# Detection of Religious Hate Speech During Elections in Karnataka

**MSVPJ Sathvik**
Raickers AI
Hyderabad
Telangana, India
msvpjsathvik@gmail.com

**Raj Sonani**
Cornell University
Ithaca
New York, USA
sonaniraj@gmail.com

**Ravi Teja Potla**
Slalom
Houston
Texas, USA
raviteja.potla@gmail.com

## Abstract

We propose a novel dataset for detecting religious hate speech in the context of elections in Karnataka, with a particular focus on Kannada and Kannada-English code-mixed text. The data was collected during the Karnataka state elections and includes 3,000 labeled samples that reflect various forms of online discourse related to religion. This dataset aims to address the growing concern of religious intolerance and hate speech during election periods, it's a dataset of multilingual, code-mixed language. To evaluate the effectiveness of this dataset, we benchmarked it using the latest state-of-the-art algorithms. We achieved accuracy of 78.61%.

## 1 Introduction

Religious tensions between Hindus and Muslims have been a sensitive issue in India, often increasing during elections(Pradhan and Mehta, 2019). In Karnataka, some political parties have been accused of spreading religious hatred to gain votes. This has led to violent incidents, communal clashes, and even loss of lives. Social media plays a major role in amplifying hate speech, as people use these platforms to express strong opinions, sometimes leading to misinformation, targeted attacks, and communal propaganda(Kumar and Gupta, 2020).

During elections, the amount of Hindu-Muslim hate speech on social media rises sharply(Narayanan et al., 2019). Many users post content that provokes religious sentiments, causing division and unrest. Despite social media companies trying to control harmful content, their existing detection systems struggle with regional languages and code-mixed text. Kannada and Kannada-English code-mixed speech make it even harder for AI models to identify hate speech accurately.

**Motivation:**Political campaigns often intensify religious, ethnic, and ideological divisions, leading to social unrest and real-world violence. Elections are a time when public opinion is highly influenced, and the spread of hate speech on social media can manipulate voters, incite communal tensions, and weaken democratic values. Unchecked hate speech can lead to misinformation, voter suppression, targeted harassment, and even violent clashes between communities. In regions like Karnataka, where religious polarization is sometimes exploited for political gains, identifying and controlling hate speech can prevent riots, protect vulnerable communities, and ensure fair and peaceful elections.

How can we detect hate speech? that too for state elections of Karnataka? Kannada is a low-resource language, and social media conversations often involve Kannada-English code-mixed text. While AI can help predict hate speech, the absence of a reliable dataset makes it difficult to develop and evaluate effective models. Although advanced AI models exist, there is no clear understanding of which model performs best for this specific task. To address this gap, we propose a novel dataset tailored for Hindu-Muslim hate speech detection in Kannada and Kannada-English text. Additionally, we benchmark this dataset using state-of-the-art large language models (LLMs) to compare their effectiveness, providing valuable insights into the most suitable AI model for detecting hate speech in regional and code-mixed languages.

Our key contributions are as follows:

1. As of our knowledge we are the first to develop a dataset for the detection of the hate speech on religious issues during elections.

2. We have benchmarked the dataset with SOTA models and presented the results and comparison.

## 2 Related Work

There are several research works focused on Kannada hate speech detection. Chakravarthi et al. (2021) introduced the Dravidian CodeMix dataset,

*Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 45–49
May 3, 2025 ©2025 Association for Computational Linguistics

which includes Kannada-English code-mixed text, allowing researchers to develop language models capable of handling mixed-script data. However, the dataset primarily focuses on sentiment classification rather than explicit hate speech detection. Patil et al. (2022) created a Kannada hate speech dataset from social media posts, demonstrating that transformer-based models like mBERT outperform traditional models like SVM and LSTMs in this task.Suryawanshi et al. (2020) built a dataset for Tamil-English code-mixed sentiment analysis, which provided valuable insights into handling mixed-language text. Extending such techniques to Kannada-English code-mixed data is crucial for improving detection models. Ramesh et al. (2023) proposed a hybrid deep learning model combining LSTMs with attention mechanisms for detecting hate speech in Tamil-English and Telugu-English code-mixed tweets. Their findings indicate that context-aware embeddings such as IndicBERT significantly improve performance,

Risch et al. (2021) explored offensive language detection using multilingual transformer models, concluding that fine-tuning models on code-mixed and regional datasets significantly enhances performance. This aligns with recent efforts to apply pre-trained multilingual models like XLM-R and IndicBERT for Kannada hate speech detection.

