Machine Learning-Based Approaches for Detecting Offensive Language in Marathi: A Review

Ritu Ranjani Singh[1], (*Assistant Professor),*
Department Of Computer Science Engineering (AIML),
Oriental Institute of Science and Technology,
Bhopal (M.P), India

Sabiha Khan[3], *(Scholar),*
Department Of Computer Science Engineering (AIML),
Oriental Institute of Science and Technology,
Bhopal (M.P), India

Mehdi Abbas Zaidi[2], (*Scholar),*
Department Of Computer Science Engineering (AIML),
Oriental Institute of Science and Technology,
Bhopal (M.P), India

Samra Iqbal[4], *(Scholar),*
Department Of Computer Science Engineering (AIML),
Oriental Institute of Science and Technology,
Bhopal (M.P), India

***Abstract*— The creation of a machine learning-based system for identifying objectionable Marathi language is the key topic of this review study. With the popularity of social media and online communication, the use of foul language has increased significantly, and its identification is essential to the survival of an online community. A dataset of Marathi text was gathered and pre-processed in Review in order to train Naive Bays, Logistic Regression, and Support Vector Machines, among other machine learning models. The Review comes to the conclusion that hostile Marathi language may be detected successfully using machine learning-based techniques, and that more research is required to enhance the performance of these models for certain forms of offensive language.**

Keywords-Offensive Language Identification, Hate Speech, Machine Learning and Deep Learning.

I. INTRODUCTION *(HEADING 1)*

The use of offensive language in online communication has become a major issue in recent years. With the growth of social media platforms and online communities, the problem has become more significant, and the need for automated systems to detect and filter offensive language has increased [13]. The Marathi language, widely spoken in India, is no exception to this trend, and the need for an automated system to detect offensive Marathi language is of utmost importance [11]. The detection of offensive language in Marathi presents a significant challenge due to the complexity and richness of the language [12]. Marathi has a complex grammar structure, and its vocabulary is extensive, making it challenging to develop a lexicon of offensive words. Moreover, the use of informal language, dialects, and slang adds further complexity to the task of detecting offensive language in Marathi. In recent years, machine learning-based approaches have shown promising results in detecting offensive language in various languages [28] [30]. These approaches use natural language processing techniques to analyze text and identify patterns that are characteristic of offensive language. In this study, we explore the effectiveness of machine learning-based approaches for detecting offensive Marathi language. The study involves collecting a dataset of Marathi text that contains both offensive and non-offensive language [29]. The dataset is pre-processed, and features are extracted to train various machine learning models [31]. The performance of these models is evaluated using metrics such as accuracy, precision, recall, and F1-score. The study also compares the performance of these models with a lexicon-based approach [27].

Offensive speech detection in social media is a crucial task. The impact of cyber bullying and offensive social media content on society's mental health is still under research, but it is undeniably negative. With the increasing number of social media users, offensive speech identification is a crucial task necessary to maintain harmony [9] [10]. In this work, we focus on offensive language detection in the low-resource Marathi language. Marathi is an Indo-Aryan language predominantly spoken in the Indian state of Maharashtra. Marathi is a rich language derived from Sanskrit and has 42 dialects. Spoken by 83 million people, it is the third-largest spoken language in India and the tenth in the world [32].

II. LITERATURE SURVEY

Onkar Litake, et. al. (2023)- In this research study, many NER systems have been deployed in English and other major languages. However, there hasn't been much work done on Hindi and Marathi languages. This study seeks to examine transformer-based deep learning-based NER solutions to Hindi and Marathi NER tasks. We benchmark for a host of monolingual and multilingual transformer-based models for

Named Entity Recognition that includes multilingual BERT, Indic BERT, XlmRoberta, mahaBERT, and others. We show that monolingual training doesn't necessarily ensure superior performance. Although Marathi monolingual models perform the best same is not true for Hindi. Moreover, we observe that the mahaBERT models even generalize well on Hindi NER datasets. It is worthwhile to investigate the poor performance of monolingual models and is left to future scope [01].

Tanmay Chavan., et. al. (2022) - This research work, the performance of various BERT models on the HASOC 2022 / MOLD v2 to observe the ability of our models to detect hate speech in Marathi. Authors fine-tune models like Mural, Maha Tweet BERT, and a domainspecific model, Maha Tweet BERT-Hateful, pre-trained on 1 million hateful data samples on both datasets. Our experiments show that the models fine-tuned on the combined dataset perform significantly better. The Maha Tweet BERT, pre-trained on 40 million Marathi tweets, outperforms all the other models. We also utilize external data sources like HASOC 2021 and Machinate Marathi hate speech detection corpus and present an effective data augmentation strategy. We observe that models fine-tuned on both datasets fail to classify some common sentences correctly. In the future, we would like to investigate the reason for this phenomenon [02].

