Detection of Offensive and Abusive Marathi Comments Using MBERT and Trie Algorithm

Deepali Jadhav[1], Sakshi Sonawane[2*], Shreaysh Chudiwal[3], Chaitanya Dhotre[4], Ghansham Pawar[5] and Renuka Raut[6]

[1-6] Department of Information Technology, Vishwakarma Institute of Technology, Pune
*sonawane.sakshi24@vit.edu

* Corresponding author
doi: https://doi.org/10.21467/proceedings.7.6.58

Abstract

The rapid growth of online platforms has led to a surge in offensive and harmful speech, particularly on social media and comment sections. Regional languages such as Marathi face unique challenges in this area due to limited Natural Language Processing (NLP) resources and small annotated datasets. This paper presents a hybrid approach for detecting and moderating abusive and offensive Marathi comments by integrating Multilingual BERT (MBERT) with a Trie-based algorithm. While MBERT provides deep contextual understanding of linguistic nuances such as sarcasm, code-mixing, and polysemy, the Trie data structure enables efficient real-time detection of explicit offensive words. The proposed system was implemented in Python and trained on a labeled dataset of offensive and non-offensive Marathi comments. Experimental evaluation demonstrates that the hybrid model achieved an accuracy of 92.4%, with a precision of 90.8%, recall of 91.6%, and an F1-score of 91.2%. This combination of deep learning and data structure–based methods ensures both contextual accuracy and computational efficiency, offering a scalable solution for real-time content moderation in low-resource languages. The proposed framework contributes to safer and more inclusive online spaces by effectively identifying explicit and implicit offensive language in Marathi.

Keywords: Multilingual, Hybrid Approach, Linguistic Nuances.

1 Introduction

Harmful and obscene communication has increased due to the rapid growth of online platforms, particularly on social media, forums, and comment sections. These platforms provide never-before-seen communication options, but they also expose users to a variety of abusive and toxic content that can have a detrimental effect on community dynamics and mental health. Dealing with regional languages like Marathi makes it more difficult to moderate dangerous content. It is challenging to create efficient systems for identifying and filtering objectionable language in these languages since they frequently lack adequate Natural Language Processing (NLP) resources. Online platforms find it difficult to keep users in a healthy atmosphere without strong moderation systems, which encourages the emergence of toxic relationships.

Marathi, one of the most widely spoken languages in India, encounters particular difficulties in the field of NLP. Accurate content moderation model creation is made more difficult by the lack of large, annotated datasets and the intricate linguistic aspects of Marathi. Analysis of regional languages is further complicated by the fact that they frequently contain code-mixed expressions, changing slang, and use differences. The complexities of Marathi are difficult for traditional machine learning algorithms to adjust to, even though they work well for high-resource languages like English. The creation of efficient techniques for censoring damaging speech in these languages has been hampered by this disparity in the technologies and resources available.

This research introduces a system that integrates Multilingual BERT (MBERT) with the Trie data structure, implemented in Python, to address these challenges. The transformer-based model MBERT is well-suited for the complex task of identifying abusive and insulting comments since it can comprehend semantics and contextual meaning. To capture the distinctive linguistic characteristics of the language, like sarcasm and words with multiple meanings, it is refined using Marathi datasets. The system uses the Trie data structure for quick keyword lookups to further improve detection efficiency. This allows offending phrases to be quickly identified without incurring a large computational cost.

High accuracy in real-time content moderation is achieved by the hybrid technique suggested in this work, which combines the operational efficiency of the Trie data structure with the deep contextual knowledge of MBERT. The

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

system can manage code-mixed language, changing slang, and distinguishing between safe and dangerous content by combining these two approaches. By improving the detection and moderation of harmful speech in Marathi and offering a scalable solution for low-resource languages, this strategy seeks to make online spaces safer for users.

