the hidden states of RNN is used as additional information in classifying the text data. BiGRU model based on word features and LTC out performed the other models with an F1-score of 0.805.

Eshan and Hasan (2017) experimented TF-IDF of unigram, bigram, and trigram features to train ML algorithms (RF, Multinomial NB, SVM with Linear, Radial Basis Function, Polynomial, and Sigmoid kernels) and evaluated Facebook dataset of Bengali abusive text. SVM with Linear kernel and trigram feature achieved the best accuracy of 76% accuracy among all the models.

ML (Linear Support Vector Classifier (LinearSVC), LR, MNB, RF) and DL (RNN with Long Short Term Memory (LSTM)) algorithms, were used to detect multi-type abusive Bengali text by Emon et al. (2019). LinearSVC, LR, and MNB models were trained with filtered non-Bengali data transformed to vectors using a CountVectorizer[3]. and RF classifier was trained with the TF-IDF vectors obtained after filtering punctuation, numerals, and emotions. For DL model, the raw dataset is stemmed and word embedding is utilized to encode the text. RNN with LSTM outperforms other algorithms with the highest accuracy of 82.20%.

Several code-mixed Tamil datasets are used in various shared tasks, such as Sentiment Analysis in Tamil (Chakravarthi et al., 2020), Hate Speech Detection in Dravidian Languages (Mandl et al., 2020), Hope Speech Detection (Chakravarthi, 2020), Offensive Language Identification (OLI) in Dravidian Languages, (Chakravarthi et al., 2021a), etc. Since code-mixed texts do not follow any grammar, Balouchzahi et al. (2021a) proposed a learning model using sub-words generated by char sequences to deal with code-mixed texts for the task of OLI in Dravidian languages (Chakravarthi et al., 2021a). They used word n-grams with sub-words and a majority voting classifier with eXtreme Gradient Boosting (XGB), LR, and MLP estimators and obtained a weighted average F1-score of 0.75.

In another experiment on code-mixed Tamil texts, Balouchzahi et al. (2021b) combined char sequences with syntactic bi-grams and tri-grams for Hope Speech Detection task (Chakravarthi, 2020) and fed a voting classifier with three ML estimators, namely: LR, XGB and MLP. The authors created a code-mixed BERT language model from

[3]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

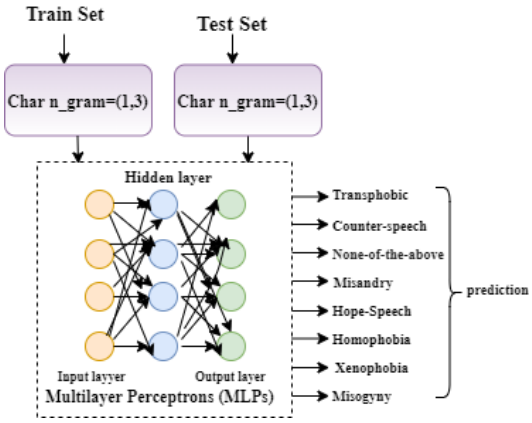

Figure 1: Framework of n-gram-MLP model

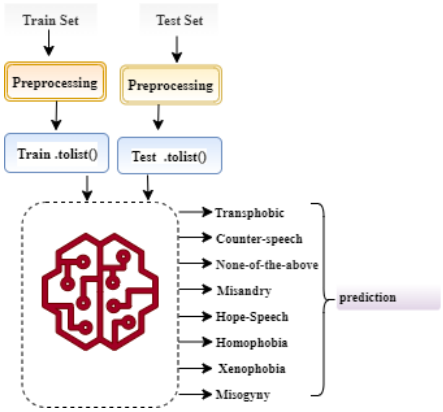

Figure 2: Framework of 1D Conv-LSTM model

scratch and obtained an average weighted F1-score of 0.54. However, in this study, the best performance was that of hard voting classifier with an average weighted F1-score of 0.59 that secured third rank in the competition.

