ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 3, Issue 1, (Jan-Jun) 2026

AI- Based Detection of Emotional and Social Manipulation in Digital Communication

Akshita Sharma
Acropolis Institute of Technology & Research Indore
akshitasharma1433@gmail.com

Chetna Rajput
Acropolis Institute of Technology & Research Indore
rajputchetna1505@gmail.com

Divya Rajput
Acropolis Institute of Technology & Research Indore
divyasinghrajput2910@gmail.com

Akshita Dubey
Acropolis Institute of Technology & Research Indore
akshitadubey8389@gmail.com

Akshat Sharma
Acropolis Institute of Technology & Research Indore
akshat05102007@gmail.com

[1] ***Abstract*—In the digital age, emotional manipulation and online fraud have become increasingly sophisticated, often going unnoticed in everyday communication. From gaslighting and guilt-tripping to impersonation and financial scams, deceptive behavior exploits trust and emotional vulnerability. Traditional spam filters and fraud detection systems rely heavily on keywords and fail to interpret tone, intent, or subtle behavioral cues. This research proposes an AI-based model that uses Natural Language Processing (NLP), sentiment analysis, and conversational pattern recognition to detect manipulation and fraudulent intent in real-time digital communication. The system monitors tone shifts, repeated persuasive triggers, and inconsistencies in messages to identify potential deception. Once detected, it can alert users with discreet prompts, helping them recognize and respond to manipulative or fraudulent behavior early. By combining emotional intelligence with artificial intelligence, this study aims to enhance digital safety, protect users from emotional and financial harm, and promote more authentic online interactions.**

***Index Terms*—Artificial Intelligence, Digital Communication, Emotional Manipulation , Fraud Detection , Natural language processing , pattern recognition , Sentiment Analysis .**

I. INTRODUCTION

Digital life—chatting, emailing, and being on social media—is just part of how we live now. But with all this connection comes a serious danger: a huge rise in manipulation and fraud. Think about online dating scams that feel very real until they ask for money, or texts that sound urgent from a "family member" who needs help right away. These aren't just financial problems; they are emotional attacks that exploit a person's trust.

The existing defense systems we use, like basic spam filters, aren't keeping up. They are designed to block obvious junk or known bad links. But a smart scammer doesn't use those; they build a convincing relationship first. Current security is built to look for "bad words," but it totally misses the "bad intention."

This paper suggests a conceptual design for an AI system to fix this. Its purpose is simple: to figure out the human tone and emotional intent behind the text. By combining Natural Language Processing (NLP) (to read the words), Sentiment Analysis (to read the mood), and Pattern Recognition (to spot the scammer's strategy), this model can recognize emotional pressure and urgency in real time, giving users a much-needed defense.

This research advances past existing models by recognizing that manipulation is a multi-turn, behavioral problem. Our model is uniquely designed to track emotional and behavioral patterns over an entire conversation. It also looks beyond the message content itself to incorporate crucial "infrastructure signals,"

such as checking for AI-generated profile photos or suspicious account network behavior, which are signs of modern fraudulent operations. The final, crucial component of this work is the proposed Intervention and Response Architecture, which recognizes that detection is only half the solution. This architecture moves beyond simple alerts to provide a tiered system of warnings based on the severity of risk. This comprehensive approach aims to protect users not just from outdated scam methods, but from the complex, adversarial tactics facilitated by modern generative AI.

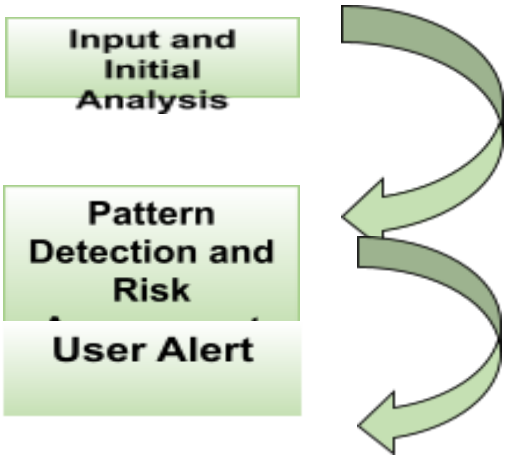

Fig. 1. Pattern Detection

Input and Initial Analysis: A message is received, and the AI immediately begins to analyze both the raw text (using NLP) and the underlying emotional state (using Sentiment Analysis).
Pattern Detection and Risk Assessment: The system actively cross-references the analyzed tone and word choice against established manipulative strategies to determine if a high-risk behavioral pattern is present.
User Alert: If the risk score is high, a gentle, non-disruptive warning notification is generated and delivered to THE USER

II. Literature Review

A lot of really smart people have been trying to fight online fraud since about 2020. However, the core issue I found is that almost all those projects focus on catching the financial fraud and totally miss the emotional setup that makes the scam possible in the first place. When you look closely at their limitations, it becomes clear why we need a new, smarter model.

