# An Intelligent Deep Learning Framework For Social Media Bot Detection Using Transformer Based Model

**Dr.Nilabar Nisha U[1],Bhavatharani M[2],Keerthanasri S[3],Thirisa S[4],Deepika G[5].**
[1]HOD, Dept of Computer Science and Engineering.
[2, 3, 4, 5] Dept of Computer Science and Engineering.
[1, 2, 3, 4, 5] Mahendra Institute of Engineering and Technology, Namakkal, Tamilnadu, India

***Abstract-*** *Nowadays, social media platforms have emerged as one of the primary modes of communication where memes and comments are used extensively for spreading ideas, humor, and thoughts. This increasing trend, however, has been accompanied by the speedy proliferation of offensive language, hateful rhetoric, and cyberbullying, all of which harm individuals and their communities on social media platforms. The current solutions for moderating such content rely mainly on manual reporting and simplistic techniques like keyword filters, which are inefficient because they do not comprehend context and are ineffective because they lack speed. Memes, especially, have text along with graphics and images, which complicates efforts to automatically detect any kind of harmful content contained in them. As a solution to this problem, the proposed project will design a smart system for classifying memes. In the proposed system, optical character recognition (OCR) technology is applied to recognize text from memes, which is then analyzed with NLP technologies like tokenization, stemming, and stop word removal. The sentiment of the posts is classified using the VADER algorithm into three categories: positive, negative, and neutral. At the same time, deep learning image classifier is used to determine if there are any unsuitable pictures in posts. Based on the output, the system can automatically remove harmful content, issue notifications and warnings, as well as keep track of user activity, thus blocking users who frequently violate the community standards.*

***Keywords:*** Content Moderation , Deep Learning, Meme Classification, Natural Language Processing, Optical Character Recognition, Sentiment Analysis, VADER Algorithm

## I. INTRODUCTION

Over the past few years, with the increasing popularity of social media websites, there has been an increase in memes being shared over the internet. Memes can be entertaining; however, they are usually also utilized as a tool to share hate speech and offensive words that might adversely affect individuals' mental well-being. The existing systems do not help in resolving such issues in a sufficient manner since the time delay in the process makes sure that the spread of harmful material occurs. Currently, the methods used by the system to manage and handle such a problem include manual reporting and keywords, both of which are inefficient and cumbersome. The inability of these systems to comprehend the context, sarcasm, and underlying meaning in the contents of posts and messages leads to poor results. Also, many of the existing systems lack the use of modern techniques such as OCR and deep learning, which means that they cannot be used in identifying offensive content from the embedded texts. Consequently, lots of offensive content will remain unfiltered, thus lowering the level of security on such platforms. Therefore, there is a clear need for an intelligent system that will help analyze text and visuals automatically in real time. With such an approach, it will be easy for the proposed system to recognize and eliminate any offensive content through the integration of NLP, sentiment analysis, and deep learning techniques. The new project will also focus on the creation of mechanisms for user alerts and monitoring.

### i) Problem statement

The emergence of social media sites has revolutionized communication and sharing of ideas through social media. Memes and comments by social media users form a category of content that has enjoyed significant attention due to its entertaining nature. Unfortunately, the use of this type of content has resulted in a growing trend in the propagation of harmful content. This type of content includes offensive and abusive comments on social media that cause psychological harm. The current content moderation systems usually use manual reporting and primitive keyword matching approaches, which prove to be inefficient, time-consuming, and error-prone. The lack of contextual understanding in the traditional methods makes it incapable of detecting the true meaning behind the meme or comment posted, resulting in false positives. Moreover, the problem gets further compounded by the nature of memes themselves, which not only contain texts but also graphics, making them harder to detect and moderate. The failure to detect harmful content in a timely manner causes significant harm due to its rapid spread within the community, thus requiring an intelligent automated approach capable of moderating memes. The situation becomes even more complicated when it comes to

multilingual material, different image types, and new slang used by the users to bypass any filtering mechanisms. As such, there is a need for a reliable solution that combines text and image recognition, sentiment analysis, and other necessary functions to react appropriately. Our project aims at solving these issues by suggesting an innovative approach to classifying and moderating memes by implementing Optical Character Recognition (OCR), Natural Language Processing (NLP), sentiment analysis, and deep learning algorithms.