3 Methodology

The first step in processing text data is to clean the text by removing the punctuation symbols, numerical data, frequently occurring words, and stopwords, as these features do not help in identifying the abusive content. Clean data is expected to improve the performance of the learning models. Two models: i) n-gram-MLP trained with char n-grams and ii) 1D Conv-LSTM model, were proposed to identify the abusive comment from native Tamil script and code-mixed Tamil texts. The framework of the proposed models are shown in Figure 1 and 2 and explanation of the models follows:

3.1 n-gram-MLP model

Many text processing projects utilize n-grams features since they are easy to implement and are scalable. A model with a larger 'n' value can store more contexts with a well-understood space-time tradeoff (Balouchzahi and Shashirekha, 2020) allowing many text processing experiments to scale up efficiently.

char n-grams in the range (1, 3) are extracted from the texts and vectorized using TfidfVectorizer[4]. These vectors are used to train MLP classifier by setting hidden layer sizes to (150, 100, 50), maximum iterations to 300, Random state to 1, activation to Relu and solver to Adam.

3.2 1D Conv-LSTM model

Keras Tokenizer[5] tokenizes the text and transforms it into a vector where the coefficient for each token could be binary, based on word count or TF-IDF. Further, the vocabulary size and maximum length of sequences are set to 60,000 and 50 respectively. "Pad_sequences" was utilized to keep all sequences at same length. The three parameters: "input dim", "output dim" and "input length" are set to 60,000 (vocabulary size), 1,000 (vector length of word) and 500 (maximum length of a sequence) respectively. Eventually, a 1D convolutional layer with 64 filtres, two pooling layers, and a relu activation function, followed by 100 fully connected LSTM layers and a soft-max output layer are used in this model to classify the given input.

4 Experiments and Results

The datasets provided by the shared task organizers contains native Tamil script (Tamil) and code-mixed Tamil (Ta-En) texts and the task is to classify the input text into different categories as shown in Table 1. Further, the table also gives the breakup of Train and Test sets for both Tamil and Ta-En datasets. The observation of data distribution reveals that both native and code-mixed Tamil datasets are imbalanced and that makes the classification task more problematic. Fox example, there are only 35, 6, and 2 samples in Homophobia, Transphobic, and not-Tamil classes respectively against 446, 149 and 95 samples in Misandry, Counter-speech and Xenophobia respectively, in the Train set of Tamil dataset. Few samples of the native script and code-mixed texts in the datasets are shown in Table 2.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[5]https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer

Label\Set	Train		Test	
	Tamil	Ta-En	Tamil	Ta-En
None-of-the-above	1296	3720	346	919
Misandry	446	830	104	218
Counter-speech	149	348	36	95
Xenophobia	95	297	29	70
Hope-Speech	86	213	11	53
Misogyny	125	211	24	50
Homophobia	35	172	8	43
Transphobic	6	157	2	40
Not-Tamil	2	-	-	-
Total	2240	5791	560	1488

Table 1: Distribution of labels in the given datasets

Language	Text	label
Tamil	தாசி மகன் சைமன் என்ற சீமான் என்ற பைத்தியகார பன்னி	Misandry
	செக்ஸ்சாஸ்திரி ஸ்கூல் ஆகா என்ன அருமையான பெயர். இனிமேல்	Homophobia
	சீமான் ஒரு தமிழர் அல்ல	Xenophobia
Ta-En	Guru murthi dhevudiyalukku porantha dhevudiya pullaiya	Misogyny
	Ama manigandan unmaitham.evaroda comments thappa peasra unga ellorukum samarpanam	Counter-speech
	Sappa nose ah udaikum alavukku	Xenophobia

Table 2: Samples of texts in the given dataset

The unlabeled Test sets shared by the organizers were used to evaluate the proposed models and the predictions were submitted to the organizers for final evaluation and ranking. As per the results in the final leaderboard of the shared task, the proposed n-gram-MLP model obtained average weighted F1-scores of 0.560 and 0.430 for Tamil and Ta-En texts respectively. Results of the proposed models on Development set and Test set are shown in Table 3 and 4 respectively. The comparison of average weighted F1-scores among the participating teams in the shared task shown in Figure 3 illustrates that the performance of the n-gram-MLP model is considerate.

5 Conclusion

This paper describes the participation of our team MUCIC in "Abusive Comment Detection in Tamil-ACL 2022" shared task. The objective of this shared task is to identify the different categories of abusive comments in native Tamil script and code-mixed Tamil texts. Among the two models, n-gram-MLP trained with n-grams and 1D Conv-LSTM model submitted for this shared task, n-gram-MLP classifier outperformed on both code-mixed Tamil and native Tamil script texts with average weighted F1-scores of 0.560 and 0.430, respectively.

