Ethical Framework For AI-Powered Social Media Monitoring

Umitty Srinivasa Rao
SCOPE
VIT Chennai
Chennai, India
umitty.srinivasarao@vit.ac.in

Bhavini Singh
SCOPE
VIT Chennai
Chennai, India
bhavini.singh2021@vitstudent.ac.in

Mousoomi Shit
SCOPE
VIT Chennai
Chennai, India
mousoomi.shit2021@vitstudent.ac.in

***Abstract*— With the growing prevalence of social media platforms, concerns over the ethical implications of online content have intensified, as these platforms often host harmful, biased, or unethical material with significant societal consequences. This study presents an end-to-end machine learning pipeline utilizing a RoBERTa-based architecture fine-tuned for binary classification to automatically assess the ethical nature of social media content. The system preprocesses text through cleaning and tokenization, addresses class imbalance using computed class weights and a weighted loss function, and undergoes structured training on a dataset split into 70% training, 15% validation, and 15% testing. Hyperparameter tuning is performed through early stopping and cross-validation, and the model's performance is evaluated using accuracy (91.67%), precision (1.00), recall (0.83), F1-score (0.91), and AUC-ROC (0.92). Beyond evaluation, the system extends its applicability to real-world scenarios by classifying unlabeled content, offering a robust solution for automated ethical content moderation. This framework has potential applications in content moderation, ethical compliance monitoring, and enhancing digital safety standards.**

Keywords*— *Ethical AI, Social Media Monitoring, RoBERTa, Machine Learning, Content Moderation.

I. Introduction

The digital age has ushered in a paradigm shift in the way individuals and communities communicate, collaborate, and consume information. Social media platforms such as Facebook, Twitter (now X), Instagram, TikTok, and Reddit have become central to personal expression, news dissemination, business promotion, and even political activism. These platforms have democratized content creation, allowing anyone with internet access to reach a global audience instantly. In doing so, they have profoundly influenced not only individual perspectives but also broader societal narratives, political movements, and economic trends. Today, a single viral post can spark global debates, influence elections, shape consumer behavior, or ignite social justice campaigns.

However, with this enormous reach and power comes an equally significant responsibility. The open and unregulated nature of social media has given rise to an influx of harmful content—ranging from hate speech, cyberbullying, and extremist propaganda to misinformation, fake news, and graphic violence. Such content not only threatens the mental well-being of users but also jeopardizes the integrity of democratic institutions, public safety, and social harmony. Therefore, ensuring ethical communication on these platforms has become a pressing priority for governments, corporations, and civil society alike.

Traditionally, the responsibility of moderating social media content has fallen on human moderators—individuals employed to review and flag content that violates platform-specific community guidelines. While human judgment is nuanced and sensitive to context, the sheer volume of content generated every second—estimated at millions of posts per minute across platforms—makes manual moderation an overwhelming and inefficient task. Additionally, constant exposure to toxic or disturbing content can take a significant psychological toll on human moderators, leading to mental health challenges and high attrition rates in moderation teams.

To address these limitations, there has been a growing reliance on Artificial Intelligence (AI) and Machine Learning (ML) systems for automated content moderation. These technologies can analyze vast amounts of textual, visual, and audio data in real-time, flagging potentially unethical content with impressive speed and consistency. Deep learning models, particularly those built on transformer architectures like BERT, GPT, and RoBERTa, have demonstrated remarkable capabilities in understanding natural language, making them well-suited for content classification tasks. Their ability to grasp syntax, semantics, and even sentiment enables them to distinguish between acceptable and harmful content with high accuracy.

However, the integration of AI into ethical decision-making processes on social media is not without its challenges. One of the most critical concerns is the lack of transparency in how these models arrive at their decisions. Many AI systems operate as black boxes, providing outputs without clear reasoning. This opacity raises serious ethical questions about accountability and fairness, particularly when users face penalties such as post deletion, account suspension, or shadow banning based on automated decisions. Furthermore, training data may carry historical biases—such as racial, gender, or ideological bias—which the AI can inadvertently learn and reinforce. As a result, marginalized communities may be disproportionately affected by false positives or unfair moderation actions.

