MULTIMODAL HATE SPEECH DETECTION WITH EXPLAINABILITY

*Atul Saju Sundaresh, *Fayas Ahamed F, *Manoj Krishna D, *Prasanth M, **Dr. Sindhu S

*UG Scholar, **Professor, Department of Computer Science and Engineering, N.S.S College of Engineering, Palakkad

PROBLEM STATEMENT

- Hate speech in social media is an increasing problem that can negatively affect individuals and society as a whole.
- Moderators on social media platforms need to be technologically supported to detect problematic content and react accordingly.
- Models provide powerful predictions while being opaque and offering little transparency, this is known as the black box problem.
- There is a lack of trust due to the hidden feature of the decisions made.
- The lack of information about when the model fails or succeeds and the inability to detect errors and correct them may cause problems.

OBJECTIVES

- Our objective is to detect hate speech with high precision using multimodal approach.
- To provide explanation on how the prediction is provided, by highlighting the important parts from both text and image given a meme.

DATASET

- First dataset used is jigsaw toxic comment classification challenge which is sourced from Kaggle and has a size of 1,59,572 entries.
- Second dataset used is Hateful Memes Dataset which is sourced from Facebook AI.Has over 8,500 multimodal examples.

TOOLS

- Tensorflow
- Python
- Word Embeddings: FastText and GloVe
- Pandas
- LIME (Local Interpretable Model-Agnostic Explanations)

