MIT-KEC-NLP@DravidianLangTech-EACL 2024: Offensive Content Detection in Kannada and Kannada-English Mixed Text Using Deep Learning Techniques

Kogilavani Shanmugavadivel[1], **Sowbarnigaa K S**[1], **Mehal Sakthi M S**[1], **Subhadevi K**[1], **Malliga Subramanian**[1]
Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{sowbarnigaak, mehalsakthi}@gmail.com
{subhadevik.22aid}@kongu.edu

Abstract

This study presents a strong methodology for detecting offensive content in multilingual text, with a focus on Kannada and Kannada-English mixed comments. The first step in data preprocessing is to work with a dataset containing Kannada comments, which is backed by Google Translate for Kannada-English translation. Following tokenization and sequence labeling, BIO tags are assigned to indicate the existence and bounds of objectionable spans within the text. On annotated data, a Bidirectional LSTM neural network model is trained and BiLSTM model's macro F1 score is 61.0 in recognizing objectionable content. Data preparation, model architecture definition, and iterative training with Kannada and Kannada-English text are all part of the training process. In a fresh dataset, the trained model accurately predicts offensive spans, emphasizing comments in the aforementioned languages. Predictions that have been recorded and include offensive span indices are organized into a database.

Keywords: Offensive Content Detection, Deep Learning, Bidirectional Long Short-Term Memory(BiLSTM), Natural Language Processing(NLP), (B stands for Beginning, I for Inside, and O for Outside)BIO Tagging.

1 Introduction

The research focused on comments that were combined Kannada and English in order to investigate language processing processes for identifying unacceptable content in multilingual situations. In the first phase, which was called Data Preparation, a rich set of unacceptable spans in Kannada comments were chosen. Subsequently, the dataset was transformed and transliterated by Google Translate, facilitating an extensive examination of patterns in the usage of foul language.

During the labeling process, text was separated into tokens, such as words or subwords, using sequencing and tokenization. The offensive spans were classified as Beginning (B), Inside (I), or Outside (O) by the innovative BIO Tagging approach. This made it possible to assess the offensive context inside the language context more precisely.

A Bidirectional Long Short-Term Memory (BiLSTM) neural network was employed in this study to detect objectionable content in comments that were written in both Kannada and English. By addressing code-mixed text and distinct linguistic patterns, the natural language processing approach improved the detection of such information in a range of linguistic situations.

The research was focused on identifying incorrect information within mixed Kannada and Kannada-English comments, which contributed to the growing interest in Dravidian languages. It recognized the value of applying specialist techniques to successfully negotiate the nuances of diverse linguistic contexts. Using keywords and knowledge from earlier research, the goal was to make NLP applications more inclusive and culturally sensitive.

After training was finished, the computer confidently moved on to new datasets, tokenizing text and correctly identifying problematic spans. This step focused on the concept's adaptability and generalizability to different language situations. The model's work was arranged into a TSV file at the Export Results phase, signifying the conclusion of the ground-breaking inquiry. The offensive content that was discovered required more examination and in-depth analysis, which was made possible by these forecasts that included offensive span indices.

This study addresses the problem of multilingualism in natural language processing by providing a systematic way to find offensive gaps in comments. It combined state-of-the-art technology with a flow diagram that offered insights and strategies for overcoming language obstacles in order to lower language barriers and promote safety and inclusivity in the online environment.

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 146–150

2 Related Works

The multilingual analysis was development (Rajalakshmi et al., 2021) of a Transformer-based technique for identifying inappropriate language in code-mixed Tamil text. However, (Arivazhagan et al., 2020) concentrated on Named Entity Recognition (NER) in Tamil, demonstrating the variety of NLP applications in this language.Further, (Srinivasagan et al., 2014) translation explored sentiment analysis in Tamil using deep learning approaches.

Along with Tamil, there have been significant efforts done in Kannada language processing. Kannada text summarization challenges were the primary focus (Jk and Nn, 2015) , (R et al., 2019) concentrated on Kannada part-of-speech tagging, highlighting the diverse range of Kannada natural language processing (NLP) problems.

Morphological study and applications of Tamil text (Rekha et al., 2010). Additionally, in 2019 (B. et al., 2019) a deep learning-based machine translation system for the English language was developed for the Indian language (Amarappa and Sathyanarayana, 2015) a multinomial Naïve Bayes (MNB) classifier was used for the Kannada Named Entity Recognition and Classification (NERC). In 2019 (Shah and Bakrola, 2019) Indo-European Neural Machine Translation System. Moreover, using Tamil tweets, (Ramanathan et al., 2019) predicted the sentimental reviews of Tamil movies.

These initiatives shows the increasing interest in and need for Dravidian language-specific methodologies. For effective NLP applications, customized methods are needed due to the challenges inherent in these languages, such as code-mixed text and unique linguistic structures.

