**BCSE209L – Machine Learning**

**J Component Report**

**A project report titled**

**Profanity Detection Model**

*By*

Reg. No: **21BCE1163** Name : **Sandip Datta**

Reg. No: **21BCE5806** Name : **Suri Venkata Rohith Kumar**

Reg. No: **21BCE5291** Name: **Sainath Chakare**

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGENEERING

*Submitted to*

**Dr. R. Rajalakshmi**

**School of Computer Science and Engineering**

****

*April 2024*

*­­­*

**DECLARATION BY THE CANDIDATE**

I hereby declare that the report titled “**Profanity Detection Model”** submitted by me to VIT Chennai is a record of bona-fide work undertaken by me under the supervision of **Dr. R. Rajalakshmi, Professor, SCOPE, Vellore Institute of Technology, Chennai.**

Signature of the Candidate(s)

**Sandip Datta**

**Suri Venkata Rohit Kumar**

**Sainath Chakare**

**ACKNOWLEDGEMENT**

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Rajalakshmi,** School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Ganesan, Dean,** School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our **Head of the Department** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

**BONAFIDE CERTIFICATE**

Certified that this project report entitled “**Profanity Detection Model**

**”** is a bona-fide work of **Sandip Datta (21BCE1163), Suri Venkata Rohit Kumar (21BCE5806) , Sainath Chakare (21BCE5291)** carried out the “**Profanity Detection Model**”-Project work under my supervision and guidance for BCSE209L - Machine Learning.

**Dr. R. Rajalakshmi**

SCOPE

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Ch. No** | **Chapter** | **Page Number** |
| 1 | Introduction | 6-7 |
| 2 | Abstract | 8 |
| 3 | Literature Survey | 9-16 |
| 4 | Proposed Methodology | 17-19 |
| 5 | Results and Discussion | 20 |
| 6 | Conclusion | 21 |
| 7 | Reference | 22-23 |
| 8 | Appendix | 24 |

**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.

**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.

**Literature Survey**

1. **Hate me, hate me not: Hate speech detection on Facebook.**

The paper presents a study on hate speech detection on Facebook in the Italian language. The authors propose a taxonomy of hate categories and annotate a corpus of Facebook comments accordingly. They develop two classifiers, one based on Support Vector Machines (SVM) and another on Long Short-Term Memory (LSTM) neural networks, to identify hate speech in the comments. The classifiers use various features, including sentiment polarity, word embeddings, and morpho-syntactic information. The results show that the classifiers perform reasonably well in distinguishing between hate and non-hate comments, with the LSTM classifier achieving slightly better accuracy. However, the classifiers struggle to differentiate between strong and weak hate speech. The authors acknowledge the challenges posed by low inter-annotator agreement and plan to refine the classifiers by considering different hate categories and improving the annotation process.

1. **Names Don't Fly: Smart Filters for Profanity Detection and Classification in User-Generated Content**

The paper proposes a pipeline approach to detect and classify inappropriate content in user-generated proper nouns submitted to the 'Send Your Names to Mars' public engagement campaign. The authors highlight the challenges of lack of negative samples, noisy labels, and the need for human-in-the-loop validation. They employ techniques like character n-gram features, data augmentation through string reversal, and a human-in-the-loop framework for SMEs to validate and correct model predictions. The pipeline includes traditional machine learning models like Support Vector Machines (SVMs) and deep learning models like Convolutional Neural Networks (CNNs). The results show that the SVM model with character n-grams and human-in-the-loop validation achieves high accuracy of 99% in classifying inappropriate names. The authors also describe the cloud-based infrastructure used for deploying the application and running predictions on large-scale data.

1. **Implementation of Anti-Profanity Words in Mobile Application Platform**

The paper describes the development of a chat profanity filtering application for mobile platforms to minimize the abuse of social media platforms in daily communication between lecturers and students at Kolej Universiti Poly-Tech MARA (KUPTM). The application implements the PurgoMalum web service to detect and filter profane words before the sender can send messages. It requires users to sign up with a unique username and secure password for authentication. The system design includes use-case and flowchart diagrams, and the development follows the Systems Development Life Cycle (SDLC) methodology. Questionnaires were distributed to lecturers and students to gather requirements, and the results showed a preference for using mobile applications for communication and the need for filtering inappropriate content. The authors conclude that the developed application can minimize the potential for users to abuse social media platforms by communicating obscenities.

