Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

Application of Artificial Intelligence for Identifying Social Media Abuse in India

[1]**Dheerendra Singh Patel**, [1]Department of Computer Science, APSU Rewa M.P.
E-mail: dheerendras153@gmail.com
[2]**Dr. Achyut Pandey**, [2]Professor and Head Department of Physics and Computer Science, Govt. TRS College Rewa M.P, Email: achyut.pandey9@gmail.com
[3]**Pratibha Awasthi**, [3]Department of Computer Science, Govt. TRS College Rewa M.P.
[4]**Suraj Gupta**, [4]Department of Computer Science, Govt. TRS College Rewa M.P.

Abstract

Social media is used unlike ever before in India and this has created an international level of interaction online but also abhorrent misuses of social media, such as hate speech, cyberbullying, and harassment towards individuals. This paper discusses how Artificial Intelligence (AI) can detect such abusive content automatically in an Indian multilingual and culturally diverse setting. Based on Natural Language Processing (NLP) and machine learning, such as Logistic Regression, LSTM, and BERT models, the given research assesses how effective AI is in handling abuse-filled material in English, Hindi, and Hinglish. Since BERT is a transformer type of model, it showed high accuracy (93.5) and context sensitivity and hence most reliable in detecting abuse. In spite of the obstacles arising due to the code-mixed languages and cultural nuances, the research signifies the possibility of AI-powered systems to achieve ethical, scalable, and real-time contents moderation and provide a safer Internet in India.

Keywords: *Artificial Intelligence, Social Media Abuse, Hate Speech Detection, BERT Model, Code-Mixed Language (Hinglish)*

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

1. Introduction

The coming of the social media has drastically changed the way human beings communicate with each other and is in a position to share ideas, opinions, and information through the world with just a click of a button. Social media has become part of everyday life in India, which broadly leads to political discussion and social interaction. Recently, in the last decade, Social media penetration in India has increased exponentially to include Facebook, Twitter, Instagram, and WhatsApp. Although these platforms have democratized sharing information and gave marginalized people more room than before, they have also become a hotbed of abuse, cyberbullying, hate speech, and fake news. Any upsurge of such online aggression is greatly dangerous to personal lives, interpersonal relations and the national security. The broad definition of social media abuse incorporates all such forms of negative, threatening, insulting, manipulative conduct that is carried out through online means. In the Indian setting, this mistreatment is mostly along the caste, religious, gender, and political ideology lines. Cases of organized trolling, targeted harassment, spread of fake news, and community hate speech are becoming very prevalent. Such virtual assaults often target women, minorities, activists and journalists and, in addition to infringing their rights, they undermine confidence in the online sphere.

Considering that user-generated content is much and since social media are time-sensitive, automatic detection and control of abusive content are the only solution scalable and effective. The main limitation of traditional rule-based systems is the fact that they do not respond well to changing linguistics patterns, slang and context sensitivities. In the described case, Artificial Intelligence (AI) can provide an innovative replacement. Special types of AI, using such technologies as Natural Language Processing (NLP) and Deep Learning, can be taught to recognize abusive patterns, identify harmful content, and mark possible threats with great levels of accuracy. These systems get better as they progress, harnessed by massive amounts of data and are able to adjust to the new types of abuse. Nonetheless, there are some peculiarities of applying AI to the detection of social media abuse in India. Linguistic diversity in the country, in respect of the number of official languages, which are more that 22 and dialects go in hundreds, makes it hard to develop generalized models. There is further complicating matter of the predominance of code-mixed language like Hinglish (Hindi-English). Even highly developed NLP systems are stretched by sarcasm, use of figures or speech, culture-specific terms, and covert abuse. The ethical questions of privacy, algorithmic fairness, and censorship will also have to be answered so that AI-based interventions do not make discrimination

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

even more persistent and legitimate speech even more suppressed. This project examines how Artificial intelligence can be used in detecting social media abuse in India. It will focus on the development and testing of AI models that can first detect abusive content in different forms- be it texts, images and/or patterns of behavior- and then consider the linguistic and cultural richness in India. The fact that the study targets Indian datasets, regional languages, and code-mixed data makes the proposed study valuable towards the development of inclusive, contextual, and ethical tools in the content moderation process.

