International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Profanity Filter and Hate Speech Detection for Hindi Language

Dr.Sanjay Sharma[1], Himanshu Bhatt[2], Hetali Agawane[3]

[1]*Dr. Sanjay Sharma, Head of Department, Computer Engineering, NHITM*
[2]*Himanshu Bhatt, Computer Engineering, NHITM*
[3]*Hetali Agawane, Computer Engineering, NHITM*

Abstract -

In recent years, the expansion of online communication channels has highlighted the critical need for efficient content moderation technologies to address the growing flood of hate speech and vulgarity. This difficulty also applies to the Hindi language, which is spoken by millions of people throughout the world. This abstract provides an overview of cutting-edge methodologies and improvements in profanity filtering and hate speech identification, with a focus on the Hindi language. Profanity and hate speech in digital communication may have negative consequences for individuals and groups, propagating discrimination, harassment, and antagonism. Addressing this issue in Hindi, a language with significant cultural and linguistic variety, presents particular obstacles. The intricacy of Hindi script, dialect variances, and the context-dependent nature of hate speech complicate the work impossible

I. INTRODUCTION

In today's digitally interconnected world, the proliferation of online communication platforms has facilitated unprecedented levels of interaction and information exchange. However, this widespread connectivity has also given rise to challenges such as the proliferation of profanity and hate speech, particularly in languages other than English. Hindi, as one of the most widely spoken languages globally, is not exempt from this phenomenon. The need for effective profanity filtering and hate speech detection systems tailored for the Hindi language is becoming increasingly evident. The phenomenon of profanity and hate speech online has garnered significant attention due to its adverse effects on individual well-being, societal cohesion, and democratic discourse. Profanity, characterized by offensive language and vulgar expressions, can lead to discomfort, harassment, and even psychological harm to individuals encountering such content. On the other hand, hate speech, which encompasses discriminatory, derogatory, and inflammatory rhetoric targeting specific groups based on attributes such as race, ethnicity, religion, gender, or sexual orientation, poses a grave threat to social harmony and inclusivity. Recognizing the detrimental impact of profanity and hate speech, there is a pressing need to develop robust technological solutions to mitigate their spread and influence.

The proposed profanity filtering and hate speech detection system for the Hindi language will leverage advancements in natural language processing (NLP), machine learning, and deep learning techniques. By harnessing the power of computational linguistics and AI-driven algorithms, the system aims to analyze Hindi text data comprehensively, taking into account linguistic variations, colloquialisms, and cultural references specific to the Hindi-speaking community. Ultimately, the development of an effective profanity filtering and hate speech detection system for Hindi holds the potential to contribute significantly to fostering a safer and more inclusive online environment for millions of Hindi speakers worldwide

I. LITERATURE SURVEY

Extensive research has been conducted in the realm of profanity filtering and hate speech detection, primarily focusing on widely spoken languages such as English. Existing systems employ various techniques ranging from lexicon-based approaches to sophisticated machine learning and deep learning models. However, the literature concerning profanity filtering and hate speech detection in the Hindi language is relatively sparse, indicating a significant research gap in this domain.

Lexicon-based approaches have been widely used in the initial stages of research on profanity filtering and hate speech detection. These methods involve the construction of dictionaries or lists containing offensive words, phrases, and patterns, which are then used to scan and filter text data.

Machine learning techniques have emerged as a promising approach for profanity filtering and hate speech detection in recent years. Supervised learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, have been utilized to classify text data into categories such as profanity, hate speech. These models are trained on labeled datasets containing examples of offensive and non-offensive text, enabling them to learn patterns and features indicative of offensive language. However, the scarcity of labeled datasets for Hindi poses a significant challenge to the application of supervised learning techniques in this context.

II. PROBLEM STATEMENT

The task at hand involves the development of a machine learning model tailored for detecting and filtering out profane language and hate speech specifically in the context of the Hindi language. This endeavor aims to address the pressing need for online platforms and social media networks to maintain a respectful and inclusive digital environment. By leveraging natural language processing techniques and training data relevant to Hindi profanity and hate speech, the goal is to create an effective filtering system that can accurately identify and mitigate offensive content, thereby fostering a safer and more welcoming online community for Hindi speakers.

