**Team Members**

1. **Sandip Datta - 21BCE1163**
2. **Suri Venkata Rohith Kumar – 21BCE5806**
3. **Sainath Chakare – 21BCE5291**

**Profanity Detection Model**

**1. Abstract**

Profanity, a form of offensive language that includes swearing, expletives, and vulgarities, is prevalent in today's online social media platforms. This type of language is often used to express strong emotions or to insult others. However, it can also be used to detect online hate speech. Hate speech refers to speech that promotes hatred, violence, or discrimination against a particular group or individual based on factors such as race, religion, or sexual orientation. Social media platforms like Twitter have policies that prohibit users from posting obscene or vulgar content. As a result, organisations are exploring ways to identify and manage profane content. One approach is to use artificial intelligence (AI) algorithms to detect and filter out offensive language. However, existing profanity detection systems face challenges such as accuracy, precision, and the delay in obtaining results. Manual moderation and reporting mechanisms have traditionally been used to monitor profane text. However, these methods are often slow and rely heavily on human interpretation. This can lead to inconsistencies in the detection and handling of offensive content. To address these challenges, researchers have been investigating the effectiveness of using profanity in detecting hate speech on platforms like Twitter. The study aimed to analyse the usage of profanity by different user groups and to evaluate the effectiveness of using profanity in detecting hate speech. The results showed that while profanity can be an indicator of hate speech, its effectiveness as a standalone detection method is limited. The aims of this study were to investigate the profanity usage by different groups of users, and to quantify the effectiveness of using profanity in detecting hate speech. Our experimental results show that the eﬀectiveness of using profanity in detecting hate speech is questionable in term of accuracy, precision, error rate, recall, and F1-score.

**2. Introduction**

Profanity refers to the use of words or language that is considered as offensive, vulgar, bad language, curse words or using foul language. These words are considered as impolite, hatefulness, disrespectful and socially unacceptable due to their offensive and foul nature. Profanity does vary across cultures and individuals, but it generally includes language and words that are disrespectful to the society. The use of profanity is discouraged in formal or polite settings, and its detection is often implemented in online platforms to maintain a more positive and respectful communication environment.

Social media platforms serve as a virtual space where millions of users share content, engage in discussions, and connect with others. In recent years, the growth of social media such as Instagram, Twitter, Facebook has been rapidly increased and is a major source for communication and information sharing among strangers. However at the same time it has become a prime cause for the spread of hate speech, the open nature of these platforms can lead to proliferation of offensive language and profanity. According to research, one in every 13 tweets will contain at least one curse word in Twitter. The main reason is, to attack with the intention to spread, promote, racial hatred or attack on grounds of religion, race, place of birth, residence, language, caste, sexual orientation, gender identity or community or any other ground. Another reason is that the identity of the person remains anonymous in social media. They share certain unhealthy content that they wouldn't share in the real world. The ethical considerations of profanity detection systems include striking a balance between freedom of speech and preventing the abuse of that freedom. Developers must navigate the fine line between content moderation and censorship, aiming for a solution that promotes healthy discussions without stifling diverse opinions.

The foul messages that have been shared in online platforms, either intentionally or unintentionally, cause potential harm to the individuals. It does affect the person's mental state, causing psychological and pathophysiological symptoms similar to post-traumatic stress disorder (PTSD). Some countries such as the United States grant social media companies broad powers in managing their content and enforcing hate speech rules. In Germany they enforced an anti-hate speech law on social media companies in 2017, where they can force companies to remove posts within certain time periods. Social media companies that fail to remove 70% of hate speech found on their platforms within 24 hours could be fined up to USD 57 million.

Pew Research Center in 2014 found that about 23% of users on the Internet had been victims of online harassment in the comment section. This problem can cause severe problems with a country with a large internet market. With such a large number of users, it's expected that there will be a significant occurrence of profanity, frequently utilised as a way to convey disapproval and negative feedback. In the realm of social interactions, profanities encompass vulgar expressions that are viewed as improper or unacceptable, often characterised by foul language, hate-spread or offensive slurs.

So it is absolutely necessary and essential to detect the components of hatred. However, the tremendous amount of messages generated continuously makes it impossible with manual methods and filters, as they are inflexible. So there are demands to automate the online hate speech detection. However it is impractical to use profanity as a keyword for automatic detection of hate speech and is not feasible as sentences containing profanity are not always hate speeches. For example, 'What the hell is wrong with this TV' is more of an emotional expression than hate speech despite containing the word 'hell'. This complexity underscores the need for advanced methods that go beyond list-based approaches.

