AI-BASED FRAMEWORK FOR AUTOMATED DETECTION AND REPORTING OF SOCIAL MEDIA VIOLATIONS

International Conference on Emerging Horizons in AI and Data Science (ICEHAIDS-2024)
PAPER ID - ADS-313
21 th December 2024
Author: Asmita Mishra
Co-Author: Anubhuti Jaiswal
Prof. Parag Sohoni

INTRODUCTION

Problem:

Social media platforms are seeing more harmful content, like hate speech, fake news, cyberbullying, and explicit material.

Key challenges include:

- Harmful content spreads quickly before it can be stopped.
- Finding a balance between free speech and keeping users safe.
- Fixing algorithmic biases in AI moderation systems.

Recent Studies:

Over 70% of users in 2024 reported encountering harmful content on social media platforms in which Cyberbullying and explicit material were the most frequent issues.

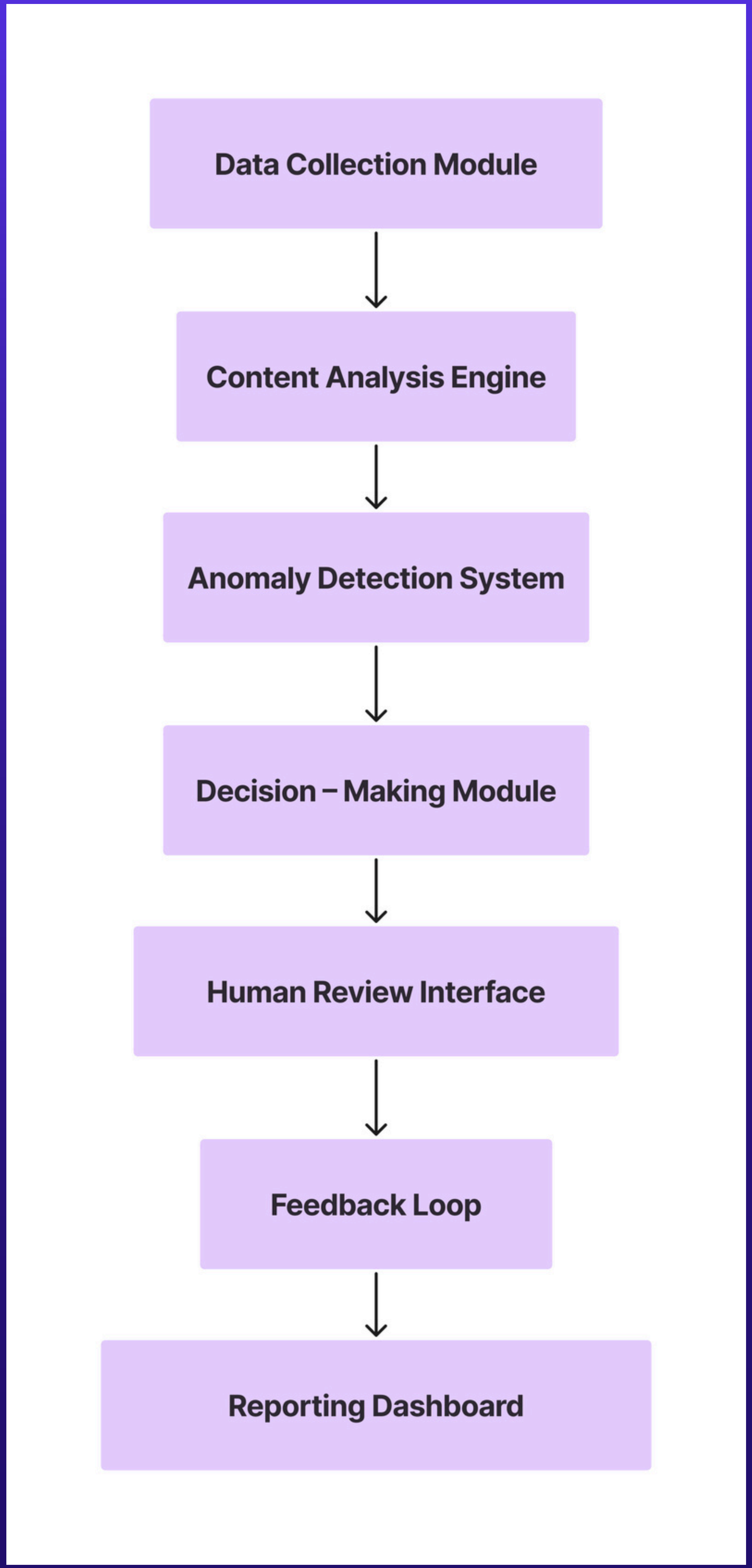
Objective:

- To develop an AI-based framework that detects and reports social media violations efficiently and ethically.
- Enhance accuracy, fairness, and transparency in moderation processes.

