# Zavira@DravidianLangTech 2024:Telugu hate speech detection using LSTM

**Z. Ahani, M. Shahiki Tash, M. T. Zamir, I. Gelbukh** and **A. Gelbukh**
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC)
Corresponding: z.ahani2023@cic.ipn.mx

## Abstract

Hate speech is communication, often oral or written, that incites, stigmatizes, or incites violence or prejudice against individuals or groups based on characteristics such as race, religion, ethnicity, gender, sexual orientation, or other protected characteristics. This usually involves expressions of hostility, contempt, or prejudice and can have harmful social consequences. Among the broader social landscape, an important problem and challenge facing the medical community is related to the impact of people's verbal expression. These words have a significant and immediate effect on human behavior and psyche. Repeating such phrases can even lead to depression and social isolation. In an attempt to identify and classify these Telugu text samples in the social media domain, our research LSTM and the findings of this experiment are summarized in this paper, in which out of 27 participants, we obtained 8th place with an F1 score of 0.68.

## 1 Introduction

While hate speech (HS) legislation varies among different countries, it is generally conceptualized as encompassing expressions of hostility or derogation directed at an individual or a group based on attributes such as race, color, national origin, sex, disability, religion, or sexual orientation (Nockleby, 2000; Jahan and Oussalah, 2023).

On platforms such as Twitter, Facebook, and various other social media outlets, hateful comments manifest as expressions containing abusive language directed towards individuals (including cyber-bullying, politicians, celebrities, or products) or specific groups (such as countries, the LGBT community, religions, genders, organizations, etc) (Badjatiya et al., 2017)

Numerous intricate challenges are currently evident in applications related to speech, vision, and text, all aimed at enhancing accuracy. The pioneering work of (Badjatiya et al., 2017). In 2017 marks the initial exploration of neural architectures for detecting hate speech. The advancement of natural language processing (NLP) (Bade, 2021) technology has spurred considerable investigation into the automated detection of textual hate speech in recent years. Notable competitions such as SemEval-2019 (Zampieri et al., 2019) and 2020 (Zampieri et al., 2020), as well as GermEval-2018 (Wiegand et al., 2018), have organized diverse events aimed at seeking improved solutions for automated hate speech detection. In response, researchers have compiled extensive datasets from various sources, fostering significant progress in the field. Numerous studies have addressed hate speech in multiple non-English languages and online communities, prompting exploration and comparison of different processing pipelines. This includes the examination of feature sets and Machine Learning (ML) methods (Tash et al., 2022; Kanta and Sidorov, 2023), encompassing supervised, unsupervised, and semi-supervised approaches, as well as various classification algorithms such as Naive Bayes, Logistic Regression (LR), Convolutional Neural Network (CNN) (Tash et al., 2023; Shahiki-Tash et al., 2023b), Long Short-Term Memory (LSTM), BERT deep learning(Yigezu et al., 2022) architectures, among others. The pervasive issue of abusive language is both common and troubling. Offensive language takes many forms, depending on the target group and the specific target, such as hate speech, cyberbullying, adult content, trolling, abuse, racism, or profanity.

In recent advancements, transformer-based models (Tonja et al., 2022), such as BERT, have significantly impacted the detection and understanding of hate speech. Hate speech, a particularly alarming category of abusive language, involves the intentional intimidation of a target group or individual with the intent of causing harm, violence, or social disruption (Husain and Uzuner, 2021; Khan et al., 2022a)



*Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 107–112

So there are subtle distinctions between different types of offensive language. The targeting of LGBT+ people with hate speech is a deep-rooted issue with far-reaching consequences, including the potential for substance abuse disorders (Shahiki-Tash et al., 2023a) and racism (Badjatiya et al., 2017). The rest of the paper is organized as follows: the related work and methodology are discussed in Section 2 and 3 respectively followed by results in Section 4.

## 2 Related work

Balouchzahi et al. address the ongoing challenge of hate speech (HS) by emphasizing the limitations of conventional identification and blocking methods. Thay advocate the development of systems that are capable of not only identifying but also profiling HS content contaminants. Using a vote classifier (VC) contributes to the hate speech broadcaster detection task organized by PAN 2021 (Bevendorff et al., 2021), which focuses on the profiles of HS broadcasters in English and Spanish on Twitter. The proposed model uses a combination of traditional character and word n-gram along with syntactic n-grams as features for classification. Using a support vector machine (SVM), logistic regression (LR) and random forest (RF) vote classifier, the models achieve commendable accuracies of 73% and 83% for English and Spanish, respectively.