METHODOLOGY

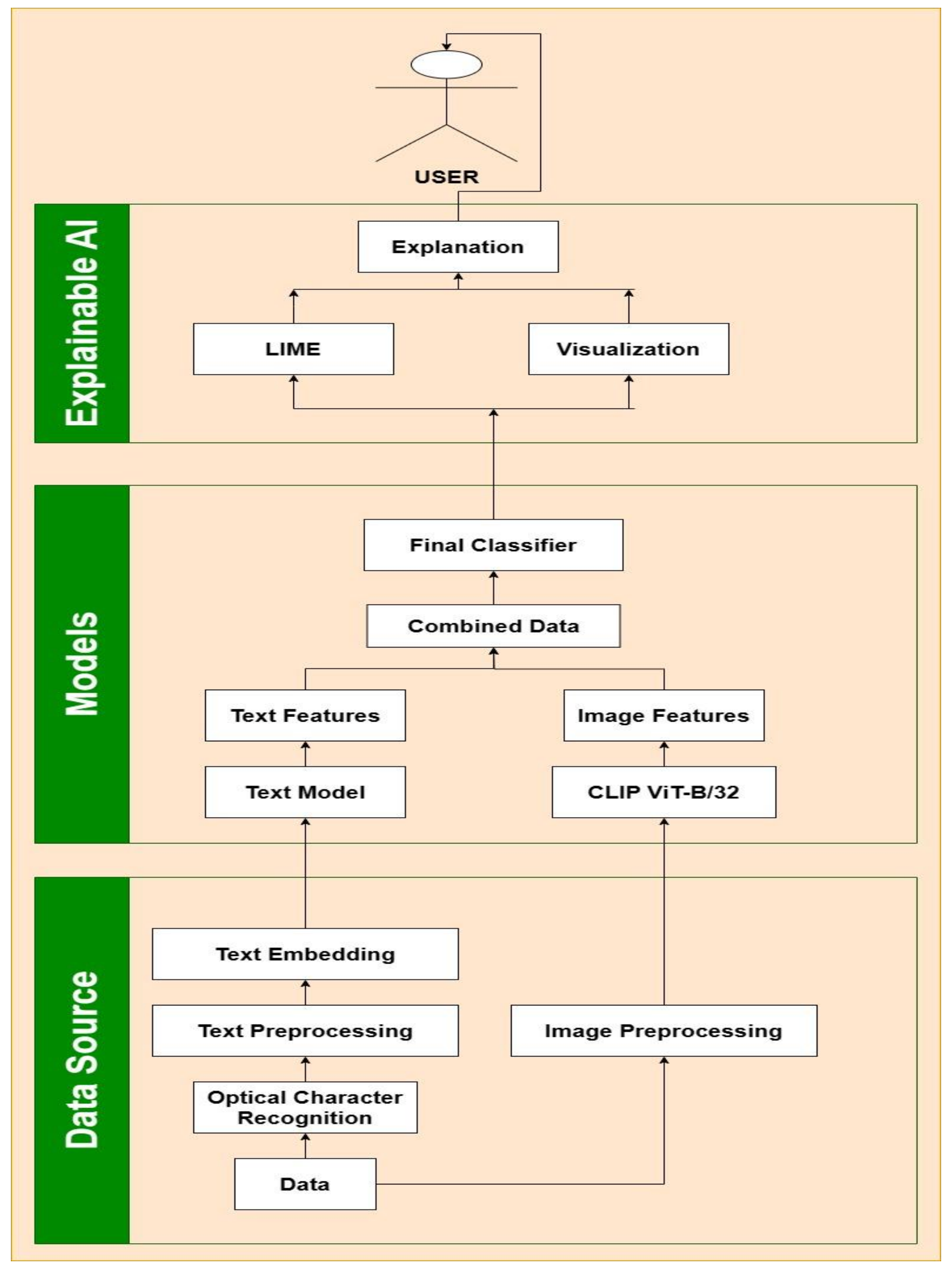


Fig 1 : General Architecture

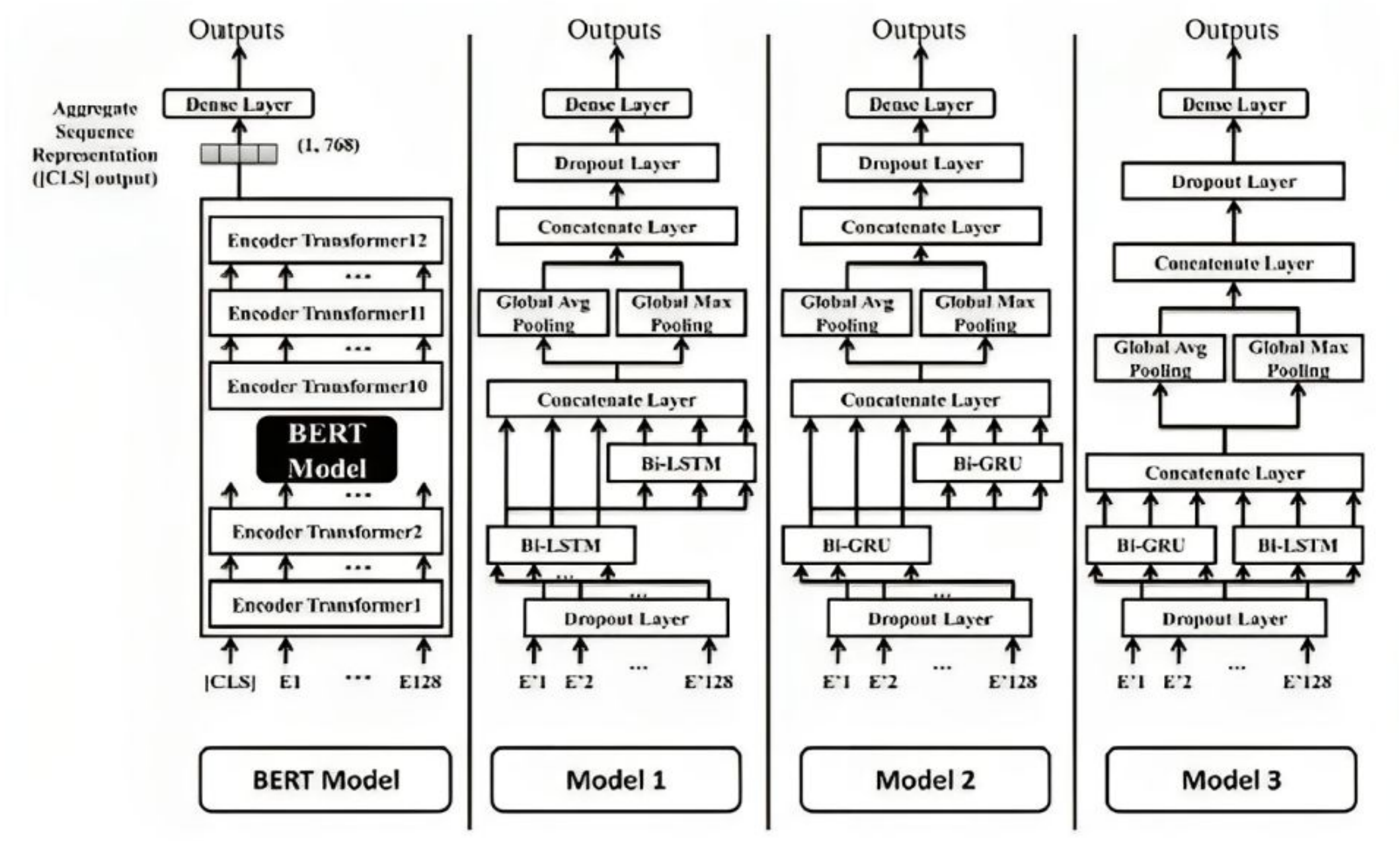