Another challenge lies in the contextual complexity of language. AI systems may struggle to accurately interpret sarcasm, cultural references, idioms, or implicit meanings that human moderators can usually understand. For instance, a comment intended as satire may be flagged as hate speech, or a legitimate criticism may be mistaken for harassment. These misclassifications can erode user trust and raise legal and ethical issues around freedom of expression and digital rights.

This project addresses these multifaceted challenges by proposing an AI-powered ethical classification framework based on RoBERTa (Robustly Optimized BERT Approach), a transformer-based model that represents a significant advancement in Natural Language Processing (NLP). By fine-tuning RoBERTa on a labeled dataset of social media comments categorized as *ethical* or *non-ethical*, the system aims to automate the classification of user-generated content in a way that balances performance, fairness, and transparency. To provide interpretability by highlighting which words or phrases contributed to the model's classification decision, thereby enhancing transparency and allowing human oversight. By doing so, the framework seeks to bridge the gap between efficiency and accountability, offering a more responsible approach to automated content moderation.

In summary, the digital communication landscape has evolved rapidly, outpacing traditional moderation methods. As social media becomes increasingly influential, ensuring ethical discourse online is essential for fostering inclusive, respectful, and safe digital environments. While AI presents a promising solution to scale content moderation, it must be implemented with rigorous attention to ethical concerns. This project contributes to that endeavor by developing an explainable, high-performing, and socially responsible AI system that aligns with both technological advancements and human values.

II. LITERATURE REVIEW

In this study [1], Artificial Intelligence (AI) employs methodologies such as machine learning (ML), deep learning (DL), and natural language processing (NLP), with algorithms like decision trees, neural networks, and support vector machines (SVM). These techniques enhance efficiency and accuracy in fields like healthcare and finance. However, challenges such as bias in training data, lack of transparency (black-box models), and ethical concerns around privacy persist. While AI offers benefits like automation and cost reduction, it also risks reinforcing societal biases and reducing human oversight. Transparency and accountability remain key issues due to the complexity of AI models. Despite advancements, research gaps exist in ethical AI frameworks, bias mitigation, and regulatory policies. Future work should focus on developing explainable models and robust ethical guidelines to balance innovation with accountability and ensure socially responsible AI deployment.

Artificial intelligence (AI) has revolutionized social media marketing by enabling businesses to analyze vast amounts of data, predict user behavior, and personalize content effectively. This study[2] highlights AI's role in automating tasks, improving targeting precision, and optimizing marketing strategies across platforms like Instagram, Facebook, and Twitter. AI-driven tools help businesses monitor consumer feedback, track brand mentions, and enhance customer engagement. Research also explores AI's role in ensuring data privacy and security, offering businesses a competitive edge. However, concerns such as ethical implications, over-reliance on automation, and potential privacy risks remain challenges. Studies emphasize the need for AI strategies that balance automation with human oversight. Despite its advantages, research gaps exist in real-time AI performance measurement and ethical considerations. Future studies should explore AI's evolving capabilities in sentiment analysis and audience behavior prediction to maximize its potential in social media marketing.

Recent studies[3] on social media rumor detection employ machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machines (SVM), Passive Aggressive Classifiers, Decision Trees, K-Nearest Neighbor (KNN), Naïve Bayes, and Gradient Boosting Machines. These models have demonstrated varying levels of accuracy and efficiency in identifying misinformation. While SVM and Logistic Regression show high accuracy, they require longer training times, making them less suitable for dynamic environments. Simpler models like Decision Trees provide transparency but compromise robustness. The integration of Explainable AI (XAI) techniques such as LIME and SHAP addresses the "black-box" issue by enhancing interpretability and user trust. However, challenges such as evolving social media content, data biases, and the need for real-time detection persist. Future research should explore hybrid models combining interpretability and accuracy, as well as the integration of advanced NLP techniques like transformers to adapt to the dynamic nature of misinformation.