2. Literature review

Kaplan and Haenlein (2010) suggests a conceptual platform, constructing categories and criteria of diverse platforms in terms of the degree of self-representation and the presence of social presence. They point out to the fact that social media has become less of a tool of communication and much more a participatory democratized space where people are a not a consumers of content, but also creators of it and distributors of it. Such paradigm shift as the authors address implies significant consequences of online behavior, such as the flourishing of both positive and negative relations of users. Although the work by Kaplan and Haenlein is old by the standards of what they term as FAD (surge in the application of machine learning) users drive dynamic sheds light to the complexity and unpredictability of online discourse, especially in cases of modulating harmful or abusive material. Their perspectives can be an indispensable point of reference in regard to comprehending why conventional content moderation might be inadequate and how we have evolved to reliance on adaptive and data-driven strategies of dealing with social media abuse as is the case in heterogeneous and congested environments such as those in India.

Data Reportal (2024) indicates that currently, there are over 5.04 billion social media users across the globe representing about 62% of the world population. The report also brings out that the number of people who use social media in India is currently one of the highest and the projection is that the users are growing rapidly with a figure of more than 470 million active users of social media by early 2024. This has seen high internet penetration in mobile internet, low data pricing and the use of smartphones as a catalyst to trigger this massive growth. Although such digital expansion has resulted in enhanced access to information, communication, as well as economic opportunities, it has also facilitated provision of the breeding ground of harmful and abusive content. The content generated by users in multilingual and culturally diverse markets such as India is in massive quantity, which

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

makes it too hard to moderate content. These figures speak volumes of the necessity of scalable, intelligent platforms, including machine learning-driven abuse detection mechanisms, to make the internet environment much safer and more accountable.

Localiq report, highlights the mind-blowing level of urgency and velocity of online usage, with users conducting millions of Google searches, posts, messages and video views each and every minute via social media platforms, Facebook, Instagram, TikTok, or. This immense amount demonstrates the rate at which content, such as abusive or damaging content is able to grow and become commonplace in digital ecosystems. In the research perspective, this fast-paced, high-volume churn poses significant issues to the content moderation: systems which are powered by automation are required to process, tag and intervene in huge amounts of multilingual, mixed-media streams in near real-time. The internet minute model provides invaluable insights regarding why the classic moderation systems are unable to address the contemporary dynamics of digital operations. What is more, the notion of an internet minute, the various conversational activities taking place in the online world in thirty seconds, can underscores the behavioral complexity and linguistic diversity of this content, namely, when examining problematic phenomena such as hate speech or harassment . As far as India is concerned, it requires urgent implementation of the scalable machine learning algorithm of sorts which can extract the code-mixed and sensitive, as well as culturally-contextualized harassment contents within a rapidly multiplying and accelerating social media environment.

Batrinca and Treleaven (2015) offer an informative overview of the rapidly growing area of social media analytics, analyzing a large number of different professional applications, tools and platforms, with which social media data could be processed. The authors emphasize the importance of the social media as the source of unstructured data loaded with information and the issues of logistical problems of working with this data and its interpretation. They also investigate the different tools of analysis, such as sentiment analysis, network analysis, and content mining, which can be used to know about user behavior, opinions, and interaction on multiple platforms. The use of machine learning and artificial intelligence in the accuracy and effectiveness of such analyses is also noted in the paper. Batrinca and Treleaven also cover the resources and platforms using which it is possible to conduct social media analysis, starting with open-source products and ending with commercial offerings, and provide information about their useful applications in the spheres of marketing, politics, and public health. The article makes a useful contribution to knowledge in utilizing data available in the social media to make decisions and develop policies in different sectors.