1. **Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus**

The implementation of anti-profanity measures in mobile applications is crucial in curbing inappropriate content dissemination and ensuring a healthier online communication environment. In their study, Razali et al. developed a chat profanity filtering application integrated with the PurgoMalum web service, aimed at detecting and filtering offensive language before messages are sent. Through the use of a systematic approach, including planning, analysis, design, implementation, testing, and maintenance phases, the application offers a secure and user-friendly platform for communication between Kolej University Poly-Tech MARA (KUPTM) lecturers and students. The findings from questionnaires distributed to lecturers and students underscored the importance of such a filtering mechanism in mobile social media platforms, emphasizing convenience and the need to avoid harmful language in online interactions. Overall, the study highlights the effectiveness of the developed application in minimizing the potential for abusive language on social media platforms, contributing to a more respectful and positive online discourse.

1. **FLOSS as a Source for Profanity and Insults: Collecting the Data**

The paper by Megan Squire and Rebecca Gazda explores the use of free, libre, and open source software (FLOSS) communities as a dataset to study profanity and insults for natural language processing (NLP) and machine learning (ML). It discusses the background literature on insult and profanity detection, as well as existing datasets and methodologies. The study collects data from FLOSS projects, including codes of conduct and profanity usage in mailing lists and IRC channels. It analyzes the language used by Linus Torvalds in the Linux Kernel Mailing List (LKML) as a case study. Additionally, it investigates gender-based insults such as "that's what she said" jokes and maternal insults. The paper concludes by suggesting future research directions in this area.

1. **Investigating the role of swear words in abusive language detection tasks. Language Resources and Evaluation.**

A comprehensive study on the role of swear words in abusive language detection tasks. The researchers developed a new benchmark Twitter corpus called SWAD (Swear Words Abusiveness Dataset) to manually annotate abusive swearing at the word level. They conducted experiments using machine learning models to predict the abusiveness of swear words in tweets and found stable results in a supervised learning setting. The study also encountered challenges in annotating the abusiveness of swear words, including cases of indirect insult, irony, and sarcasm. Overall, the study provides valuable insights into the complex nature of swearing in online communication and its implications for abusive language detection tasks.

1. **Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science 7**

The paper discusses the urgent need for automatic hate speech detection due to the increasing prevalence of offensive and harmful content on social media platforms. It emphasizes the challenges of detecting hate speech, offensive language, and abusive language, and the importance of generalizability in hate speech detection models. The survey paper aims to provide a comparative summary of existing research on generalizability in hate speech detection, analyze the main obstacles to generalizable hate speech detection, and suggest future research directions to address these obstacles. The survey paper contributes to the literature by providing a systematic analysis of the challenges and existing attempts to address them in hate speech detection. It is relevant to researchers working on hate speech and abusive language detection, as well as those involved in other types of offensive or harmful language detection. The paper is structured to offer a comprehensive overview of the current state of hate speech detection and closely related fields, providing a valuable resource for researchers and practitioners in this area.

1. **Hate Speech in social media: An Exploration of the Problem and its proposed Arrangement in India.**

The importance of national integration and the challenges it faces due to untrustworthy activities of citizens in a multicultural and multi-religious country like India. It highlights the vulnerability of India to various issues that pose a threat to national integration, especially in the age of digital access where most untrustworthy activities occur on social media platforms. The abstract also mentions the efforts made by the constitution of India and various legislations to contain hate speech on online media platforms, but notes that the prevalence of hate speech has increased despite legal restrictions, leading to societal issues.

1. **Profanity use in online communities. Conference on Human Factors in Computing Systems - Proceedings.**

The challenges associated with detecting and managing inappropriate and objectionable user-generated web content. It highlights the limitations of current list-based profanity detection systems, emphasizing their susceptibility to circumvention and their inability to adapt to evolving profane slang. Additionally, these systems are criticized for offering a one-size-fits-all solution that does not consider domain, community, and context-specific needs. The passage proposes a shift towards more contextually nuanced profanity detection systems and suggests that social settings have their own normative behaviors, with acceptability varying between different communities. The paper aims to provide evidence of the shortcomings of current profanity detection systems through an analysis of comments from a social news site and to evaluate the cases in which they fail.

