International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 01 | Jan – 2026 DOI: 10.55041/ISJEM03334
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Multilingual Hate Speech Detection using Meta-Transfer-Learning

Kushal Sinha
Computer Science and Engineering
PES Institute of Technology
Karnataka, India
kushaljay4723@gmail.com

Mohammed Basim Alam
Computer Science and Engineering
PES Institute of Technology
Karnataka, India
basim4jee@gmail.com

Nandana Aravind
Computer Science and Engineering
PES Institute of Technology
Karnataka, India
nanukutty2003@gmail.com

Priyanshu Hiranyareta Surbhi Choudhary Computer Science and Engineering Dept. of CSE
PES Institute of Technology PES Institute of Technology Karnataka, India Karnataka, India
priyanshu5555555@gmail.com surbhichoudhary@pes.edu

Abstract—The fast expansion of online platforms created a record-setting rise in hate speech, undermining safe and wel- coming online spaces. This project contributes a multilingual hate speech system based on meta-learning and transfer learning to achieve improved detection for languages. We employ a pre-trained mBERT model for the English language and fine-tune it to the low-resource language of Hindi, Hinglish, Marathi and Bengali using meta-learning and transfer learning methods. With the incorporation of these practices, the system generalizes remarkably well to a new language given limited labeled exam-ples, with impressive accuracy, scalability, and versatility. The proposed method greatly increases detection rates, reduces bias, and promotes healthy digital interactions ultimately leading to an improved safer online space. Besides, the framework of the system supports easy adaptation to other languages with small corpora.

INTRODUCTION

Multilingual Hate Speech Detection System solves the in- creasing problem of hate speech detection in various languages by combining the use of meta-learning and transfer learning techniques within advanced AI. The system bases its work with a pre-trained mBERT model for the English language as an initial point from which to build for low-resource languages such as Hindi, Hinglish, Marathi and Bengali. The adaption is realized through meta-learning, which amplifies the capacity of the model to learn with few samples, and transfer learning, which ensures efficient knowledge transference from the source language to the target language.

By this method, the system can generalize effectively, be less dependent on massive labeled data, and scale effectively in resource-poor environments. The model can mitigate bias, effectively classify hate speech in language and culture en- vironments, and facilitate constructive digital communication. The system design also provides the capability to scale to future languages, thus making it an effective and efficient multilingual hate speech detector.

RELATED WORKS

Dynamics of Online Abuse

Online aggression in the shape of context-dependent hate and threats towards individuals or groups based on discrimi- natory feelings is a common occurrence in the online world. Freedom of expression is often called upon to justify tolerance of offending speech, but hate speech is riddled with dangers for society. Repeated exposure to insulting speech not only entrenches discrimination but also reinforces prejudice against outgroups, generating polarizing sentiments in communities. Moreover, offline violent attacks also generate hateful dis- course online, creating a two-way relationship between of- fline and online aggression. With the rapid progression of multilingual online

media, hate speech becomes language- and culture-transcendent, prompting the research community to create focused efforts in the development of solutions to combat online abuse effectively across languages and contexts.

Research on Abusive Speech

The study on detecting abusive speech has evolved significantly since its beginnings. Early research was primarily focused on analyzing user-generated content from a lexical and syntactic perspective to identify abusive behavior. However, with the advent of social media, hate speech has appeared in different forms and contexts, particularly in multilingual settings. Realizing the nature of the complexity of hate speech, researchers have moved away from traditional machine learn- ing methods to more sophisticated deep learning and meta- learning methods. Waseem and Hovy (2016) initiated the early research with the launch of the racism-annotated and sexism- annotated dataset, to which Davidson et al. (2017) responded in exploring the difference between hateful and offensive tweets. All these early experiments used linguistic features like n-grams, part-of-speech tags, emotion lexicons, and tf-idf

