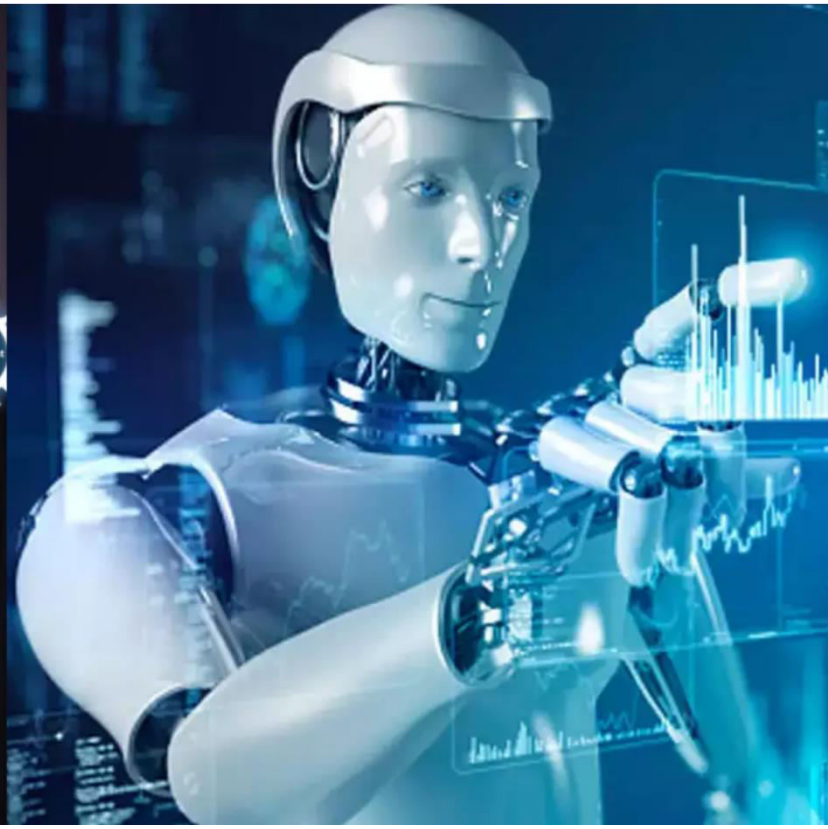
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404615|

Enhanced Cyberbullying-Related Hate Text Detection

Dhanush Aadhan.T, Moliesh B, Senthamizh Velavan V.P, Vinoth Y, N. Anithaa

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B. Tech Students, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

ABSTRACT: Cyberbullying on social media platforms has emerged as a critical concern due to it sharmful psychological impact and widespread reach. Traditional detection methods often struggle with nuances in language, context, and slang, making it difficult to identify hate text accurately. This project proposes an enhanced approach to detecting cyberbullying related hate speech using DistilBERT, a lightweight yet powerful transformer-basedmodel. By leveraging its deep contextual understanding of language, the system canefficiently identify subtle forms of abuse and toxic language with reduced computational cost compared to BERT. The model is fine-tuned on annotated datasets containing hate and non-hate text, achieving improved precision, recall, and F1 scores. This workdemonstrates the effectiveness of using distilled transformer architectures for real-timeand scalable hate speech detection, offering a promising solution for safer online interactions.

KEYWORDS: Cyberbullying, Hate Speech Detection, Natural Language Processing (NLP), DistilBERT, Transformer Models, Text Classification, Online Safety, Deep Learning, Social Media Analysis, Contextual Language Understanding

