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

Manikandan Ravikiran [†,*], **Ratnavel Rajalakshmi** [⊕], **Bharathi Raja Chakravarthi** [‡]
Anand Kumar Madasamy [⋆], **Sajeetha Thavareesan** [⊖]
[†] Georgia Institute of Technology, Atlanta, Georgia
[⊕] Vellore Institute of Technology, Chennai, India
[⋆] National Institute of Technology Karnataka Surathkal, India
[‡] School of Computer Science, Univeristy of Galway, Ireland
[⊖] Eastern University, Srilanka
mravikiran3@gatech.edu
bharathi.raja@insight-centre.org

Abstract

Effectively managing offensive content is crucial on social media platforms to encourage positive online interactions. However, addressing offensive contents in code-mixed Dravidian languages faces challenges, as current moderation methods focus on flagging entire comments rather than pinpointing specific offensive segments. This limitation stems from a lack of annotated data and accessible systems designed to identify offensive language sections. To address this, our shared task presents a dataset comprising Kannada-English code-mixed social comments, encompassing offensive comments. This paper outlines the dataset, the utilized algorithms, and the results obtained by systems participating in this shared task.

1 Introduction

Addressing offensive content holds immense importance for various parties engaged in content moderation, such as social media companies and individuals (Subramanian et al., 2022; Chinnaudayar Navaneethakrishnan et al., 2023). Typically, moderation methods involve either human moderators reviewing content to flag offensive material or the use of semi-automated and automated tools employing basic algorithms and predefined block lists (Jhaver et al., 2018). Despite the appearance of content moderation as a straightforward decision between allowing or removing content, this process is complex (Swaminathan et al., 2022). This complexity is amplified on social media platforms due to the overwhelming volume of content, making it challenging for human moderators (Kumaresan et al., 2022; Chakravarthi, 2022b,a). With the continuous rise in offensive social media content, particularly offensive comments and statements, there's a preference for semi-automated and fully automated content moderation approaches (Ravikiran et al., 2022; Chakravarthi, 2023; Chakravarthi et al., 2023a).

Kannada, an ancient Dravidian language, holds a significant historical legacy (Narasimhacharya, 1990). Predominantly spoken in the Indian state of Karnataka, Kannada serves as the official language in the state, carrying cultural significance that extends beyond regional boundaries (TNN, 2010). With the emergence of digital communication platforms, code-switching has also found its way into Kannada discourse, especially in informal online exchanges. This blending of languages and linguistic variations within social media has led to the integration of code-switched content in discussions, including offensive content, impacting the nature of online conversations in Kannada-speaking communities.

Despite recent advancements in natural language processing (NLP), addressing offensive code-mixed content in Dravidian languages, including Kannada, remains challenging due to limitations in available data and tools (Sitaram et al., 2019). However, there has been a noticeable surge in research focused on offensive code-mixed texts in Dravidian languages (Chakravarthi, 2020;

[*] Work done during graduate school

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 43–48

Chakravarthi et al., 2023a,b), although few of these studies concentrate on pinpointing the specific segments within a comment that render it offensive (Ravikiran and Annamalai, 2021; Ravikiran et al., 2022). Identifying these specific segments could significantly aid content moderators and semi-automated tools that prioritize the detection and categorization of offensive content. The existing body of research on identifying offensive spans primarily stems from the works of Ravikiran and Annamalai (2021). Post this there are multiple iterations of shared tasks focusing on offensive span identification in Tamil Ravikiran and Annamalai (2021); LekshmiAmmal et al. (2022); Rajalakshmi et al. (2022); Ravikiran et al. (2023). However to date there are no works in code-mixed Kannada language. To address this gap, we introduced the first phase of code-mixed social media text in Kannada, encompassing offensive segments. We invited participants to develop and submit systems under two distinct settings for this collaborative task. Our CodaLab website[1] will remain open to encourage further research in this domain.

2 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.

2.1 Subtask 1: Supervised Offensive Span Identification

With provided comments and labeled offensive spans used for training, the systems were tasked with detecting these offensive segments within the comments in the test dataset. This challenge could be addressed through supervised sequence labeling, involving training on the given posts that contain verified offensive spans. Alternatively, it could be tackled as rationale extraction by employing classifiers trained on other datasets of posts manually marked for offensive content classification, even in the absence of specific span annotations.