vectors with traditional classifiers like Logistic Regression and Support Vector Machines. Still, variations in labeling schemes across datasets posed a major challenge to generalizability and replicability. Thus, Founta et al. introduced a strong labeling mechanism to avoid overlaps between classes of abusive language, offering a stronger benchmark for future research. The advent of big data facilitated the uptake of data- intensive methods, including deep learning and graph-based models. For instance, Pitsilis et al. proved that LSTMs can excel at English abusive tweet classification, whereas Zhang et al. attained exceptional success through the convolutional and gated recurrent networks combination for hate speech detec- tion across various datasets. More recently, transformer-based models like BERT, mBERT, and XLM-R have transformed the field, performing better than previous methods in multilingual and cross-lingual hate speech detection. Even with these developments, multilingual hate speech detection is confronted with specific challenges. Structural linguistic heterogeneity, cultural variation, and the unavailability of labeled data in low- resource languages prevent traditional models from scaling. Meta-learning has been a potential solution to address these challenges. Through knowledge transfer from high-resource to low-resource languages, meta-learning enables the adaptation of flexible and efficient models to generalize across domains and languages. For example, recent work has been able to use meta-learning methods to fine-tune multilingual transformer models for hate speech detection with state-of-the-art perfor- mance on a variety of datasets. With hate speech increasingly taking hold in multilingual online forums, scalable, context- aware, and inclusive detection systems must be developed. Meta-learning, in combination with advances in deep learning and multilingual pre-trained models, offers a sound platform for addressing these challenges and facilitating attempts to stem online hostility across the globe

DATASETS

The datasets used for this project comprise the major languages spoken in India. Each of these datasets contains examples of normal and abusive speech written by people on different social media platforms. A brief description of the datasets is provided below and also in Table I. The numbers corresponding to each language in Table I represent the number of sentences of each class present for that language.

English : A large fraction of offensive speech datasets available on web are in the English language. Among these, we selected the Hate Speech Detection curated Dataset from Kaggle by Wendyelle´ A. Alban.

Hindi : The dataset used for Hindi is written in Devana- gari Hindi from Kaggle by Hari Thapliyaal.

Bengali : Bengali is spoken by a significant population in the eastern and northeastern parts of our country. The dataset was created by crawling Facebook Posts and YouTube comments.

Marathi : For Marathi, the dataset was taken from GitHub, uploaded by ravirajoshi

PROPOSED METHODOLOGY

This paper proposes a multilingual hate speech model that can manage the issue of low-resource languages with fewer examples available. The approach utilizes mBERT as the source English model and initializes regional languages through Model-Agnostic Meta-Learning (MAML) and transfer learning techniques. The framework is able to quickly learn from new languages with few labeled examples and thus is an efficient and useful tool to detect hate speech in numerous linguistic situations.

Base Model: mBERT

mBERT is a transformer multilingual language model pre- trained on 104 languages with a masked language modeling (MLM) objective. It is composed of 12 transformer encoder layers with self-attention and is optimized for a maximum of 512 tokens. Fine-tuning mBERT involves adding a fully connected layer on top of its architecture, with the output being that of the CLS token, the encoding of the input sentence. mBERT has been demonstrated to perform well in abusive speech detection tasks and has been demonstrated to outperform current baselines and become a state-of-the-art method for multilingual NLP tasks

Meta-Learning Framework: MAML

MAML enables fast adaptation to new target languages from few labeled examples. Its model-agnostic design enables it to be used with a broad variety of machine-learning architectures. MAML consists of two stages:
Inner Loop: The model is trained on a support set of examples with labels for a task and thus gets to learn its parameters for the particular task.
Outer Loop: The parameters are used with a test set of queries obtained from the same task. The gradients of the test are employed to update the shared initialization parameters of the meta-learner.

Transfer Learning for Low-Resource Languages

Transfer learning plays an important role in enhancing hate speech detection performance for low-resource languages. Rather than training models from scratch, pre-trained mul- tilingual models such as mBERT and MuRIL (Multilingual Representations for Indian Languages) offer contextualized word representations fined-tuned over smaller regional lan-guage datasets. This enhances detection with a much better accuracy, lowers training time, and enables generalization across languages with limited labeled data. Pre-trained models transfer knowledge for linguistically related languages, and it becomes cost-effective to acquire new languages with little supervision.

System Workflow

The model is first trained on a support set of labeled instances in the auxiliary languages and English to have proper parameter initialization. The model is then fine-tuned on a target language using only a small labeled data set.

The model then predicts the target language on an unlabeled query set. For regional Indian languages, MuRIL (Multilingual Representations for Indian Languages) offers a supplement to mBERT with pre-trained Indian language and translitera- tion representations. This facilitates interoperability between languages and facilitates fine-tuning for low-resource regional languages.