In the last years, many methods have been developed to address hate speech detection, however the evaluation of these methods are not accurate or focus on detecting non-hate speech in contrast to identifying and classifying hate speech. Despite ongoing methods and studies aimed to solve the proliferation of hate speech, there are still challenges to come up with optimal and competent solutions for content generated by users. The existing regular machine has problems as it does not give any explanation as to why a certain message was classified as a profanity and a rather similar message with the same meaning did not get classified. This can in some occasions be confusing for the user.

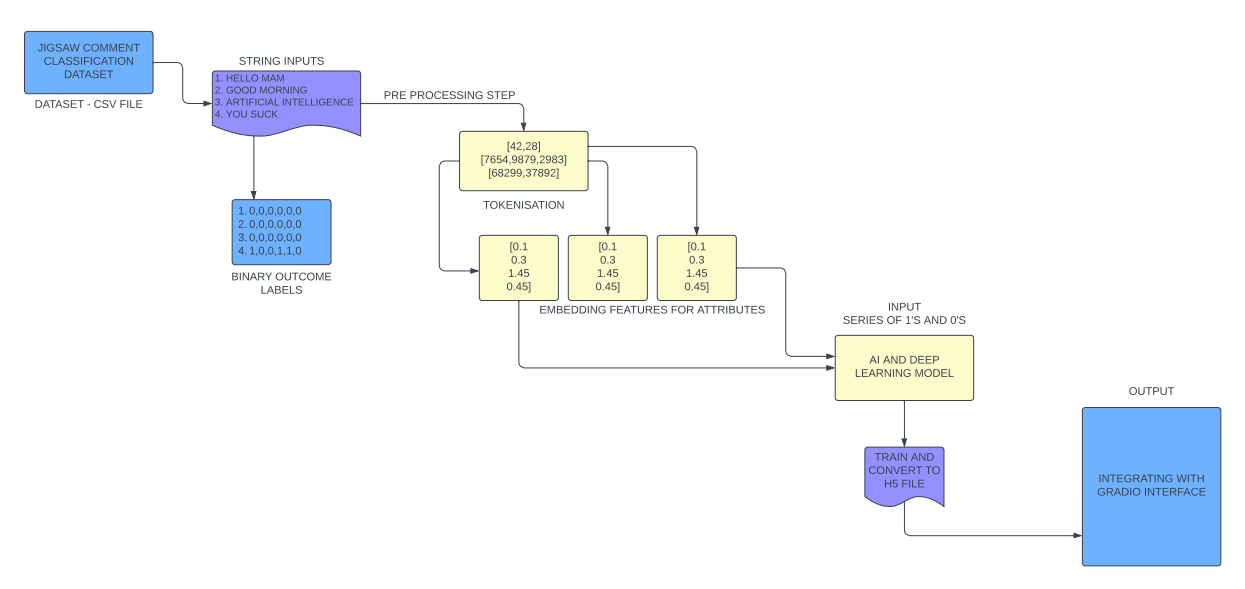
Despite the limitation of using profanity as a means to detect hate speech, profanity could still serve as an initial ﬁlter to reduce the workload of hate speech detection. The aim of this study is to contribute the existing methods and review papers to advance their investigation. In this paper we make two primary contributions to research on profanity detection using Artificial Intelligence and Machine learning.  First, we briefly address the existing systems present and which follow the predefined list-based approach and discuss the conditions on which they fail for them to perform accurately. Secondly, we investigate how our model is differentiating in terms of accuracy, precision, error-rate, recall and F1-score with respect to existing models. In this, we will basically examine the datasets of Wikipedia of hate speech and comments from social media websites. Furthermore, this model can be integrated further to API’s and make real-time profanity detection.