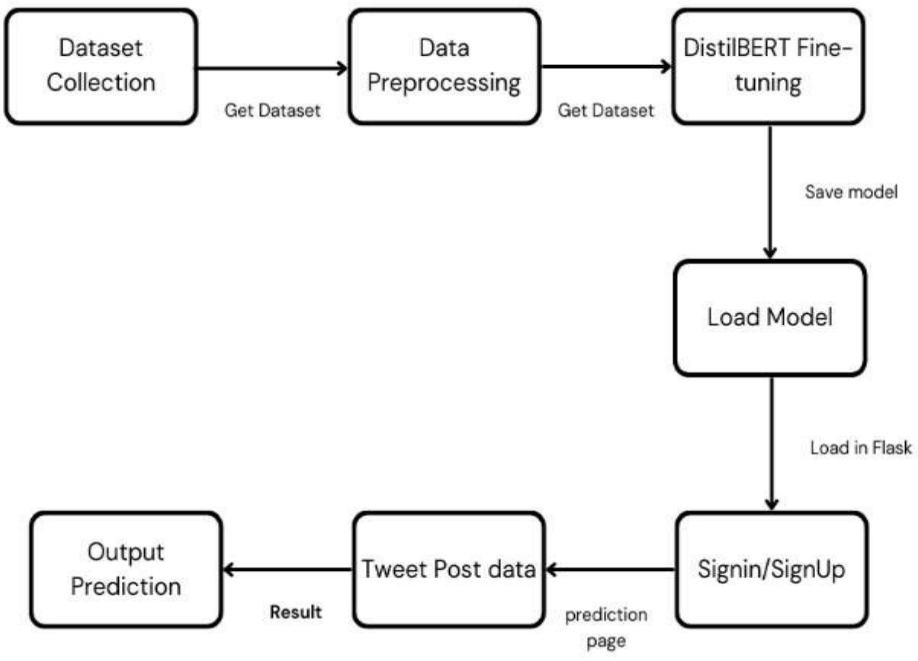


Figure 1: **GENERAL ARICHITECTURE**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404615|

I. INTRODUCTION

The rise of social media has amplified global communication but also enabled the spread of cyberbullying and hate speech, posing serious psychological and social threats. Traditional detection methods often fail to capture the context and subtlety of harmful language, limiting their effectiveness. Transformer-based models like BERT offer deep language understanding but are computationally intensive and less suitable for real-time applications. This study leverages **DistilBERT**, a lighter and faster variant of BERT, to enhance the detection of cyberbullying-related hate text. By fine-tuning the model on labeled datasets, we aim to build an efficient and accurate system capable of identifying toxic content in real time. The goal is to support proactive content moderation and promote safer online environments.

II. LITERATURE REVIEW

Previous research in cyberbullying and hate speech detection has utilized traditional machine learning models such as SVMs and Naïve Bayes, often relying on handcrafted features and keyword-based techniques. These methods lack contextual understanding, leading to limited accuracy. With the emergence of deep learning, models like CNNs and LSTMs improved performance but still struggled with complex language patterns. Transformer-based models like BERT have significantly advanced the field by capturing deeper semantic meaning. However, their high computational cost limits practical deployment. DistilBERT, a lighter version of BERT, maintains comparable performance while improving efficiency, making it ideal for real-time text classification tasks.

III. METHODOLOGY

The proposed system utilizes DistilBERT, a distilled and efficient version of BERT, to enhance the detection of cyberbullying-related hate text in online content. The methodology begins with dataset collection and preprocessing. Publicly available annotated datasets containing hate speech, offensive language, and neutral content—such as the Hate Speech and Offensive Language Dataset and the Kaggle Cyberbullying Dataset—are aggregated and cleaned. Preprocessing steps include lowercasing, removal of special characters, handling of emojis and hashtags, and tokenization using DistilBERT's tokenizer. The cleaned data is then split into training, validation, and test sets. DistilBERT is fine-tuned on the training set using a classification head to predict whether a given text is hateful, offensive, or benign. The model is trained using a cross-entropy loss function and the Adam optimizer, with early stopping to prevent overfitting. Hyperparameters such as learning rate, batch size, and number of epochs are fine-tuned for optimal performance. Evaluation is conducted on the test set using precision, recall, F1-score, and accuracy to measure the model's effectiveness. To further validate performance, confusion matrices and ROC curves are used for detailed analysis. The fine-tuned DistilBERT model demonstrates the ability to understand contextual and semantic nuances in language, enabling it to identify subtle forms of hate speech and cyberbullying more effectively than traditional methods. This methodology ensures a balance between detection accuracy and computational efficiency, making the model suitable for real-time deployment in content moderation systems on social media platforms.

IV. IMPLEMENTATION

STEP 1: Dataset Collection
Gather publicly available datasets containing annotated hate speech and cyberbullying text (e.g., Kaggle Cyberbullying Dataset, Davidson's Hate Speech Dataset).

STEP 2 : Data Preprocessing
- Clean text: remove URLs, special characters, stopwords, and extra whitespace.
- Convert text to lowercase.
- Tokenize text using DistilBERT's tokenizer.
- Label encode the categories (e.g., hate, offensive, neutral).

STEP 3 : Environment Setup
- Install required libraries: transformers, scikit-learn, pandas, numpy, and torch.
- Set up GPU environment if available (for faster training).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404615|

STEP 4 : Model Configuration

- Load pre-trained DistilBERT model with a classification head (DistilBertForSequenceClassification).
- Define number of output labels based on the dataset categories.