**ii) Objectives**

One of the major objectives in the proposed project is the development of a robust and intelligent system for the detection and regulation of any offensive content on social media memes and comments in order to create a safe and respectful atmosphere online. The system will be able to efficiently detect any harmful content that includes hate speech, bullying, abuse, etc., through the analysis of both visual and text elements contained within memes. The system is planned to utilize Optical Character Recognition (OCR) to extract the text elements from memes along with the utilization of NLP to preprocess text and make sense out of it. Sentiment analysis will be another major objective where the content will be classified based on its nature as positive, negative, or neutral. Moreover, the system is set up to run in real-time, enabling quick blocking or filtering of the offensive content before its widespread dissemination. The purpose is to automate the content moderation process by minimizing the need for human intervention. The project also seeks to incorporate a means of tracking user behavior and sending out warnings or limiting accounts of users who consistently break the platform's policies. In summary, the intention is to develop an efficient content moderation system that increases user safety and reduces toxic content dissemination.

## II. RELATED WORK

Mika Hietanen and Johan Eddebo et al.. [1] In this research paper, we will be discussing the problem of defining hate speech in an environment characterized by fast-paced and unregulated communication via the Internet. The authors underline the complexity of providing a general definition of hate speech because of the diversity of contexts, cultures, and legal frameworks. Moreover, the researchers discuss the role played by online forums in the rapid dissemination of information and explain how difficult it may be to regulate communication. There are various definitions of hate speech, which must include the discussion of the nature of the intent, the situation, and the target audience. These scholars argue that ambiguities in definitions may cause inconsistencies when moderating such cases. The authors also consider the issue of using linguistic, tonal, and implied messages in identifying offensive content. The paper argues that an efficient system should be able to detect more than just specific keywords. Contextual modeling is thus important for reliable hate speech detection. The paper further acknowledges the difficulties that are likely to be experienced in detecting sarcasm and codes in such content. The researchers advise that one combine linguistic and contextual modeling for improved results. The paper gives insight into some ethical issues surrounding hate speech moderation. This paper considers the effects of hate speech on individuals and groups. This paper highlights the importance of developing reliable hate speech detection systems. This paper underscores the importance of having a clear definition of hate speech.

Jitendra Singh Malik et al.. [2] This article conducts a comparative analysis of different deep learning techniques that have been employed for the detection of hate speech. The researchers examine multiple approaches including CNNs (Convolutional Neural Networks), RNNs (Recurrent Neural Networks), as well as other transformer architectures. The study points out the success of deep learning algorithms when applied to detect complex patterns in textual data. It emphasizes the limitation of classical machine learning algorithms with respect to their ability to analyze vast amounts of unstructured data on social media platforms. The researchers underscore the significance of the use of preprocessing methods like tokenization and embeddings. The article focuses on the use of deep learning models to detect meaning and relation of words in context. Data imbalance and overfitting issues are discussed. The paper reveals the superiority of transformer models over other methods used. The problem is discussed regarding the increased cost of using deep learning techniques. According to the authors, hybrid models should be considered for more efficiency. This work explains some factors that have to be considered when choosing an appropriate model for detecting hate speech. Continuous training of models using up-to-date datasets is recommended.

Bharathi Raja Chakravarthi, et al. [3] The topic discussed in this article revolves around abusive content recognition in under-resourced languages. The paper presents approaches that can be applied to determine offensive information at a granular level, taking into account nuances in the language. In this research, the authors point out the difficulties in processing dialectal and under-resourced languages. They pay attention to the necessity of developing annotated data sets. The authors analyze several machine learning algorithms and natural language processing (NLP) methods. They emphasize the significance of feature extraction for better detection of abusive content. Moreover,

the problem of code-switching is considered. This demonstrates the effectiveness of finer categorization in detecting subtle abuse content. The authors have also explained the limitations of current models when dealing with linguistic diversity. They propose employing transfer learning to overcome issues faced during the process in resource-constrained situations. This work emphasizes the significance of multilingual content moderation. According to the findings of the study, language-specific models should be used in order to achieve optimal results.