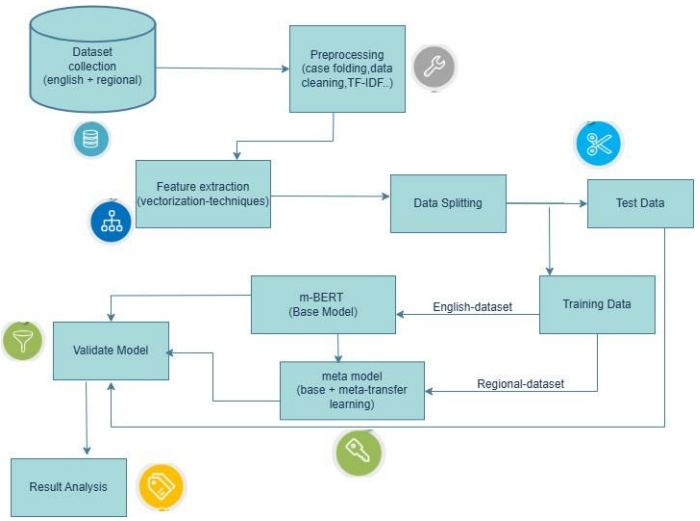

Fig. 1. System flow diagram

Real-Time Hate Speech Detection

The proposed framework can be deployed as a real-time social media hate speech detection API. Using the MAML- driven mBERT model, the system can make accurate pre- dictions even in zero-shot settings—detecting hate speech in languages where there are no pre-existing labeled datasets. The solution facilitates safer online spaces and responsible communication through effective detection and minimization of offensive content.

ETHICAL CONSIDERATIONS AND BIAS MITIGATION

The development of hate speech detection models has several ethical issues like training data bias, content misclas- sification, and over-censorship. For fairness and accuracy, the following were observed:

Dataset Bias Mitigation

Balanced Data Sampling: Data was chosen to avoid over- representation of certain languages, dialects, or view- points.
Diverse Annotation Practices: Annotated data was used to reduce subjectivity in hate speech classification and minimize cultural bias

Explainability and Model Interpretability

The mBERT system uses the attention visualization tech- niques for model decision explanation facilitation.

The system outputs confidence scores for all classifications, allowing for human moderation of borderline cases.

Preventing Over-Censorship

A human-in-the-loop methodology guarantees that censored content is checked prior to being censored automatically.

The model separates explicit hate speech from provoking but not hateful conversation, minimizing content removal that is unnecessary.

RESULTS AND DISCUSSION

The performance of the proposed meta-learning model was evaluated across various low-resource languages. The models were trained with a batch size of 16, an inner loop learning rate of 2 x 10^-5, and an outer loop learning rate of 1 x 10^-5. Training was for 100 epochs, followed by which models were evaluated on key metrics of accuracy, precision, and recall. The meta-learning approach was found to outperform fine-tuned models across almost all measures. Specifically, for small- resource languages like Marathi, the difference in performance was even sharper.

Languges	Accuracy
English	91%
Hindi	76%
Bengali	80%
Marathi	82%
Hinglish	70%

TABLE I
TABLE SHOWING THE RESPECTIVE ACCURACIES(ROUNDOFF)

CONCLUSION AND FUTURE WORK

The system, as proposed, is especially appropriate to tackle the problem of multilingual hate speech detection, especially in low-resource languages with small amounts of labeled data. The combination of meta-learning algorithms with the strength of mBERT, the system becomes highly accurate, flexible, and scalable across diverse linguistic settings. The application of MAML ensures quick adaptation to new target languages even with a small amount of labeled data, which makes the approach especially viable for global use. In addition, the ability of the system to generalize well adds to its utility in practical applications, such as hate speech moderation on social media platforms and building safer online communities.

Expanding to More Languages Today, the system shows encouraging results in many languages, but there are still a huge number of underrepresented and regional languages waiting to be addressed. In the future, efforts will be directed towards adding more languages, particularly those with special linguistic structures or sparse digital resources, by making use of larger multilingual datasets and fine-tuning models such as MuRIL or other regional embeddings.

Integrating Multimodal Data Hate speech is not limited to text but frequently takes place using a combination of text, images, audio, and video. Subsequent versions of this system will add multimodal data analysis to identify offending content in images (such as memes), speech (such as videos or voice messages), and videos. Application of technologies such as

vision transformers for image processing and voice-to-text for audio will make the moderation platform more comprehensive.

Refining Real-Time Feedback Mechanisms Giving users timely and accurate responses is the core of effective moder- ation. Sophisticated means of giving real-time warnings with explanations for suspected content will be a topic of future study to make the moderation process more transparent and usable. Adaptive algorithms will be developed to enhance responsiveness of the system without reducing accuracy.

