Frontiers in Health Informatics www.healthinformaticsjournal.com
ISSN-Online: 2676-7104
2024; Vol 13: Issue 5 Open Access

Multilingual Hate Speech Identification in Hindi, English and Marathi with Deep Learning Techniques

Brahmanaidu K[1], Vishnuvardhan B[2]

[1]Department of Artificial Intelligence and Data Science, Faculty of Science and Technology, (IcfaiTech), ICFAI Foundation for Higher Education(IFHE),Hyderabad-501203, India&Research Scholar, JNT University, Hyderabad Email: brahmanaidu@ifheindia.org.

[2]Senior Professor, Dept. of CSE, JNT University, Hyderabad.

Cite this paper as: Brahmanaidu K, Vishnuvardhan B (2024) Multilingual Hate Speech Identification in Hindi, English and Marathi with Deep Learning Techniques. *Frontiers in Health Informatics, 13* (5), 106-115

Abstract

Given the extensive adoption of the social media platforms, detecting hate speech and offensive language is essential for maintaining community unity and managing digital content. While existing research primarily concentrates on single-language training within a multilingual context, our study introduces a model trained across multiple languages. Our hybrid methodology incorporates mBERT and MuRILBERT models customized for English, Hindi, and Marathi. The model exhibited exceptional performance in English and demonstrated moderate efficacy in Hindi and Marathi.
Keywords: mBERT, MuRILBET, Hate Speech, Deep Learning, Multilingual

1. Introduction

Hate speech (HS) is typically defined as any form of communication that belittles or targets an individual or a collective based on characteristics like race, color, ethnicity, gender, sexual orientation, nationality, religion, or other distinguishing traits. With the substantial volume of user-generated content present on the Twitter platform, the issue of identifying and potentially counteracting the dissemination of hate speech has grown to be of paramount importance. This is particularly relevant in the context of combatting instances of misogyny and xenophobia.

In this endeavor, our objective is to initially recognize potential propagators of hate speech on Twitter, thereby taking a preliminary step towards impeding the proliferation of such harmful discourse among online users. Adhering to the guidelines established by Twitter, tweets are expected to refrain from engaging in threats or harassment directed at individuals due to attributes like ethnicity, gender, religion, or any other defining factor. Notably, YouTube also imposes restrictions on content that fosters violence or animosity towards specific individuals or groups, extending its purview to include age, caste, and disabilities. The surge in the dissemination of information across online platforms has sparked a compelling incentive to delve into the automated identification of hate speech, prompting an urgent need for exploration.

Drawing upon the diversity in national hate speech legislation, the intricate task of defining boundaries in the ever-evolving cyberspace, the growing necessity for societal players as well as individuals to articulate their viewpoints and counter opposing arguments, lag in the manual oversight by internet administrators, the proliferation of hate speech in the digital realm has gained renewed momentum. This trend consistently presents a multifaceted challenge to policymakers and the research community alike. Capitalizing on advancements in natural language processing (NLP) technology, a significant body of research has been dedicated to the automated identification of hate speech within textual content in recent years. Noteworthy competitions such as SemEval-2019 [37] and SemEval-2020 [38], along with GermEval-2018 [39], have organized diverse events aimed at discovering improved solutions for the automated detection of hate speech.

In this context, scholars have compiled extensive datasets from numerous origins, thereby fueling research endeavors within this domain. Many of these investigations have extended their focus to encompass hate speech across multiple languages and online communities. As a result, this has prompted the examination and comparison of diverse processing pipelines, encompassing choices of feature sets and Machine Learning (ML) techniques, such as supervised, unsupervised, and semi-supervised methods. A range of classification algorithms, including Naive Bayes, Logistic Regression (LR), Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), BERT deep learning architectures, and others, have been explored. It is widely recognized that the efficacy of the automated text-based approach for detection has its limitations, underscoring the need for ongoing research in this arena. Additionally, the diverse array of technologies, application domains, and contextual factors mandates the continual updating of advancements in this field, ensuring that researchers are equipped with a holistic and global perspective on the subject of automatic hate speech (HS) detection. Building upon the foundation of existing survey papers within this realm, the present paper adds to this objective by furnishing an up-to-date and structured examination of the literature concerning the automated identification of textual hate speech