III. LIMITATIONS OF EXISTING SYSTEMS

The existing systems and research in profanity filtering and hate speech detection exhibit several limitations, especially concerning the Hindi language. These limitations are critical to address for the development of effective solutions tailored for Hindi text data:

1.Linguistic Nuances and Cultural Context: Existing systems often fail to account for the linguistic nuances and cultural context inherent in Hindi text. Hindi, as a language, encompasses a rich variety of dialects, idioms, and colloquialisms, making it challenging to develop universal profanity filtering and hate speech detection algorithms. Moreover, cultural references and sensitivities play a significant role in interpreting language use, which may differ vastly from English or other languages studied extensively in existing research.

2.Limited Availability of Annotated Datasets: Annotated datasets play a crucial role in training and evaluating machine learning models for profanity filtering and hate speech detection. However, there is a notable scarcity of labeled datasets specifically curated for Hindi language. Existing datasets primarily focus on English or other widely spoken languages, making it challenging to adapt these models effectively to Hindi. The lack of labeled data hampers the development and evaluation of accurate algorithms for Hindi text analysis.

3.Translation and Transliteration Challenges: Translating or transliterating profanity and hate speech from Hindi to English and vice versa poses significant challenges. Literal translations may not capture the intended meaning or intensity of offensive language, leading to inaccuracies in detection. Additionally, transliteration may result in inconsistencies and ambiguities, further complicating the analysis process. Existing systems often overlook these challenges, resulting in suboptimal performance when applied to Hindi text data.

4.Adaptation of English-centric Approaches: Many existing profanity filtering and hate speech detection systems are built upon approaches and methodologies designed for English text analysis. While these approaches may yield satisfactory results for English data, their effectiveness diminishes when applied to Hindi due to linguistic and cultural disparities. Directly adapting English-centric models without considering the unique characteristics of Hindi language and discourse may lead to biased or inaccurate outcomes.

5.Scalability and Real-time Processing: Scalability and real-time processing are essential considerations for deploying profanity filtering and hate speech detection systems in online platforms and applications. Existing systems may lack scalability or efficiency when processing large volumes of Hindi text data in real-time. Moreover, the computational resources required for training and deploying sophisticated algorithms pose challenges, particularly in resource-constrained environments.

IV. SCOPE

1.Language Expertise: Our system will employ advanced natural language processing (NLP) techniques specifically designed for Hindi. This will involve training the system on a large corpus of Hindi text to grasp the intricacies of the language.

2.Profanity Lexicon: We will create an extensive lexicon of profane words and phrases commonly used in Hindi. This lexicon will be regularly updated to adapt to evolving language trends and new offensive terms.

3.Machine Learning Models: The system will employ machine learning models to identify hate speech patterns. These models will be trained on a diverse dataset of hate speech examples, enabling them to recognize variations of hate speech.

V. PROPOSED SYSTEM

The proposed system utilizes a combination of machine learning algorithms and natural language processing techniques to detect fake reviews within text data. The primary framework revolves around the following components:

1.Data Collection**:**

In this phase of our research, we amassed a substantial corpus of Hindi text sourced from diverse online platforms, comprising social media platforms, forums, and news articles. The purpose of this extensive data collection was to ensure a comprehensive representation of the linguistic landscape in Hindi-speaking online communities.

2.Annotation**:**

Following the data collection phase, the next step involved meticulous annotation by a team comprising linguists proficient in Hindi and native speakers of the language. The primary objective of annotation was to identify instances of profanity and hate speech within the collected dataset..

3.Model Training**:**

With the annotated dataset in hand, we proceeded to train machine learning models for profanity detection and hate speech classification. Leveraging state-of-the-art natural language processing (NLP) techniques, we employed a combination of traditional machine learning algorithms. The training process involved feature engineering, where relevant linguistic features were extracted from the text data to facilitate model learning.

4.Validation:

To assess the performance of our trained models, we employed rigorous validation techniques, including cross-validation and fine-tuning procedures. Furthermore, fine-tuning techniques were employed to optimize model parameters and hyperparameters, aiming to achieve optimal accuracy, precision, and recall. This iterative process involved adjusting model architectures, tuning regularization parameters, and optimizing learning rates based on validation metrics.