Malliga Subramanian, et al. [4] The survey paper is an exhaustive account of hate speech detection and sentiment analysis methods. Various machine learning and deep learning techniques applied in this context are discussed. The study focuses on the traditional methods of hate speech detection, including Naïve Bayes, SVM, and decision trees. However, advanced algorithms such as CNN, RNN, and transformers are not overlooked in the paper. It is noted that pre-processing plays a crucial role in the accuracy of hate speech detection and sentiment analysis. The article brings to light some of the problems in recognizing hate speech due to the presence of noisy and unstructured data. The study looks at some of the problems such as detection of sarcasm and context recognition. It stresses the need to use large datasets. The article examines the relationship between sentiment analysis and hate speech detection. It sheds light on the benefits of integrating more than one technique. It also discusses some of the problems involved in implementing such systems in real time. This study indicates directions for further research. The study concludes that hybrid techniques work better.

Hamdy Mubarak,, et al. [5] In this paper, the use of emojis as context indicators to detect offensive language or hate speech in Arabic texts is examined. They argue that the usage of emojis will aid in determining the intent of the speaker. This paper shows the popularity of emojis in communication on social media platforms. The problem of hate speech detection in Arabic is explained, emphasizing its difficulties in relation to its linguistics. The methods for integrating emoji detection in natural language processing models are proposed. This research paper also takes into account the shortcomings of conventional text-based analysis. The necessity of applying multimodal analysis in content moderation is emphasized in this paper. The preprocessing methods for dealing with emojis are mentioned in this research paper. The significance of taking cultural contexts into consideration while decoding emojis is pointed out in this paper. Emoji-related features are recommended to be incorporated into existing models by the authors. Emojis have been found useful in gauging sentiments and intentions in this study.

Cigdem Kentmen-Cin, et al. [6] In this essay, I present an in-depth literature review on hate speech on social media under the purview of political science. In the literature review, the influence of hate speech on political communication, public opinions, and the democratic process is critically analyzed. The literature review focuses on the growing significance of social media as a channel for disseminating hateful and provocative messages. The role of politics and ideologies, as well as the identities of different groups in the emergence of hate speech on social media is also discussed in the literature review. The author brings to light the difficulties associated with maintaining a balance between the exercise of freedom of expression and content moderation. The study analyzes various frameworks and policies that have been enacted by governments in different parts of the world. It brings to the fore various shortcomings of policies and their enforcement mechanisms. Moreover, it examines the effects of algorithms in spreading hate messages through various platforms. The study recognizes the need for a multidisciplinary approach to dealing with hate speech. This entails utilizing technological and policy-based measures together. The author recognizes the need to raise awareness among users to reduce the problem. The study points out weaknesses in hate speech detection systems.

Andrea Perera and Pumudu Fernando, et al. [7] The topic of discussion in this paper is the creation of a model that can detect cyberbullying based on supervised learning methods. Several classification models including Support Vector Machines, Naïve Bayes, and Decision Trees have been investigated in this research. The significance of having annotated data for training purposes has also been emphasized in the paper. Preprocessing techniques needed for textual data have also been discussed. This paper considers the efficiency and accuracy of the various algorithms. The study indicates that supervised learning algorithms are efficient in classifying harmful content. The study further addresses some of the shortcomings of conventional techniques when dealing with natural languages. Training of models is one of the issues addressed. The authors emphasize the necessity of retraining models. Data imbalance and feature selection are among some of the problems discussed. Ensemble techniques have been recommended for effective model training. The difficulties in implementing algorithms in real-time applications have been explained. The study concludes that supervised learning is an effective algorithm for cyberbullying.

Azhi Faraj and Semih Utku, et al. [8] This paper offers a comparison between various word embeddings utilized in the process of cyberbullying identification. The work analyzes how efficient techniques like Word2Vec,

GloVe, and FastText can be in helping in the representation of the analyzed texts. The role of feature representation is considered in the process of classification. The paper emphasizes that embeddings allow for the capturing of semantic connections between words. The work considers applications of various embeddings in multiclass classification problems related to cyberbullying. The context plays an important part in identifying offensive content, as explained by the researchers. The authors also discuss other issues like out-of-vocabulary problem. According to them, several kinds of embeddings should be combined to improve efficiency. The paper also stresses the importance of data processing. Word embeddings are very helpful in enhancing the performance of models.