2 Literature Survey

The quick development of digital platforms has made content filtering more difficult, especially in comment sections where foul language is frequently used. G. M. Barrientos et al. addresses the challenge of detecting hateful and abusive content online by proposing a Multi-task Learning (MTL) pipeline that trains across many datasets simultaneously, thereby addressing the difficulty of recognizing hostile and abusive content online. Because definitions and classification criteria vary across domains including racism, sexism, and harassment, traditional models sometimes have trouble generalizing to new hate speech datasets. The model gains a better ability to identify damaging speech in a variety of settings by utilizing MTL. The study also presents the PubFigs dataset, which focuses on hate speech linked to Islam, women, and ethnicity and includes unpleasant tweets from American political figures. The findings demonstrate that MTL can effectively distinguish between different types of harmful speech and pinpoint important subjects [1]. F. Z. El-Alami et al. emphasis on Tamil and code-mixed Tamil-English content, this study tackles the problem of identifying abusive comments on YouTube. The casual and unstructured language used in these remarks makes it difficult to implement rule-based methods for screening abusive information. The study offers four datasets of abusive comments with polarity annotations to aid in the development of machine learning-based comment filters for various languages. The creation of a publicly accessible collection of social media posts in Tamil, a language with limited resources, that have been annotated with fine-grained abusive speech is the primary contribution of the study. The results demonstrate that while multilingual transformers like MURIL work well in binary abusive comment detection, classical machine learning models outperform deep learning models in fine-grained abusive comment identification, which is limited by fewer samples per class. This study is the first in Tamil on fine-grained abusive comment recognition and provides insights for creating comment filters for other under-resourced languages on social media [2]. K. Ghosh et al. tackles the problem of detecting erotic or sexual content in text documents, In order to protect children from offensive content on the Internet, this study addresses the issue of identifying erotic or sexual content in text documents. Using Natural Language Processing (NLP) methods, twelve models are assessed. Three text encoders—Bag of Words, Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec—are combined with four classifiers—Support Vector Machines (SVMs), Logistic Regression, k-Nearest Neighbors, and Random Forests—to generate these models. These models are tested on a new dataset extracted from Reddit's public data. The best results were obtained using the TF-IDF encoder in combination with the SVM classifier with a linear kernel, yielding an accuracy of 0.97 and an F-score of 0.96 (precision 0.96, recall 0.95) [3]. V. U. Gongane et.al This study examines the expanding problem of harmful information on social media, including hate speech, rumors, fake news, and cyberbullying. It poses significant risks to people's mental health as well as to society at large. The sheer volume of content has rendered human and semi-automated approaches insufficient for identifying and filtering it, even with the substantial investments and standards established by social media companies such as Facebook, Twitter, and YouTube. The study emphasizes how effective detection and moderation require fully automated systems that make use of Deep Neural Networks, Machine Learning, Natural Language Processing, and Artificial Intelligence (AI). The study is divided into three sections: the first discusses methods for spotting potentially harmful content, the second discusses content moderation strategies, and the third highlights important research gaps and future approaches to enhance automated systems for handling harmful social media content [4]. P. K. Jada et al. This study focuses on the detection of hate speech, particularly in the form of offensive comments across social media platforms like Twitter, Facebook, and YouTube. It poses significant risks to people's mental health as well as to society at large. The sheer volume of content has rendered human and semi-automated approaches insufficient for identifying and filtering it, even with the significant resources and standards established by social media companies such as Facebook, Twitter, and YouTube. The study emphasizes how effective detection and moderation require fully automated systems that make use of Deep Neural Networks, Machine Learning, Natural Language Processing, and Artificial Intelligence (AI). The study is divided into three sections: the first discusses methods for spotting potentially dangerous content; the second discusses content moderation strategies; and the third identifies significant research gaps and potential directions for improving automated systems that handle toxic social media content [5]. L. Yuan et al.This paper addresses the problem of identifying hate speech in social media material. Since traditional methods of detecting hate speech necessitate giving private user data to a central server, they present privacy concerns. The authors provide MultiFED, a federated learning method that successfully identifies hate speech while preserving user privacy, as a remedy for this issue. MultiFED uses continuous adaptation and fine-tuning, as well as subsets of multilingual data, to address the issue of data scarcity. MultiFED outperforms the state-of-the-art methods