Model	Language	w_F1-score	m_F1-score
MLP	Ta-En	0.64	0.28
	Tamil	0.56	0.33
1D Conv-LSTM	Ta-En	0.54	0.29
	Tamil	0.60	0.27

Table 3: Macro F1-score(m_F1-score) and Weighted F1-score(w_F1-score) F1-score on Development set

Language /Metric	w_F1-score	m_F1-score	Rank
Ta-En	0.560	0.290	6
Tamil	0.430	0.120	10

Table 4: Macro F1-score(m_F1-score) and Weighted F1-score(w_F1-score) F1-score on Test set

References

Maaz Amjad, Noman Ashraf, Alisa Zhila, Grigori Sidorov, Arkaitz Zubiaga, and Alexander Gelbukh. 2021a. Threatening Language Detection and Target Identification in Urdu Tweets. *IEEE Access*, 9:128302–128313.

Maaz Amjad, Alisa Zhila, Grigori Sidorov, Andrey Labunets, Sabur Butt, Hamza Imam Amjad, Oxana Vitman, and Alexander Gelbukh. 2021b. UrduThreat@ FIRE2021: Shared Track on Abusive Threat Identification in Urdu. In *Forum for Information Retrieval Evaluation*, pages 9–11.

MD Anusha and HL Shashirekha. 2020. An Ensemble Model for Hate Speech and Offensive Content Identification in Indo-European Languages. In *FIRE (Working Notes)*, pages 253–259.

Noman Ashraf, Arkaitz Zubiaga, and Alexander Gelbukh. 2021. Abusive Language Detection in YouTube Comments Leveraging Replies as Conversational Context. *PeerJ. Computer science*, 7:e742.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021a. MUCS@DravidianLangTech-EACL2021: COOLI-Code-Mixing Offensive Language Identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021b. MUCS@LT-EDI-EACL2021:

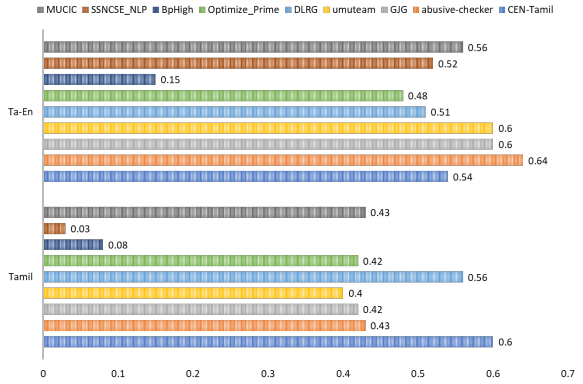

Figure 3: Comparison of average weighted F1-scores of the participating teams

CoHope-Hope Speech Detection for Equality, Diversity, and Inclusion in Code-Mixed Texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 180–187, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi and H L Shashirekha. 2021. LA-SACo: A Study of Learning Approaches for Sentiments Analysis in Code-Mixing Texts. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 109–118, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi and HL Shashirekha. 2020. Puner-Parsi ULMFiT for Named-Entity Recognition in Persian Texts. In *Congress on Intelligent Systems*, pages 75–88. Springer.

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.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander Gelbukh. 2021. Sexism Identification using BERT and Data Augmentation-EXIST2021. In *International Conference of the Spanish Society for Natural Language Processing SEPLN 2021, IberLEF 2021*.

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, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. 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.

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, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021a. 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*, pages 133–145, Kyiv. 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. 2021b. Dataset for identification of homophobia and transophobia in multilingual YouTube comments. *arXiv preprint arXiv:2109.00227*.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Maeve Duggan. 2017. Online Harassment 2017.

Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, and Tanni Mittra. 2019. A Deep Learning Approach to Detect Abusive Bengali Text. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE.

Shahnoor C Eshan and Mohammad S Hasan. 2017. An Application of Machine Learning to Detect Abusive Bengali Text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael

Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

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.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative Studies of Detecting Abusive Language on Twitter. *arXiv preprint arXiv:1808.10245*.

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*, pages 29–32.

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.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection using Multi-level Classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

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.

Ellen Spertus. 1997. Smokey: Automatic Recognition of Hostile Messages. In *Aaai/iaai*, pages 1058–1065.