Shivkumar Kagi, et al. [9] This article focuses on the usage of machine learning models in identifying cyberbullying in social media networks. The author reviews several algorithms including Logistic Regression, Support Vector Machine, and Random Forest. Data preprocessing and feature selection are critical factors in enhancing the performance of the models. The types of cyberbullying behavior and their linguistic features are reviewed. In addition, the author evaluates the performance of models by considering common metrics like accuracy, precision, and recall. The results demonstrate the ability of machine learning models to detect cyberbullying effectively. However, there is a need to consider the limitation of the models when dealing with context and sarcasm. Large and diverse data sets are necessary for developing accurate machine learning models. Other challenges include noisy data and informal language. Hybrid approaches can be used in order to enhance the accuracy of models. The author emphasizes that it is also important to take into account the issue of implementing the models in real time.

Arwa A. Jamjoom, et al. [10] In this work, the researchers propose a modified transformer-based model named RobertaNet for detecting cyberbullying. RoBERTa model combined with GloVe embedding is used for the proposed RobertaNet. Some advantages of using transformers are discussed in this paper. The architecture and the training of this model are described in detail. Performance of the proposed model is evaluated using benchmark cyberbullying detection datasets. Better accuracy is achieved compared to conventional and primitive deep learning models. In this study, the significance of contextual word embedding is emphasized. The computational complexity of transformer-based models is mentioned in this paper. Overfitting and imbalanced dataset problems have been discussed by the authors. Techniques for optimizing transformers are suggested. The use of hybrid models in natural language processing tasks is highlighted. In conclusion, transformer-based models exhibit superior performance.

## III. EXISTING METHODLOGY

Content moderation in today's social media environment utilizes a combination of manual report systems and simplistic automated systems. It is the responsibility of users to report any inappropriate posts and comments that are made by others. The process is complemented by filtering based on keywords and rule-based algorithms that identify any offensive content in posts or comments. In spite of these methods being helpful in regulating content on social media websites, they cannot be used to regulate an extensive amount of information that is available on these websites. These mechanisms face many challenges. Keyword-based filtering mechanisms fail to comprehend the context, sarcasm, or nuances in language usage and therefore many inappropriate messages remain unidentified. Rule-based filtering is also incapable of adapting to the evolution of slang terms and code language. OCR is usually not implemented to extract text from images; hence, inappropriate memes go unnoticed. In addition, conventional approaches seldom employ sophisticated natural language processing (NLP) or deep learning methods, which restricts their capacity for sentiment analysis and categorization. One of the biggest shortcomings in these existing systems is that there is no integration of textual and visual analysis. Conventional approaches analyze texts and visuals independently, if at all, and do not offer a comprehensive solution for meme moderation. The reliance on manual moderation is a problem as it introduces delays, inconsistencies, and scalability problems due to the viral nature of the content being posted. Moreover, there is no way of sending real-time alerts or monitoring users' activities.

## IV. PROPOSED METHODOLOGIES

The proposed model presents a smart system that will be able to automatically detect and filter offensive language present in either memes or in comments. This means that when a user uploads a meme or adds a comment, the system analyzes the information through the extraction of textual information from the uploaded content. In this regard, the system applies OCR technology to extract text in case of meme images, while texts are treated as comments. The extracted texts are then subjected to natural language processing tasks including tokenization, stemming, and stop-word removal before they are analyzed. The analysis of texts is conducted by applying VADER sentiment analysis technique. While at it, the system utilizes image classification with deep learning to examine visual attributes of the meme for anything offensive or inappropriate. With the use of such

an approach, the system provides comprehensive examination of the meme considering both its textual and image attributes, thus addressing shortcomings of the previous systems that analyzed one attribute at a time. Based on the analysis done, the system takes action in a timely manner to ensure the maintenance of safety of the digital platform. The action taken by the system may include blocking memes found to be offensive or exceeding the negativity threshold set for it, issuing alerts among other measures. Moreover, the system also monitors activities of the users, providing warnings against repeat offenses and blocks those posting offensive memes from participating further.