2. Literature

Monolingual uses single language data and detects hate speech spreaders. [1, 2, 3, 42] Employed word frequency and linguistic characteristics as feature descriptors in the context of K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms. [4] Employed lexicon, theme based and semantic features and developed a rule based system to predict hate speech. [5] used logistic regression on character based n-grams to detect hate speech. [6] used hybrid CNN on character and word2vec feature representation. [7] Employed SVM on lexical, semantic context and word embedding representations to detect hate speech. [8] Employed SVM with RBF kernel on n-grams, skip-grams, hierarchical word cluster representation to detect hate speech. [9, 44] used biLSTM, LSTM and CNN on word embedding feature to detect hate speech. [10] used multinomial Naive Bayes classifier on bag-of-words features and detected hate speech in Turkish language, [11] Employed SVM on n-grams features and detected hate speech in Arabic language. [12] used RNN based GRU on word embedding Aravec and detected hate speech in Arabic language. [13] employed SVM on lexicon and bag-of-means features and detected hate speech in Arabic language. [14, 43] employed perceptron on skip-gram and character tri-grams features and detected hate speech in German language. [15] used XLM-RoBERT, mBERT models to detect hate speech on code mixed Tamil-English, Malayalam-English language datasets. [16] used ALBERTO to detect hate speech on Italian language dataset. [17] Employed XLM to detect hate speech on Turkish, Arabic, Danish and Greek language datasets.

Hate speech is a global phenomenon, and enhancing the diversity of available resources is crucial for the progress of automated systems. Collaborative initiatives, exemplified by events like SemEval 2020 [18], HASOC 2020 [19], Evalita 2018 [20], and HateEval 2019 [21], have played a significant role in driving progress in multilingual hate speech research. Furthermore, recent advancements in transformer-based large multilingual Language Models (LMs) like mBERT [22] and XLMR [23], pre-trained on over 100 languages, have contributed to the development of state-of-the-art classifiers even in resource-scarce languages.

Prior studies in the realm of multilingual hate speech detection have encompassed various facets, including (i) resource expansion through dataset creation [18], [24], shared tasks, and workshop organization, (ii) cross-lingual transfer learning utilizing multilingual shared embeddings and pre-trained Language Models [25], [26], (iii) incorporation of supplementary features from relevant domains [27], such as emotion and sentiment analysis, and (iv) the application of data augmentation techniques [26], [28], [29], which may involve external services like translation APIs for supervised training.

3. Methodology

In this section, we present the proposed methodology for detection of hate speech from the multilingual tweets.

3.1 Data Preprocessing

Prior to implementing the proposed model, the tweets in the text data are cleaned by removing URLs, punctuation, stop words, and emojis. Also the tokens are moved to root word by stemming. Finally, all tokens are converted to lowercase.

3.2 Data Representation

The way data is represented is crucial for detecting hate speech. In our proposal, we use word embeddings for this purpose. Word embeddings convert text into vectors, capturing the semantics and context of words in the text data. GloVe, a well-known word vector representation, uses word-word co-occurrence from a corpus and is a pre-trained, unsupervised learning technique. Word2Vec is another model that creates vectors based on surrounding words using CBOW and Skip-gram models. FastText, developed by Facebook, incorporates subword information to represent word vectors. BERT, a popular technique based on transformer architecture, also provides word vectors. MuRIL offers multilingual representations for Indian languages, trained on a corpus of 17 Indian languages.

3.3 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a robust pre-trained language model created by Google. Unlike traditional models that process text in a single direction (either left-to-right or right-to-left), BERT analyzes the entire sequence of words simultaneously, which helps it grasp the context of a word based on its surrounding words. This bidirectional method allows BERT to capture more comprehensive linguistic information and relationships within the text.