1. **Detecting hate speech on the world wide web.**

The paper presents an approach to detecting hate speech in online text, focusing on abusive speech targeting specific group characteristics. It discusses the challenges of defining hate speech and the need for consistent language models to distinguish different forms of hate speech based on stereotypes. Additionally, the paper outlines the collection and annotation of a hate speech corpus, and describes pilot classification experiments focused on anti-semitic speech, achieving an accuracy of 94%, precision of 68%, and recall of 60%. The authors also examine interlabeler agreement, labeling quality, and provide an error analysis from the classification experiments.

1. **Method of Profanity Detection Using Word Embedding and LSTM. Mobile Information Systems.**

The authors analyze previous work on this subject and point out the shortcomings of typical methods that rely on rule-based systems or predetermined dictionaries to capture contextual subtleties. In more recent research, machine learning methods like support vector machines and n-grams have been investigated; however, these methods can have difficulty with words that are not in the dictionary and require a lot of feature engineering. The authors suggest a unique approach that blends word embeddings and an LSTM neural network model for profanity detection in order to overcome these shortcomings. The LSTM model is well-suited to accurately depict the sequential and contextual aspects of language, while word embeddings, notably GloVe, express words as dense numerical vectors that encode semantic associations. The authors go over the procedures for gathering data, preparing it, training it, and assessing the LSTM model. The outcomes demonstrate the efficacy of using word embeddings and LSTM networks for profanity detection, since the suggested method outperforms a number of baseline approaches in terms of accuracy, precision, recall, and F1-score.

**12. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review**

The writers then go into the several machine learning methods—supervised, unsupervised, and semi-supervised—that have been used to detect hate speech. They compare the effectiveness of newer developments in deep learning architectures, such as Transformers, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), with the performance of traditional machine learning algorithms, such as Support Vector Machines (SVMs), Logistic Regression, and Decision Trees.   
  
The paper discusses the main obstacles to hate speech detection, including the difficulty of identifying hate speech, the necessity for multilingual and cross-cultural methods, the imbalance and lack of data, and the complexity of language. The authors also cover potential biases and ethical issues in the creation and application of such systems.

**13. Filtering Offensive Language in Online Communities using Grammatical Relations. 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference**

The authors acknowledge the shortcomings of conventional methods that rely on predefined dictionaries or keyword matching. These methods frequently have trouble capturing the context and subtleties of offensive language, which can result in a high false positive rate or the failure to identify more nuanced types of abuse. In order to overcome these drawbacks, the authors suggest a novel approach that examines the grammatical relationships between words in a phrase as opposed to concentrating only on individual words. The main thesis is that offensive language frequently has distinctive grammatical patterns that can be utilized as characteristics that differentiate it from other types of discourse. The authors go into great detail about their methodology, which is extracting grammatical relations from the text through parsing, building a feature vector based on the frequency and occurrence of these relations, and training a support vector machine (SVM) classifier to determine whether a given text contains offensive language. Based on a dataset of user-generated content from an online gaming community, the authors analyze their method. The results show that the grammatical relations-based method performs much better than baseline approaches that employ n-grams or word-based characteristics.