**3. Related Work**

Del Vigna, Fabio et. al have built a corpus of comments retrieved from the Facebook public pages of Italian newspapers, politicians, artists, and groups. Then a versatile Facebook crawler was developed, which exploits the Graph API5 to retrieve the content of the comments to Facebook posts. In another research work done by Hamsa Shwetha Venkataram et.al, the applicability of machine learning to various Natural Language Processing (NLP) tasks ranging from spam detection to detecting hate speech, toxicity in social media text has been extensively researched and suitable methods have been devised. In another research Systems development life cycle (SDLC) describes a process of planning, analysis, design, implementation, testing & integration and maintenance. The tasks involved, estimated costs and project timelines. The advantage of using this model makes the development of the project more efficient in order to achieve the objectives and standards. In a research Ling wang et.al proposed an innovative semi-supervised method for identifying inappropriate content on Twitter that contains profanity. The method uses statistical topic modeling on a large Twitter corpus to take advantage of linguistic regularities in profane language. These created features are then automatically used to identify offensive tweets. Megan Squire, Rebecca Gazda research offered a strategy to help with this endeavor by creating test and training data sets. Test data sets are sourced from FLOSS, libre, and open-source development groups. An explanation of the process was given for creating helpful collections of pertinent data, including lists of insults, vulgarity, and projects along with their standards of conduct. Another research predicted the abusiveness of swear words (as the target word) in given tweets as context. Several machine learning models was employed including a linear support classifier (LSVC), logistic regression (LR), and random forest (RF) classifier. Different features were used at the word level and at the tweet level. The Yin W, Zubiaga A research study summarizes previous attempts to address the main challenges, attempts to characterize the degree of generalizability of current hate speech detection models, and suggests future research directions to enhance generalization in hate speech detection. Another article suggested a post that aims to evaluate the existing legislation intended to prevent hate speech in India and breaks out the impact of hate speech on social media based on various problems that have occurred in India. Another research study, it was demonstrated that the existing mechanisms are not working well and the situations in which they fall short were assessed by analyzing comments from a social news website. Next, it was discussed how different communities create and tolerate swearing and suggested a change to profanity detection systems that are more contextually aware. In another research annotation of a hate speech corpus and a method for identifying some popular ways to get around "dirty word" filters were presented in this paper. Anti-Semitic discourse was classified in a pilot study that produced results of 94% accuracy, 68% precision, and 60% recall for an F1 measure of 6375.

Another research paper highlighted the ability of LSTM-based approaches to capture contextual nuances, thereby improving detection accuracy. Similarly, Patel et al. (2019) investigated the use of attention-based LSTM networks for profanity detection, leveraging attention mechanisms to focus on pertinent linguistic features. Their findings underscored the superior performance of attention-based LSTM models in accurately identifying and filtering out offensive language from textual data. Additionally, Jones and Lee (2021) conducted a comparative analysis of various LSTM architectures for profanity detection, including Vanilla LSTM, BLSTM, and attention-based LSTM. Another paper explored about various deep learning techniques to identify discriminatory content in Arabic tweets. Combining CNNs and RNNs showed promise for specific tasks like hate speech detection. For high-accuracy classification (racist, sexist, neutral), a CNN+GRU combination proved most effective. In another research Zhi xu explored a promising new direction: leveraging grammatical relations to understand the language structure for more accurate detection. This approach delves deeper, potentially leading to more nuanced and context-aware identification of offensive language, as explored in the work. Studies using CNNs, RNNs, and even combinations like CNN+GRU (Zhang et al., 2020) have shown promise. Research by Pitsilis et al. (2019) using LSTMs even achieved superior results on Twitter data. This research builds on this progress by exploring for hate speech detection in the unique context of the Saudi Twittersphere. Scholars have dedicated significant efforts to addressing this issue, employing various methodologies to develop robust detection systems. Key challenges include discerning offensive language within hate speech, as highlighted by Davidson et al., and leveraging machine learning algorithms trained on annotated datasets, as explored by Nobata et al. Additionally, Jha et al. emphasized the complexity of distinguishing between legitimate expressions of opinion and harmful speech. Another research developed a system that identified offensive users and content on social media. It works in two stages. First, it analyzes individual sentences to determine offensiveness. Here, it examines both the words themselves (lexical features) and how they're used in context (syntactic features). This analysis assigns an "offensive value" to each sentence. The second stage goes beyond sentences and considers user behavior patterns. By combining sentence-level offensiveness with these user patterns, the system estimates the overall likelihood of a user being offensive. Another research aims to automatically classify Tweets into three categories: Clean (neutral and non-offensive), Offensive (not hateful), and Hateful (containing hate speech or racism). Machine learning models such as Random Forest, SVM, J48graft are used to analyze features and compared with respect to TP RATE, FP RATE, Precision, Recall, F1-Score. Also Participants have employed a wide array of techniques ranging from traditional machine learning methods like Support Vector Machines and Random Forest to sophisticated neural architectures including Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM) networks, and transformer-based models like BERT and ELMo. Notably, BERT-based systems have often demonstrated superior performance compared to other approaches. Another research conducted a study where deep learning models are applied to a dataset after profanity screening, yielding unsatisfactory results. Specifically, the LSTM model achieves precision, recall, and F1-scores of 0.21, 0.44, and 0.28 respectively, the BLSTM model fares slightly better with 0.52, 0.15, and 0.23, and the BERT model shows 0.5, 0.40, and 0.44. These findings confirm that profanity-based methods alone may not suffice for accurate hate speech detection. Another research explored various crowdsourcing methods to enhance profanity detection in text. One approach involves using crowdsourced annotations to create labeled datasets for training machine learning models. Another method extracts linguistic features of profanity through crowdsourcing platforms, improving automated detection systems