Built on transformer architecture, BERT utilizes self-attention mechanisms to determine the significance of different words in a sentence. This capability enables BERT to excel in various NLP tasks, including text classification, question answering, and named entity recognition, achieving state-of-the-art accuracy.

A key feature of BERT is its pre-training on a large corpus of text, followed by fine-tuning for specific tasks. During pre-training, BERT employs two main objectives: masked language modeling (MLM) and next sentence prediction (NSP). MLM involves masking some words in a sentence and training the model to predict them, while NSP involves predicting whether two sentences are consecutive in a document. BERT's proficiency in understanding context and semantics has made it a favored choice for numerous NLP applications, significantly improving performance benchmarks in the field

3.4 MuRIL

MuRIL (Multilingual Representations for Indian Languages) is a pre-trained language model developed by Google, specifically designed for understanding and processing various Indian languages. It supports a broad range of Indian languages, making it highly versatile for multilingual tasks. Trained on a corpus that includes 17 Indian languages such as Hindi, Tamil, Telugu, Bengali, and others, as well as English, MuRIL can effortlessly handle code-mixed language inputs and multilingual text. Built on the transformer architecture, MuRIL uses self-attention mechanisms to assess the importance of different words in a sentence, enabling it to grasp the context and semantics of words within a sentence and capture deeper linguistic relationships.

3.5 Proposed System Architecture

In our proposal, we focus on detecting hate speech in tweets written in English, Hindi, and Marathi. To detect hate speech in tweets written in English, Hindi, and Marathi, we utilize embeddings from two pre-trained language models, mBERT and MuRILBERT. The steps involved are as follows:

1 Extract Embeddings: mBERT for English Tweets: Tokenize and extract embeddings using mBERT.

2 MuRILBERT for Hindi and Marathi Tweets: Tokenize and extract embeddings using MuRILBERT.

3. Pooling: Reduce the dimensionality of the extracted embeddings by averaging them across the sequence length.

4 Stack Embeddings: Concatenate the pooled embeddings from mBERT and MuRILBERT to create a combined feature vector.

5 Classification Layer: Feed the combined embeddings into a fully connected layer for hate speech detection. The architecture is presented in the Fig1