3 Problem and System Description

The dataset was shown in Table 1. The objective was to determine, given the supplied text, the range of spans that related to offending material. In the sample, the text "Tik tok Shata adds to offensiveness" matched a character offset between 8 and 13. As part of the shared effort on offensive span identification, the details of an offensive span annotation dataset were made available [1].

3.1 Dataset Description

One file had span annotations, while the other contained a shared task dataset that was open to the public. There were 444 unannotated cases in the testing dataset and 1800 offensive span samples in the training dataset.

Sample Text	Spans
Tik tok shata	[8,9,10,11,12,13]
Ade old same story	[0,1,2,3,4,5,6,7]
Bindu gowda lofer avlu	[12,13,14,15,16,17,18]
Eppa all fake	[5,6,7,8,9,10,11,12,13]

Table 1: Example of the dataset

4 Development Pipeline

The proposed system pipeline employed in this study was shown in Figure 1. Our pipeline consisted of four modules: (a) preparation of the data; (b) tokenization and sequence labeling; (c) training of the model; and (d) predictions on test data. After that, the predictions were exported and stored in a TSV file. This description was correct in every way.

4.1 Data Preparation Phase

Standardizing a group of unwanted spans in Kannada remarks was the main objective of the project's data preparation. Standardizing the offensive span annotations required pre-processing in order to give consistency for further analysis. The code sample demonstrated a practical technique for managing and enhancing the raw dataset, laying the groundwork for language normalization. With the use of language recognition and translation algorithms, the project attempted to give a consistent representation of text data, especially when converting Kannada comments into English. This stage was important because it laid the groundwork for an offensive span identification strategy that was both coherent and linguistically consistent. Making the dataset more appropriate for additional machine learning model training and evaluation was the ultimate goal.

4.2 Tokenization and Sequence Labeling

In order to assess Kannada comments with problematic spans during the Tokenization and Sequence Labeling step, the project employs a thorough technique. Two objectives are involved: first, tokenize the text into meaningful units like as words or subwords; second, assign a BIO tag to each token designating where it falls within the offensive spans (B stands for Beginning, I for Inside,

[1](Ravikiran et al., 2022)

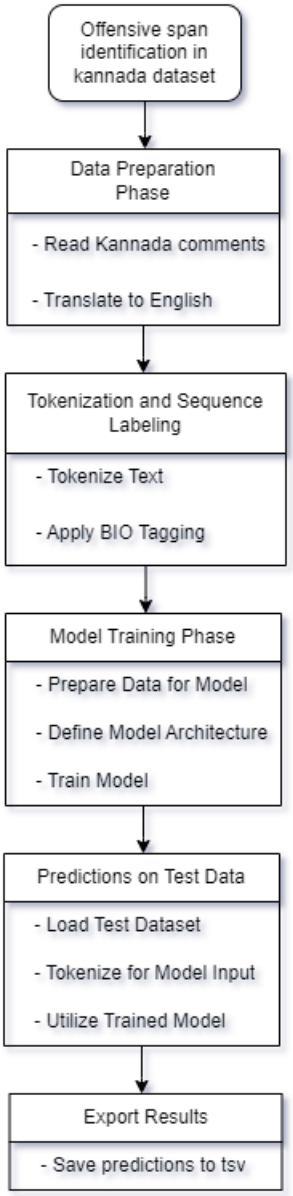

Figure 1: Proposed system pipeline

and O for Outside). This step involves creating a meticulously planned Python script that uses a bespoke function to carry out the BIO tagging strategy and the NLTK package for word tokenization. Using the dataset, the code functions flawlessly, guaranteeing that every token is categorized correctly based on how close it is to potentially dangerous text spans. This initial stage, which enables the creation of annotated papers, is essential for later phases.

4.3 Model Training phase

During the Model Training Phase, the project moves forward from the preliminary stages to the deployment of a Bidirectional Long Short-Term Memory (BiLSTM) neural network with the purpose of detecting offensive spans in Kannada comments. The main goal is to provide a numerical representation of the tokenized text data that can be fed into the model. Using a tokenizer, the textual content is transformed into numerical sequences. For model compatibility, the relevant BIO tags denoting offending span labels are one-hot encoded and encoded. The dataset is cleverly divided into training and testing subsets to facilitate in-depth model evaluation, guaranteeing strong performance on both known and unknown data. The definition of the neural network architecture, which has an embedding layer for word recognition, is crucial to this step.