Tharindu Ranasinghe, et.al. (2022) – In this research work presented, Marathi continues to be an underresourced language for many NLP-related tasks. Sub track aimed to encourage the development of ML models that were capable of performing well, regardless of having a limited amount of available data. The results of Subtrack 3 indicate that traditional approaches can still be effective when given a small training set. Nevertheless, ensemble architectures and cross-lingual transfer learning would likely surpass the performance of traditional approaches if said training sets were supplemented with additional instances in Marathi or Hindi. Authors suspect these to become common approaches within future iterations of this shared-task [03].

Abhishek Velankaret. al. (2022) - The presented a hate speech dataset containing 25000 distinct samples equally distributed in 4 classes. This is the first major dataset in the domain of hate speech. We also provide the binary version of the dataset of over 37500 samples. We further perform experiments to obtain baseline results on various deep learning models like CNN, LSTM, BiLSTM, and transformer-based BERT models such as Indic BERT, mBERT and RoBERTa. The dataset is also evaluated on monolingual Marathi BERT models like Maha BERT, Maha ALBERT, and Maha RoBERTa. For CNN and LSTM based models, the non-trainable fast text mode outperforms its trainable counterpart in both binary and 4class classification. In transformer-based models, Maha BERT and Maha RoBERTa give the best results in binary and 4-class classification respectively [04].

Abhishek Velankar et. al. (2022) - In this research work, a comparison between monolingual and multilingual transformer-based models, particularly the variants of BERT. We have evaluated these models on hate speech detection and text classification datasets. We have used standard multilingual models namely m BERT, indic BERT, and xlm-RoBERTa for evaluation. On the other hand, we have used Marathi monolingual models trained exclusively on large Marathi corpus i.e. Maha BERT, Maha AlBERT, and Maha RoBERTa for comparison. The Maha AlBERT model performs the best in the case of simple text classification whereas Maha RoBERTa gives the best results for hate speech detection tasks. The monolingual versions for all the datasets have outperformed the standard multilingual models when focused on single language tasks. The monolingual models also provide better sentence representations. However, these sentence representations do not generalize well across the tasks, thus highlighting the need for better sentence embedding models [05].

Saurabh Gaikwad, et. al. (2021) - In this research work, The widespread presence of offensive language on social media motivated the development of systems capable of recognizing such content automatically. Apart from a few notable exceptions, most research on automatic offensive language identification has dealt with English. To address this shortcoming, we introduce MOLD1, the Marathi Offensive Language Dataset. MOLD is the first dataset of its kind compiled for Marathi, thus opening a new domain for research in low resource IndoAryan languages. Author present results from several machine learning experiments on this dataset, including zero-short and other transfer learning experiments on stateofthe-art cross-lingual transformers from existing data in Bengali, English, and Hind [06].

Disha Gajbhiye, et. al. (2021) - This research work, Hate speech content has become a significant issue in today's world. Hate speech detection is an automated task of detecting textual content that contains discriminatory language regarding a person or group based on who they are, their race, gender, caste, etc. In this paper, we discuss the models submitted by our team, Mind Benders, for Marathi subtask A, for "Hate Speech and Offensive Content Identification in English and IndoAryan Languages (HASOC)" at Forum for Information Retrieval Evaluation. A training and test dataset in Marathi language containing 1874 and 625 tweets, respectively, were shared by the HASOC organizers. Using these datasets, we propose an approach to automatically classify the tweets into two categories: "NOT" (Non-HateOffensive) and "HOF" (Hate and Offensive). The classification models developed are applied to the test dataset. They are

experimented with to predict the categories of respective test data [07].

Atharva Kulkarni, et. al.(2021) - This research work as a result, Sentiment analysis is one of the most fundamental tasks in Natural Language Processing. Popular languages like English, Arabic, Russian, Mandarin, and also Indian languages such as Hindi, Bengali, Tamil have seen a significant amount of work in this area. However, the Marathi language which is the third most popular language in India still lags behind due to the absence of proper datasets. In this paper, we present the first major publicly available Marathi Sentiment Analysis Dataset - L3CubeMahaSent. It is curate using tweets extracted from various Maharashtra personalities' Twitter accounts [08].

III. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that deals with the interaction between computers and humans using natural language. The goal of NLP is to enable machines to understand, interpret, and generate human language [21].

NLP involves developing algorithms and models that can process human language in a way that is similar to how humans process it. This includes tasks such as language translation, sentiment analysis, text classification, speech recognition, and text summarization [23].

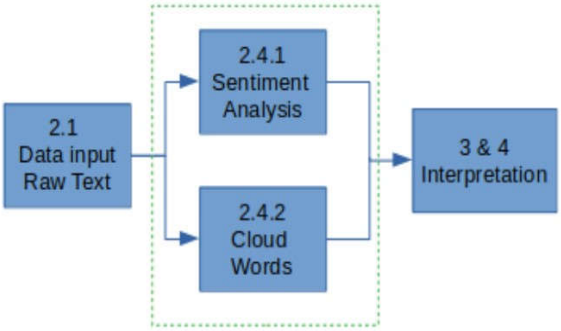

Fig .1 Natural Language Processing

NLP involves a wide range of techniques and approaches, including rule-based systems, statistical models, and deep learning methods. Some popular NLP frameworks and libraries include NLTK, spaCy, and Transformers [25].

NLP has many real-world applications, such as chat bots, virtual assistants, sentiment analysis of customer feedback, and language translation. With the increasing amount of data generated in the form of text, NLP is becoming an increasingly important field in AI research and development. NLP is amongst the most complicated techniques in the world of artificial intelligence, text mining findings are inputs for NLP [17, 18]. NLP's capacity is that humans can speak words. It is the method of converting natural language output (spoken or written) into usable results. NLP is an exciting challenge because it requires computer and human interaction to implement it [15]. NLP is a field of computer studies concerned with studying and understanding the link between computers and the human language [19]. These tools help developers design practical tech applications. Several areas of interest have been established in NLP. Therefore, the core areas' most important activities concentrate on mining named persons, extracting knowledge from texts, translating texts between languages, summarizing written works, inferring answers by inference algorithms, and classifying and clustering papers [20]. It is common practice to discuss theoretical ideas in an academic setting. NLP is a subset of data science that uses dynamic mathematical computations and statistics. In the previous, naive Bayes, k-nearest neighbors, hidden Markov structures, conditional random fields (CRFs), decision trees, random forests, and support vector machines were used by great ML tools [20].

A. Semantic Analysis

In NLP, it is studied how to use NLP strategies to users' emotions and decide what users are expressing through them. Culture may affect this area differently, and This could be misinterpreted if it has taken too literally. "This new gadget is bad!" Although it was evident that the title alludes to the user's dislike of the gadget, the title might endorse the gadget to a particular age group of the community. The sentiment analysis will determine the time at which you express your opinion. To gather statements on a time axis can provide a better insight into peoples' feelings, Facebook and Twitter both provide challenges and opportunities for social movements. On the positive side, it allows people to express and express themselves freely [22]. The records can be carefully observed for a specified time to study trends. The data will provide a preponderance of the evidence that support the researcher's hypothesis. With the advent of the Internet, many fields of research have chosen to gather data from the web. Companies like Google, 22 YouTube, and Amazon know how to customize the content for the customer's best interests. Depending on objective metrics such as social media likes, the number of consumers, and sales. So, there is little data to study this topic. It is challenging because of a: using different languages on one topic or blog, b: using non-standard words that cannot be found in a dictionary, and c: using emoji and symbols. These are questions relevant to both the emotion and sentiment analysis domain [24]. There is a need to provide social scientists and psychiatrists the necessary vocabulary and tools to analyse the web's content and get the necessary data. This work intends to advance this area of study. The paper describes the following [26]:

- Define NLP and explain its scope of application.

- To explain NLP in traditional and statistical ways are Great things you have to accomplish.
- The concept of sentiment analysis gained significant attention in recent times.