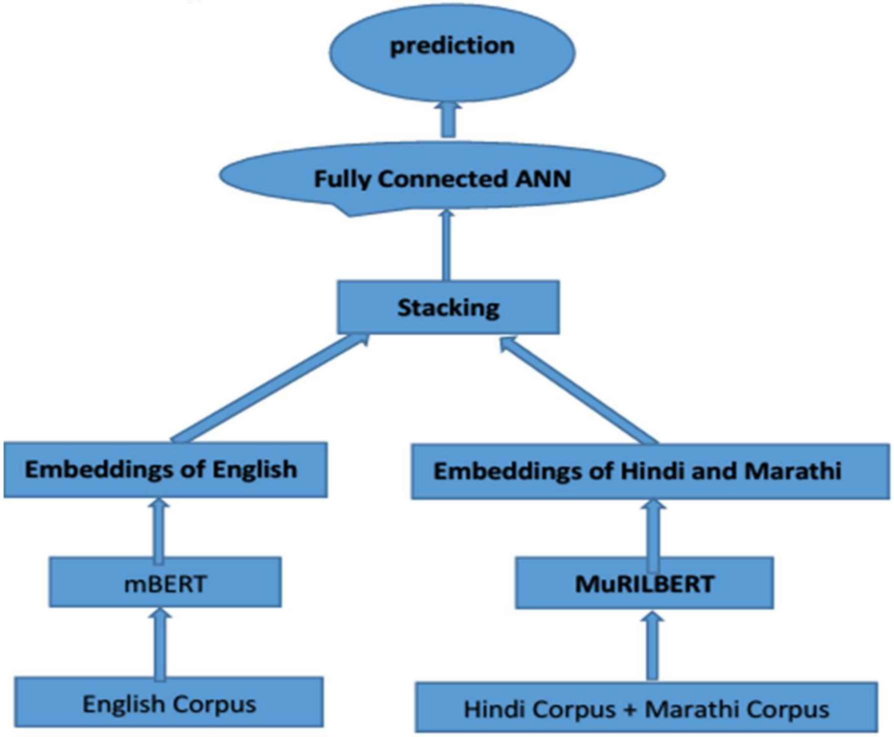

Fig1: Architecture for hate speech detection

4. Implementation and Result Analysis

4.1 Dataset Description

We have considered HASOC 2019, HASOC 2020, and HASOC 2021 datasets. The details of the corpus are presented in Table. 1, Table. 2, and Table. 3 for English, Hindi and Marathi languages respectively. The total English, Hindi and Marathi language corpus sample sizes are 14556, 13540, and 1874 respectively.

Year	Samples Size
2019	7005
2020	3708
2021	3843

Table. 1. Yearwise English Corpus

Year	Samples Size
2019	5983
2020	2963
2021	4594

Table. 2. Yearwise Hindi Corpus

Year	Samples Size
2021	1874

Table. 3. Marathi Corpus

4.2 Implementation

We implemented the model in Python using the pre-trained mBERT (bert-base-multilingual-cased) and MuRILBERT(google/muril-base-cased) models. The proposed model achieved an accuracy of 76.16% on the combined dataset. When evaluated on individual testing datasets, the model attained accuracies of 81.24% for English, 76.56% for Hindi, and 70.68% for Marathi.The results are presented in the Table. 4.

Testing Language	Training Accuracy	Testing Accuracy
English	76.16	81.24
Hindi	76.16	76.56

Testing Language	Training Accuracy	Testing Accuracy
English	76.16	81.24
Marathi	76.16	70.68

Table. 4. Accuracy of the Proposed Model

4.3 Result Analysis

The proposed model, which uses pre-trained mBERT and MuRILBERT models, shows different accuracy levels across various datasets. It achieved an overall accuracy of 76.16% on the combined dataset of English, Hindi, and Marathi tweets. This suggests that while the model performs reasonably well across diverse languages, there is still room for improvement. The model attained the highest accuracy of 81.24% for English tweets, indicating that mBERT, which is extensively trained on multilingual data including English, effectively captures the nuances of English text. For Hindi tweets, the model achieved an accuracy of 76.56%, reflecting MuRILBERT's capability to handle Hindi, despite potential linguistic or contextual challenges. The model's accuracy for Marathi tweets is 70.68%, the lowest among the three languages. This lower performance could be due to factors such as less representation of Marathi in the training corpus or the language's inherent complexity, indicating a need for further fine-tuning or more training data. The significant differences in accuracy across languages highlight the importance of language-specific nuances in hate speech detection. While the model is robust for English, its effectiveness decreases for Hindi and further for Marathi, likely due to the complexity of the languages, script and syntax diversity, or variations in hate speech expressions across languages.

5. Comparison Analysis

In the existing literature, researchers have employed various methodologies to detect hate speech in multilingual HASOC (Hate Speech and Offensive Content) data. These methodologies often utilize pre-trained language models such as mBERT, MuRILBERT, XLM, and mDistilBERT, which have been applied to individual language datasets. Table 5 summarizes the results of these comparative studies.

The prevailing trend indicates that existing methodologies generally outperform our proposed model. This observation can be attributed to the fact that previous approaches typically focus on training and testing within a single language. This approach simplifies the challenge by allowing for more effective fine-tuning of models to the specific linguistic characteristics as well as nuances of each language. Consequently, these methods achieve higher accuracy rates within their respective language domains.

Model	Trained Language	Accuracy
mBERT	English	80.80
mBERT	Hindi	80.50
mBERT	Marathi	88.84

mBERT+MuRIL	English+Hindi+Marathi	76.16

Table. 5. Comparison of Results

Conversely, our approach adopts a multilingual framework, which inherently introduces additional complexity. By addressing multiple languages simultaneously, our model encounters diverse linguistic structures and cultural contexts. While this broader scope offers advantages in terms of generalizability and inclusivity across languages, it also poses challenges in achieving the same level of fine-tuning and specificity as single-language models.