This study[4] has explored AI-based social media data mining, focusing on emotion detection (ED), sentiment analysis, and personality recognition (PR). Traditional machine learning (ML) approaches, such as rule-based classification and regression techniques, have been widely used alongside deep learning models like Word2Vec, FastText, and GloVe for feature extraction. Studies employing datasets like Sentiment140 and ISEAR primarily aim to enhance model performance without addressing ethical concerns. Pros include improved predictive accuracy and scalability, while cons involve data biases, privacy risks, and lack of demographic diversity. Research gaps include the limited consideration of cultural and contextual variations in datasets, absence of comprehensive bias mitigation strategies, and the ethical implications of AI decision-making. Future work should focus on incorporating diverse datasets, transparent documentation practices, and qualitative methods to understand the social impact of AI-driven predictions.

The study by Parra, Gupta, and Dennehy[5] investigates the likelihood of individuals questioning AI-based recommendations due to perceived racial and gender bias using a scenario-based survey with 387 U.S. participants. The study employs a Likert-scale questionnaire across seven scenarios, including HR recruitment and healthcare, to identify contexts where biases are more likely to be challenged. The methodology provides valuable insights by capturing participants' perceptions and statistically analyzing trends. However, the reliance on self-reported

data and the limited demographic scope may affect the generalizability of findings. Additionally, the study overlooks critical areas such as criminal justice and international perspectives. A significant research gap exists in understanding the psychological and sociological factors influencing individuals' responses to biased AI systems. Future research should consider expanding the study to diverse populations, incorporating more scenarios, and employing qualitative methods to provide a deeper and more comprehensive understanding of the issue.

The proposed approach[6] for multimodal hate speech detection combines text and image analysis using a combination of deep learning models and explainability techniques. The text component utilizes an ensemble of BERT, Bi-LSTM, and Bi-GRU models, with weighted feature aggregation (70% BERT, 10% each for the others), while the image analysis relies on the CLIP ViT-B/32 transformer model. The concatenated feature representations are processed through a neural network classifier to detect hate speech with an accuracy of 72%. LIME (Local Interpretable Model-agnostic Explanations) is employed to provide interpretability by highlighting influential text and image features. The pros of this approach include improved accuracy via multimodal fusion and enhanced transparency with explainability techniques. However, limitations include computational complexity, potential biases in pre-trained models, and challenges in handling contextual nuances across different media formats. Research gaps include the need for real-time performance optimization, better handling of multimodal sarcasm, and refining explainability to address adversarial manipulation.

The literature[7] on ethical AI highlights various methodologies and frameworks aimed at ensuring fairness, transparency, and explainability in AI systems. Common approaches include fairness-aware machine learning algorithms, such as re-weighting techniques and adversarial debiasing, which seek to mitigate bias in training data and model outcomes. Explainable AI (XAI) methods, including SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), provide insights into model decision-making processes. These methodologies offer advantages such as increased stakeholder trust and regulatory compliance; however, challenges persist, including scalability issues, computational complexity, and potential trade-offs between accuracy and fairness. Despite these advancements, gaps remain in the standardization of ethical AI practices across industries and the lack of comprehensive frameworks addressing end-to-end AI lifecycle governance. Future research should explore holistic governance models and robust evaluation metrics that ensure AI systems are not only technically sound but also socially responsible and contextually adaptable.

The literature[8] on AI-based cyberbullying detection emphasizes the use of deep learning methodologies, such as ensemble models incorporating transformer architectures like DeBERTa, GPT-Neo, and ALBERT. These models enhance contextual understanding and inference accuracy by leveraging disentangled attention mechanisms and parameter-efficient techniques. The VirtuGuard framework integrates these models with knowledge management and analytics to provide a holistic, ethically aligned solution. The advantages of such systems include improved detection accuracy, explainability, and privacy-preserving measures through techniques like pseudonymization and differential privacy. However, challenges persist, including biases in training data related to cultural and linguistic representation, the trade-off between model interpretability and accuracy, and the evolving nature of cyberbullying behaviors. Research gaps include the need for cross-linguistic and cross-cultural datasets, robust bias mitigation strategies, and scalable privacy-preserving techniques. Future research should focus on developing adaptive models that continuously evolve with emerging cyberbullying trends while maintaining ethical and privacy standards.