In the BiCHAT (Khan et al., 2022a) research, an innovative deep learning(Ahani et al., 2024) model, combining BiLSTM with deep CNN and hierarchical attention, is employed to acquire tweet representations for the detection of hate speech. The proposed model undergoes a process of mining, training, and evaluation using three benchmark datasets from Twitter. These datasets include HD1, introduced by (Founta et al., 2018; Bade and Afaro, 2018), HD2, derived from the Kaggle [1] competition dataset, and HD3 with statistics presented in Table 1, provided by (Davidson et al., 2017; Bade and Seid, 2018). The F1-score outcomes (HD1=0.88, HD2=0.91, and HD3=0.75) demonstrate superior performance compared to the State-of-the-Art (SOTA) methods (Khan et al., 2022b; Roy et al., 2020; Ding et al., 2019).

In the publication (Badjatiya et al., 2017), an examination was conducted on 16,000 tweets employing three neural network models (CNN and BOWL, LSTM) and various methodologies, including GBDT, TF_IDF, and Random Embedding. The dataset originates from the (Waseem and Hovy, 2016). The study demonstrated that combining embeddings acquired from deep neural network models with gradient-boosted decision trees yields the highest accuracy values. Specifically, the combination of LSTM+Random Embedding+GBDT achieved an F1-score of 0.930.

In this study(Waseem and Hovy, 2016), the method is based on a dataset of 16,000 tweets collected by (Waseem and Hovy, 2016) and colleagues. This dataset, which includes a total of 136,052 tweets, was annotated by the researchers, and 16,914 tweets were specifically flagged. Of these, 3,383 tweets containing sexual content were identified, originating from 613 users. Additionally, 1,972 tweets were flagged for racist content and contributed by 9 users, while the remaining 11,559 tweets were deemed non-sexist or racist. The analysis of hate speech comments included a thorough review of the features used, with the aim of determining those that yielded the most effective detection performance. Notably, examination of the features influencing hate speech recognition in the dataset revealed that, despite potential variations in geographic distribution and word length, these factors did not consistently improve performance and rarely outperformed personality-level features. An exception to this trend can be seen with gender-related characteristics, as detailed in Table 2.

## 3 Methodology

in this section, we summarize the data set used in this task and the proposed methodology in detail. LSTM networks prove advantageous in binary text classification tasks, such as hate speech detection, due to their inherent ability to capture contextual dependencies and long-range dependencies in sequential data. Imbalanced datasets, on the other hand, might lead the model to be skewed towards the majority class, potentially hindering its performance in identifying instances of the minority class, such as hate speech, and affecting overall classification accuracy.

### 3.1 Dataset

The dataset, generously provided by Hold Telugu for the Telugu language, consists of two separate datasets for educational purposes. The first dataset contains 4000 tweets for training, while the second

[1] www.kaggle.com

Table 1: Statistics of the datasets

| Datasets | Hate tweet | Normal tweet | Total |
|---|---|---|---|
| HD1 (Relatively balanced) | 2615 | 5385 | 8000 |
| HD2 (Unbalanced) | 1421 | 10579 | 12000 |
| HD3 (Unbalanced) | 1430 | 4162 | 5592 |

Table 2: F1 achieved by using different features sets

| | char n-grams | +gender | +gender +loc | word n-grams |
|---|---|---|---|---|
| F1 | 73.89 | 73.93 | 73.62 | 64.58 |

dataset contains 500 tweets for testing (B et al.; Priyadharshini et al., 2023)

Table 3: Data set samples

| Tweets | Label |
|---|---|
| Adhi Show na lanjala kompana | Hate |
| Papam erry flower ayipoindu | Hate |
| Valla dhagara bochu vunttundi | Hate |
| West Godavari lo adii jarigindhi | Non-hate |
| Venakala unnonni adugu cheptadu | Non-hate |
| turning thisukuna vadihi | Non-hate |

### 3.2 Embedding Layer

The model begins with an embedding layer, a fundamental component in natural language processing tasks. The ‘Embedding‘ layer is responsible for converting the input text data into a dense vector representation. In this case, each word in the vocabulary is represented as a vector of 32 dimensions ("embedding_vector_length"). This vector representation allows the model to capture semantic relationships between words and enables better understanding of the textual data.

### 3.3 LSTM Layer

Following the embedding layer, the model incorporates an LSTM layer. LSTMs are a type of recurrent neural network (RNN) designed to address the vanishing gradient problem, making them effective for sequence modeling tasks. The LSTM layer with 100 units captures long-range dependencies and temporal patterns in the input sequences. The ‘dropout‘ and ‘recurrent_dropout‘ parameters are introduced to mitigate overfitting by randomly dropping connections during training.