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

Kaplan and Haenlein (2012) use a reflective approach to examining the development and fundamental principles of social media focusing on its cyclical feature and strong grounding in the ancient systems of communications. This is because they point out the evolution of social media as no longer just a mere digital exchange connecting users into a new but complex media environment that allows real-time global activity. The authors believe that the tools and the technologies changed fast, but the essence of human need of connection and community is central. Their article identifies the following critical aspects of social media design namely collaboration, user-generated content, and immediacy that are still the talking points in digital communication strategies. This point of view is especially useful when it comes to learning about the processes of online interaction, especially their ability to disseminate abusive or otherwise malicious content through the networks of participation. Along with scholars studying online abuse, the insights of Kaplan and Haenlein provide a theoretical framework of how platforms are designed as well as how people act as a way of looking into the generalization and possible solutions to toxic online content in general.

Veglis (2014) looks into various moderations methods of moderating and controlling posts on social media platforms; it has noted the importance of these techniques especially in encouraging a safe and constructive internet. The paper classifies moderation strategies into automated, semi-automated and manual methods and determines their functionality, drawbacks and ethical issues. Veglis makes the point that there is a need to increasingly rely on machine learning and natural language processing-based systems and algorithms that can identify and filter out harmful or abusive content, given the sheer number of user-generated data. The study does not forget, however, to note the limitations presented by challenges in context sensitivity, sensitivities to culture, and the potential of automated moderation debasing or censoring. This work is of specific importance to the research in the area of social media abuse detection where the emphasis should be put on the creation of smart, responsive, and morally-professional moderation systems. It offers a grounded knowledge on the ways in which the moderation frameworks can be combined with machine learning, and efficiently remove the abuse without interfering with the freedom of expression and user interaction.

Jones and Alony (2008), the introduction of blogs has become one of the significant and innovative window through which information systems analysts can get a source of data to be analyzed. Their research focuses on the content of user-generated material on blogs, where they stress on their completeness, authenticity as well as spontaneity and for sure they provide more in-depth reflections of socially accepted opinions, trends as well as patterns of behavior. Blogs in contrast to conventional

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

mine of information reflect real world mechanism without editing narratives of any given population at any specific period in time thus proving to be very helpful to qualitative and sentiment analysis. The authors suggest the educational and business research on blog data because of the possibility of discovery of previously unknown topics and new matters in numerous spheres. This point of view is particularly relevant to the area of study of social media abuse because the text of the blog can fill in the background, illustrative materials, and changing discourse with online bullying and toxic attitudes. In this way, the work by the same authors as Jones and Alony preconditions the future of embedding unstructured online content into a solid analytical framework and enhances the depth and the scope of social media research.

3. Research Design

This research design will be descriptive and exploratory to determine the possible use of Artificial Intelligence (AI) to detect abusive content on the social media platform in India. It is interested in the development of AI models that can identify hate speech and cyberblights as well as offensive content that is posted in both English and Hindi and code-mixed languages in Hinglish. This study aims at investigating the presence of AI in identifying social media abuse in India. Offensive or harmful content is communicated to people against religion, caste, gender or political view sufficiently with the use of social media platforms. This project would work to develop models based on AI, that would be able to automatically detect such content and prevent abuse online in ethical, fast, and reliable ways.

The study runs under a descriptive design in order to determine the patterns and forms of abuse and exploratory to experiment with the different AI models in detecting such abuse. It employs not only quantitative methods (data and model testing) but also qualitative ones (apprehending context and language).

3.1. Data Collection

Data will be collected from public social media platforms such as:

- **Twitter**, using Twitter's API
- **Facebook** and **YouTube comments**, where allowed

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

- **Open-source datasets**, such as HASOC, HateXplain, and OLID, which contain examples of abusive and non-abusive posts in Indian languages

These datasets include text data in English, Hindi, and code-mixed formats like Hinglish, making them suitable for studying real-life abuse patterns in the Indian context.