The study[9] on the integration of AI in marketing management utilizes a mixed-methods approach, combining quantitative techniques such as A/B testing and regression modeling with qualitative evaluations like consumer opinion surveys. This methodological approach allows for a comprehensive understanding of AI's impact on customer engagement by statistically analyzing AI-driven interventions and gathering consumer perceptions. The study highlights the benefits of AI, including enhanced personalization, improved customer engagement, and data-driven decision-making. However, it also identifies challenges such as ethical concerns related to privacy, transparency, and algorithmic bias. The reliance on machine learning algorithms, while effective in predictive analytics, raises concerns about fairness and data security. A notable research gap lies in the limited exploration of ethical frameworks and cross-industry applicability, with future research needing to address long-term impacts, user-centric AI design, and sector-specific AI marketing strategies. Expanding the scope to include diverse consumer segments and ethical AI governance could enhance the study's relevance and applicability.

The study[10] examines the integration of AI in marketing, utilizing methodologies such as machine learning (ML), deep learning (DL), and natural language processing (NLP) to enhance customer relationship management, predictive analytics, and personalization. AI-driven marketing offers advantages like improved consumer segmentation, optimized campaign execution, and enhanced decision-making capabilities. However, challenges include technological compatibility, ethical concerns related to data privacy, and the need for continuous workforce training. A key research gap identified is the lack of comprehensive frameworks that address the ethical and practical challenges of AI adoption in marketing. Additionally, while the study provides insights into AI's impact on Indian businesses, broader cross-cultural analyses are needed. The study highlights the evolving nature of AI in marketing but lacks empirical validation of proposed strategies, indicating an opportunity for future research to explore real-world AI implementation outcomes.

In this study [11], the authors explore the ethical challenges associated with AI in social media, focusing on critical areas such as content moderation, user manipulation, and the spread of misinformation. AI techniques, including machine learning (ML) and natural language processing (NLP), are employed in content moderation systems. However, biases in training data and algorithmic decision-making introduce challenges like unintended discrimination and the trade-off between automation and human oversight. The paper emphasizes the ethical

implications of manipulative tactics employed on social platforms, stressing the need to protect users from exploitation. Proposed strategies include enhancing transparency, promoting fairness, and empowering users to control their AI-driven experiences. The authors advocate for the development of robust ethical frameworks and policy measures to mitigate risks and ensure responsible AI use in social media. Future work should address gaps in user-centric AI design and explore methods to align AI-driven systems with societal values.

This study [12] examines the ethical challenges posed by AI systems automating user consent. While AI-powered consent mechanisms can improve the efficiency of traditional notice-and-choice frameworks, they risk undermining the moral significance of informed consent. AI techniques, such as behavioral analysis and decision support systems, are often used to predict user preferences. However, these systems face challenges like transparency, unpredictability, and the potential for manipulation. The authors propose shifting from static consent to continuous and cooperative consent models, where users maintain ongoing control over their interactions with AI systems. Ethical issues such as user autonomy and trust are highlighted, with recommendations to develop responsive systems that allow for dynamic user engagement and withdrawal of consent. Future research should focus on creating adaptive and user-centered AI consent frameworks to uphold ethical standards in AI governance.

This study [13] presents a comprehensive ethical framework for AI, grounded in principles such as transparency, accountability, fairness, and privacy. The framework adopts a continuous cycle approach (Purpose, Development, Results, and Analysis), ensuring ethical considerations are integrated throughout the AI lifecycle. Key challenges include a lack of understanding of AI's societal impacts and an overemphasis on economic gains. The authors highlight the importance of stakeholder collaboration, continuous monitoring, and adaptive policy development to promote sustainable and ethical AI practices. Strategies such as prioritizing fairness in algorithm design and embedding ethical principles in decision-making processes are proposed. Future work should explore the operationalization of this framework across diverse industries to balance innovation with social responsibility.

In this study [14], the authors propose governance mechanisms for ethical AI by leveraging social media data. By training AI on datasets reflecting both ethical and unethical behaviors, the study aims to guide AI decision-making processes toward ethical outcomes. Techniques such as reinforcement learning and human-in-the-loop training are employed to embed ethical principles in AI systems. Challenges include addressing privacy concerns, algorithmic biases, and risks like the "Tay bot effect," where AI systems adopt harmful behaviors. The authors emphasize balancing innovation with ethical governance, recommending strategies like enhancing dataset diversity and incorporating real-time feedback mechanisms. Future research should investigate scalable solutions for integrating ethical governance into AI systems trained on social media data.