Tell the reader briefly about how NLP concepts and ideas can be applied to mental health issues and sentiment analysis

IV. CHALLENGES

Offensive Marathi language detection using machine learning can be a challenging task due to several factors:

- **Data Availability**: One of the biggest challenges in developing an offensive Marathi language detection system is the lack of a large, diverse dataset of offensive Marathi language. A model is only as good as the data it is trained on, and if there is not enough data available, it can be difficult to accurately train a model [28].
- **Complex Language Structure:** Marathi language has a complex structure with multiple variations of words and sentence formations. Offensive language can be constructed in different ways, and therefore it can be challenging to identify the exact structure of the sentence that may contain offensive language.
- **Cultural and Regional Variations:** Marathi language is spoken in different regions and cultures, and the understanding of what constitutes offensive language can vary significantly. A model trained on one region's dataset may not perform well when applied to another region.
- **Constantly Evolving Language**: Language and the usage of language are constantly evolving, with new words and expressions entering the language all the time. A model trained on a static dataset may not be able to adapt to new language variations and trends, making it less effective over time.
- **Variations in Language**: Marathi language is spoken in various regions and countries, and there are often differences in the language depending on the region. The differences can include variations in grammar, pronunciation, and vocabulary. This can make it challenging to detect offensive language accurately.
- **Sarcasm and Irony:** Marathi language has a rich tradition of sarcasm and irony, which can make it difficult for automated systems to distinguish between actual offensive content and sarcastic comments.
- **Contextual Understanding:** In Marathi language, the context of a sentence is critical in determining whether it is offensive or not. The same word can have different meanings depending on the context. Therefore, understanding the context of a sentence is necessary to detect offensive language accurately.
- **Limited Training Dat**a: There is a limited amount of publicly available annotated training data for offensive language detection in Marathi, which makes it challenging to develop accurate machine learning models [32].
- **User Behavior:** Offensive language in Marathi can be highly dependent on user behavior, culture, and regional differences. Therefore, detecting offensive language accurately requires an understanding of these factors [31].
- **Morphological Complexity:** Marathi language has a complex morphology, with a large number of inflections and grammatical cases. The use of inflections and cases changes the meaning of the word, making it challenging to detect offensive language accurately [29].

Overall, the challenges of offensive Marathi language detection using machine learning require a combination of advanced natural language processing techniques, a diverse and constantly updated dataset, and a deep understanding of the cultural and linguistic nuances of Marathi language to develop an accurate and effective model [30].

V. CONCLUSION

In conclusion, detecting offensive language in the Marathi language using machine learning is a complex task due to the language's diverse dialects, informal speech, and cultural nuances. Despite these challenges, this review highlights that machine learning techniques, such as natural language processing (NLP) approaches, have shown significant potential in this domain. The application of advanced models like BERT, MahaTweet BERT, and other transformer-based architectures has yielded promising results, particularly when fine-tuned on datasets like HASOC and MOLD. Additionally, the integration of semantic analysis, contextual understanding, and sentiment analysis further enhances the accuracy of detecting offensive language.

Key methods proposed to address these challenges include building larger and more diverse datasets, particularly focusing on informal speech, dialects, and evolving language patterns. Leveraging models pre-trained on large Marathi corpora, like Maha BERT and its variants, has demonstrated superior performance in tasks like hate speech detection and text classification. Furthermore, the study of user behavior, sarcasm, irony, and morphological complexity requires ongoing dataset updates and model retraining to ensure high accuracy over time. Ultimately, this review suggests that while offensive language detection in Marathi is a challenging task, the use of advanced machine learning techniques and continuous improvements in data collection

and model refinement offer a promising path forward in automating the detection of offensive content.

VI. REFRENCES

[1] Onkar Litake, Maithili Sabane, Parth Patil, Aparna Ranade, and Raviraj Joshi "Mono Versus Multilingual BERT: A Case Study in Hindi and Marathi Named Entity Recognition" ,volume 540, 24 February 2023.

[2] Tanmay Chavan, Shantanu Patankar, Aditya Kane, Omkar Gokhaleand Raviraj Joshi "A Twitter BERT Approach for Offensive Language Detection in Marathi" 20 Dec 2022.

[3] Tharindu Ranasinghe, Kai North, Damith Premasiri and Marcos Zampieri "Overview of the HASOC Subtrack at FIRE 2022: Offensive Language Identification in Marathi" 18 Nov 2022.

[4] Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi "L3CubeMahaHate: A Tweet-based Marathi Hate Speech Detection Dataset and BERT models" 22 May 2022.

[5] Abhishek Velankar, Hrushikesh Patil, Raviraj Joshi "Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi" 19 Apr 2022.

[6] Saurabh Gaikwad1 , Tharindu Ranasinghe2 , Marcos Zampieri1 , Christopher M. Homan "Cross-lingual Offensive Language Identification for Low Resource Languages: The Case of Marathi" 8 Sep 2021

[7] Disha Gajbhiye, Swapnil Deshpande, Prerna Ghante, Abhijeet Kale and Deptii Chaudhari "Machine Learning Models for Hate Speech Identification in Marathi Language" 2021.