In summary, while existing methodologies excel in accuracy within individual languages due to their focused training paradigms, our multilingual approach aims to broaden the applicability of hate speech detection across diverse linguistic environments, albeit with current performance gaps compared to single-language counterparts.

.

6. Conclusion

The proposed model, leveraging pre-trained mBERT and MuRILBERT models, yields promising results in detecting hate speech across tweets in English, Hindi, and Marathi. Achieving an overall accuracy of 76.16%, the model demonstrates strong performance across various linguistic inputs. The highest accuracy of 81.24% for English tweets highlights mBERT's effectiveness in capturing English language nuances. However, performance declines for Hindi (76.56%) and further for Marathi (70.68%), indicating challenges in processing these languages.

The significant variation in accuracy across languages emphasizes the importance of addressing language-specific nuances in hate speech detection. On the other hand, model's robust performance in English, contrasted with lower accuracies in Hindi and Marathi, suggests the need for further refinement. Factors such as language complexity, script and syntax diversity, and variations in hate speech expressions likely contribute to these discrepancies.

References

1. 1.K. Dinakar, B. Jones, C. Havasi, H. Lieberman, R. Picard, Common sense reasoning for detection, prevention, and mitigation of cyberbullying, ACM Transactions on Interactive Intelligent Systems (TiiS) 2 (2012) 1–30.
2. Warner, W., Hirschberg, J., 2012. Detecting hate speech on the world wide web, in: Proceedings of the second workshop on language in social media, Association for Computational Linguistics. pp. 19–26.
3. S. Agarwal, A. Sureka, Using knn and svm based one-class classifier for detecting online radicalization on twitter, International Conference on Distributed Computing and Internet Technology, Springer. (2015) 431–442.
4. N.D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230.
5. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N., 2015. Hate speech detection with comment embeddings, in: Proceedings of the 24th international conference on world wide web, pp. 29–30.

6. Park, J.H., Fung, P., 2017. One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206.
7. Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C., 2018a. Inducing a lexicon of abusive words–a feature-based approach.
8. S. Malmasi, M. Zampieri, Challenges in discriminating profanity from hate speech, Journal of Experimental & Theoretical Artificial Intelligence 30 (2018) 187–202.
9. Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, Semantic Web 10 (2019) 925–945.
10. Özel, S.A., Saraç, E., Akdemir, S., Aksu, H., 2017. Detection of cyberbullying on social media messages in turkish, in: 2017 International Conference on Computer Science and Engineering (UBMK), IEEE. pp. 366–370.
11. A. Alakrot, L. Murray, N.S. Nikolov, Towards accurate detection of offensive language in online communication in arabic, Procedia computer science 142 (2018) 315–320.
12. Albadi, N., Kurdi, M., Mishra, S., 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere, in: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM. pp. 69–76.
13. Alshehri, A., El Moatez Billah Nagoudi, H.A., Abdul-Mageed, M., 2018. Think before your click: Data and models for adult content in arabic twitter, in: TA- COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety, p. 15.
14. Jaki, S., De Smedt, T., 2019. Right-wing german hate speech on twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518.
15. Sai, S., Sharma, Y., 2020. Siva@ hasoc-dravidian-codemix-fire-2020: Multilingual offensive speech detection in code-mixed and romanized text. FIRE (Working Notes).
16. M. Polignano, V. Basile, P. Basile, M. de Gemmis, G. Semeraro, Alberto: Modeling italian social media language with bert, IJCoL. Italian Journal of Computational Linguistics 5 (2019) 11–31
17. Wang, S., Liu, J., Ouyang, X., Sun, Y., 2020. Galileo at semeval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. arXiv preprint arXiv:2010.03542.
18. M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H.Mubarak,L.Derczynski,Z.Pitenis,andC ¸C ¸o ̇ltekin,“SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020),” in *SemEVal*, Dec. 2020, pp. 1425–1447.
19. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel, “Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages,” in *FIRE*, 2019, pp. 14–17.
20. E. Fersini, D. Nozza, and P. Rosso, “Overview of the evalita 2018 task on automatic misogyny identification (ami),” in *EVALITA@CLiC-it*, 2018.
21. V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, “SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter,” in *SemEval*, 2019, pp. 54–63.
22. J.Devlin,M.-W.Chang,K.Lee,andK.Toutanova,“Bert:Pre-trainingof deep bidirectional transformers for language understanding,” in *NAACL*, 2019, pp. 4171–4186.
23. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzma ́n, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsu- pervised cross-lingual representation learning at scale,” in *ACL*, 2020, pp. 8440–8451.
24. T. Mandl, S. Modha, A. Kumar M, and B. R. Chakravarthi, “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*, ser. FIRE 2020, 2020, p. 29–32.