### 3.4 Dense Layer and Sigmoid Activation

The LSTM layer is followed by a dense layer with a single output unit. This dense layer acts as a classifier for binary sentiment classification, with a sigmoid activation function applied to produce probabilities. The sigmoid activation function is well-suited for binary classification tasks as it squashes the output values between 0 and 1, representing the likelihood of the input belonging to the positive class (hate speech) or negative class (non-hate speech).

### 3.5 Model Loading and Compilation

The model is then loaded with pre-trained weights saved during training, specifically the weights that achieved the best performance on the validation set. This practice ensures that the model used for evaluation is the one that demonstrated the highest generalization ability during training.

The model is compiled using binary cross-entropy loss, which is suitable for binary classification problems, and the Adam optimizer, a popular choice for training neural networks. The evaluation metrics include loss and accuracy, providing insights into the model’s performance on the test data.

### 3.6 Evaluation on Test Data

Finally, the model is evaluated on a separate test dataset ("X_test" and "y_test"). The "model.evaluate" method computes the loss and accuracy of the model on the test data, providing a quantitative measure of its generalization performance. The obtained accuracy is then printed as a percentage, offering a clear indication of how well the model is able to classify hate speech in unseen textual data.

In summary, this methodology section describes the architecture and training process of an LSTM-based hate speech detection model, emphasizing the role of embedding, LSTM, and dense layers in capturing intricate patterns in text data. The model’s evaluation on a distinct test set ensures a

robust assessment of its real-world performance.

## 4 Result

During the sharing task competition that focused on detecting hate and offensive language in Telugu mixed code text, our main goal was to determine the F1-score for the given data set. Using the previously trained LSTM model, we fed the entire test data into the model and obtained a prediction that yielded significant results. In a single performance evaluation, we scored an admirable 0.68%, placing 8th out of 27 participating teams. For an overview of the results achieved by all participating teams, please refer to Table 4, which provides a detailed insight into the performance metrics and points earned by each participant in the competition.

Table 4: Results of the participants in Telugu Hate speech

| Team | Run | F1-score (macro) | Rank |
|---|---|---|---|
| **Sandalphon** | 1 | 0.7711 | 1 |
| **Selam** | 2 | 0.7711 | 1 |
| **Kubapok** | 1 | 0.7431 | 3 |
| **DLRG1** | 1 | 0.7101 | 4 |
| **DLRG** | 1 | 0.7041 | 5 |
| **CUET_Binary** | 2 | 0.7013 | 6 |
| **CUET_OpenNLP** | 1 | 0.6878 | 7 |
| **Zavira** | 1 | 0.6819 | 8 |
| **IIITDWD-zk_lstm** | 2 | 0.6739 | 9 |
| **lemlem** | 1 | 0.6708 | 10 |
| **Mizan** | 1 | 0.6616 | 11 |
| **byteSizedLLM** | 1 | 0.6609 | 12 |
| **pinealai** | 1 | 0.6575 | 13 |
| **IIITDWD_SVC** | 2 | 0.6565 | 14 |
| **MUCS** | 3 | 0.6501 | 15 |
| **Lemlem-eyob** | 2 | 0.6498 | 16 |
| **Tewodros** | 2 | 0.6498 | 16 |
| **Fida** | 2 | 0.6369 | 18 |
| **Lidoma** | 1 | 0.6151 | 19 |
| **MasonTigers** | 1 | 0.5621 | 20 |
| **Habesha** | 1 | 0.5284 | 21 |
| **MasonTigers** | 1 | 0.4959 | 22 |
| **CUET_DASH** | 3 | 0.4956 | 23 |
| **Fango** | 1 | 0.4921 | 24 |
| **Tayyab** | 1 | 0.4653 | 25 |

## 5 limitations

1. The research grapples with a limitation arising from the exclusion of hyperparameter tuning in the experimental setup. Optimizing hyperparameter configurations is pivotal for refining the performance of machine learning models, and the absence of such optimization in our experiments may impact the overall efficacy of our approach.

2. Another constraint in our methodology arises from the absence of experiments specifically tailored to address the issue of imbalanced datasets. Tasks related to hate speech detection commonly face challenges with imbalances between the instances of different classes. Exploring strategies like oversampling, undersampling, or employing specialized algorithms for imbalanced datasets could be considered to enhance the model's capability in handling such distribution challenges.