METHODOLOGY

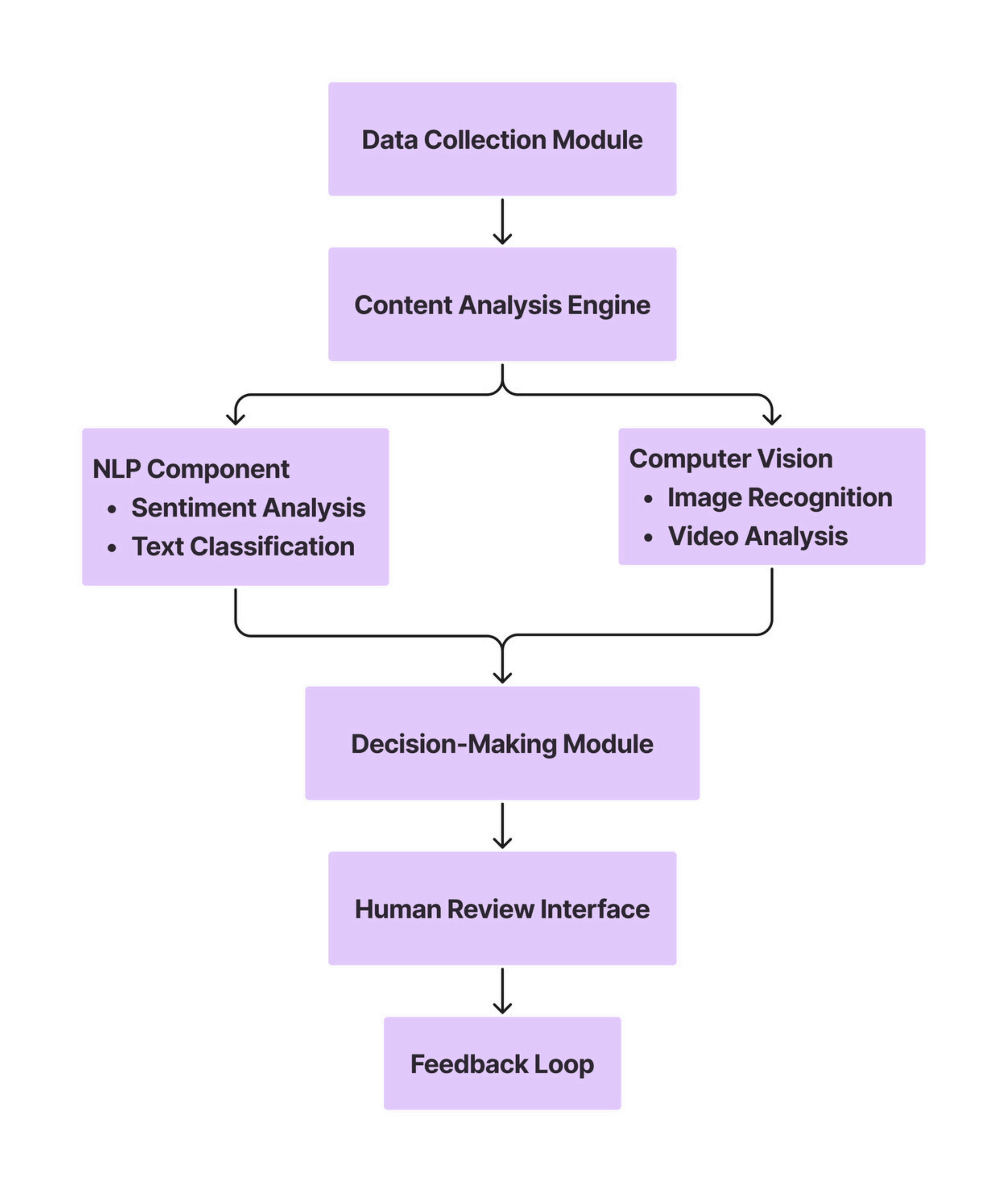
COMPONENTS OF THE FRAMEWORK:

- **Data Collection and Preprocessing:** Gather data from social media, datasets, and reports. Clean text and standardized multimedia.
- **Content Analysis Engine:** Apply NLP (sentiment, entity recognition, violation detection) with BERT. Use CNNs for object detection and video motion analysis.
- **Anomaly Detection System:** Use ML for pattern and anomaly detection.
- **Decision-Making Module:** Recommend flagging or removal actions based on model outputs.
- **Human Review Interface:** Moderators validate flagged content.
- **Feedback Loop:** Refine ML models using moderator feedback.
- **Reporting Dashboard:** Display real-time analytics on flagged content and system performance.



METHODOLOGY

MULTI-MODAL PROCESSING PIPELINE:



- The proposed model operates as a multi-pipeline system, enabling parallel processing of text, images, videos, and user behavior.
- Each pipeline specializes in a specific content type (e.g., NLP for text, computer vision for images/videos) to optimize performance and scalability.
- Results from all pipelines are integrated in the decision-making module for cohesive and accurate actions.

Tools Required: TensorFlow, PyTorch, OpenCV, Amazon Rekognition, and WebPurify

RESULTS...

Experiments Performed:

A comparative analysis was performed using dataset of size:
Approximately 1 million entries.

- 500,000 text posts categorized into hate speech, misinformation, and cyberbullying.
- 200,000 images, including explicit and violent content.
- 100,000 videos containing explicit or illicit activity-related content.

Performance Metrics:

We have derived Accuracy, Precision, Recall, and F1-Score from confusion matrices for comparison with existing state-of-the-art models such as YOLOv7, BERT, and Faster R-CNN.

Confusion Matrix:

A confusion matrix is a table used to evaluate a classification model's performance by showing the number of correct and incorrect predictions, helping to assess how well the model distinguishes between classes.

RESULTS

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Confusion Matrix for Proposed Model:

	Predicted Positive	Predicted Negative
Actual Positive	80 (TP)	20 (FN)
Actual Negative	10 (FP)	90 (TN)