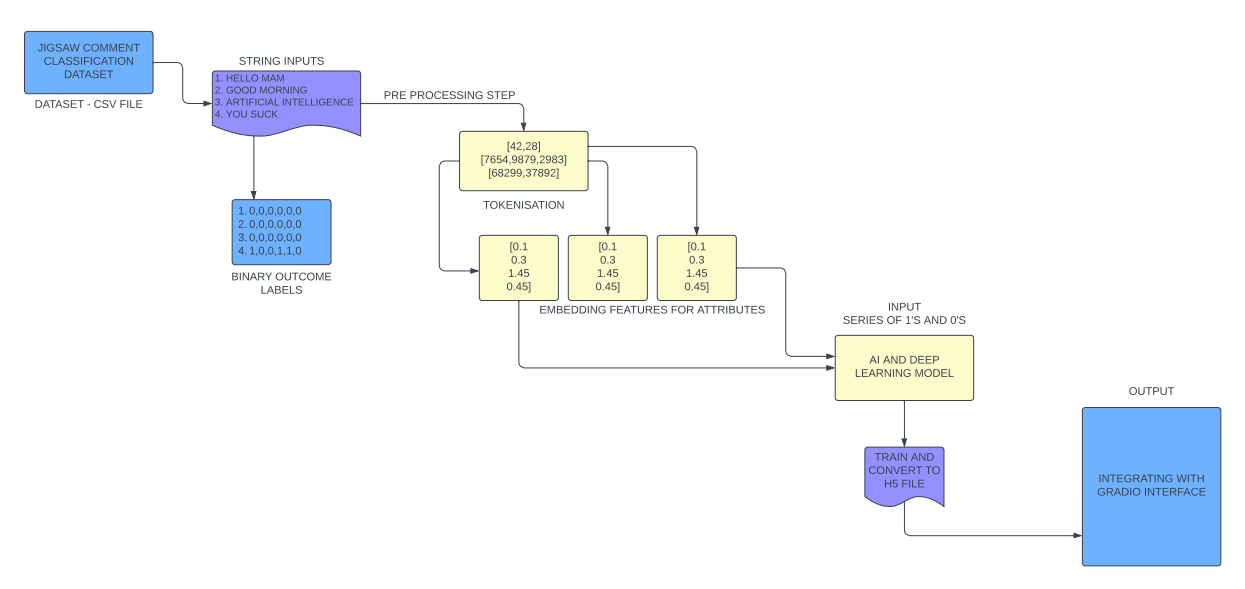
**14.A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere. Applied Sciences.**

The authors begin by giving a summary of earlier studies on the detection of hate speech, emphasizing the drawbacks of conventional machine learning approaches that depend on manually created features and lexicon-based strategies. They contend that these methods frequently fall short of capturing the complex and contextual aspects of hate speech, especially in contexts with a diversity of languages and cultures. The authors suggest a deep learning-based approach to overcome these difficulties, which makes use of neural networks' representational capability to extract important characteristics from the textual input. They specifically use an architecture known as a Convolutional Neural Network (CNN), which has been trained on a dataset of tweets from Saudi Arabia that have been flagged as hate speech. The model architecture and training procedure, as well as the phases involved in data collection and preparation, are all explained by the writers. Along with outlining the model architecture and training procedure, the authors also cover data collecting and preparation procedures. The suggested CNN model is evaluated and it is found to be superior in terms of accuracy, precision, recall, and F1-score compared to numerous baseline techniques, such as Support Vector Machines and Recurrent Neural Networks.

**15. Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media.**

The authors thoroughly examine a sizable dataset of tweets that have been annotated, investigating the traits and linguistic elements that differentiate hate speech apart from other abusive language. According to their findings, hate speech can take on subtler and contextual forms in addition to being expressed through overt vulgarity or explicit slurs. This highlights the drawbacks of oversimplified keyword-based strategies and the requirement for more advanced machine learning methods. The authors then go over their suggested automated hate speech detection technique, which combines characteristics that are specifically designed to capture the syntactic and semantic patterns of hate speech with a supervised learning algorithm (logistic regression). The authors evaluate their approach rigorously using the twitter dataset to show its benefits over baseline models, especially in terms of lower false positive rates. But the authors also note that automated hate speech detection has intrinsic difficulties and possible traps, such as the possibility of enhancing biases in the training set or unintentionally suppressing speech that is considered appropriate.