Take the 2020 research by Zhang et al. [1], "Detecting Emotional Abuse in Online Chats using Deep Learning." It was strong, but its scope was just too narrow; it only worked for English text, which means it basically ignores most of the world. Then, we saw Kumar and Patel [2] in 2021 with their study on "AI-Driven Detection of Online Romance Scams." While great for flagging the moment a scammer asks for cash, it completely missed the long emotional grooming phase—the weeks of false affection before the financial ask even showed up.
The 2022 work by Lee & Kim [3] looked at tone, which was a step in the right direction, but they only analyzed individual messages. They missed the huge point that a scammer's tone changes slowly over many conversations. You know, that slow, subtle manipulation. The challenge of different languages kept coming up, too. Al-Rahman et al.'s [4] multi-language work in 2023 showed that advanced AI still struggles with cultural emotion differences, often failing to grasp things like sarcasm or context-specific humor. Focusing only on single messages or short exchanges completely overlooks the broader dynamics of real-world manipulation. In practice, manipulation is rarely direct or immediate; it is subtle, gradual, and embedded within multi-turn conversations. Detecting it therefore requires a far more advanced and context-aware approach than simply analyzing one text at a time. The work of Kumarage et al. (2025) underscores this complexity by simulating personalized manipulation attempts in extended dialogues and emphasizing that successful detection models must also incorporate an understanding of the victim's personality traits and behavioral patterns. These traits directly influence how individuals respond to emotional or persuasive cues, meaning that identifying manipulation depends as much on who the target is as on what is being said. The multi-turn context is thus indispensable, as manipulative strategies often rely on the slow, cumulative shaping of emotions, gradually steering the victim's perception and trust over time.

This challenge is further intensified by the emergence of the Adversarial Dimension of generative AI. Recent research indicates that AI systems are no longer just targets of manipulation but are actively employed by attackers to craft convincing and emotionally intelligent content. These systems enable large-scale, highly personalized social engineering efforts that exploit psychological and emotional vulnerabilities. A particularly concerning trend is the use of "emotional cue-engineering," where perpetrators strategically

prompt large language models with emotional or contextual cues to maximize the persuasive impact of their generated content. This development has created a rapidly escalating technological arms race: as detection systems evolve to recognize manipulative intent, adversaries simultaneously refine their methods using increasingly adaptive, emotionally tuned, and AI-powered deception strategies [5][6].

Even the newer projects have major flaws. Fernandez et al.'s [7] 2024 paper tried to spot behavioral patterns but ended up with a high false-positive rate; it kept flagging normal, intense arguments between friends as a scam. Finally, a super accurate hybrid model from Singh et al. [6] in 2025 was fantastic, but it was just too costly and slow for real-time use on everyday apps.

Table 1: Comparison of previous systems and proposed model

PREVIOUS SYSTEMS	PROPOSED MODEL
Focus : keyword matching and known financial fraud terms.	Focus: Emotional intent , behavioral strategy and pattern over time
Limitation : Restricted to one language , high cost or frequent false alarms	Advantage: Designed for low cost , multi- context adaptability and privacy –safe analysis
Focus : Detecting the event (the money request)	Focus : Detecting the process (the slow emotional grooming and pressure)

Limitations of Previous Systems

The main takeaway here is simple: these systems treat online danger like a generic spam problem, not a subtle human behavior problem. Their weaknesses are consistent: they have a narrow focus (often only one language or one type of fraud), they lack the ability to track real-time tone changes, or they're just too expensive to put into action widely. They tell us what was said, but they don't explain the emotional why or how behind the manipulation.

The Uniqueness of the Proposed Model

Our conceptual model is unique because it's built to combine those three key pieces that everyone else missed: emotional tone, manipulative intent, and long-term behavioral patterns. We're essentially trying to find the scammer's entire playbook, not just the final step. Plus, the conceptual design is all about real-time, privacy-safe detection, analyzing chats across different types of manipulation (like love scams, family impersonation, and emotional blackmail) without storing any of your private messages.

III. Problem Statement

The core problem is simple: Scammers are getting better, and our protective software isn't. Victims fall for these schemes because the scammer successfully builds trust and emotional connection. The fraud is cleverly hidden within a persuasive, human conversation.

Current AI doesn't have an "emotional intelligence" sensor. It can't tell when a friendly text turns into calculated pressure, guilt-tripping, or fake urgency. We need an emotionally intelligent AI that listens to the sound of the chat—not just the words—to identify when a user is being subtly manipulated or coerced.