Fig 2 : Detailed Description of Text Model

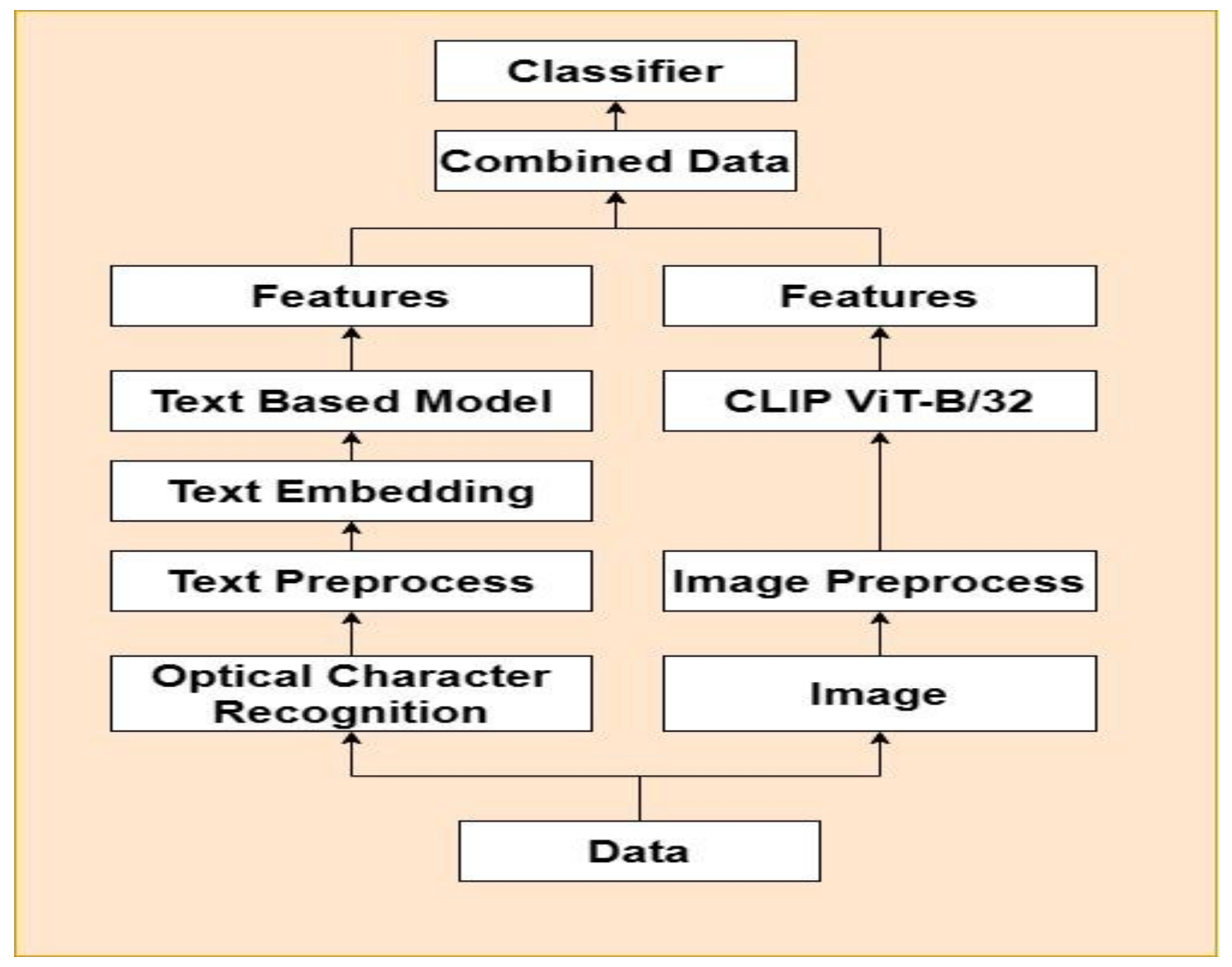


Fig 3 : Multi-modal Module