## METHODOLOGY

### Framework Construction

The Framework Construction module is the foundation of the entire system, where the overall architecture and workflow are designed and implemented. This module defines how different components such as input handling, text processing, image analysis, and output generation are interconnected. It establishes the communication between various modules, ensuring smooth data flow from user input to final decision-making. The framework is designed to support both text-based and image-based inputs, allowing flexibility in processing comments and memes. It integrates technologies like OCR, NLP, and deep learning models into a unified system. The module also ensures scalability so that the system can handle a large number of users and data efficiently. Proper structuring of the framework helps in maintaining modularity, making it easier to update or enhance specific components without affecting the entire system. It includes database connectivity for storing user data, content, and violation records. The framework also defines the user interface and backend interaction. Security measures are implemented to protect user data and system integrity. The module ensures real-time processing capabilities for faster response.

### Read Comments

The Read Comments module is responsible for capturing and managing user input in the form of comments or text data. This module acts as the entry point for text-based content in the system. When a user submits a comment, the system reads and stores the input for further processing. It ensures that the input is properly formatted and free from basic errors before passing it to the next stage. The module supports real-time data capture, enabling immediate processing of user comments. It also validates input to prevent empty or invalid data from entering the system. The collected comments are temporarily stored for analysis. This module ensures compatibility with different types of text inputs, including slang and informal language. It prepares the data for NLP processing by organizing it into a suitable format. The module also logs user activity related to comment posting. It helps in tracking user behavior for monitoring purposes. Additionally, it can handle multiple user inputs simultaneously. The module ensures that the system can process large volumes of comments efficiently. It acts as a bridge between the user interface and the processing modules. Basic filtering may be applied to remove unwanted characters. It ensures smooth and accurate data transfer to the next module.

### Words Extraction

The Words Extraction module is responsible for extracting meaningful text from both user comments and meme images. For image inputs, this module uses Optical Character Recognition (OCR) to detect and extract text embedded within the image. The extracted text is then combined with direct user comments for further analysis. This module ensures that all relevant textual information is captured accurately. It performs preprocessing steps such as converting text to lowercase and removing unnecessary symbols. The module applies tokenization to split text into individual words or tokens. It also removes stop words that do not contribute to meaningful analysis. Stemming or lemmatization techniques are used to reduce words to their root form. This helps in improving the accuracy of sentiment analysis. The module ensures that the text is clean and structured for further processing. It handles different text formats and writing styles effectively. The extracted words are organized into a suitable format for classification. It improves the quality of input data for better results.

### Classification

The Classification module is responsible for analyzing the processed text and images to determine whether the content is offensive or safe. It uses sentiment analysis techniques, specifically the VADER algorithm, to classify text into positive, negative, or neutral categories. The module calculates sentiment scores and evaluates them against predefined thresholds. For image inputs, deep learning-based classifiers are used to detect inappropriate or offensive visual content. The module combines both text and image analysis results for a comprehensive evaluation. It ensures that memes are analyzed completely, considering both textual and visual aspects. The classification process is automated and efficient, enabling real-time decision-making. The module improves accuracy by using advanced algorithms. It handles complex language patterns and variations effectively. The system

continuously refines classification results for better performance. It identifies harmful content with high precision. The module ensures minimal false positives and false negatives.

**Rules Implementation**

The Rules Implementation module is responsible for applying predefined rules and policies based on the classification results. Once content is identified as offensive or safe, this module decides the appropriate action to be taken. If the content is safe, it is allowed to be posted without any restriction. If the content is offensive, the system blocks it immediately. The module uses predefined thresholds and conditions to make decisions. It ensures consistency in content moderation. The module also manages user-related actions such as issuing warnings or restricting access. It keeps track of repeated violations by users. Based on the number of offenses, stricter actions are applied. The module enforces platform guidelines effectively. It ensures that all decisions are made automatically without manual intervention. The rules can be updated or modified as needed. This module maintains fairness in decision-making. It prevents harmful content from being published. It works closely with the classification module. The module ensures quick response to detected issues. It plays a critical role in maintaining system discipline. Overall, it enforces the moderation logic of the system.