**Proposed Methodology**

****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.

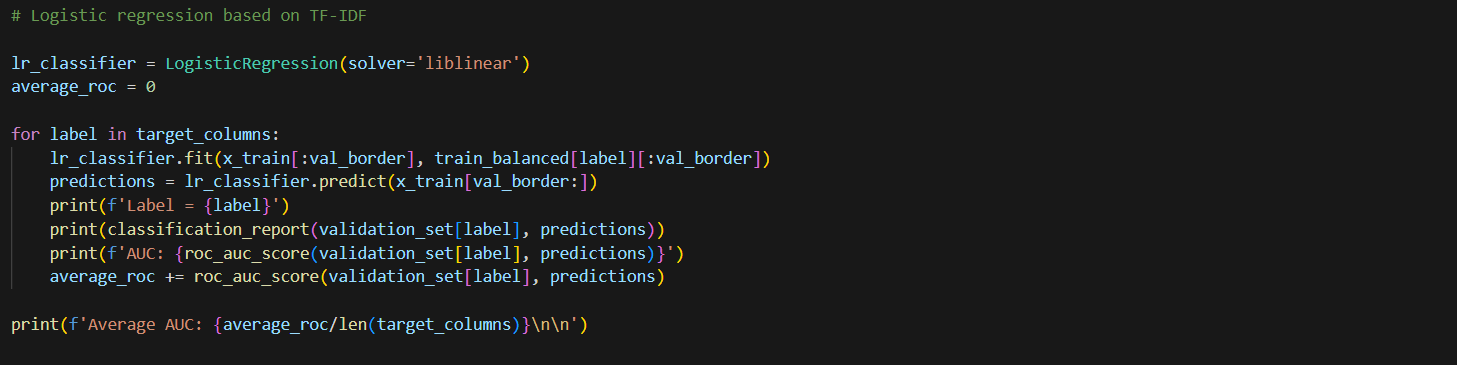
**SVM Classifier Model**

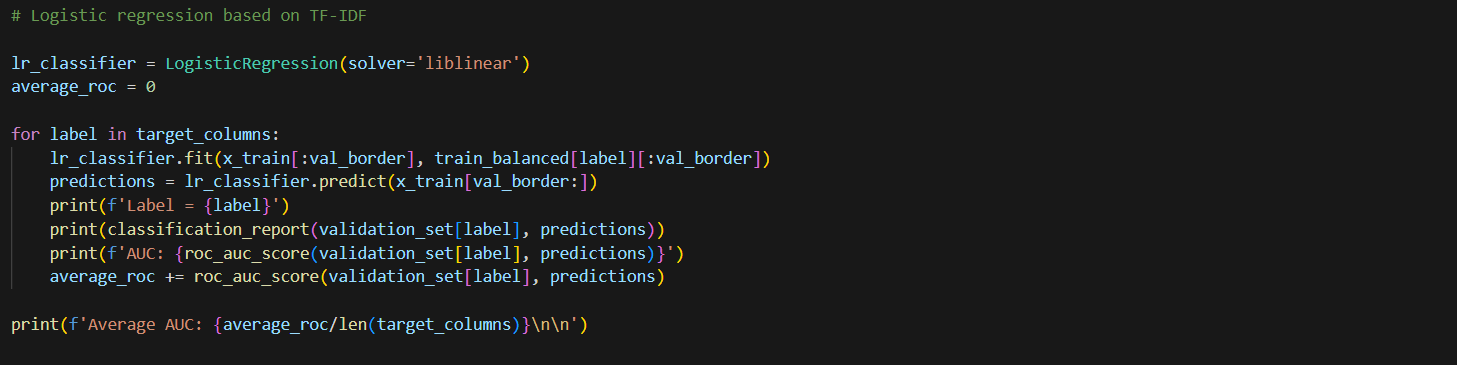
Support Vector Machines (SVMs) are a class of supervised learning algorithms that can be used for both classification and regression tasks. In the context of our profanity detection system, we will employ the Support Vector Classifier (SVC) to separate the comments into profane and non-profane categories. 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.