# 3 Methodology

## 3.1 Data Collection and Annotation

For this study, we collected data from Twitter using the Twitter API, ensuring that the dataset includes real-time social media conversations in both Kannada and Kannada-English code-mixed text. The tweets were filtered based on keywords, hashtags, and engagement metrics to capture a diverse set of opinions and discussions related to Hindu-Muslim hate speech.

To ensure high-quality annotations, we employed a team of three native Kannada-speaking annotators who were responsible for labeling the dataset. Each annotator was provided with an Excel sheet containing the collected data, and they were instructed to label each text instance as either hate speech on religion (1) or non-hate speech (0). To maintain annotation consistency and reliability, each data point was labeled by at least two annotators, ensuring that disagreements could be reviewed and resolved.

Table 1: Statistics of the Dataset

| Metrics | label 0 | label 1 | Total/Overall |
|---|---|---|---|
| Data Size | 1542 | 1633 | 3175 |
| Number of Words | 38103 | 42294 | 80397 |
| Words per data point | 24.71 | 25.90 | 25.32 |

To measure the Inter-Annotator Agreement (IAA), we calculated pairwise agreement scores between annotators. The agreement scores were as follows: I(1,2) = 87.2%, I(2,3) = 89.6%, and I(1,3) = 86.1%, demonstrating strong agreement among the annotators. These high agreement values indicate that the dataset is well-annotated and reliable, making it suitable for training and evaluating AI models for religious hate speech detection.

### Data Annotation Guidelines

To ensure high-quality labeling, three native Kannada-speaking annotators followed strict annotation guidelines. Each text sample was labeled based on the presence of religious hate speech. The annotators were given with the data points and were given to label each and every data point accordingly.

**Labeling Criteria:**

- **Hate Speech on Religion (1)**: Texts containing explicit or implicit hate against a religion (Hindu or Muslim), including:
  - Direct insults, slurs, or abusive language targeting a religious group.
  - Calls for violence, discrimination, or exclusion based on religion.
  - Misinformation or stereotypes that incite hatred.
  - Provocative statements intending to cause communal disharmony.
- **Not Hate Speech (0)**: Texts that do not contain religious hate speech, including:
  - Political discussions without religious targeting.
  - Opinions, sarcasm, or humor that do not promote hatred.
  - Discussions about religion without intent to incite violence.

If a sample was unclear, annotators referred to context and intent. Disagreements between two

Table 2: Overview of the dataset

| Text | Label[0/1] |
|---|---|
| Deshada Hindu matthu Muslim galu yendu ondu ge barolla avaru olage namma rashtrada vairi Muslim galu namma desh dushman | 1 |
| Muslim galu hamesha Hindu mele droh madtare ivaru namma desh ke khatara ivaranna olage vittu namma rashtrada surakshya dodda aapathige nuggi bidutte | 1 |
| Kannada naadu belongs to all religions we must promote peace and harmony society flourishes when we stand together respecting all faiths and backgrounds equally | 0 |
| Muslim galu matthu Hindu galu seridre matra naadu balutte, yava ondu dharmavannu dushisuvudu sari alla, namma sahane namma balavagi iruvudu. | 0 |
| Religious harmony is important for a strong nation, Hindus and Muslims must coexist peacefully, respecting each other's traditions, ensuring equality, and spreading love, not hate. | 0 |
| Muslims get out of our country this land belongs to Hindus they are a threat to our nation they should leave immediately we cannot trust them anymore | 1 |

annotators were resolved through discussions. Misinformation leading to potential hate was labeled as hate speech.

### 3.2 Analysis

Table 1 presents the statistical analysis of the dataset, 3,175 data points with 1,542 labeled as non-hate speech (0) and 1,633 as hate speech (1), ensuring a balanced distribution for training AI models. The dataset contains 80,397 words, with hate speech samples contributing 42,294 words and non-hate speech 38,103 words (47.4%), hence the dataset is balanced. Table 2 represents the overview of the dataset which displays few examples from the proposed dataset.