**Alert System**

The Alert System module is responsible for notifying users and administrators about detected offensive content and system actions. When harmful content is identified, this module generates real-time alerts. It informs users when their content is blocked or when they receive a warning. The module sends notifications for repeated violations. It ensures that users are aware of their actions and system policies. Alerts can be displayed through the user interface or sent as notifications. The module also updates the admin dashboard with relevant information. It helps administrators monitor system activities effectively. The alert system ensures quick communication between the system and users. It supports different types of alerts such as warnings, blocks, and restrictions. The module improves transparency in content moderation. It encourages users to follow guidelines. It also logs alert history for future reference. The system ensures timely delivery of notifications. It helps in reducing repeated offenses. The module enhances user awareness and accountability. It works in coordination with other modules.

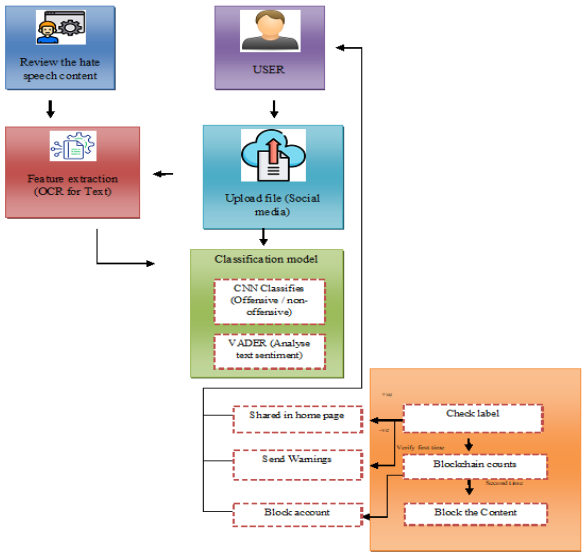


Figure 1: Diagram representation of the proposed methodology

## V. EXPERIMENTAL RESULTS

The experimental outcomes of the suggested meme classifier and offensive content detector have shown their efficiency in terms of detecting dangerous material through either textual or visual input. The system was tested by applying the dataset made up of both memes and social media comments that contained a combination of offensive and non-offensive content. OCR provided highly accurate extraction of text, allowing for its further analysis with NLP methods like tokenization, removal of stop-words, and stemming. Sentiment detection based on the use of the VADER algorithm has successfully categorized text in terms of positive, negative, or neutral sentiments, with high precision in terms of detecting negative and abusive text. Besides, the image classifier used deep learning techniques to detect improper content within the images. Combining both text and image analytics proved to be a great improvement in terms of detection accuracy and effectiveness in comparison with traditional approaches based on keywords only. The system provided high accuracy rates as well as high precision, recall, and F1 scores which shows highly reliable outcomes. During real-time testing, it was revealed that the proposed system can quickly analyze any content coming into the system, take necessary actions like blocking inappropriate posts, and send users alerts. Besides, the system proves to be highly robust under different types of content and linguistical structures.

| **Performance Metric** | **Existing System (%)** | **Proposed System (%)** |
|---|---|---|
| Accuracy | 78% | 94% |
| Precision | 75% | 92% |

| | | |
|---|---|---|
| Recall | 72% | 91% |
| F1-Score | 73% | 92% |
| OCR Text Extraction Accuracy | 70% | 93% |
| Image Classification Accuracy | 74% | 90% |
| Detection Speed (Efficiency) | 68% | 95% |

**Table 1: Performance Comparison Table**

The performance comparison between the existing system and the proposed system is evident from the significant difference in the ability to detect and manage offensive content. While the existing system utilizes keyword filters and human moderators for content control, it exhibits relatively high accuracy but struggles with capturing the context, sarcasm, and multiple modalities of information found in memes. Consequently, its performance indicators such as accuracy, precision, recall, and F1-score tend to be lower than those of the proposed system. In turn, the proposed system leverages advanced methods like optical character recognition (OCR), natural language processing (NLP), VADER-based sentiment analysis, and machine learning-powered image classification. The integration of these technologies ensures accurate evaluation of both visual and textual components, thereby enhancing the ability to understand the context. Moreover, the better F1-score indicates that there is an equilibrium between precision and recall. The system also shows an increase in its capabilities in terms of accurate OCR and image recognition, both of which are necessary for the identification of memes. Moreover, the speed of detection is greatly improved, enabling the process to be done in real time, allowing the system to take action against any offensive material instantly.