The significance of this step in the research is that it marks the transition from data preparation to the application of sophisticated neural network architectures for offensive span recognition in Kannada literature. The careful coordination of the tokenizer makes sure that the model absorbs and comprehends the nuances of the language, and the encoding of BIO tags that follows establishes the basis for the neural network's ability to recognize and anticipate problematic spans. With the aid of the previously indicated iterative training technique that makes use of the BiLSTM architecture and fine-tuning parameters in Figure 2, the model is able to capture complex patterns in the Kannada comments dataset. The project's potential impact on multilingual content moderation is highlighted by its alignment with wider objectives and incorporation of deep learning and natural language processing technology.The adaptability of the model architecture makes it a flexible tool for resolving the problems caused by offensive language, enhancing the capacity to identify and reduce instances of hazardous content and promoting a safer online environment.

4.4 Predictions On Test Data

During the Predictions on Test Data phase, the study extends its offensive span identification capabilities to a fresh dataset of Kannada or Kannada-English statements. This step includes loading the test dataset, tokenizing the comments using the pre-established tokenizer, and padding the sequences to the maximum length allowed by the model. The trained BiLSTM model incorporates intricate patterns from the linguistic environment to provide offensive span predictions. In order to identify violating spans, the post-processing stage evaluates the model's predictions; the findings are then recorded in a new column within the test dataset. The project's flexibility in adjusting to new data is highlighted in this phase, showcasing its potential for reliable and sensitive problematic material identification in multilingual settings.

5 Results

The goal of the study was to create a technique for finding unacceptable stretches in multilingual literature by concentrating on mixed Kannada and English comments. Google Translate was used to train a Bidirectional LSTM neural network model for translation, tokenization, and sequence labeling on annotated data. Table 2 displays the BiLSTM model's macro F1 score of 61.0 and accuracy of 0.9759, demonstrating its capacity to identify problematic content. Next, using a Test dataset with an emphasis on multilingual comments, the algorithm predicted problematic spans.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 128)]             0

 embedding (Embedding)       (None, 128, 100)          500000

 dropout (Dropout)           (None, 128, 100)          0

 bidirectional (Bidirection  (None, 128, 256)          234496
 al)

 time_distributed (TimeDist  (None, 128, 3)            771
 ributed)

=================================================================
Total params: 735267 (2.80 MB)
Trainable params: 735267 (2.80 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 2: BiLSTM Model Architecture

E	TL	TA	VL	VA
1	0.2144	94.18	0.0807	96.36
2	0.0763	96.36	0.0786	96.44
3	0.0705	96.73	0.0794	96.22
4	0.0651	97.20	0.0782	96.41
5	0.0594	97.59	0.0779	96.48

Table 2: BiLSTM Model Accuracy

Epoch (E)
Training Loss (TL)
Training Accuracy (TA)
Validation Loss (VL)
Validation Accuracy (VA)

6 Conclusion

A multilingual approach was employed in the study to pinpoint problematic spans in Kannada and Kannada-English literature. The Bidirectional LSTM model's macro F1 score of 61.0 was made possible by extensive pre-processing of the data. Because of its flexibility, the model was able to be used with new datasets to predict offensive spans in situations with a mixture of Kannada and English as well as pure Kannada. An important advancement in natural language processing was made by this work, which tackled the issue of culturally acceptable material filtering in online communication.

References

S. Amarappa and S. Sathyanarayana. 2015. Kannada named entity recognition and classification (nerc) based on multinomial naïve bayes (mnb) classifier. *International Journal on Natural Language Computing*, 4.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Premjith B., M. Kumar, and Soman Kp. 2019. Neural machine translation system for english to indian language translation using mtil parallel corpus: Special issue on natural language processing. *Journal of Intelligent Systems*, 28.

Geetha Jk and Deepamala Nn. 2015. Kannada text summarization using latent semantic analysis. pages 1508–1512.

Swaroop R, Rakshit S, Shriram Hegde, and Sourabh U. 2019. Parts of speech tagging for kannada. pages 28–31.

Ratnavel Rajalakshmi, Yashwant Reddy, and Lokesh Kumar. 2021. DLRG@DravidianLangTech-EACL2021: Transformer based approachfor offensive language identification on code-mixed Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 357–362, Kyiv. Association for Computational Linguistics.

Vallikannu Ramanathan, Meyyappan Thirunavukkarasu, and S.M. Thamarai. 2019. Predicting tamil movies sentimental reviews using tamil tweets. *Journal of Computer Science*, 15:1638–1647.

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.

R. Rekha, M. Kumar, V. Dhanalakshmi, Soman Kp, and Rajendran Sankaravelayuthan. 2010. A novel approach to morphological generator for tamil. pages 249–251.

Parth Shah and Vishvajit Bakrola. 2019. Neural machine translation system of indic languages - an attention based approach. pages 1–5.

K. Srinivasagan, S. Suganthi, and N. Jeyashenbagavalli. 2014. An automated system for tamil named entity recognition using hybrid approach. pages 435–439.