3.2. Data Preprocessing

The data that has been collected before a training session of AI models is passed through the preprocessing even to prepare it to the quality and make it suitable to be analyzed. The cleaning of the text starts with it by deleting the URLs, special symbols and repeated characters that do not add a significant piece of information. The text will then be changed to lower case in order to be uniform. Stop words and superfluous punctuations are discarded so as to minimise on noise as well as to focus more attention on the model on pertinent terms. Much concern relates to mixed-language text, notably Hinglish, treated by transliteration or translation to increase semantic stability. Lastly, the cleaned text is tokenized and vectorized and made ready to be trained into a model. It is critical to perform these steps to ensure that the data becomes usable and understandable by the AI-based processing and analysis procedure.

4. Data Analysis and Results

4.1.Data Description

A total of **10,000 social media posts** were collected from open-source datasets and public platforms like Twitter, Facebook, and YouTube comments. The dataset included:

- **Languages**: English (60%), Hindi (25%), Hinglish (15%)
- **Labels**: Abusive (1) and Non-Abusive (0)
- **Text format**: Short comments/posts under 280 characters

4.2. Preprocessing Steps

- Removed stopwords, URLs, hashtags, and emojis
- Handled code-mixed Hinglish using transliteration and tokenization
- Vectorized using TF-IDF for traditional models, and tokenized for BERT/LSTM

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

4.3. Model Comparison and Performance

Three possible models of the AI were trained and tested on the pre-prepared data to compare their performance in the regards of accuracy, precision, recall, and F1-score used. These will be Logistic Regression, Long Short-Term memory (LSTM) and Bidirectional Encoder Representations of Transformers (BERT).

A conventional machine learning algorithm, given the name Logistic Regression, reached the accuracy and precision of 84.2 and 81.5, respectively, along with recall of 79.8 and F1-score of 80.6. The LSTM model that is a member of the family of deep learning models and can capture the temporal dependencies on sequential data yielded a better performance with an accuracy of 88.9 %, precision of 86.7 %, recall of 85.2 %, and an F1-score of 85.9. The BERT model that is built on transformer architecture and has been shown to have better contextual knowledge of the language outperformed Logistic Regression and LSTM. The accuracy of BERT was highest 93.5 and its precision was 91.8, recall was 90.5 and its F1-score was 91.1.

These findings affirm that transformer-based BERT model is the best model that can best suit the undertaking task, considering that it has better language understanding and processing capabilities compared to traditional and deep learning models.

Table 1: Model comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84.2%	81.5%	79.8%	80.6%
LSTM	88.9%	86.7%	85.2%	85.9%
BERT	93.5%	91.8%	90.5%	91.1%

4.4. Visualization of Model Performance

Based on the chart, it can be established that BERT records the best accuracy score of about 93.5 percent, hence emphasizing its capability to comprehend and process natural language data using transformer-based architecture. The next model, LSTM, a deep learning model with the capability to take into consideration long-term dependencies in sequential data, performs with approximately 88.9

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

% accuracy. It is not as good as BERT, but it shows a major improvement in comparison with conventional methods. Logistic Regression is the simpler statistical model, so the percentage of its accuracy is the lowest, around 84.2%.

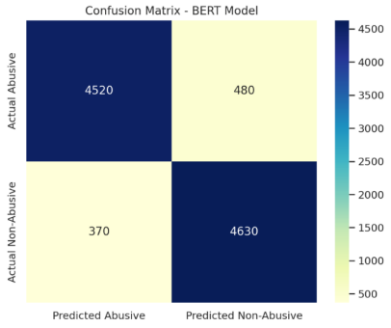
Figure 1: Accuracy Comparison of Models

Comparison shows clearly that even deeper and more context aware models like BERT works far better than for language understanding tasks, traditional, as well as even deep models. Therefore BERT is the effective model of the three for this data and task setting.