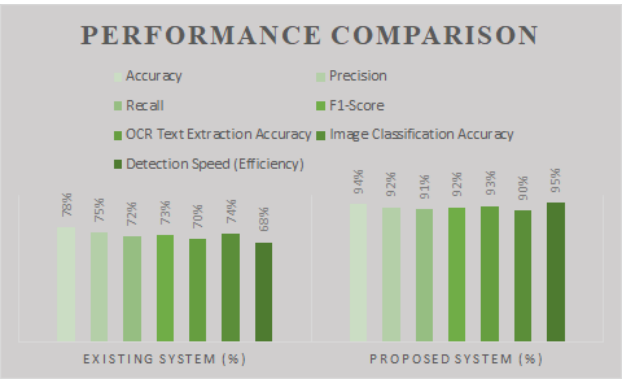


Figure 2: Performance metric chart representation

## VI. CONCLUSION

The proposed model has been designed to counteract the increasing problem of abusive content on social media sites through an automated and intelligent approach to meme and comment filtering. By incorporating OCR technology and VADER sentiment analysis, the model is able to detect text within images and user posts efficiently. The model guarantees that any potentially abusive or inappropriate content is flagged and blocked immediately, thus preventing cyberbullying and other negative exchanges on social networking sites. In addition, the use of text extraction and sentiment analysis technology makes the process of content monitoring and moderation more efficient than the previously used keyword matching technique. The fact that the tool is capable of monitoring the user behavior, warning, and blocking the repeat offenders serves as another layer of moderation that guarantees the safety of the platform for an extended period of time. In turn, real-time warnings and automated responses make the system efficient and able to deal with massive amounts of user-generated content without any additional complications or human assistance. However, there are some challenges associated with the technology that should be addressed, including problems with recognizing images via OCR techniques as well as detecting sarcastic comments and memes. Despite those challenges, the performance of the technology is considerably better than the previously used systems.

## REFERENCES

[1] Hietanen, Mika, and Johan Eddebo. "Towards a definition of hate speech With a focus on online contexts." Journal of Communication Inquiry 47.4 (2023): 440-458.

[2] Malik, Jitendra Singh, et al. "Deep learning for hate speech detection: a comparative study." International Journal of Data Science and Analytics 20.4 (2025): 3053-3068.

[3] Chakravarthi, Bharathi Raja, et al. "Detecting abusive comments at a fine-grained level in a low-resource language." Natural Language Processing Journal 3 (2023): 100006.

[4] Subramanian, Malliga, et al. "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models." Alexandria Engineering Journal 80 (2023): 110-121.

[5] Mubarak, Hamdy, Sabit Hassan, and Shammur Absar Chowdhury. "Emojis as anchors to detect arabic offensive language and hate speech." Natural Language Engineering 29.6 (2023): 1436-1457.

[6] Kentmen-Cin, Cigdem. "Hate speech on social media: A systemic narrative review of political science contributions." Social Sciences 14.10 (2025): 610.

[7] Perera, Andrea, and Pumudu Fernando. "Cyberbullying detection system on social media using supervised

machine learning." Procedia Computer Science 239 (2024): 506-516.

[8] Faraj, Azhi, and Semih Utku. "Comparative analysis of word embeddings for multiclass cyberbullying detection." UHD Journal of Science and Technology 8.1 (2024): 55-63.

[9] Kagi, Shivkumar. "Cyberbullying detection using machine learning." Journal of Scientific Research and Technology (2025): 148-157.

[10] Jamjoom, Arwa A., et al. "Robertanet: Enhanced robertatransformer based model for cyberbullying detection with glove features." IEEE access 12 (2024): 58950-58959.