25. S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, “A deep dive into multilingual hate speech classification,” in *ECML/PKDD*, 2020. [9] E. W. Pamungkas and V. Patti, “Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multi-lingual lexicon,” in *ACL-Student research workshop*, 2019, pp. 363–370.

26. I. Markov, N. Ljubesˇic´, D. Fisˇer, and W. Daelemans, “Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection,” in *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021, pp. 149–159.

27. T. Wullach, A. Adler, and E. Minkov, “Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech,” 2021.

28. I.Bigoulaeva,V.Hangya,andA.Fraser,“Cross-lingualtransferlearning for hate speech detection,” in *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, 2021, pp. 15–25.

29. T. Ranasinghe and M. Zampieri, “Multilingual offensive language iden-tification with cross-lingual embeddings,” in *EMNLP*, 2020, pp. 5838–5844.

30. M. Artetxe and H. Schwenk, “Massively multilingual sentence embed- dings for zero-shot cross-lingual transfer and beyond,” *TACL*, vol. 7, pp. 597–610, 2019.

31. A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Je´gou, “Word translation without parallel data,” *arXiv preprint arXiv:1710.04087*, 2017.

32. A. Jiang and A. Zubiaga, “Cross-lingual capsule network for hate speech detection in social media,” in *HYPERTEXT*, 2021, p. 217–223.

33. H. Sohn and H. Lee, “Mc-bert4hate: Hate speech detection using multi- channel bert for different languages and translations,” in *ICDMW*, 2019, pp. 551–559.

34. P. Saha, B. Mathew, P. Goyal, and A. Mukherjee, “Hatemonitors: Language agnostic abuse detection in social media,” *CoRR*, vol. abs/1909.12642, 2019.

35. S. Mishra, “3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages,” in *FIRE*, 2019.

36. D. Nozza, “Exposing the limits of zero-shot cross-lingual hate speech detection,” in *ACL-IJCNLP*, Aug. 2021, pp. 907–914.

37. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R., 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.

38. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç., 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). arXiv preprint arXiv:2006.07235.

39. Wiegand, M., Siegel, M., Ruppenhofer, J., 2018b. Overview of the germeval 2018 shared task on the identification of offensive language.

40. Jahan, Md Saroar, and Mourad Oussalah. "A systematic review of Hate Speech automatic detection using Natural Language Processing." *Neurocomputing* (2023): 126232.

41. Awal, Md Rabiul, et al. "Model-agnostic meta-learning for multilingual hate speech detection." *IEEE Transactions on Computational Social Systems* (2023).

42. Brahma Naidu K, Vishnuvardhan B, Adi Narayana Reddy K "Enhancing Hate Speech Detection with Integrated Content-Based and Stylistic Features" J. Electrical Systems 20-7s (2024): 3660-3666.

43. Adi Narayana Reddy, K., Laskari, N. K., Shyam Chandra Prasad, G., & Sreekanth, N. (2022). Fusion-Based Celebrity Profiling Using Deep Learning. In Intelligent System

Design: Proceedings of INDIA 2022 (pp. 107-113). Singapore: Springer Nature Singapore.

44. L. Lakshmi, K. Dhana Sree Devi, K. Adi Narayana Reddy, Suresh Kumar Grandhi, Sandeep Kumar Panda. "WOMT: Wasserstein Distribution based minimization of False Positives in Breast Tumor classification using Deep Learning", IEEE Access, 2023.