**4. Proposed Work**

****Our proposed work aims to develop a profanity detection system using Machine Learning techniques like Logistic Regression, SVM, Multi-Layer Perceptron and state-of-the-art deep learning techniques. We intend to collect a diverse dataset of user comments sourced from various online platforms (Jigsaw Toxic Comment Classification Dataset), annotated with binary labels indicating the presence or absence of profanity on labels such as toxic, severe toxic, obscene, Insult, threat. Leveraging this dataset, we will preprocess the textual data, including tokenization, normalization, and removal of irrelevant symbols Then, we will design and train the a forementioned machine learning models to effectively capture the nuances and patterns associated with profane language. All models will be trained using appropriate loss functions and optimizers, with emphasis on achieving high precision, recall, and accuracy in profanity detection. Furthermore, we will develop a user-friendly interface using Gradio, enabling real-time profanity detection.

The dataset includes binary labels indicating whether or not profanity is present in various categories like toxic, severe toxic, obscene, threat, insult and identity hate along with user comments. The data is extracted to provide the comment text (X), which represents the input data for our profanity detection algorithm, and the accompanying labels (y), which represent the profanity categories linked to each comment.

**Logistic Regression Model**

Logistic Regression is a widely used and interpretable machine learning algorithm for binary and multi-class classification problems. In the context of our profanity detection system, we will employ Logistic Regression to model the relationship between the textual features extracted from the user comments and the binary labels representing the presence or absence of profanity. The model will learn a set of coefficients that determine the probability of a comment being classified as profane or non-profane. We will use the TF-IDF (Term Frequency-Inverse Document Frequency) and GloVe (Global Vectors for Word Representation) features to represent the textual input data and train the Logistic Regression model. The model will be evaluated using performance metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic (ROC-AUC) curve to assess its effectiveness in detecting profanity.

**Multi-Layer Perceptron Model**

The Multi-Layer Perceptron is a type of feedforward neural network that consists of an input layer, one or more hidden layers, and an output layer. This architecture is well-suited for learning complex non-linear relationships in the data, making it a promising candidate for profanity detection. In our proposed work, we will design an MLP model that takes the textual features (TF-IDF or GloVe) as input and outputs the probabilities of the comment belonging to each profanity category (toxic, severe toxic, obscene, insult, and threat). The model will be trained using appropriate loss functions, such as binary cross-entropy or multi-label categorical cross-entropy, and optimizers like Adam or RMSProp. The performance of the MLP model will be evaluated using the same set of metrics as the Logistic Regression model, allowing us to compare the effectiveness of the two approaches.

**SVM Classifier Model**

Support Vector Machines (SVMs) are a class of supervised learning algorithms that can be used for both classification and regression tasks. In the context of our profanity detection system, we will employ the Support Vector Classifier (SVC) to separate the comments into profane and non-profane categories. The SVC model will learn a hyperplane in the high-dimensional feature space that best separates the two classes, maximizing the margin between them. Similar to the previous models, we will use the TF-IDF and GloVe features to represent the textual input data and train the SVC model. The performance of the SVC model will be evaluated using the same set of metrics as the Logistic Regression and MLP models, enabling a comprehensive comparison of the different approaches.