RESULT & ANALYSIS

- The proposed Multimodal model achieved a validation accuracy of 72%, outperforming models using only text (65.65%) or images (68.53%).

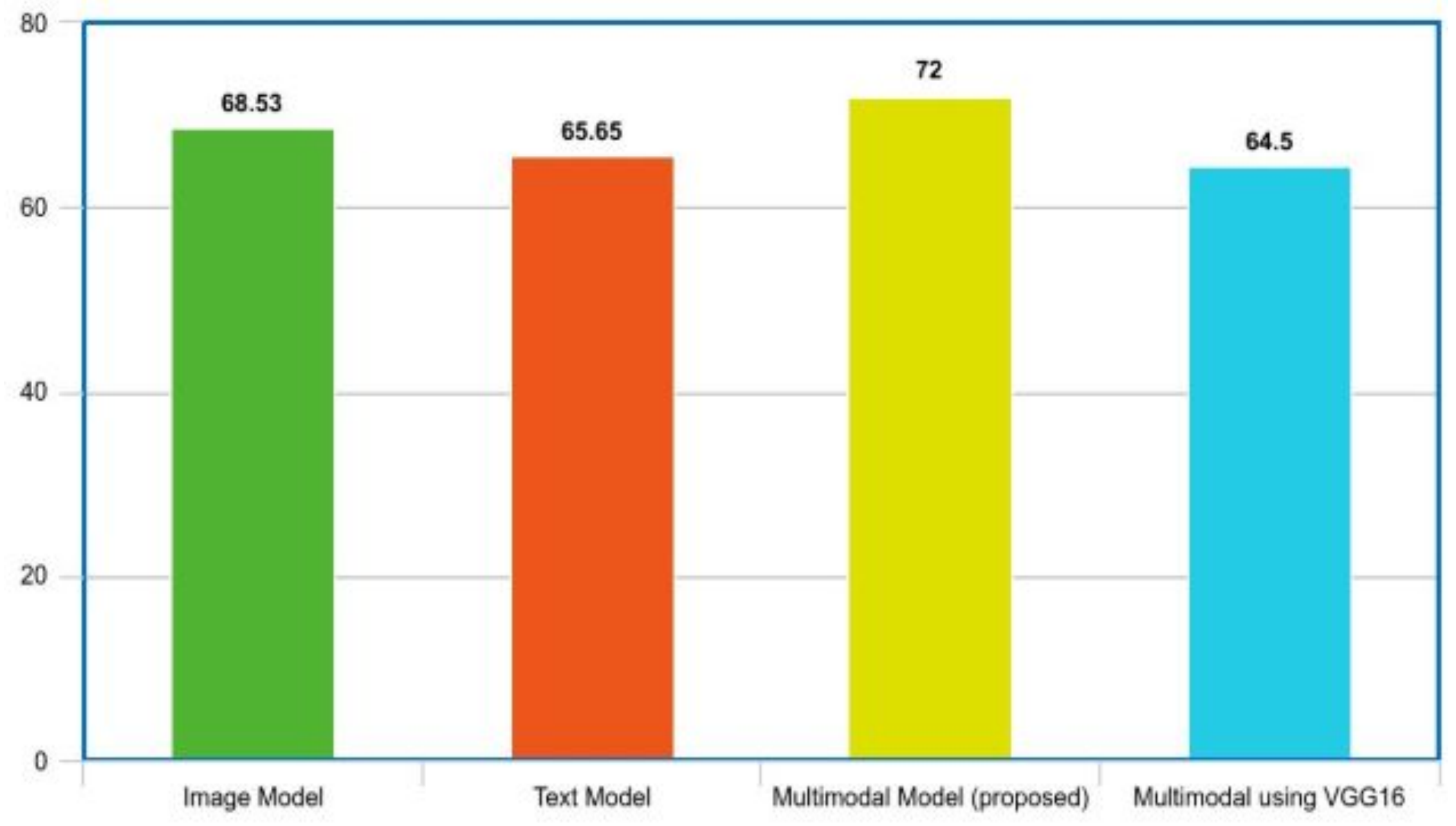
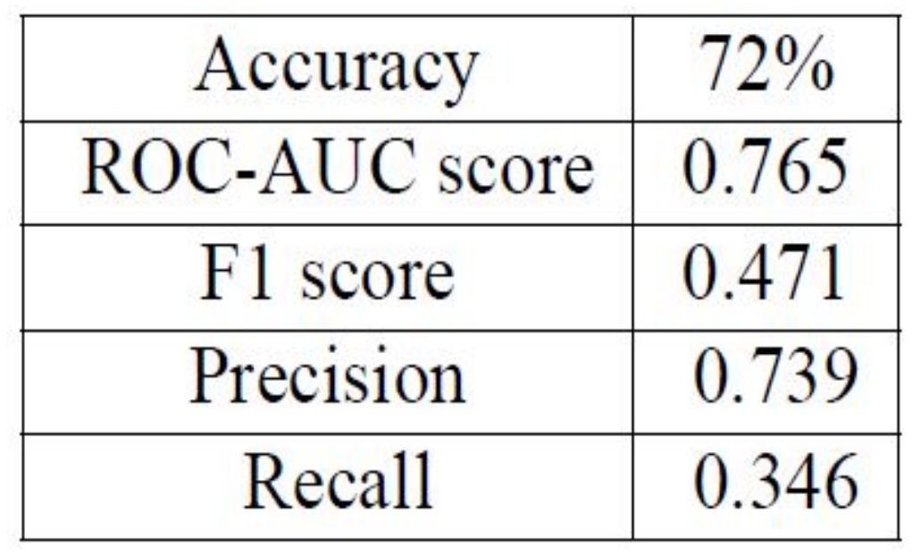