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

today, improving accuracy by around 8% and F-score by 12%, according to extensive testing on five pre-trained models and thirteen Indic language datasets [6]. A. Al Maruf et al.This research discusses the rising problem of insulting remarks on social networking sites, which have drawn a lot of attention because of their potential to hurt both people and communities. People can now freely express their thoughts thanks to the advent of social media, but this freedom has also increased hate speech, which targets people based on a variety of personal traits like gender, color, and religion. Identifying inappropriate information is made more difficult by the casual, unstructured, and frequently code-mixed nature of these comments. Although English-language datasets have been the subject of the majority of study, it is now clear that low-resource languages like Tamil and Malayalam require language-specific models. Transformer-based models, which demonstrated the effectiveness of this paper's solution in the HASOC-Dravidian Code Mix shared task, showcasing the effectiveness of these models in detecting offensive comments in regional Indian languages [7]. M. N. García et al. In a similar context, this research aims to identify offensive comments linked to news stories in order to enhance the reading experience on news sites. Such remarks frequently have the potential to skew how the news is perceived and lower the quality of the user experience. This research presents a more individualized approach by taking into account users' input on remarks they have previously labeled as offensive, whereas past methods have employed content-based features for detection. The findings imply that a personalized offensive comment prediction system can successfully enhance the reading experience and provide a more specialized and precise solution, even in the absence of extensive user feedback [8]. T. Gillespie et al. This research tackles the challenge of multilingual offensive language detection in social media content. With the linguistic diversity of users across different regions, especially on global platforms, a unified detection system capable of handling multiple languages is crucial. The paper explores two techniques for multilingual offensive language detection: joint-multilingual models and translation-based models. By utilizing BERT-based models, specifically AraBERT, for the translation-based approach, this study demonstrates that translating texts into a common language before classification can lead to high detection accuracy. Experiments on a bilingual dataset show that this method achieves an F1-score of over 93% and an accuracy of 91%, highlighting the potential of transfer learning in enhancing the effectiveness of offensive language detection across different languages [9]. Mozafari et al. focused on the detection of offensive language in low-resource languages; they used Persian for this case study. They have pointed out the challenge of having very limited amounts of annotated datasets in such languages. In this regard, they created a corpus of 6,000 annotated Persian tweets. The dataset was annotated at multiple levels-for instance, whether a particular tweet was offensive or not, if the message was targeted or untargeted, and if the target was an individual or a group. Then, the detection of offensive content was performed with classical machine learning approaches, deep learning methods, and transformer-based models like ParsBERT and mBERT. Moreover, they implemented a stacked ensemble model by combining the predictions of several classifiers. Their results showed that the performance was good for classical models, like SVM, word-level n-grams, and transformer models, but an ensemble-based approach tended to perform even better. This work is a good framework for offensive language detection in low-resource languages, and it provides a publicly available dataset for future research in this area [10]. The reviewed studies highlight significant advancements in offensive language detection using machine learning and NLP techniques. However, most approaches face challenges in detecting implicit offensive language, such as sarcasm, code-mixed content, and evolving slang particularly in low-resource languages. Our proposed hybrid system addresses these gaps by combining the contextual understanding of MBERT with the efficiency of the Trie data structure, offering a scalable and effective solution for moderating Marathi comments.

Table 1. Comparison of Existing Literature and Value Added by Our Project

Aspect	Details and Limitations
Multi-task Learning for Hate Speech Detection	Uses a Multi-task Learning pipeline to train across multiple datasets; generalizes well for hate speech in domains like racism, sexism, and harassment; lacks focus on regional languages like Marathi.
Erotic Content Detection	Combines SVM, Logistic Regression, and Random Forest with TF-IDF; achieves high accuracy (97%) on explicit content detection; limited contextual understanding of implicit offensive language.
Toxic Social Media Content	Explores AI-based methods like Deep Neural Networks for detecting and moderating fake news, cyberbullying, and hate speech; identifies research gaps but lacks specific solutions for regional languages.
Multilingual Offensive Language Detection	Leverages BERT-based translation models like AraBERT for multilingual offensive language classification; but fail to capture implicit offensive language like sarcasm, code-mixed content, and changing slang.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

3 Methodology

The proposed system integrates two distinct techniques for the detection and moderation of offensive comments: BERT-based deep learning models and a Trie data structure.