### 3.3 Baselines

We conducted experiments on the proposed dataset using various pre-trained language models and large language models (LLMs), including: (i) GPT-4o(OpenAI, 2023), (ii) Gemini(DeepMind, 2023), (iii) LLaMA 3(Touvron et al., 2023), (iv) Kannada-BERT(Khanuja et al., 2021), (v) IndicBERT(Kakwani et al., 2020), and (vi) Multilingual-BERT(Devlin et al., 2018).

For baseline experimentation, we implemented the few-shot prompting technique, where eight training samples were selected from the dataset. These examples were provided as context to guide the LLMs in classifying the input text.

The dataset was randomly split into 80% for training and 20% for testing. Pre-trained models were fine-tuned for five epochs with a learning rate of 0.01, while other parameters were kept at default settings. GPT variants were fine-tuned using the OpenAI API key, while BERT-based models were fine-tuned on Google Colab (free GPU version). Few-shot prompting was executed without GPU on Google Colab, whereas LLaMA models were fine-tuned using NVIDIA GPUs with CUDA support.

Table 3: Test results

| Model | Precision | Recall | Accuracy |
|---|---|---|---|
| LSTM | 58.42 | 60.19 | 59.32 |
| Bi-LSTM | 62.71 | 67.02 | 65.14 |
| CNN Bi LSTM | 64.12 | 70.15 | 68.22 |
| m-BERT | 68.35 | 72.14 | 71.48 |
| Kannada BERT | 70.92 | 75.34 | 73.59 |
| Indic BERT | 72.54 | 76.92 | 75.36 |
| Gemini 2.0 | 75.18 | 79.11 | 77.51 |
| LLAMA 3 | 76.24 | 77.61 | 77.02 |
| GPT-4o | 78.36 | 79.81 | 78.61 |

## 4 Experimental Results and Discussion

Table 3 presents the test results of several models across precision, recall, and accuracy. In general, older models like LSTM, Bi-LSTM, and CNN Bi LSTM have lower accuracy compared to the more advanced models. Among these, CNN Bi LSTM achieves the highest accuracy but still falls short of the more recent models.

When it comes to transformer-based models, m-BERT shows a noticeable improvement over the earlier models. Its accuracy is higher than the LSTM-based models, indicating that transformer architectures tend to perform better for the given task. Kannada BERT and Indic BERT further outperform m-BERT, with Indic BERT reaching the highest accuracy among the BERT-based models.

The highest accuracy scores are seen in the latest models Gemini 2.0, LLAMA 3, and GPT-4o.

These models significantly surpass the accuracy of the previous ones, with GPT-4o achieving the highest accuracy overall. This suggests that the most recent developments in transformer-based models, particularly those like GPT-4o, are highly effective for the task and represent a significant leap in performance.

**Real time usecases:**

The models trained on this dataset can be applied to various real-time scenarios to help manage and control hate speech online, particularly during sensitive times like elections.

1. **Election Commission:** The model can assist the Election Commission in identifying and removing posts that spread religious hate during election periods. This helps maintain a peaceful and unbiased environment, ensuring that elections remain fair and free from divisive content. For example, if a social media post targets a particular religious group with inflammatory remarks, the system can flag it for removal, promoting a more respectful and impartial electoral process.

2. **Social Media Platforms:** Social media companies can use this model to monitor and regulate content that may negatively influence teenagers and children. As young people are more susceptible to harmful content, the system can help identify and restrict posts that spread religious intolerance or hate speech. For instance, if a user posts a hate-filled comment targeting a minority religious group, the platform could use the model to detect it and either warn the user or remove the post to protect younger audiences.

3. **Government Agencies and Law Enforcement:** The model could be used by government bodies or law enforcement agencies to track and prevent the spread of hate speech across public forums, particularly during sensitive times like political unrest or elections. By detecting harmful content early, agencies can take proactive measures to prevent violent outbreaks or social division. For example, it could help identify extremist posts before they escalate into offline actions.