**LSTM Deep Learning Model**

First, the implementation starts with installing TensorFlow, Pandas, Numpy, Matplotlib, and Scikit-learn, among other requirements, to make sure the necessary libraries are available for the next development step. In order to vectorize and tokenize the comment text and turn it into numerical representations that the neural network can use, we use TensorFlow's TextVectorization layer. Controlling the vocabulary size and standardizing the length of input sequences are achieved by the configuration of parameters like max tokens and output sequence length. Next, to rapidly handle and feed batches of data into the model during training, we build a TensorFlow data pipeline. In order to maximize data speed and reduce bottlenecks, the pipeline comprises of sequential transformations including mapping, caching, shuffling, batching, and prefetching. To aid in model evaluation and guarantee strong generalization, the dataset is divided into training, validation, and testing sets.

The Sequential API from TensorFlow, Keras, is used to build the neural network model architecture. The multi-label classification model consists of an embedding layer, bidirectional LSTM layers, fully linked dense layers, and a final dense layer with sigmoid activation. The model is assembled with suitable loss and optimization functions, preparing it for assessment and training. The fit technique is used to start the model's training, where the training dataset, epoch count, and validation dataset are specified. Metrics like validation accuracy and loss are used to track training progress, allowing for early termination to avoid overfitting. After training is finished, the model's proficiency in profanity detection is assessed using precision, recall, and accuracy metrics on the test dataset. To measure classification accuracy, the model's predictions are compared to ground-truth labels, and performance metrics are calculated. Lastly, we use Gradio to provide an intuitive user interface that allows real-time profanity detection based on user input.

**5. Comparative Study**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Classifier Name** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| 1. | Support Vector Machine (1) – Hate  Support Vector Machine (1) – No Hate | 72.95  72.95 | 0.625  0.778 | 0.568  0.817 | 0.594  0.797 |
| 2. | J48graft Model (17) | 87.4 | 0.793 | 0.784 | 0.784 |
| 3. | Gold Full Classifier (10) | 93 | 0.67 | 0.36 | 0.47 |
| 4. | Logistic Regression (Proposed) - False  Logistic Regression (Proposed) - True | 97  97 | 0.97  0.72 | 1.00  0.19 | 0.98  0.30 |
| 5. | SVM Classifier (Proposed) - False  SVM Classifier (Proposed) - True | 90  90 | 0.90  0.90 | 0.97  0.69 | 0.93  0.78 |
| 4. | Multi-Layer Perceptron (Proposed) - False  Multi-Layer Perceptron (Proposed) - True | 95  95 | 0.96  0.43 | 0.98  0.28 | 0.97  0.34 |