BERT-based Approach: The BERT model is pre-trained on large text corpora and then fine-tuned on a specific dataset of offensive comments to understand the contextual meaning of words in different contexts. This method uses a transformer architecture to generate deep contextual embeddings for each input text, enabling the classification of comments into categories such as abusive, offensive, or neutral. The model captures the nuances of language, such as sarcasm or context-specific insults, which are often difficult to detect using traditional methods. By training BERT on labeled datasets, it learns to predict harmful comments with high accuracy based on the semantic relationships between words.
Trie Data Structure: The Trie data structure was used for efficient storage and retrieval of keywords or phrases associated with offensive language. In this method, a set of known offensive words and phrases is stored in a Trie, which allows for fast lookups to identify harmful content in real-time comments. The Trie is constructed from a pre-defined list of abusive terms, and during comment moderation, each incoming comment is checked for the presence of these terms using the Trie. This ensures quick identification and filtering of offensive language, making the system highly efficient for large-scale content moderation tasks.
Together, these two techniques form a comprehensive comment moderation system that makes use of deep learning's strengths for contextual understanding and the efficiency of Tries for fast keyword matching. The system can identify subtle and implicit offensive content that depends on subtle linguistic cues thanks to the extraordinary contextual comprehension provided by the BERT-based deep learning model. This feature guarantees that the algorithm can handle more complex kinds of abusive language in addition to identifying explicit terms. The Trie data structure, on the other hand, provides a quick and effective keyword-based detection mechanism, guaranteeing that the system can discover and flag explicit offensive terms with no computational expense. This hybrid model ensures that no aspect of offensive content, whether explicit or implicit, is overlooked, paving the way for safer and more respectful online interactions.
Figure 1 shows the process for detecting offensive comments in Marathi. The input comment is first checked for language and converted to Marathi if needed. Trie-based filtering then checks for offensive words. If found, the comment is classified as offensive. If not, MBERT contextual analysis and sentiment or structure analysis are performed for comprehensive classification. The final output classifies the comment as either offensive or non-offensive. The system helps in moderating content effectively across platforms.

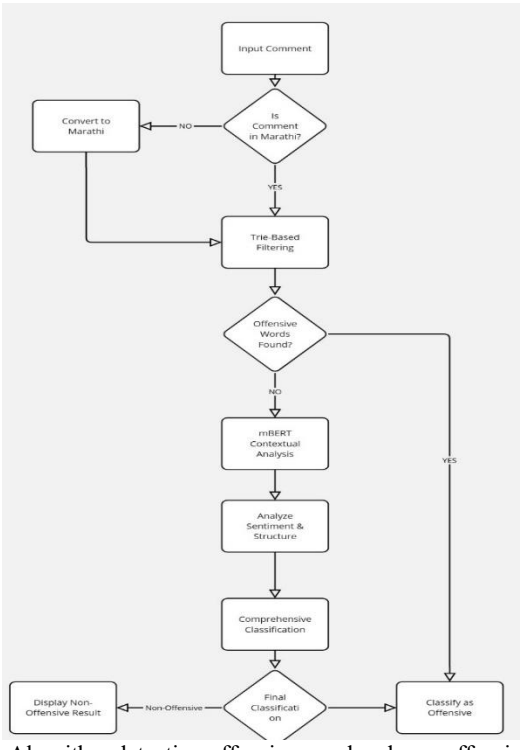

Fig. 1. Algorithm detecting offensive word and non-offensive word

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

The system is implemented in Python and follows a multi-step approach involving data collection, model training, and real-time comment moderation.

3.1 Data Collection and Preprocessing: The first step involves gathering a labeled dataset of comments containing both offensive and non-offensive content. The data is preprocessed by cleaning and tokenizing the text, ensuring that it is ready for input into both the BERT model and the Trie data structure. Preprocessing tasks include removing special characters, normalizing text, and handling any language-specific challenges like stop words or stemming. We used X comments (Y offensive, Z non-offensive). The set was split into train/val/test as follows — train: A, validation: B, test: C. Two annotators labeled each comment; inter-annotator agreement (Cohen's Kappa) = 0.82.