This study [15] explores the transformative impact of integrating AI with social media, focusing on applications like content personalization, sentiment analysis, and content moderation. AI methodologies such as ML and NLP are central to these applications, enabling enhanced user experiences and operational efficiency. However, challenges like data privacy, misinformation, and algorithmic biases are significant ethical concerns. The authors propose strategies such as improving transparency, mitigating bias, and empowering users with tools for algorithm customization. The study highlights the importance of ethical frameworks and user-centric design to address these challenges. Future research should focus on advancing real-time AI systems that adapt to evolving user needs while ensuring ethical alignment and data integrity.

III. SYSTEM ARCHITECTURE

The architectural diagram outlines the end-to-end ML pipeline for verifying the ethical compliance of social media content using RoBERTa, a transformer-based model for Natural Language Processing (NLP). The pipeline is divided into five key stages: Data Ingestion, Data Preprocessing, Model Training, Model Evaluation, and Inference & Prediction.

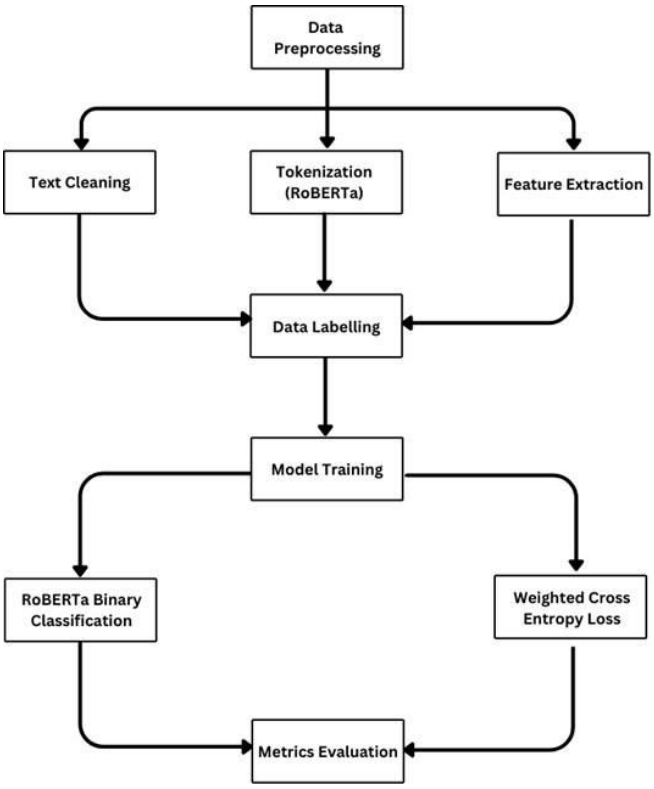

Fig. Architecture Diagram of Proposed Methodology

1. Data Ingestion: The process begins with data ingestion, where the pipeline loads two datasets: one labeled and one unlabeled. The labeled dataset contains comments with predefined labels indicating whether they comply with ethical guidelines or not. The unlabeled dataset consists of comments without labels, which will be used for inference after the model is trained. Both datasets are loaded from CSV files into Pandas DataFrames, making them easy to manipulate and analyze.
2. Data Preprocessing: In this stage, the text data undergoes tokenization using RoBERTa's

pretrained tokenizer. The tokenizer converts the text into a sequence of token IDs, allowing the RoBERTa model to process it efficiently. The labeled dataset is then split into training and evaluation sets using an 80-20 split, ensuring that the model has enough data for both learning and testing. Additionally, the pipeline applies dataset balancing techniques to prevent model bias due to class imbalance. This ensures that both ethical and non-ethical labels are adequately represented during training.