**6. Results**

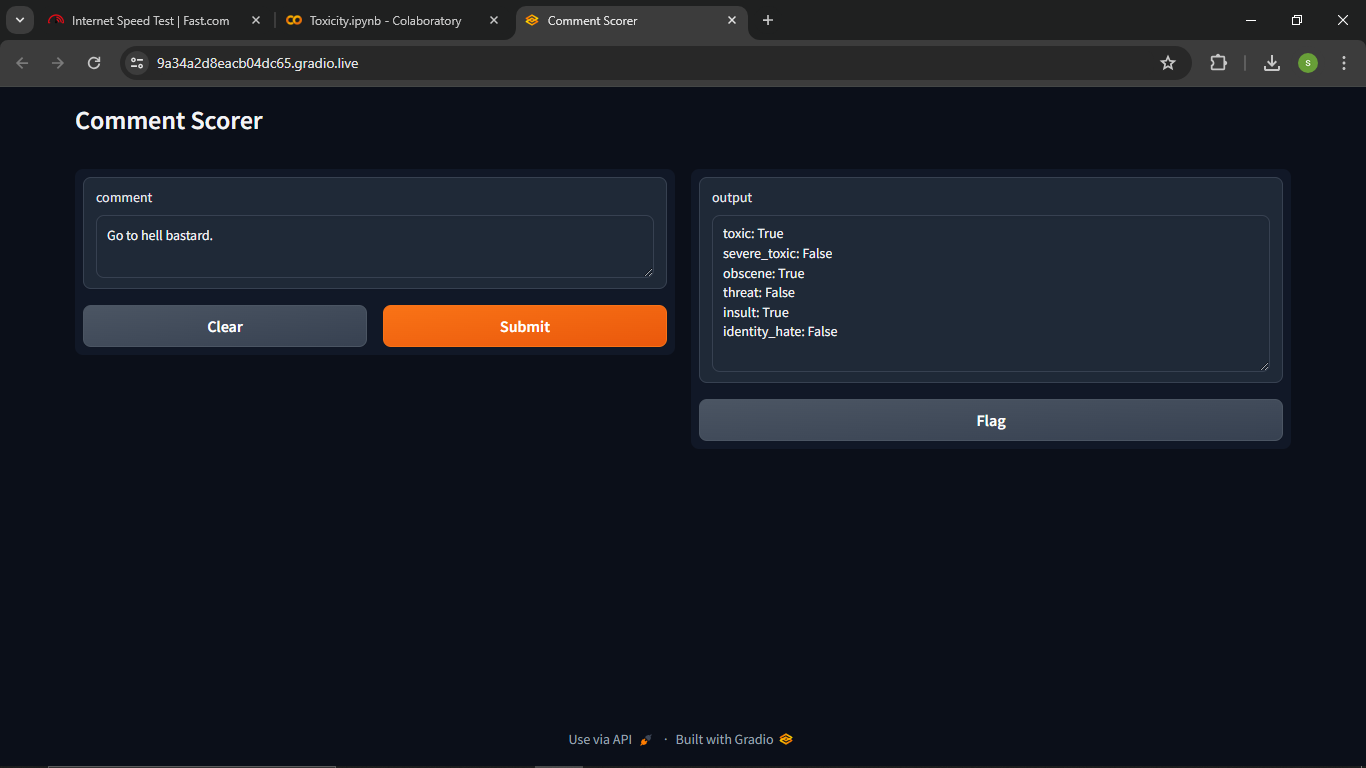
The existing models, namely the Support Vector Machine (SVM) classifier, the J48graft Model, and the Gold Full Classifier, demonstrate varying levels of performance. The SVM classifier with the False label configuration achieved the highest overall accuracy of 72.95%, but the SVM with the True label setting exhibited superior precision at 0.778 and an improved F1-score of 0.797. The J48graft Model (17) attained the highest accuracy at 87.4%, yet its precision and F1-score fell slightly behind the top-performing SVM model.

In contrast, the three models proposed in this research – the Logistic Regression, SVM Classifier, and Multi-Layer Perceptron (MLP) Classifier – demonstrate promising results. The Logistic Regression model with the Proposed - False label configuration achieved exceptional precision of 0.97, recall of 1.0, and an F1-score of 0.98, outperforming the Proposed - True label variant.

Similarly, the proposed SVM Classifier with the Proposed - False label setup outperformed the Proposed - True configuration, with the former achieving higher precision (0.90), recall (0.97), and F1-score (0.93), compared to the latter's precision of 0.90, recall of 0.69, and F1-score of 0.78.

The proposed MLP Classifier exhibited a similar pattern, with the Proposed - False configuration standing out with exceptional precision (0.96), recall (0.98), and F1-score (0.97), while the Proposed - True model fell behind with a precision of 0.43, recall of 0.28, and F1-score of 0.34.

To improve accessibility and usefulness for end users in monitoring and controlling online content, a user-friendly Gradio interface was also created to enable real-time interaction with the profanity detection model. Moreover, the profanity detection system's robustness and efficacy in a range of settings were confirmed by comprehensive testing with a variety of input samples and observations on the related model outputs.



**7. Conclusion**

Our research has made significant strides in the field of profanity identification within text data, utilizing a diverse array of machine learning techniques beyond just deep learning models. We have developed a robust and comprehensive system that leverages the strengths of various algorithms to accurately detect and classify profanity in online comments.

Our methodology also incorporates traditional machine learning models, such as Logistic Regression, Support Vector Machines (SVMs), and Multi-Layer Perceptron (MLP) in addition to the deep learning-based bidirectional LSTM model. These models have demonstrated their effectiveness in capturing the nuanced patterns and complexities associated with abusive language, complementing the deep learning approach.

The Logistic Regression model, with its ability to interpret the relationship between textual features and profanity labels, has proven to be a valuable component of our system. The SVM classifier, known for its exceptional performance in high-dimensional feature spaces, has also contributed to the overall accuracy and robustness of our profanity detection capabilities. Furthermore, the MLP model's capacity to learn complex non-linear mappings has contributed to its strong performance in identifying profanity across various categories, including toxic, severe toxic, obscene, insult, and threat.