Enhancing Robustness Against Evasion Techniques Users tend to use innovative methods, including misspellings, synonyms, or code-switching between languages, to evade moderation systems. Improvements in the future will con- centrate on developing a strong framework to manage such adversarial examples with the help of adversarial training and dynamic language models that are capable of responding to such modifications in real-time.

Cross-Domain Adaptation Beyond social media, hate speech detection is also critical in other domains such as education, online gaming, and public forums. Future research will be to deploy this system in various contexts and domains so that it can be useful in detecting harmful content across platforms and audiences.

REFERENCES

N. Papadopoulos et al., "Multilingual hate speech detection using meta-learning techniques," *IEEE Transactions on Artificial Intelli- gence*, vol. 3, no. 4, pp. 320–334, July 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10322364

R. Salim and M. Ahmad, "Transformers for low-resource hate speech detection," in *Proceedings of the 2023 International Conference on Artificial Intelligence (ICAI)*, vol. 1, pp. 412–420, 2023. [Online].
Available:
https://ieeexplore.ieee.org/document/10075749

T. Zhang, Y. Xu, and P. Chen, "Adversarial training for multilingual offensive language classification," *IEEE Computational Intelligence Magazine*, vol. 18, no. 2, pp. 56–65, April 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10306709

A. Fernandez and L. Martinez, "Advances in multilingual hate speech detection," *Information*, vol. 12, no. 1, pp. 5–15, Jan. 2021. [Online].
Available:https://www.mdpi.com/2078-2489/12/1/5

P. Gupta and K. Sharma, "Hate speech and bias detection: Challenges and opportunities," in *CEUR Workshop Proceedings*, vol. 3395, pp. 78–85, Oct. 2023. [Online]. Available: https://ceur-ws.org/Vol-3395/T7- 8.pdf

M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-lingual few-shot hate speech and offensive language detection using meta learning," *IEEE Access*, vol. 10, pp. 14880–14896, 2022.

G. V. Suresh, B. R. Chakravarthi, and J. P. McCrae, "Meta-learning for offensive language detection in code-mixed texts," in *Proc. Forum Inf. Retr. Eval.*, Dec. 2021, pp. 58–66.

A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and
I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proc. 10th ACM Conf. Web Sci.*, New York, NY, USA, Jun. 2019, pp. 105–114, doi: 10.1145/3292522.3326028.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, Geneva, Switzerland, 2017, pp. 759–760, doi: 10.1145/3041021.3054223.

L. Gao and R. Huang, "Detecting online hate speech using context aware models," in *Proc. Recent Adv. Natural Lang. Process. Meet Deep Learn. (RANLP)*, Varna, Bulgaria, Nov. 2017, pp. 260–266, doi: 10.26615/978- 954-452-049-6 036.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, Perth, WA, Australia, Apr. 2017, pp. 759–760.

T. Putri, "Analisis dan deteksi hate speech pada sosial Twitter berbahasa Indonesia," M.S. thesis, Dept. Comp. Sci., Indonesia Univ., Indonesia, 2018.

J. Sachdeva, K. K. Chaudhary, H. Madaan, and P. Meel, "Text based hatespeech analysis," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Coimbatore, India, Mar. 2021, pp. 661–668.

J. Koza, "Genetic programming: On the programming of computers by means of natural selection," *Stat. Comput.*, vol. 4, no. 2, pp. 87–112, Jun. 1994.

A. Esparcia-Alca´zar, A. Eka´rt, S. Silva, S. Dignum, and A. Uyar, "Genetic programming," in *Proc. EuroGP*, Istanbul, Turkey, Apr. 2010,
pp. 7–9.

O. de Gibert, N. Perez, A. Garc´ıa-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," in *Proc. 2nd Workshop Abusive Lang. Online (ALW2)*, 2018, pp. 11–20.

T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," in *Proc. 3rd Workshop Abusive Lang. Online*, 2019, pp. 25-35.

E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

K.E. Abdelfatah, G. Terejanu, A.A. Alhelbawy, "Unsupervised detection of violent content in arabic social media," *Comput. Sci. Inf. Technol. (CS IT)*, (2017), pp. 1-7.

E.A. Abozinadah, A.V. Mbaziira, J. Jones, "Detection of abusive ac- counts with arabic tweets."

M.P. Akhter, Z. Jiangbin, I.R. Naqvi, M. Abdelmajeed, M.T. Sadiq, "Automatic detection of offensive language for urdu and roman urdu," *IEEE Access*, 8 (2020), pp. 91213-91226.

Albadi, N., Kurdi, M., Mishra, S., "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.