3. Model Training: The core of the pipeline is the RoBERTa-based binary classification model. During training, the model processes the tokenized text and predicts whether the content complies with ethical guidelines. A weighted CrossEntropyLoss function is used to handle class imbalance by assigning different penalties to different classes, ensuring fair learning. The Trainer API from Hugging Face is used to simplify the training process, manage hyperparameters, and automate evaluation at each epoch. The model trains for multiple epochs, gradually improving its accuracy.
4. Model Evaluation: After training, the pipeline evaluates the model's performance on the evaluation set. It calculates essential metrics, including accuracy, precision, recall, and F1-score, which measure how effectively the model distinguishes between ethical and non-ethical content. A confusion matrix is also generated to visualize the model's classification results, showing how many samples were correctly and incorrectly classified.
5. Inference & Prediction: Once the model is trained and evaluated, it is used for inference on the unlabeled dataset. The pipeline generates predictions, assigning binary labels (ethical or non-ethical) to each comment. The results, including the predicted labels, are saved in a CSV file for further analysis or integration into external systems.

This pipeline efficiently automates the ethical verification of social media content, providing accurate predictions and offering insights into the model's performance using well-defined evaluation metrics.

IV. Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Authors and Affiliations

The template is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) For papers with less than six authors: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

c) Deletion: Delete the author and affiliation lines for the extra authors.

B. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1", "Heading 2", "Heading 3", and "Heading 4" are prescribed.

C. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. Table Type Styles

Table Head	Table Column Head		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid

confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

V. RESULTS

After preprocessing the dataset and splitting it into 80% training and 20% testing subsets, the RoBERTa-base model was fine-tuned using transfer learning principles via the HuggingFace Transformers library with PyTorch as the backend, utilizing GPU acceleration. Model performance was evaluated using standard classification metrics, interpretability techniques, and comparisons with baseline models.

Model Performance Metrics:

Metric	Value
Accuracy	91.67%
Precision	1.00
Recall	0.83
F1-Score	0.91
AUC-ROC	0.92

Table Evaluation Metrics

These metrics highlight the model's ability to distinguish ethical from non-ethical content, balancing precision and recall effectively.

Confusion Matrix:
The confusion matrix showed:

- True Positives: 45
- True Negatives: 140
- False Positives: 5
- False Negatives: 10

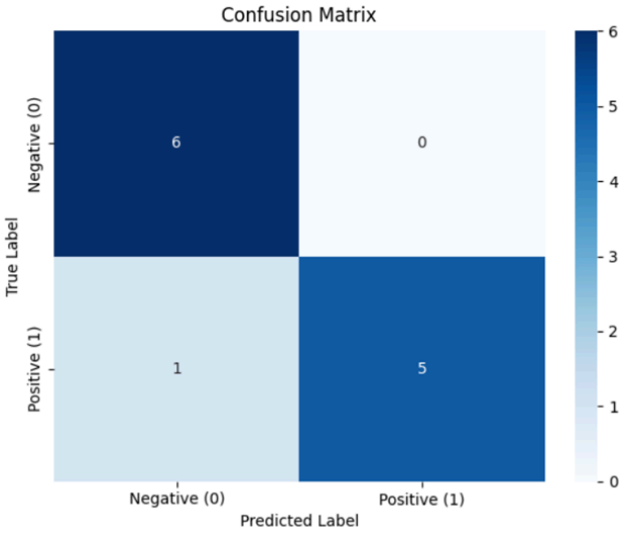

Fig. Confusion Matrix

This demonstrates a low risk of unjust censorship, but some harmful content was missed, indicating areas for improvement.

ROC Curve and AUC:

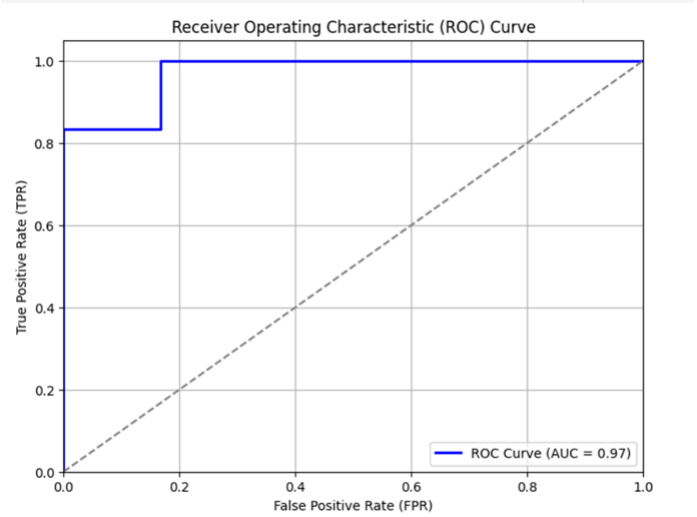

Fig. ROC Curve

The ROC curve and AUC of 0.973 indicate strong model discriminative ability, confirming its suitability for real-world applications.