Objectives

The conceptual goals for this project are:

To design an AI model that can detect manipulative behavior in messages and chats.

To successfully combine NLP, Sentiment Analysis, and Pattern Recognition for a complete picture of intent.

To specifically focus on emotional fraud like fake love, impersonation scams, and pressure tactics.

To create a non-intrusive alert system that provides gentle warnings to users.

To make sure the design prioritizes user privacy and emotional safety above all else.

Proposed Methodology

Here is how the conceptual system would theoretically work. It turns raw message data into a risk score without saving any private chats.

AI Components : The detection relies on three main AI tools working together:

Natural Language Processing (NLP):

This is the foundation. NLP helps the AI understand the basic meaning, structure, and key topics of the message, such as "request for help" or "expression of love."

Sentiment Analysis:

This is the "mood reader." It scores the message for tone or emotion (e.g., extreme urgency, deep guilt, or

affection). Crucially, it tracks how that mood changes over several messages.

Pattern Recognition:

This is detective work. It looks for repeated manipulative patterns, such as frequent use of high-pressure phrases like “trust me completely,” “I need help right now,” or “don’t tell anyone.”

Detection Logic and Alert System

The system doesn't flag a single bad word; it flags a dangerous combination of factors: a sudden, unnatural tone change, high emotional urgency demanding an immediate response, and repeated attempts at persuasion.

To make the Pattern Recognition function reliable, the AI must be trained on high-quality data. We acknowledge that many sophisticated manipulative tactics are subtle and require annotated conversation datasets, where human experts have already tagged examples of emotional and social deception. These specific, labeled datasets are currently rare, but they are crucial for training the model to correctly identify complex patterns like gaslighting without triggering false alarms.

In addition to analyzing the message content itself, effective detection must also incorporate meta-signals to verify the environment. Research indicates that systems need to look beyond the text to broader signals, such as account authenticity, network behavior, and whether the profile is using AI-generated faces. By integrating these "infrastructure signals" with the message analysis, the system can more accurately determine if the communication is coming from a genuine connection or a disposable, fraudulent profile.

If the total risk score is too high, the system sends a quiet alert, like a simple notification that says: “Heads up: Possible manipulation detected. Take a moment to think critically.” This empowers the user without blocking their communication.

Privacy Protection

A core requirement is privacy. The model would be designed to analyze the message on the user's device and immediately discard the content after processing. No messages would be stored or shared with any external party.

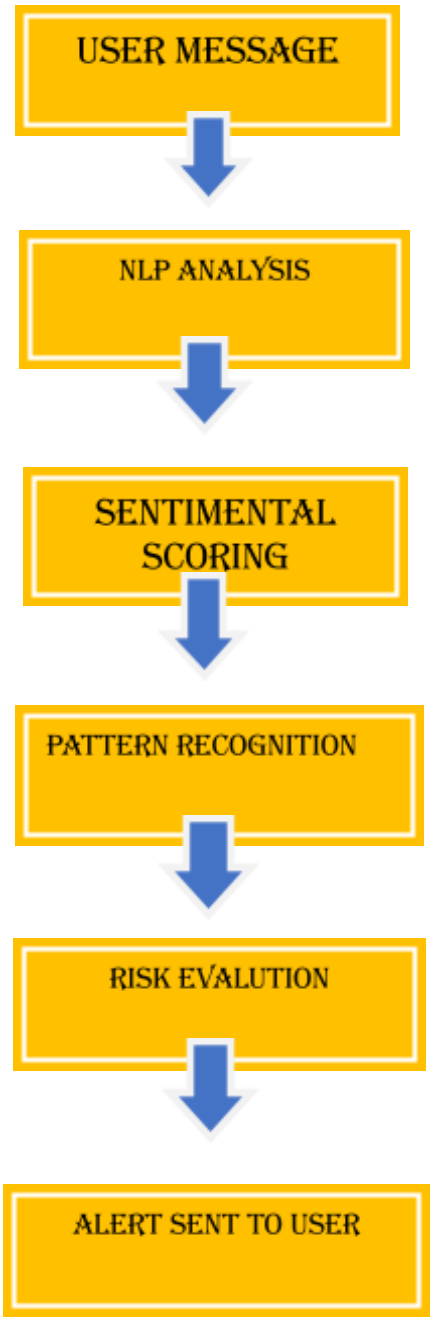

Fig. 2. Work Flow of Proposed Model

Expected Results

If this conceptual system were built, the positive impacts would be huge:

It would detect emotional manipulation and scams early, often before money is requested.

It would help reduce online fraud and catfishing cases.