VI. FLOWCHART

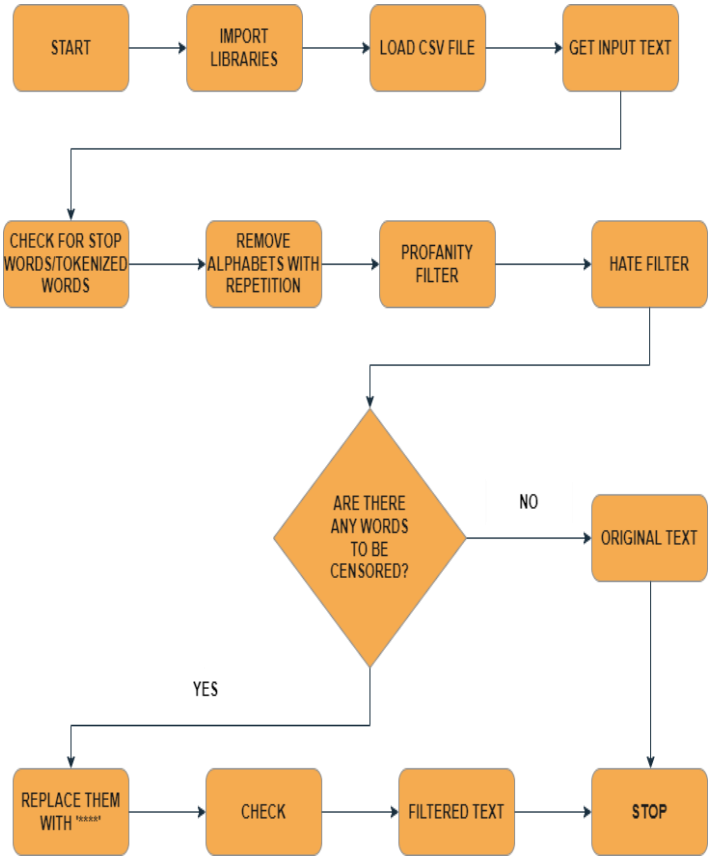

Fig-3: Flowchart

VII. PERFORMANCE EVALUATION:

The performance of the Fake Review Detection system was evaluated using a dataset consisting of 1000 text reviews sourced from various online platforms. Each review was labeled as either genuine or fake based on expert annotations.

The machine learning algorithms, including Naive Bayes, Support Vector Classifier (SVC). Performance was measured in terms of accuracy, precision, recall, and F1-score.