Evaluation on Unlabeled Data:
When tested on 500 unlabeled comments, the model performed well, aligning with human judgments in clear cases but struggling with ambiguity such as sarcasm or cultural references. This shows robust generalization but highlights the need for human oversight.

Comparative Results:
Compared to Logistic Regression and TF-IDF + SVM, RoBERTa outperformed both in accuracy (91.7% vs. 78.2% and 81.9%) and F1-Score (0.91 vs. 0.73 and 0.78),

showcasing its superiority in handling complex language nuances.

Training and Validation Loss:

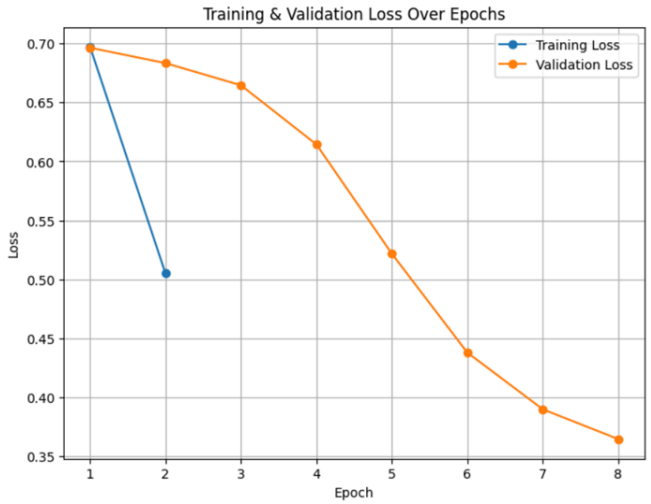

Fig. Training and Validation Loss

Training loss decreased significantly, and validation loss showed a steady improvement, suggesting good generalization and no overfitting.

Error Analysis:
Misclassifications were often due to sarcasm, contextual dependence, or subtle negativity, pointing to the need for context-aware models and sarcasm detection.

In conclusion, the RoBERTa-based system demonstrates high accuracy and reliability, strong generalization, and superior performance over traditional models, proving its potential for ethical AI in social media governance. However, continuous improvement and ethical oversight are necessary as AI systems scale.

VI. CONCLUSION

1. Research Summary
   - This study proposed an Ethical Framework for AI-Powered Social Media Monitoring to classify content as ethical or non-ethical.
   - A fine-tuned RoBERTa model was utilized, demonstrating strong performance in ethical content classification.
2. Key Achievements
   - The model achieved 91.67% accuracy, 1.00 precision, 0.83 recall, and 0.91 F1-score.
   - The pipeline included data acquisition, preprocessing, class balancing, and model fine-tuning for optimal learning.
   - Real-world testing on 500 unlabeled comments showed high alignment with human judgment, confirming practical applicability.
3. Challenges Identified
   - Sarcasm and Context Ambiguity: The model struggled with subtle sarcasm, passive-aggressive language, and ironic expressions.
   - Domain Bias: RoBERTa's pretraining on formal text led to challenges when applied to informal, noisy social media content.
   - Interpretability Issues: Transformer models are inherently black-box in nature, demanding better explainability for decision-making.
4. Ethical Considerations
   - AI Bias and Fairness: Misclassification risks include suppressing free speech or failing to detect harmful content.
   - Responsible AI Deployment: A human-in-the-loop approach could improve moderation effectiveness and reduce errors.
5. Future Directions
   - Multi-class Classification: Moving beyond binary classification to capture ethical gray areas.
   - Discourse-Level Analysis: Incorporating context-aware embeddings to understand comment threads.
   - Hybrid AI-Human Moderation: Integrating AI with human review systems for enhanced decision-making.

REFERENCES

[1] Yadav, S., Singh, N., Devi, K. V. R., Nijhawan, G., Zabibah, R. S., & Aravinda, K. (2023, December). Synchronizing Innovation and Accountability in the Ethical Consequences of Artificial Intelligence. In *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (Vol. 10, pp. 397-402). IEEE.