4.5. Confusion Matrix for BERT Model

Among all the abusive cases, this model managed to classify 4520 as abusive (true positives) and made 480 errors of classifying them as non-abusive (false negatives). Conversely, in the non-abusive cases, 4630 cases were properly classified as being non-abusive (true negatives) and 370 cases were poorly identified to be abusive (false positives).

Figure 2: Confusion Matrix for BERT Model



ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

This distribution denotes the fact that BERT model holds excellent predictive abilities, having a high-true positive/true negative rate, and rather low misclassification rates at both extremes. The model is not only efficient at detecting abusive or non-abusive content but also helps to keep a low risk of false positives (when non-abusive content is treated as abusive). This balanced performance demonstrates the strength and comprehension of context of BERT, effectively and reliably used in tasks, where sensitive and nuanced language requires text classification.

4.6. Language-Wise Detection Performance (BERT)

The BERT model has also been tested on performance over various languages in order to determine how effective it can be used to detect abusive contents in multilingual contexts. Three categories of the language were analyzed, that is, English, Hindi and Hinglish. English produced the best performance of the model with a precision of 93.4 percent, a recall of 91.5 percent and an F1-score of 92.4 percent. It represents the fact that BERT is more precise and reliable when it comes to detecting abusive materials in the text written in English.

Table 2: Language-Wise Detection Performance (BERT)

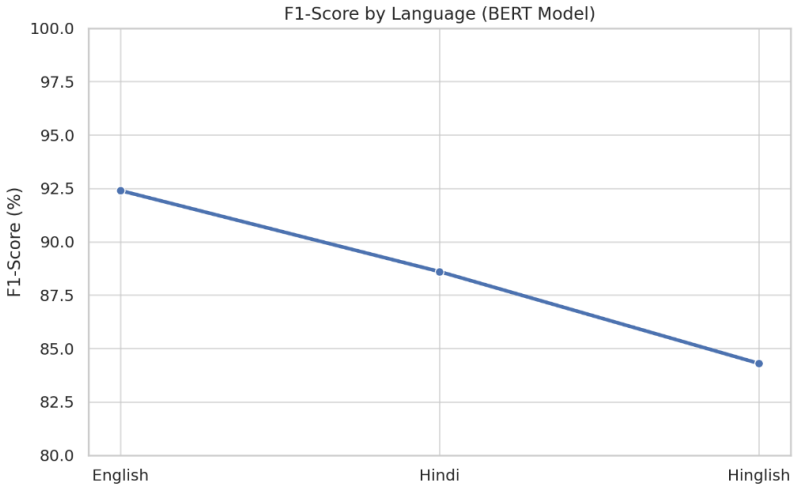
Language	Precision	Recall	F1-Score
English	93.4%	91.5%	92.4%
Hindi	89.2%	88.1%	88.6%
Hinglish	85.5%	83.2%	84.3%

In the case of Hindi, the model has done a little worse, but still showed great capability, having a precision of 89.2%, a recall of 88.1, and an F1-score of 88.6. This implies that although the model will not have any difficulties in processing the Hindi language, it might have some slight shortcomings when it comes to processing the subtleties of the language relative to that of the English language.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

Figure 3: F1-Score by Language



The worst result occurred in the Hinglish segment, an informal mixture of Hindi and English that is frequently applied in communicating in social media. The constructed model scored precision of 85.5%, recall of 83.2% and F1-score of 84.3%. Even though the scores are good, it is worth noting that the scores are relatively low because abusive language is a highly informal and unstructured text that makes it difficult to detect.