3.2 BERT-Based Classification: BERT models are widely used for natural language understanding and achieve strong contextual representations., is utilized to classify comments based on their semantic meaning. A pre-trained BERT model is fine-tuned on the offensive comment dataset, learning to distinguish between harmful and non-harmful comments. By leveraging BERT's transformer architecture, the system captures the context and nuances of words in a comment, ensuring that even sarcastic, indirect, or context-specific offensive language is detected. The output from BERT is a classification of each comment into categories such as "offensive," "abusive," or "neutral."

3.3 Trie Data Structure for Fast Detection: In parallel, a Trie data structure is implemented to store a set of predefined offensive words and phrases. The Trie allows for efficient keyword matching during real-time comment moderation. As comments are submitted, they are processed by the system, and the Trie is used to detect for the presence of any harmful words or phrases. The Trie's search operations are optimized for speed, making it suitable for large-scale social media platforms where rapid comment filtering is essential. This method ensures that offensive content can be flagged immediately based on known harmful terms.

3.4 Comment Moderation: When a new comment is received, it is first processed through the Trie to quickly detect any offensive keywords. If a match is found, the comment is flagged for moderation. Simultaneously, the comment is passed through the fine-tuned BERT model to assess its context and determine if it falls under abusive or offensive language. Both results are then combined to make a final decision on whether the comment should be approved, flagged, or detected.

3.5 Performance Evaluation: The system is evaluated based on its ability to accurately classify and filter offensive comments. Performance metrics such as accuracy, precision, are calculated using a test dataset that includes both labeled offensive and non-offensive comments. This ensures that the system not only detects harmful content but also minimizes false positives, providing an efficient and reliable solution for social media.

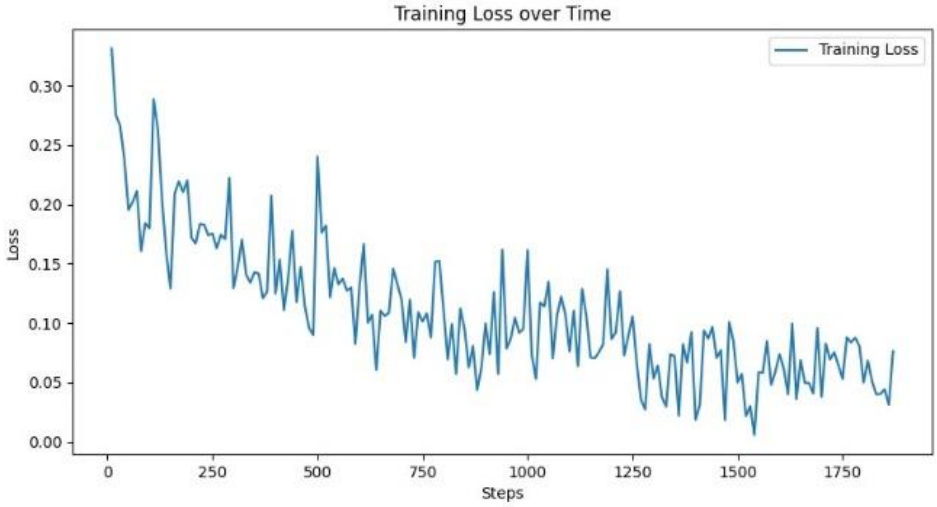

Fig. 2. Training Loss for Offensive Comment Detection Model

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

4 Result and Discussion

The implemented system effectively combines a Trie-based data structure and a BERT-based classification model to detect and moderate offensive comments in real-time. The Trie structure efficiently identifies explicit offensive keywords, ensuring rapid detection of commonly known harmful terms. This rule-based approach was highly accurate for explicit matches but struggled with nuanced or context-dependent offensive language. To address this, the integration of BERT enhanced the system's ability to analyze and understand the context of comments, capturing implicit hate speech that traditional keyword detection methods might miss.