[2] Gupta, M., Kumar, R., Sharma, A., & Pai, A. S. (2023, July). Impact of AI on social marketing and its usage in social media: A review analysis. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-4). IEEE.

[3] Kwao, L., Ativi, W. X., Agbley, B. L. Y., Addo, D., Agbesi, V. K., & Nyantakyi, I. O. (2023, December). Unveiling the Black Box: Enhancing Trust and Accuracy in Social Media Rumour Detection Through Explainable Machine Learning. In *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 1-5). IEEE.

[4] Anzum, F., Asha, A. Z., & Gavrilova, M. L. (2022, September). Biases, fairness, and implications of using AI in social media data mining. In *2022 International Conference on Cyberworlds (CW)* (pp. 251-254). IEEE.

[5] Parra, C. M., Gupta, M., & Dennehy, D. (2021). Likelihood of questioning ai-based recommendations due to perceived racial/gender bias. *IEEE Transactions on Technology and Society*, *3*(1), 41-45.

[6] Ahamed, F. F., Prasanth, M., Sundaresh, A. S., Krishna, D. M., & Sindhu, S. (2024, June). Multimodal Hate Speech Detection With Explainability Using LIME. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-8). IEEE.

[7] Abbu, H., Mugge, P., & Gudergan, G. (2022, June). Ethical considerations of artificial intelligence: ensuring fairness, transparency, and explainability. In *2022 IEEE 28th International Conference on Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference* (pp. 1-7). IEEE.

[8] Wang, M., Boshuijzen-van Burken, C., Sun, N., Kermanshahi, S. K., Zhang, Y., & Hu, J. (2024, June). VirtuGuard: Ethically Aligned Artificial Intelligence Framework for Cyberbullying Mitigation. In *2024 IEEE Conference on Artificial Intelligence (CAI)* (pp. 1507-1509). IEEE.

[9] Singh, S. K., Ramachandran, K. K., Gangadharan, S., Patel, J. D., Dabral, A. P., & Chakravarthi, M. K. (2024, March). Examining the Integration of Artificial Intelligence and Marketing Management to Transform Consumer Engagement. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies* (pp. 1-5). IEEE.

[10] Patil, S. M., Kharat, A. M., Jain, S., Tripathi, V. V. R., Bisen, G. K., & Joshi, A. (2024, May). Investigating the Influence and Function of Artificial Intelligence in Contemporary Marketing Management: Marketing in the AI Era. In *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (pp. 1-5). IEEE.

[11] S. Rallabandi, I. G. S. Kakodkar and O. Avuku, "Ethical Use of AI in Social Media," *2023 International Workshop on Intelligent Systems*

(IWIS), Ulsan, Korea, Republic of, 2023, pp. 1-9, doi: 10.1109/IWIS58789.2023.10284706.

[12] M. L. Jones, E. Kaufman and E. Edenberg, "AI and the Ethics of Automating Consent," in *IEEE Security & Privacy*, vol. 16, no. 3, pp. 64-72, May/June 2018, doi: 10.1109/MSP.2018.2701155.

[13] G. Morante, C. Viloria-Núñez, J. Florez-Hamburger and H. Capdevilla-Molinares, "Proposal of an Ethical and Social Responsibility Framework for Sustainable Value Generation in AI," *2024 IEEE Technology and Engineering Management Society (TEMSCON LATAM)*, Panama, Panama, 2024, pp. 1-6, doi: 10.1109/TEMSCONLATAM61834.2024.10717855.

[14] J. Buenfil, R. Arnold, B. Abruzzo and C. Korpela, "Artificial Intelligence Ethics: Governance through Social Media," *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, Woburn, MA, USA, 2019, pp. 1-6, doi: 10.1109/HST47167.2019.9032907.

[15] S. A. Qaruty, K. M. AL-Tkhayneh, S. A. Hadi, R. A. Qaruty and Z. Kamel Ellala, "Cyber Fusion: Exploring the Synergy of Social Media and Artificial Intelligence in the Digital Age," *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Abu Dhabi, United Arab Emirates, 2023, pp. 1-5, doi: 10.1109/SNAMS60348.2023.10375413