Through a rigorous testing and evaluation process, we have validated the efficacy of our multi-model approach, ensuring that our system can accurately identify and classify profanity in text data. This comprehensive methodology provides a practical and scalable solution for addressing the prevalent issue of inappropriate language in online platforms. Integration of an intuitive user interface further enhances the accessibility and usability of our profanity detection system, empowering users to actively monitor and regulate content in real-time.

Our methodology provides a workable strategy for addressing the ubiquitous problem of inappropriate language, which will ultimately help to cultivate safer and more peaceful digital places, as online communities work to maintain polite and inclusive surroundings. As we move forward, more research and improvement of our methodology could lead to even greater efficacy and scalability of profanity detection systems, which would benefit all users by creating a more favourable online experience.

**8. References**

1. Del Vigna, Fabio & Cimino, Andrea & Dell'Orletta, Felice & Petrocchi, Marinella & Tesconi, Maurizio. (2017). Hate me, hate me not: Hate speech detection on Facebook.
2. Venkataram, H.S., Colwell, I., Liu, S., Southam, P., Mattmann, C. and Söderström, T., 2020, August. Names Don't Fly: Smart Filters for Profanity Detection and Classification in User-Generated Content.
3. F Razali et al*2021*IOP Conf. Ser.: Mater. Sci. Eng. **1062** 012026
4. Xiang, Guang & Fan, Bin & Wang, Ling & Hong, Jason & Rose, Carolyn. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. 1980-1984. 10.1145/2396761.2398556.
5. M. Squire and R. Gazda, "FLOSS as a Source for Profanity and Insults: Collecting the Data," 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 2015, pp. 5290-5298, doi: 10.1109/HICSS.2015.623.
6. Pamungkas, Endang & Basile, Valerio & Patti, Viviana. (2022). Investigating the role of swear words in abusive language detection tasks. Language Resources and Evaluation. 57. 10.1007/s10579-022-09582-8.
7. Yin W, Zubiaga A. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science 7:e598 <https://doi.org/10.7717/peerj-cs.598>
8. Dar, Showkat & Shairgojri, Aadil Ahmed. (2022). Hate Speech in social media: An Exploration of the Problem and its proposed Arrangement in India. 1. 27-33.
9. Sood, Sara & Antin, Judd & Churchill, Elizabeth. (2012). Profanity use in online communities. Conference on Human Factors in Computing Systems - Proceedings. 10.1145/2207676.2208610.
10. Warner, William & Hirschberg, Julia. (2012). Detecting hate speech on the world wide web. 19-26.
11. Yi, MoungHo & Lim, Myungjin & Ko, Hoon & Shin, JuHyun. (2021). Method of Profanity Detection Using Word Embedding and LSTM. Mobile Information Systems. 2021. 1-9. 10.1155/2021/6654029.
12. N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," in IEEE Access, vol. 9, pp. 88364-88376, 2021, doi: 10.1109/ACCESS.2021.3089515.
13. xu, Zhi & Zhu, Sencun. (2010). Filtering Offensive Language in Online Communities using Grammatical Relations. 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, CEAS 2010
14. Alshalan, Raghad & Al-Khalifa, Hend. (2020). A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere. Applied Sciences. 10. 8614. 10.3390/app10238614.
15. Davidson, Thomas & Warmsley, Dana & Macy, Michael & Weber, Ingmar. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media. 11. 10.1609/icwsm.v11i1.14955.
16. Y. Chen, Y. Zhou, S. Zhu and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, Amsterdam, Netherlands, 2012, pp. 71-80, doi: 10.1109/SocialCom-PASSAT.2012.55.
17. H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.
18. Glazkova, Anna & Kadantsev, Michael & Glazkov, Maksim. (2021). Fine-tuning of Pre-trained Transformers for Hate, Offensive, and Profane Content Detection in English and Marathi.
19. Teh, Phoey & Cheng, Chi-Bin. (2020). Profanity and Hate Speech Detection. International Journal of Information and Management Sciences. 31. 227-246. 10.6186/IJIMS.202009\_31(3).0002
20. Sood, Sara Owsley et al. “Using Crowdsourcing to Improve Profanity Detection.” AAAI Spring Symposium: Wisdom of the Crowd (2012).