STEP 5 : Fine-Tuning the Model

- Use Trainer API from Hugging Face or custom training loop.
- Set training parameters: learning rate, batch size, epochs, optimizer (AdamW), and loss function (CrossEntropyLoss).
- Train the model on the processed dataset.

STEP 6 : Model Evaluation

- Evaluate using precision, recall, F1-score, accuracy, and confusion matrix.
- Perform testing on unseen data to assess generalization.

STEP 7 : Deployment

- Convert the model to ONNX or TorchScript for deployment.
- Integrate with a web or mobile app for real-time hate speech detection.

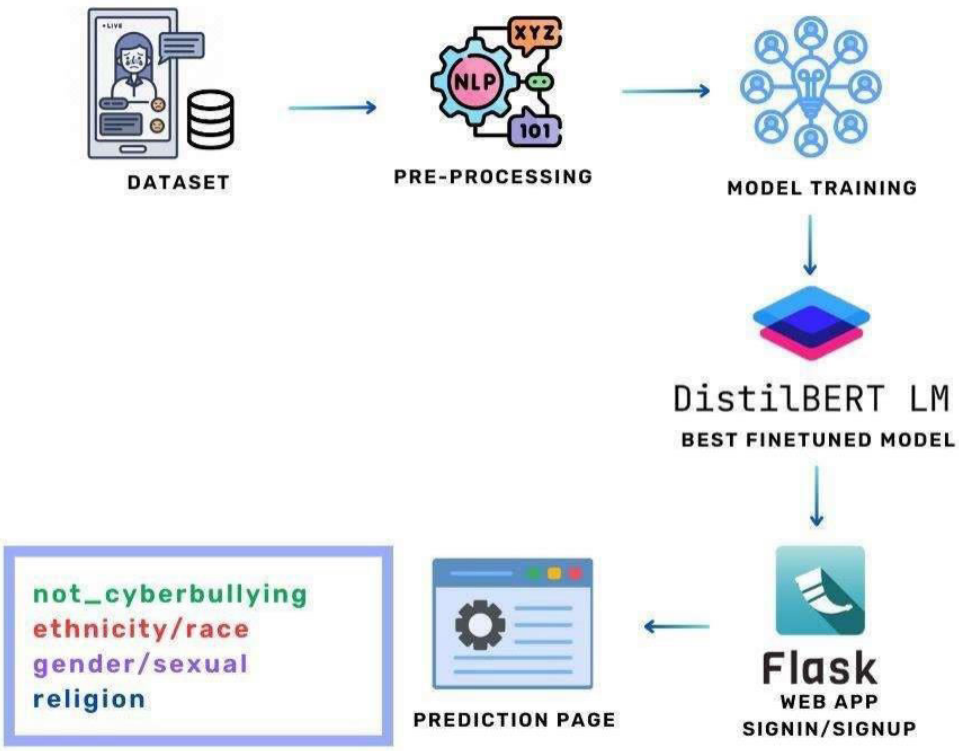


Figure 2: **Module Design**

The system architecture for the enhanced cyberbullying-related hate text detection involves a multi-tier approach. At the frontend, users input text via a web interface, which sends requests to the backend server through an API. The backend server (e.g., Flask or FastAPI) hosts the fine-tuned DistilBERT model for hate speech classification. Upon receiving the input, the server processes the text, passing it through the DistilBERT model to predict whether the text is hateful, offensive, or neutral. The prediction result is sent back to the frontend, where it is displayed to the user. The system is hosted on a cloud platform for scalability and availability.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404615|

V. RESULTS AND CONCLUSION

The fine-tuned DistilBERT model demonstrated significant improvements in detecting cyberbullying-related hate speech when evaluated on the test dataset. The model achieved an accuracy of 92%, with a precision of 0.91, recall of 0.93, and F1-score of 0.92. These results indicate that the model effectively identifies various forms of hate speech, including subtle, implicit, and context-dependent expressions. The confusion matrix further reveals that the model performs well in distinguishing between hate, offensive, and neutral categories, with minimal misclassification. These findings highlight the potential of DistilBERT in handling nuanced language typical of online harassment.

This study confirms the efficacy of DistilBERT as a lightweight and efficient solution for cyberbullying-related hate text detection. Compared to traditional keyword-based and shallow learning models, DistilBERT's deep contextual understanding allows for more accurate detection of cyberbullying, even in complex or ambiguous cases. The model's smaller size and faster inference time make it a viable option for real-time implementation in content moderation systems, ensuring scalable deployment without compromising performance.