## 6 Conclusion

Hate speech that incites violence or prejudice against individuals or groups based on different characteristics is an important challenge in contemporary society. The damaging effects of such expressions, including hostility and prejudice, go beyond immediate social consequences and can contribute to deep psychological effects such as depression and social isolation.

This research deals with the important issue of hate speech in the context of Telugu mixed code text on social media platforms. Using Natural Language Processing (NLP), specifically using short-term memory (LSTM) neural networks, we aimed to identify and classify hate speech in Telugu text samples. In a competitive environment of 27 participants, our LSTM-based model achieved eighth place with an F1 (large) score of 0.68

The significance of our research lies in the effective application of NLP techniques to combat hate speech in multilingual contexts, contributing valuable insights and solutions to a pervasive social problem. The balanced dataset, consisting of 4000 training tweets and 500 test tweets, provides a strong foundation for training and evaluating the model's performance.

Our findings underscore the potential of advanced technologies, such as deep learning models, in addressing complex social issues. The results of the competition presented in Table 4 show the relative performance of different teams and show the effectiveness of different approaches in detecting hate speech.

## Ethics Statement

We affirm our commitment to ethical research practices and compliance with ACL guidelines in conducting and presenting our study. No ethical concerns or conflicts of interest arose during the course of this research.

## Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

## References

Zahra Ahani, Moein Shahiki Tash, Yoel Ledo Mezquita, and Jason Angel. 2024. Utilizing deep learning models for the identification of enhancers and super-enhancers based on genomic and epigenomic features. *arXiv preprint arXiv:2401.07470*.

Premjth B, Bharathi Raja, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Prashanth Karnati, Sai Rishith Reddy Mangamuru, and booktitle = Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages month = March year = 2024 address = Malta publisher = European Chapter of the Association for Computational Linguistics Chandu, Janakiram". Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu).

Girma Yohannis Bade. 2021. Natural language processing and its challenges on omotic language group of ethiopia. *Journal of Computer Science Research*, 3(4):26–30.

Girma Yohannis Bade and Akalu Assefa Afaro. 2018. Object oriented software development for artificial intelligence. *American Journal of Software Engineering and Applications*, 7(2):22–24.

Girma Yohannis Bade and Hussien Seid. 2018. Development of longest-match based stemmer for texts of wolaita language. *vol*, 4:79–83.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, and Grigori Sidorov. 2021. Hssd: Hate speech spreader detection using n-grams and voting classifier.

Janek Bevendorff, BERTa Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, et al. 2021. Overview of pan 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 419–431. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Yunxia Ding, Xiaobing Zhou, and Xuejie Zhang. 2019. Ynu_dyx at semeval-2019 task 5: A stacked bigru model based on capsule network in detection of hate. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 535–539.

Antigoni 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 *Proceedings of the international AAAI conference on web and social media*, volume 12.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Selam Kanta and Grigori Sidorov. 2023. Selam@ dravidianlangtech: Sentiment analysis of code-mixed dravidian texts using svm classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179.

Shakir Khan, Mohd Fazil, Vineet Kumar Sejwal, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, Ashraf Kamal, and Abdul Rauf Baig. 2022a. Bichat: Bilstm with deep cnn and hierarchical attention for hate speech detection. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4335–4344.

Shakir Khan, Ashraf Kamal, Mohd Fazil, Mohammed Ali Alshara, Vineet Kumar Sejwal, Reemiah Muneer Alotaibi, Abdul Rauf Baig, and Salihah Alqahtani. 2022b. Hcovbi-caps: hate speech detection using convolutional and bi-directional gated recurrent unit with capsule network. *IEEE Access*, 10:7881–7894.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Bharathi Raja andS Malliga andCN SUBALALITHA Priyadharshini, Ruba andChakravarthi, Premjith andMurugappan Abirami S V, Kogilavani andB, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. 2020. A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, 8:204951–204962.

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*.

Moein Shahiki-Tash, Jesús Armenta-Segura, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Lidoma at hope2023iberlef: Hope speech detection using lexical features and convolutional neural networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), colocated with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS. org*.

M Shahiki Tash, Z Ahani, Al Tonja, M Gemeda, N Hussain, and O Kolesnikova. 2022. Word level language identification in code-mixed kannada-english texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 25–28.

Moein Tash, Jesus Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Lidoma@ dravidianlangtech: Convolutional neural networks for studying correlation between lexical features and sentiment polarity in tamil and tulu languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–185.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Word level language identification in code-mixed kannada-english texts using deep learning approach. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 29–33.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.