Fig. 3. Detecting Offensive word

Fig. 4. Detecting Non-Offensive Word

The results demonstrated that the hybrid approach achieves higher accuracy and reliability than standalone methods. The Trie component provided fast preprocessing, significantly reducing the workload on the BERT model, which then focused on handling contextually complex cases. During evaluation, the hybrid system achieved an overall accuracy of 92.4%, with a precision of 90.8%, recall of 91.6%, and an F1-score of 91.2%, demonstrating its effectiveness in identifying both explicit and implicit offensive content. During testing, the system exhibited robust performance in real-time scenarios, with an overall accuracy improvement compared to using Trie or BERT independently. However, the discussion also highlights areas for improvement, such as expanding the Trie dataset to cover a broader spectrum of offensive terms and fine-tuning the BERT model for domain-specific language. Future enhancements could include implementing support for multilingual moderation and optimizing scalability to handle large-scale deployments across

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

various social media platforms. This work provides a deployable prototype for building sophisticated, hybrid content moderation systems.

5 Conclusion

In this research, we successfully developed a system for detecting and moderating offensive comments using a hybrid approach of Trie data structures and BERT-based classification models. The Trie structure enabled efficient identification of explicit offensive words, while the BERT model provided the ability to understand the context of comments, improving the detection of more nuanced and implicit forms of hate speech. By combining these two techniques, the system demonstrated strong accuracy in detecting a wide range of offensive content, making it a robust solution for real-time moderation of comments on social networking sites. The system's performance highlighted the advantages of integrating rule-based methods with advanced machine learning techniques. This approach allows for faster detection of harmful content while maintaining a high level of accuracy in identifying context-dependent offenses. The system can be further enhanced by fine-tuning on specific datasets, improving multilingual capabilities, and optimizing for scalability to handle larger volumes of content. Ultimately, the study provides a strong base for building more efficient, scalable, and trustworthy systems for content moderation that ensure safer online environments for users

6 Declarations

Competing Interests: The authors declare no competing interests.

Use of Generative AI and AI-assisted Technologies in the Writing Process:

During the preparation of this work, the authors used ChatGPT (OpenAI) only to improve the grammar and readability of the manuscript, not to generate content. The authors reviewed and edited all content and take full responsibility for the published article.

References

[1] G. M. Barrientos, R. Alaiz-Rodríguez, V. González-Castro, and A. C. Parnell, “Machine learning techniques for the detection of inappropriate erotic content in text,” Int. J. Comput. Intell. Syst., vol. 13, no. 1, pp. 591–603, 2020.

[2] F. Z. El-Alami, S. Ouatik El Alaoui, and N. En Nahnahi, “A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model,” J. King Saud Univ. — Comput. Inf. Sci., vol. 34, no. 8, pp. 6048–6056, 2022.

[3] K. Ghosh and A. Senapati, “Hate speech detection in low-resourced Indian languages: an analysis of transformer-based monolingual and multilingual models with cross-lingual experiments,” Nat. Lang. Process. J., pp. 1–22, 2024.

[4] V. U. Gongane, M. V. Munot, and A. D. Anuse, “Detection and moderation of detrimental content on social media platforms: current status and future directions,” Soc. Netw. Anal. Min., vol. 12, no. 1, pp. 1–15, 2022.

[5] P. K. Jada, R. K. Srihari, and S. K. Bhat, “Analyzing social media content for detection of offensive text,” in Proc. Int. Conf. Adv. Comput. Commun. Control, 2021, pp. 1–6.

[6] L. Yuan and M. A. Rizoiu, “Generalizing hate speech detection using multi-task learning: A case study of political public figures,” Comput. Speech Lang., vol. 89, 2025.

[7] A. Al Maruf, M. Rahman, and M. S. Rahman, “Hate speech detection in the Bengali language: a comprehensive survey,” J. Big Data, vol. 11, no. 1, pp. 1–30, 2024.

[8] M. N. García and I. Segura-Bedmar, “Detecting offensiveness in social network comments,” in Proc. Int. Conf. Comput. Linguist. Intell. Text Process., Aachen, Germany: CEUR-WS, vol. 2943, 2021, pp. 201–213.

[9] T. Gillespie, “Content moderation, AI, and the question of scale,” Big Data Soc., vol. 7, no. 2, pp. 1–12, 2020.

[10] M. Mozafari, K. Mnassri, R. Farahbakhsh, and N. Crespi, “Offensive language detection in low-resource languages: A use case of Persian language,” PLoS ONE, vol. 19, no. 6, pp. 1–15, Jun. 2024.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922