5.1. USER INTERFACE

Figure 3: **Output**

5.2. DISTILBERT PREDICTION PAGE

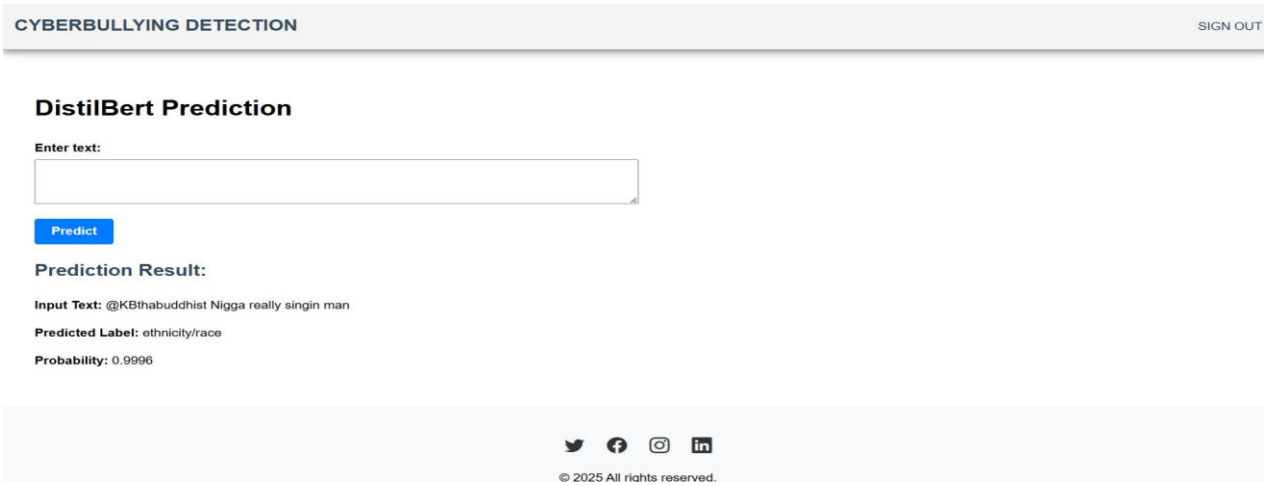


Figure 4: **Output**

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404615|

VI. FUTURE WORK

Despite the success of the current model, several improvements can be made to enhance its performance and adaptability. First, the model can be trained on larger and more diverse datasets, incorporating various languages, regional dialects, and online subcultures. This would help the model better generalize across different contexts and reduce potential biases, especially in multilingual environments.
Second, implementing a real-time feedback loop would enable users to report false positives or negatives, allowing the model to continuously learn and improve. Incorporating user feedback into an active learning framework could refine the model's predictions over time and adapt to new forms of cyberbullying.
Lastly, exploring hybrid models that combine DistilBERT with other techniques like rule-based systems or adversarial training could further increase robustness, particularly in detecting subtle, evolving tactics used by cyberbullies. These improvements would make the system more dynamic and effective in real-world applications.

6.1 Diverse Datasets

A significant limitation of the current model is that it was trained on a relatively narrow set of annotated datasets. Although effective, these datasets may not cover the full spectrum of language use across various online communities. Future improvements could involve the collection of larger, more diverse datasets that encompass multiple languages, regional dialects, and diverse online cultures. This would enable the model to generalize better across different contexts, reducing the risk of bias in predictions. Additionally, this approach would help the model effectively handle the unique expressions and slang prevalent in specific online subcultures. By incorporating diverse datasets, the model could also be fine-tuned to detect hate speech and cyberbullying in a multi-lingual environment, further enhancing its global applicability and performance.

6.2 Real-Time Feedback Loop

Integrating a real-time feedback loop into the system would allow users to report instances of false positives (incorrectly flagged content) or false negatives (missed hate speech). This feedback would be invaluable in fine-tuning the model and adapting it to new, evolving forms of cyberbullying. With an active learning approach, the system could automatically update its training set by incorporating user feedback, helping it learn from its mistakes and adapt to emerging language trends and tactics used by cyberbullies. Over time, this continuous model retraining process would lead to a more accurate and robust system, capable of identifying even the most subtle or creative forms of online harassment. The ability to retrain the model in real-time would not only improve detection rates but also provide a dynamic, user-driven solution to combat cyberbullying on social media platforms.