4. **Media and News Outlets:** News organizations could use this model to monitor and manage the spread of biased or harmful religious narratives in the media, especially during election seasons when the risk of divisive rhetoric is higher. By detecting inflammatory language, media outlets can avoid amplifying hate speech or biased content in their reports, ensuring a more balanced and responsible approach to news coverage.

5. **Educational Institutions:** Schools and universities can use the model to monitor online discussions, forums, or social media groups where students interact. This helps maintain a safe and inclusive environment by identifying and addressing harmful content related to religion, fostering respectful discourse among young people. For example, a student might post hateful comments about another religion in an online forum, and the system can flag it for review by administrators.

## 5 Conclusion and Future Work

The proposed hate speech detection model has significant potential to mitigate religious hate speech in real-time, particularly during elections and on social media platforms. By using advanced natural language processing (NLP) techniques, this model helps various stakeholders, including election commissions, social media companies, law enforcement agencies, and educational institutions, to identify and control harmful speech. The real-world applications discussed demonstrate the necessity of such models in maintaining a fair, unbiased, and safer online environment. Future work includes developing advanced algorithms for detecting hate speech in Kannada and other Dravidian languages while also expanding the scope to explore related issues such as caste politics and other forms of social discrimination.

## Limitations

This study primarily focuses on text-based hate speech detection, which means it does not account for other modalities such as images, videos, or audio, where hate speech can also be prevalent. Additionally, the scope of this research is centered around elections, providing valuable insights into political discourse but not extending to other important social contexts such as caste-based discrimination or general communal hate speech outside election periods. However, this focused approach allows for a deeper and more precise understanding of election-related hate speech, laying the groundwork for future research to expand into multimodal analysis and broader societal issues.

## Ethics Statment

We strongly oppose any potential misuse of this dataset, such as training models to generate hate speech or promote religious discrimination. Our sole aim is to detect religious hate speech during elections and help mitigate its spread on social media, fostering a safer and more inclusive online environment.

## References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Jishnu Jose, Thomas Mandl, Mitesh M. Kumar, and Elizabeth Sherly. 2021. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. In *Proceedings of the 13th International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 714–722.

Google DeepMind. 2023. Gemini 2.0: Scaling and advances in large language models. *Google AI Blog*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhruva Kakwani, Anoop Kunchukuttan, Shiva Meena Golla, Pushpak Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Simran Khanuja, Sandipan Dandapat, Ritesh Kumar, Sunayana Sitaram, K. P. Soman, and Anup Kumar. 2021. Mahanlp: Towards indic language understanding using bert models for hindi, marathi, and kannada. *arXiv preprint arXiv:2106.07469*.

Rajesh Kumar and Anjali Gupta. 2020. The role of social media in spreading religious hate speech during elections in india. *International Journal of Communication and Society*, 5(2):220–235.

Vidya Narayanan, Vladimir Barash, Bence Kollanyi, Lisa-Maria Neudert, and Philip Howard. 2019. News and information overload in the indian elections: The case of whatsapp. *Computational Propaganda Project, Oxford Internet Institute*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aniket Patil, Raghav Sharma, and Anil Kumar. 2022. Kannada hate speech detection using transformer-based models. *Journal of Computational Linguistics and Artificial Intelligence*, 8(2):102–118.

Pratyush Pradhan and Sanjay Mehta. 2019. Religious polarization and electoral politics in india. *Indian Journal of Political Science*, 80(3):345–362.

S. Ramesh, V. Kumar, P. Srinivasan, and A. Iyer. 2023. Hybrid deep learning model for hate speech detection in tamil-english and telugu-english code-mixed text. *International Journal of Computational Linguistics and NLP*, 10(2):45–62.

Julian Risch, Anke Stoll, and Ralf Krestel. 2021. Offensive language detection exploiting multilingual transformer models. *Natural Language Processing Journal*, 35(4):567–589.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Karun Arora, Elizabeth Sherly, and John P. McCrae. 2020. A dataset for sentiment analysis of code-mixed tamil-english text. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 2459–2468.

Hugo Touvron, Louis Martin, Kevin Stone, Pierre Albert, Amjad Almahairi, Ross Taylor, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.