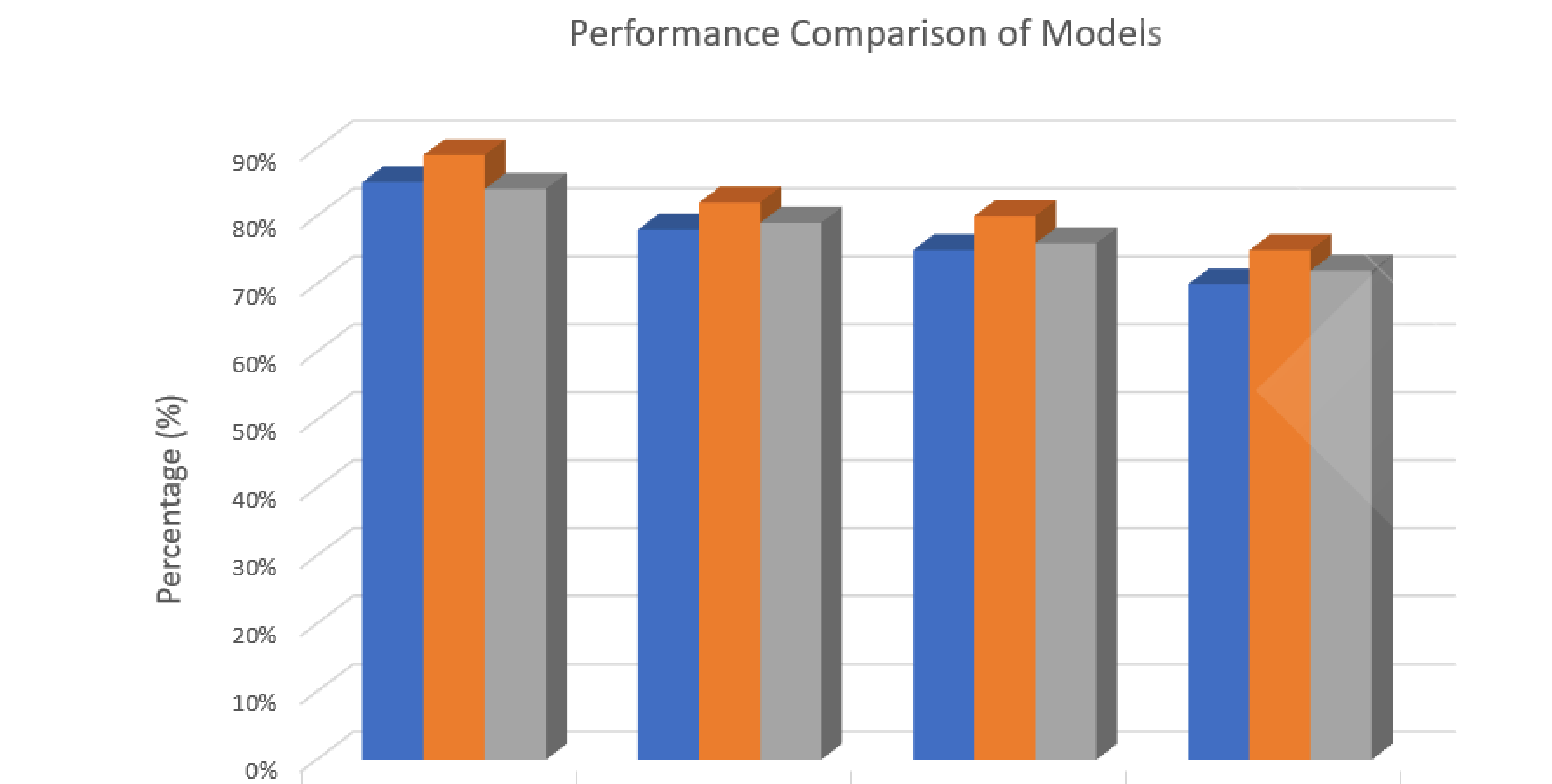
Using this Confusion Matrix, we can calculate the Performance Metrics:

- **Accuracy:** 85%
- **Precision:** 89%
- **Recall:** 80%
- **F1-Score:** 84%.
- **Scalability:** Achieved <100ms latency during high activity simulations.

RESULTS...

BAR CHART REPRESENTATION OF PERFORMANCE COMPARISON MODELS

Existing Model A: YOLOv7
Existing Model B: BERT
Existing Model C: Faster R-CNN



	Proposed Model	Existing Model A	Existing Model B	Existing Model C
Accuracy (%)	85%	78%	75%	70%
Precision (%)	89%	82%	80%	75%
F1-Score (%)	84%	79%	76%	72%

DISCUSSION

KEY INSIGHTS:

- **Multi-Modal Analysis:** Combines text, images, and videos for better detection accuracy.
- **Reduces Algorithmic Bias:** Continuous feedback and oversight minimize algorithmic biases.
- **Builds Trust:** Enhances user confidence through transparent AI decisions.
- **Speeds Up Review:** AI reduces manual moderation, improving efficiency.
- **Outperforms Existing Methods:** Offers greater accuracy, scalability, and efficiency than current solutions.

CONCLUSION

SUMMARY:

- Developed an AI framework for detecting harmful content in text, images, and videos.
- Achieves higher accuracy and efficiency than existing methods.
- Minimizes algorithmic bias with feedback and human oversight.
- Builds trust through transparent decisions and faster reviews.

FUTURE SCOPE:

- Further optimize the system to handle large-scale platforms with diverse content types.
- Improve the AI's ability to detect emerging forms of harmful content.
- Develop algorithms that are more globally inclusive, considering diverse cultural contexts for better content moderation.

REFERENCES...

- Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy.
- Gonzalez, M., & Garcia, A. (2020). Cultural Sensitivity in AI: A Framework for Ethical AI Development.
- Kleinberg, J. et al. (2018). Human Decisions and Machine Predictions.
- Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating Bias in Machine Learning.
- Zhang, B., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning.

THANK YOU

...

Any Questions?