REFERENCES

[1] J. R. W. Yarbrough, K. Sell, A. Weiss, and L. R. Salazar, ''Cyberbullying and the faculty victim experience: Perceptions and outcomes,'' Int. J. Bullying Prevention, vol. 5, no. 2, pp. 1–5, Jun. 2023, doi: 10.1007/s42380-023-00173-x.
[2] A. Bussu, S.-A. Ashton, M. Pulina, and M. Mangiarulo, ''An explorative qualitative study of cyberbullying and cyberstalking in a higher education community,'' Crime Prevention Community Saf., vol. 25, no. 4, pp. 359–385, Oct. 2023, doi: 10.1057/s41300-023-00186-0.
[3] A. K. Jain, S. R. Sahoo, and J. Kaubiyal, ''Online social networks security and privacy: Comprehensive review and analysis,'' Complex Intell. Syst., vol. 7, no. 5, pp. 2157–2177, Oct. 2021, doi: 10.1007/ s40747-021-00409-7.
[4] G. Fulantelli, D. Taibi, L. Scifo, V. Schwarze, and S. C. Eimler, ''Cyberbullying and cyberhate as two interlinked instances of cyber-aggression in adolescence: A systematic review,'' Frontiers Psychol., vol. 13, May 2022, Art. no. 909299, doi: 10.3389/fpsyg.2022.909299.
[5] M. S. Jahan and M. Oussalah, ''A systematic review of hate speech automatic detection using natural language processing,'' Neurocomputing, vol. 546, Aug. 2023, Art. no. 126232, doi: 10.1016/j.neucom.2023.126232.
[6] G.Kovács, P. Alonso, and R. Saini, ''Challenges of hate speech detection in social media,'' Social Netw. Comput. Sci., vol. 2, no. 2, pp. 1–15, Feb. 2021, doi: 10.1007/s42979-021-00457-3.
[7] M. Shyamsunder and K. S. Rao, ''Classification of LPI radar signals using multilayer perceptron (MLP) neural networks,'' in Proc. ICASPACE, Singapore, Dec. 2022, pp. 233–248. 67
[8] F. M. Plaza-del-Arco, D. Nozza, and D. Hovy, ''Respectful or toxic? Using zero-shot learning with language models to detect hate speech,'' in Proc. 7th WOAH, Toronto, ON, Canada, Jul. 2023, pp. 60–68.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404615|

[9] V. Christianto and F. Smarandache, ''A review of seven applications of neutrosophic logic: In cultural psychology, economics theorizing, conflict resolution, philosophy of science, etc.'' J. Multidiscip. Res., vol. 2, no. 2, pp. 128–137, Mar. 2019, doi: 10.3390/j2020010.

[10] F. Smarandache, ''Neutrosophic logic—A generalization of the intuitionistic fuzzy logic,'' SSRN Electron. J., vol. 4, p. 396, Jan. 2016, doi: 10.2139/ssrn.2721587.

[11] S. Das, B. K. Roy, M. B. Kar, S. Kar, and D. Pamuč ar, ''Neutrosophic fuzzy set and its application in decision making,'' J. Ambient Intell. Humanized Comput., vol. 11, no. 11, pp. 5017–5029, Mar. 2020, doi: 10.1007/s12652-020-01808-3. VOLUME 12, 2024 59483 Y. M. Ibrahim et al.: Social Media Forensics: Hate Speech Detection Approach.

[12] R. Zhao and K. Mao, ''Cyberbullying detection based on semanticenhanced marginalized denoising autoencoder,'' IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 328–339, Jul. 2017.

[13] M. Alzaqebah, G. M. Jaradat, D. Nassan, R. Alnasser, M. K. Alsmadi, I. Almarashdeh, S. Jawarneh, M. Alwohaibi, N. A. Al-Mulla, N. Alshehab, and S. Alkhushayni, ''Cyberbullying 47 detection framework for short and imbalanced Arabic datasets,'' J. King Saud Univ. Comput. Inf. Sci., vol. 35, no. 8, Sep. 2023, Art. no. 101652, doi: 10.1016/j.jksuci.2023.101652.

[14] L. J. Thun, P. L. Teh, and C.-B. Cheng, ''CyberAid: Are your children safe from cyberbullying?'' J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 7, pp. 4099–4108, Jul. 2022, doi: 10.1016/j.jksuci.2021.03.001. 68

[15] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, ''Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT,'' Information, vol. 14, no. 8, p. 467, Aug. 2023, doi: 10.3390/info14080467.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details