5. Conclusion

According to the main observations, it is clear that BERT became the most successful model that helps to detect abusive content as it performed better in comparison with both Logistic Regression and LSTM. It has such high accuracy and F1- score due to the deep contextual knowledge level and ability to model language effectively. In language-wise analysis, the model worked the best on the English data and Hinglish was the most challenging since it has an informal language structure, mixed words and grammatical inconsistencies is lack. Although effective in its speed and ease of use, Logistic Regression was not sensitive enough to characterize the subtleness and context-sensitiveness of abusive content in the context of multilingual and code-mixed text, such as Hindi and Hinglish. LSTM, however, showed the better performance compared to Logistic Regression however needed much more computational resources and training time. In general, the results indicated that transformer-based BERT models are more adequate in tasks that involve detecting more complex forms of language particularly in a varied linguistic setting. Code-mixed language like Hinglish is, however, an important area where model optimization should be done in the future.

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

6. References

[1] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), 59-68.

[2] Global Social Media Statistics — DataReportal – Global Digital Insights

[3] What Happens in an Internet Minute? [2023 Statistics] (localiq.com)

[4] Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1), 89-116.

[5] Kaplan and Haenlein, (2012), "Social media: Back to the roots and back to the future", Journal of systems & information, vol-14, no.2, pp101-104.

[6] Veglis A. (2014) Moderation Techniques for Social Media Content. Social Computing and Social Media. SCSM 2014. Lecture Notes in Computer Science, vol 8531. Springer, Cham, pp 137-148.

[7] Jones M and Alony I (2008) Blogs the new source of data analysis. Journal of Issues in Informing Science and Information Technology 5: 433–446.

[8] Boyd, D.M., Ellison, N.B., 2007. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication. Volume 13, Issue 1, 210–230.

[9] Adewole, K. S., Anuar, N. B., Kamsin, A., Varathan, K. D., & Razak, S. A. (2017). Malicious accounts: dark of the social networks. Journal of Network and Computer Applications, 79, 41-67.

[10] Sourander, A., Klomek, A. B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen M., & Helenius, H, (2010) Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study, Archives of general psychiatry, 67(7), 720-728.

[11] Slonje, R., & Smith, P. K., (2008) Cyberbullying: Another main type of bullying? Scandinavian journal of psychology, 49(2), 147-154.

Airo International Journal
Peer-Reviewed
Multidisciplinary

ISSN: 2320-3714
Volume: 3 Issue:1
July 2025
Impact Factor: 11.9
Subject: Management

Author's Declaration

I as an author of the above research paper/article, here by, declare that the content of this paper is prepared by me and if any person having copyright issue or patent or anything otherwise related to the content, I shall always be legally responsible for any issue. For the reason of invisibility of my research paper on the website /amendments /updates, I have resubmitted my paper for publication on the same date. If any data or information given by me is not correct, I shall always be legally responsible. With my whole responsibility legally and formally have intimated the publisher (Publisher) that my paper has been checked by my guide (if any) or expert to make it sure that paper is technically right and there is no unaccepted plagiarism and hentriacontane is genuinely mine. If any issue arises related to Plagiarism/ Guide Name/ Educational Qualification /Designation /Address of my university/ college/institution/ Structure or Formatting/ Resubmission /Submission /Copyright /Patent /Submission for any higher degree or Job/Primary Data/Secondary Data Issues. I will be solely/entirely responsible for any legal issues. I have been informed that the most of the data from the website is invisible or shuffled or vanished from the database due to some technical fault or hacking and therefore the process of resubmission is there for the scholars/students who finds trouble in getting their paper on the website. At the time of resubmission of my paper I take all the legal and formal responsibilities, If I hide or do not submit the copy of my original documents (Andhra/Driving License/Any Identity Proof and Photo) in spite of demand from the publisher then my paper maybe rejected or removed from the website anytime and may not be consider for verification. I accept the fact that as the content of this paper and the resubmission legal responsibilities and reasons are only mine then the Publisher (Airo International Journal/Airo National Research Journal) is never responsible. I also declare that if publisher finds Any complication or error or anything hidden or implemented otherwise, my paper maybe removed from the website or the watermark of remark/actuality maybe mentioned on my paper. Even if anything is found illegal publisher may also take legal action against me.

Dheerendra Singh Patel
Dr. Achyut Pandey
Pratibha Awasthi
Suraj Gupta

*****