[8] Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Josh "L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset" 22 Apr 2021.

[9] K. E. Riehm, K. A. Feder, K. N. Tormohlen, R. M. Crum, A. S. Young, K. M. Green, L. R. Pacek, L. N. La Flair, R. Mojtabai, Associations Between Time Spent Using Social Media and Internalizing and Externalizing Problems Among US Youth, JAMA Psychiatry 76 (2019) 1266–1273.

[10] L. P. Dinu, I.-B. Iordache, A. S. Uban, M. Zampieri, A Computational Exploration of Pejorative Language in Social Media, in: Proceedings of EMNLP, 2021.

[11] M. Yao, C. Chelmis, D.-S. Zois, Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media, in: Proceedings of WWW, 2019..

[12] J. Shetty, K. N. Chaithali, A. M. Shetty, B. Varsha, V. Puthran, Cyber-Bullying Detection: A Comparative Analysis of Twitter Data, in: N. N. Chiplunkar, T. Fukao (Eds.), Advances in Artificial Intelligence and Data Engineering, Springer Singapore, Singapore, 2021, pp. 841–855

[13] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, A. Alothaim, Online extremism detection in textual content: A systematic literature review, IEEE Access 9 (2021) 42384–42396.

[14] I. Kwok, Y. Wang, Locate the Hate: Detecting Tweets against Blacks, in: Proceedings of AAAI, 2013

[15] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, W. Daelemans, A Dictionary-based Approach to Racism Detection in Dutch Social Media, in: Proceedings of TA-COS, 2016

[16] Lee, S.H.; Chan, C.S.; Wilkin, P.; Remagnino, P. Deep-plant: Plant identification with convolutional neural networks. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 452–456.

[17] segmentation and multiclass support vector machine. In Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, ON, Canada, 30 April–3 May 2017; pp. 1–4.

[18] M. L. Williams, P. Burnap, A. Javed, H. Liu, S. Ozalp, Hate in the Machine: Anti-Black and AntiMuslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime, The British Journal of Criminology 60 (2019).

[19] A. John, A. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, K. Hawton, Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review, Journal of Medical Internet Research 20 (2018)

[20] A. Akins, Facebook's Oversight Board Overrules 4 Hate Speech, Misinformation Takedowns, SNL Kagan Media and Communications Report (2021).

[21] K. Dinakar, R. Reichart, H. Lieberman, Modeling the Detection of Textual Cyberbullying, in: Proceedings of ICWSM, 2011.

[22] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting Offensive Language in Social Media to Protect Adolescent Online Safety, in: Proceedings of ASE, 2012.

[23] J. Salminen, H. Almerekhi, M. Milenkovic, S.-g. Jung, A. Jisun, H. Kwak, B. J. Jansen, Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media, in: Proceedings of ICWSM, 2018.

[24] A. Bellmore, Learning from Bullying Traces in Social Media, in: Proceedings of NAACL, 2012.

[25] M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, Improving Cyberbullying Detection with User Context, in: Proceedings of ECIR, 2013.

[26] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive Language Detection in Online User Content, in: Proceedings of WWW, 2016.

[27] K. Nugroho, E. Noersasongko, Purwanto, Muljono, A. Z. Fanani, Affandy, R. S. Basuki, Improving Random Forest Method to Detect Hatespeech and Offensive Word, in: Proceedings of ICOIACT, 2019.

[28] B. Wang, Y. Ding, S. Liu, X. Zhou, YNU_Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language, in: Proceedings of FIRE, 2019.

[29] [22] M. A. Bashar, R. Nayak, QutNocturnal at HASOC 2019: CNN for Hate Speech and Offensive Content Identification in Hindi Language, in: Proceedings of FIRE, 2019.

[30] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, HateMonitors: Language Agnostic Abuse Detection in Social Media, in: Proceedings of FIRE, 2019.

[31] H. Hettiarachchi, T. Ranasinghe, Emoji Powered Capsule Network to Detect Type and Target of Offensive Posts in Social Media, in: Proceedings of RANLP, 2019.

R. Kumar, B. Lahiri, A. K. Ojha, A. Bansal, ComMA at HASOC 2020: Exploring Multilingual Joint Training across different Classification Tasks, in: Proceedings of FIRE, 2020.