Fig 4 : Accuracy Chart

- ROC-AUC score of 0.765 indicates strong ability to distinguish hate speech and F1 score of 0.471 suggests a moderate balance between precision and recall.

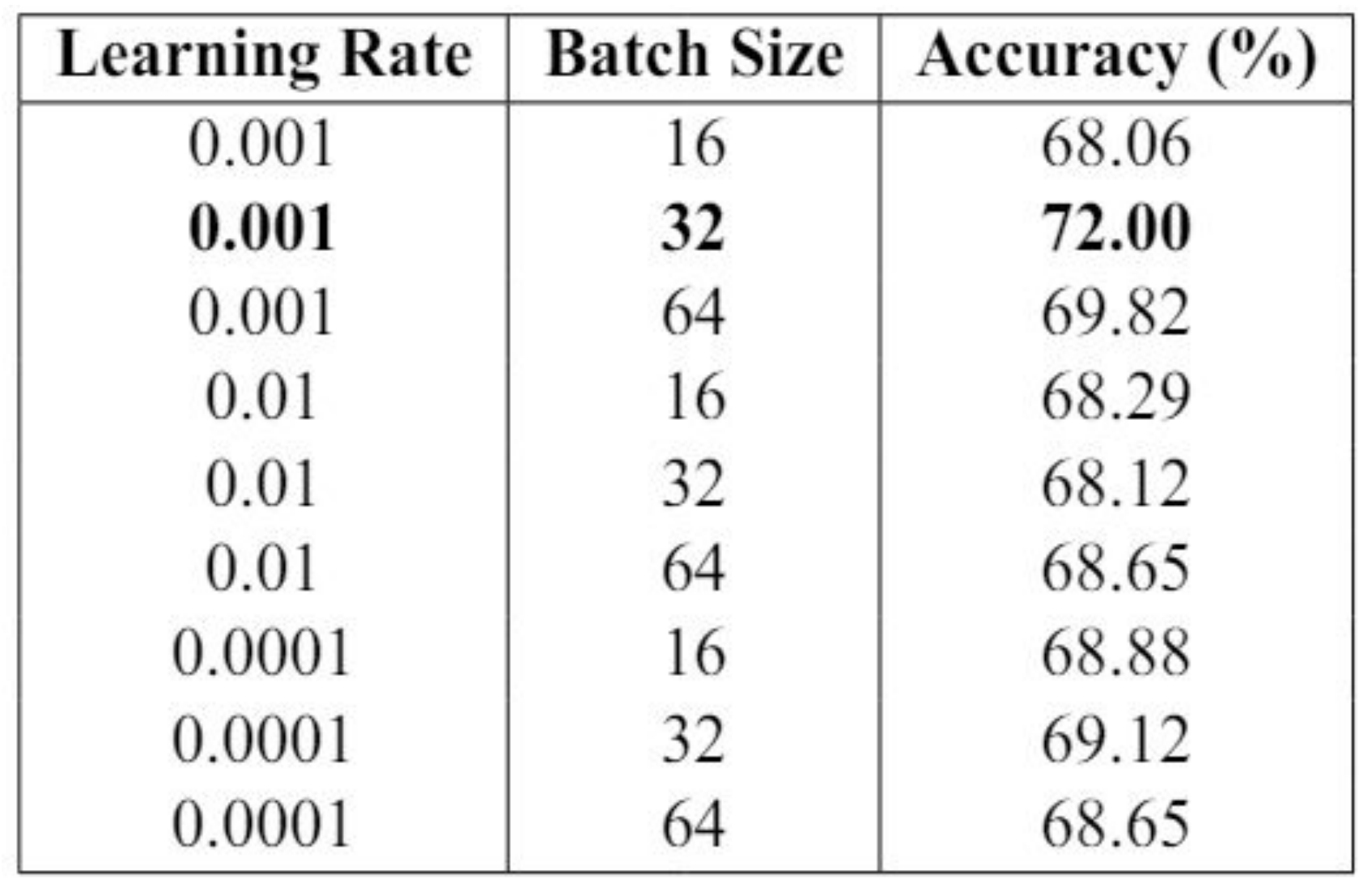
Table 1 : Performance Analysis



Accuracy	72%
ROC-AUC score	0.765
F1 score	0.471
Precision	0.739
Recall	0.346

- Images significantly influence hate speech perception. Same text with different images resulted in varying hate classifications.
- LIME Analysis provided insights into textual elements affecting model predictions.
- Learning rate of 0.001 and batch size of 32 achieved the highest accuracy (72.00%).

Table 2 : Accuracy Results for Different Learning Rates and Batch Sizes



Learning Rate	Batch Size	Accuracy (%)
0.001	16	68.06
0.001	**32**	**72.00**
0.001	64	69.82
0.01	16	68.29
0.01	32	68.12
0.01	64	68.65
0.0001	16	68.88
0.0001	32	69.12
0.0001	64	68.65

CONCLUSION

- The proposed Multimodal model offers a significant advancement in tackling hateful memes online.
- Combining text and image analysis allows for nuanced understanding of the harmful nature of multimodal memes.
- Training on the Facebook Hateful Meme Dataset ensures the model's relevance to real-world situations

FUTURE WORKS

- Generative AI: Utilize generative models to create variations of text and images, improving model adaptation to new hate content.
- Video Analysis : Expand the created model for video analysis and censoring.
- Model Enhancement and Scalability : Optimize the model for scalability and enhancement to handle diverse online content.

REFERENCES

[1] Mazari, Ahmed Cherif et.al., 2023, “BERT-based ensemble learning for multiaspect hate speech detection”, Cluster Computing, pp: 1-15, Springer.

[2] Nandini, D., Schmid, U, 2023, “Explaining Hate Speech Classification with Model Agnostic Methods.”, arXiv preprint, arXiv:2306.00021

[3] Mehta, Harshkumar et.al., 2022, “Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)”, Algorithms, Vol: 15, pp: 291, MDPI.

[4] Christian Meske Enrico Bunde, 2022, ”Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection”, Information Frontiers, Vol: 25, pp: 743–773, Springer.

[5] Malhotra, Shivani, et.al., 2021, ”Bidirectional transfer learning model for sentiment analysis of natural language.”, Journal of Ambient Intelligence and Humanized Computing, Vol:12, pp: 10267–10287, Springer.

[6] Modha, Sandip et.al., 2020, “Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance”, Expert Systems with Applications, Vol: 161, pp: 113725, Elsevier.