**Results and Discussion**

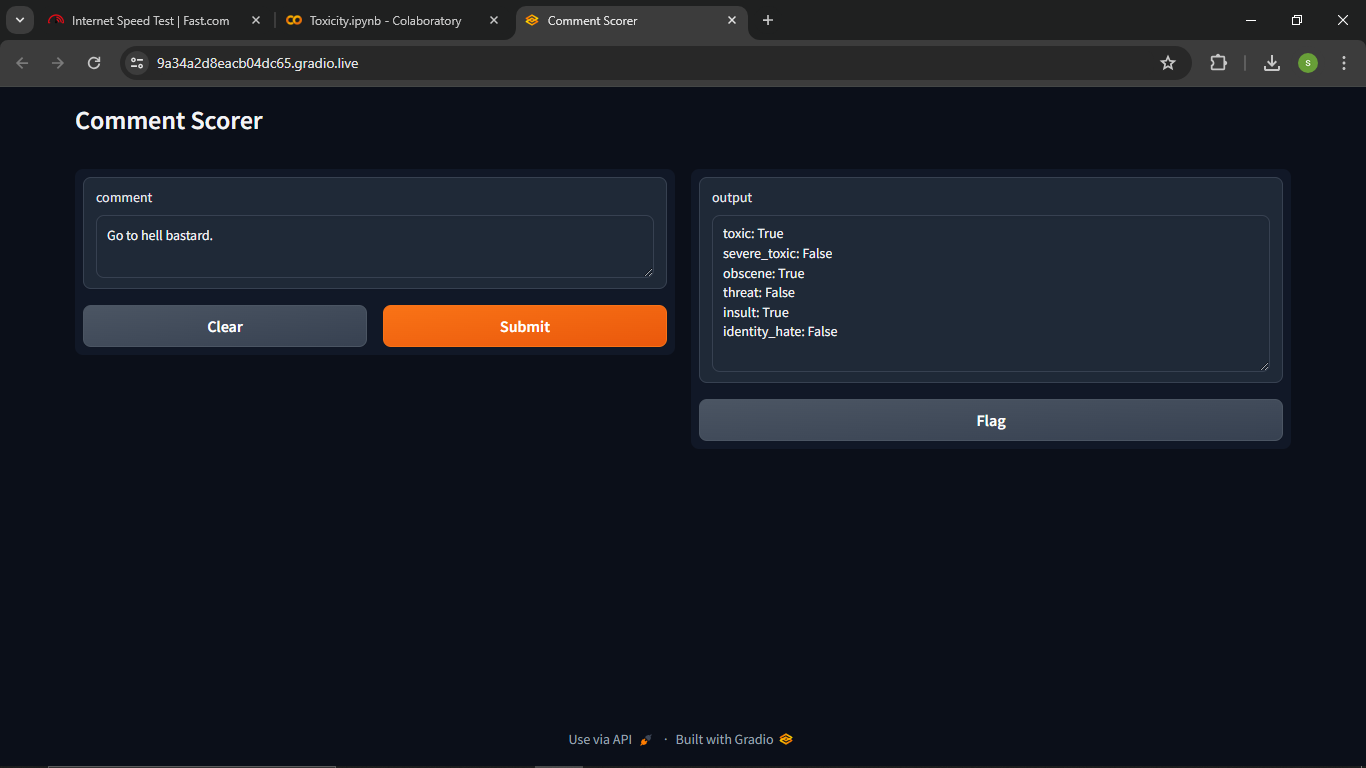
The existing models, namely the Support Vector Machine (SVM) classifier, the J48graft Model, and the Gold Full Classifier, demonstrate varying levels of performance. The SVM classifier with the False label configuration achieved the highest overall accuracy of 72.95%, but the SVM with the True label 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.



**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.

**Reference**

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).

From Kaggle and githubs also we have taken so many references and youtube helped us a lot to learn and the project.

[**https://github.com/topics/logistic-regression**](https://github.com/topics/logistic-regression)

[**https://github.com/topics/mlp-classifier**](https://github.com/topics/mlp-classifier)

[**https://github.com/topics/machine-learning-projects**](https://github.com/topics/machine-learning-projects)

**APPENDIX**

**Implementation / Code**

**(Copy the code here)**

**\*\* Attention : Team Lead \*\***

* Create a shared folder in your G-drive
* Upload the code (ipynb and converted html/pdf) with necessary documentation and dataset in your Team Lead’s G-drive.
* Give Edit Access to [rajalakshmi.r@vit.ac.in](mailto:rajalakshmi.r@vit.ac.in)
* If the dataset is too big, share the link of that dataset too.

G-DRIVE LINK: <https://drive.google.com/drive/folders/1TvzR4alqgEpHhLUaiLfGDLwePeDeNvrb?usp=sharing>