3 Dataset

For this shared task, we build on top of the dataset from earlier work of Ravikiran and Annamalai (2021), which originally released 1801 code-mixed Kannada-English comments with 1641 offensive spans. We released this dataset to the participants during training phase for model development. No additional data were released for development/validation purposes. Meanwhile for testing we extended introduced new additional annotated comments. To this end, the dataset of Hande et al. (2021) was used. From this we selected 444 comments for testing purpose. The test data had multiple instances where the offensive parts were completely not present. Such comments would help in identification of model biases in predciting spans if any.

Building on prior investigations (Ravikiran et al., 2023), we established span-level annotations for this fresh selection of 444 test comments. Employing the same procedures and guidelines for annotation, including measures to maintain anonymity, we introduced a explanation regarding offensive contents in the data, offering the option to abstain from the annotation process if deemed necessary. To ensure precision, each annotation underwent scrutiny by one or more annotation verifiers before amalgamating them through hard voting to form a standardized gold test set. Overall, concerning the 444 comments, we achieved a Cohen's Kappa inter-annotator agreement of 0.61.

4 Competition Phases

4.1 Training Phase

In the training phase, the train split with 1801 comments, and their annotated spans were released for model development. Participants were given training data and offensive spans. Participants were also emphasized on cross-validation by creating their splits for preliminary evaluations or hyperparameter tuning. In total, 45 participants registered for the task and downloaded the dataset.

4.2 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.

[1] https://codalab.lisn.upsaclay.fr/competitions/16090

5 System Descriptions

Overall we received only a total of 14 submissions from 7 teams All these were only for subtask 1. No submissions were made for subtask 2.

5.1 The SELAM Submission

Selam Submission used large language models composing one of the BERT or RoBERTA models. The methods showed the best result of 81.18% in F1.

5.2 The MIT_KEC_NLP Submission

MIT_KEC_NLP submission preprocessed data using custom stop word removal. These processed sentences are used converted to form TF-IDF which were used to train ensemble of multiple models. These final ensemble showed the result of 61.05% in F1.

5.3 The BYTESIZED_LLM Submission

BYTESIZED_LLM team utilized embeddings generated from a large open dataset, encompassing 100,000 comments. Following this Bi-LSTM model was trained to predict token level labels on test set. The final F1 obtained was 33.02%.

5.4 The CUET_RUN2 Submission

CUET_RUN2 used text preprocessing involving punctuation removal without any addition of more training data. This prepared data was was used for BERT finetuning with supervised method with L3-cube Kannada model to achieve result of 31.84% in F1.

5.5 The DLRG_3 Submission

DLRG_3 used a Bi-LSTM architecture with results of 21.92% in F1.

5.6 The MLG Submission

MLG team used an inception layer based CNN with kernel sizes 3, 5, 7, 9 and 11 with prediction of character level offensiveness probability. The output span is created by taking all the characters with higher probability of being offensive and multiplying with a mask to ensure that output does not exceeds the original sentence length. Finally the result obtained was 23.65% in F1.

5.7 The TEAMKUBOK Submission

TEAMKUBOK employed preprocessing with changing character level spans to word level spans. They fine tuned four pretrained language models and their predictions were averaged for the first occurrence of the offensive span of all the models and the last occurrence of the offensive span of all the models. Between these spans are returned as final output. The final result obtained was 12.94% in F1.

6 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 benchmarking of overall results. Finally we present remark on the approaches used and the analysis of the results from these submitted systems.

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 444 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.

The overall results of benchmarked systems are as shown in Table 1.

Table 1: Official rank and F1 score (%) of the 3 participating teams that submitted systems. The baselines benchmarks are also shown.

TEAM NAME	F1	RANK
SELAM	81.18	1
MIT_KEC_NLP	61.05	2
BYTESIZED LLM	33.02	3
CUET_RUN2	31.84	4
MLG	23.65	5
DLRG_3	21.92	6
TEAMKUBOK	12.94	7

Overall, the shared task showed higher level engagement compared to earlier iterations with members beyond Indian subcontinent showing interest in in obtaining datasets, and seeking potential baseline codes for the project. Infact many of the participants wanted earlier submission window open and have multiple runs to be submitted. To this end, we allowed maximum of three submission runs and selected the best. Moreover we received total of 12 different runs with variety of results and many interest unexplored approaches. Table 1 shows the scores and ranks of two teams that made their submission. SELAM (section 5.1) was ranked first,

followed by the rest of the teams with lowest result of 12.94% by TEAMKUBOK (section 5.7) using ensemble of four language models. There is a large gap between the methods especially in top three after which we find the results to spread within 35% F1.