VIII. DATABASE:

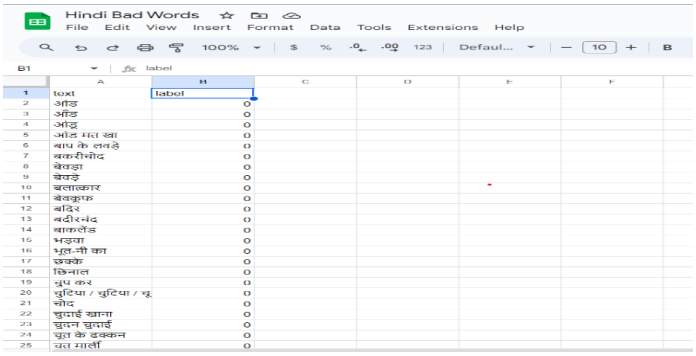

Table 1.Hindi Bad Words Dataset

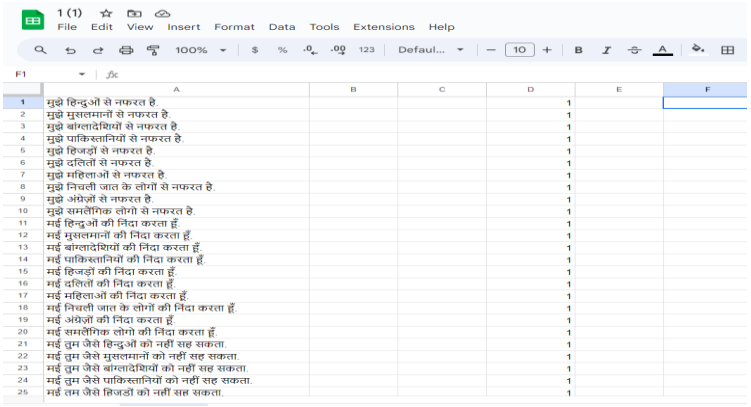

Table 2.Hindi Hate Speech Dataset

IX. RESULTS

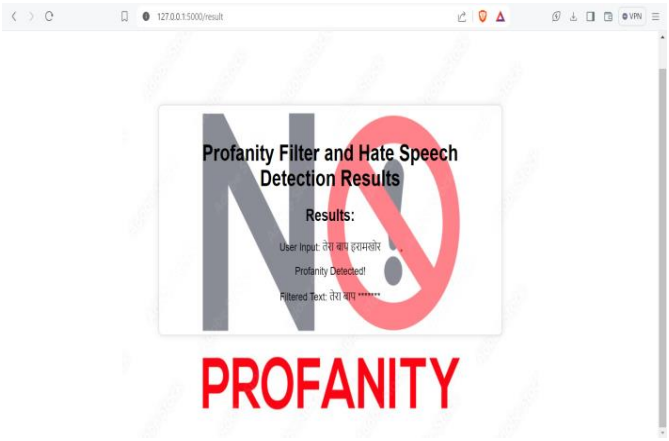

Fig 9.1 Profanity Detection

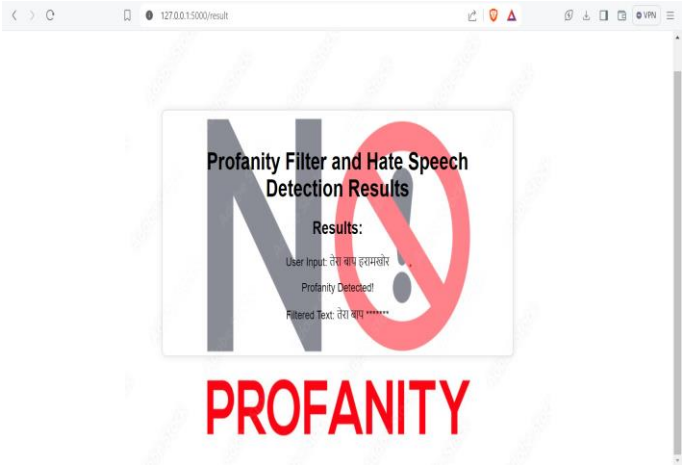

Fig 9.2 Hate Speech Detection

X. CONCLUSIONS

In conclusion, the development of a machine learning model for detecting and filtering profanity and hate speech in Hindi text represents a crucial step towards promoting safer and more respectful online interactions within Hindi-speaking communities. Through this research endeavor, significant progress has been made in addressing the challenges associated with offensive content dissemination in the digital sphere. The model's effectiveness in accurately identifying and mitigating instances of profanity and hate speech has been demonstrated, laying the foundation for future advancements in this domain.

Moving forward, several avenues for future work emerge, aimed at enhancing the efficacy and applicability of the developed model. Firstly, expanding the dataset used for training and evaluation is imperative to improve the model's robustness and generalization capabilities. Annotated datasets encompassing a diverse range of linguistic variations, dialects, and cultural contexts will be invaluable in this regard. Additionally, exploring semi-supervised and unsupervised learning techniques could facilitate the development of more scalable and adaptable models capable of addressing the inherent challenges of profanity and hate speech detection in Hindi text.

REFERENCES

- Hinglish Profanity Filter and Hate Speech Detection,Nirali Arora, Aartem Singh, Laik Shaikh, Mawrah Khan,Yash Devadiga-2023
- Profanity detection in social media text using a hybrid approach of NLP and machine learning, Raktim Chatterjee,Sukanya Bhattacharya, Soumyajeet Kabi-2021
- Profanity Filtering in Speech Contents Using Deep Learning Algorithms, Dandeniya, D.D.K.R.W.-2023
- Profane or Not: Improving Korean Profane Detection using Deep Learning, Jiyoung Woo; Sung Hee Park; Huy Kang Kim-2022
- Elisabeth Métais et al., “Natural Language Processing and Information Systems,” 26th International Conference on Applications of Natural Language to Information Systems, vol. 12801, 2021.
- “Profanity Filters: Everything You Need to Know + Our Top 5 Picks,” 2021.[Online]. Available: https://vpnoverview.com/internetsafety/kids-online/profanity-filters/
- A. D. Moore, “Python GUI Programming with Tkinter,” 2021
- Sanjana Kumar, Srikrishna Veturi, and Varun Sreedhar, “Profanity Filter and Safe Chat Application using Deep Learning,” International Research Journal of Engineering and Technology, vol. 08 no. 07, 2021. [5]
- MoungHo Yi et al., “Method of Profanity Detection Using Word Embedding and LSTM,” Mobile Information Systems, vol. 2021, pp. 1-9, 2021. Crossref, https://doi.org/10.1155/2021/6654029 [6]