Throughout this shared task we can see the trend to shift more towards language specific pretrained language models. Especially top three systems all employ language models. Meanwhile explainable AI method finds its place inside the rank list with ensemble of simple classifiers. At the same team few teams employed significant preprocessing indeed leading to improvement in results. Besides, we also see that Bi-LSTM methods are still there in the overall list.

6.1 Analysis

Table 1 illustrates the comprehensive outcomes, showcasing the peak achievement of 81.18% by Team SELAM. The subsequent best performance, standing at approximately 61.05%, is notably trailing by roughly 20% from MIT_KEC_NLP, while BYTESIZED LLM lags further behind by a significant margin of 50% in F1 scores. Subsequently, the remaining five systems display closely competitive results. A noticeable trend among the lower-ranking four models reveals a tendency to overestimate (bias) the presence of offensive spans, primarily due to limited generalization. Furthermore, ensemble language models, particularly TEAMKUBOK, exhibit a stark overfitting issue, displaying an F1 score of 12.94%. Notably, we find that in the test set, deliberate inclusion of non-offensive samples aimed to distinctly benchmark the models' performances, has impacted the scores of several models.

7 Conclusion

In this research, we initiated a first shared task focused on identifying offensive spans within code-mixed Kannada-English text. Unlike our previous attempt, we worked with 2k+ social media comments that were annotated to pinpoint offensive sections. Among 45 registered participants, 7 teams submitted their systems. We detailed their approaches in our study and discussed their respective outcomes. Notably, a strategy that employed pretrained language models and explainable AI have shown the best results. Conversely, the LSTM model performed notably worse particularly displaying sensitivity to offensiveness. We've made the data and related information publicly accessible to support future investigations. Looking ahead, our plan involves revisiting the identification of offensive spans within a multitask framework, encompassing various forms of offensiveness alongside the identification of offensive language spans for kannada.

Ethics Statement

In this paper, we discuss the shared task organized around identifying offensive spans in Kannada-English text. To achieve this, we've introduced a novel dataset tailored for both model refinement and diagnostic purposes. Notably, our data collection process didn't involve human participants, eliminating the need for ethical board approval. All datasets utilized in this study are accessible under licenses permitting sharing and redistribution. Our aim is to encourage the development of NLP systems using these datasets, fostering a deeper understanding of offensive spans. This, in turn, could significantly enhance the identification of offensive language across various platforms, carrying considerable societal implications. When appropriately used, these models and datasets hold promise for elevating the quality of discussions on social media channels. However, it's crucial to acknowledge potential biases in the models and the datasets themselves. Our analysis might lean in certain directions due relatively small dataset so used for evaluation. To counteract this to some degree, we have considered offensive content aimed at underrepresented communities, aiming to minimize potential biases and negative repercussions.

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 by Manikandan Ravikiran. 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).

References

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. 2022a. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi. 2022b. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023a. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023b. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Subalalitha Chinnaudayar Navaneethakrishnan, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi, Lavanya Sambath Kumar, and Rahul Ponnusamy. 2023. Findings of shared task on sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 18–21, New York, NY, USA. Association for Computing Machinery.

Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. 2021. Hope speech detection in under-resourced kannada language. *ArXiv*, abs/2108.04616.

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.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Hariharan LekshmiAmmal, Manikandan Ravikiran, and 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*, pages 75–78, Dublin, Ireland. Association for Computational Linguistics.

Ramanujapuram Narasimhacharya. 1990. History of kannada language (readership lectures).

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*.

Ratnavel Rajalakshmi, Mohit More, Bhamatipati Shrikriti, Gitansh Saharan, Hanchate Samyuktha, and Sayantan Nandy. 2022. DLRG@TamilNLP-ACL2022: Offensive span identification in Tamil usingBiLSTM-CRF approach. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 248–253, Dublin, Ireland. Association for Computational Linguistics.

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.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha S, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on offensive span identification fromCode-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 261–270, Dublin, Ireland. Association for Computational Linguistics.

Manikandan Ravikiran, Ananth Ganesh, Anand Kumar M, R Rajalakshmi, and Bharathi Raja Chakravarthi. 2023. Findings of the second shared task on offensive span identification from code-mixed Tamil-English comments. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 52–58, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Krithika Swaminathan, Bharathi B, Gayathri G L, and Hrishik Sampath. 2022. SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/transphobia detection in multiple languages using SVM classifiers and BERT-based transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 239–244, Dublin, Ireland. Association for Computational Linguistics.

TNN. 2010. Indiaspeak: English is our 2nd language: India news - times of india. TOI.