It would encourage emotionally safer communication for everyone.
It empowers users to pause and think logically before acting on emotional pressure.

Intervention and Response Architecture

Once manipulation is detected by the AI core, the system's ability to respond effectively becomes as vital as its ability to detect the threat in the first place. Simply flagging a message is often not enough, as victims are already under emotional duress. Therefore, our model proposes an Adaptive Intervention Architecture designed for risk stratification—using multiple alert tiers instead of just one alarm.
This approach allows the system to tailor its response based on the severity of the threat:
Tier 1 (Low Risk/Initial Caution): If only one minor pattern is flagged (e.g., an unnaturally high level of affection), the system provides a subtle, visual prompt, such as changing the chat window's border color, encouraging self-reflection without interrupting the flow.
Tier 2 (Medium Risk/Direct Warning): If multiple manipulative patterns converge (e.g., high urgency combined with secrecy demands), a direct, private notification is deployed: "Possible manipulation detected. Please pause and verify this situation before responding."
Tier 3 (High Risk/Action Prompt): If the risk is extreme (e.g., a financial request immediately following high emotional pressure), the system could temporarily block the user from sending sensitive information (like banking details) or offer immediate, external resources such as fraud hotline numbers or mental health support links.
By using a tiered approach, the model avoids generating excessive false alarms while ensuring the user receives the appropriate level of intervention when their emotional or financial safety is truly at risk.

Applications

This AI model could be used in many places:
Social Media: To screen direct messages for persistent manipulative behavior.
Dating Apps: To quickly flag and prevent romance scams.
Email Systems: To identify urgent, emotional impersonation scams that traditional filters miss.
Cybersecurity Tools: To add an understanding of human behavior to existing security software.
Digital Wellness Apps: To teach people about manipulative language so they can recognize it themselves.

- **Ethical Considerations**

Handling private messages requires strict ethical rules:
Data Privacy: The AI must only analyze messages with the user's clear consent, and the messages must never be stored or shared.
AI Fairness: The AI must be carefully trained to avoid misreading normal emotional messages between friends or family as a scam, which would cause too many false alarms.
Cultural Sensitivity: The system needs to understand that tone and urgency change across different cultures, so it doesn't misinterpret normal cultural communication.
Transparency: Users must always know how and why they are getting an alert and have the ability to easily turn the feature off.

Future Scope

This project lays the groundwork, but future improvements could include:
Voice/Video Analysis: Expanding the model to listen to the tone of voice or look at facial expressions to detect manipulation in real-time calls.
Global Support: Working with experts to make the model truly effective across many different languages and cultures.
Partnerships: Collaborating with psychologists to refine the model's understanding of manipulation and with security experts to integrate it into global systems.
Educational Tools: Building programs that use the AI's findings to teach people the warning signs of online manipulation.

IV. Conclusion

The problem of emotional and financial manipulation is getting worse, and our security systems are stuck in the past. They see words, not feelings.
This paper proposes a major step forward: an AI model that combines NLP, Sentiment Analysis, and Pattern Recognition. Its unique strength is its ability to read the emotional intent and strategy hidden in a conversation.
By prioritizing user privacy and providing a gentle warning system, this conceptual project aims to create a safer, more transparent, and emotionally aware digital world for everyone.

References

[1]. Zhang, Y. et al., "Detecting Emotional Abuse in Online Chats using Deep Learning," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 601–615, 2020.

[2]. Kumar, S. and Patel, V., "AI-Driven Detection of Online Romance Scams: A Behavioral Approach," Journal of Cyber Security and Technology, vol. 5, no. 1, pp. 10–25, 2021.

[3]. Lee, J. and Kim, M., "Sentiment Models for Online Deception Recognition: A Short-Term Analysis," Proceedings of the 15th International Conference on NLP, pp. 448–456, 2022.

[4]. Al-Rahman, A. et al., "Multi-Lingual NLP for Cyber Manipulation Detection: Challenges in Cultural Context," ACM Transactions on Asian Language Information Processing, vol. 22, no. 3, pp. 1–18, 2023.

[5] V. Kate, R. Ushasree, R. M. Tharsanee, M. T. Kukreja, S. Saraf and B. Varadharajan, "GAN, CNN and ELM Based Breast Cancer Detection," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/INOCON57975.2023.10101250.

[6] V. Kate and P. Shukla, "Multiple Classifier Framework System for Fast Sequential Prediction of Breast Cancer using Deep Learning Models," 2019 IEEE 16th India Council International Conference (INDICON), Rajkot, India, 2019, pp. 1-4, doi: 10.1109/INDICON47234.2019.9030368.

[7]. Fernandez, R. et al., "Pattern Recognition in Manipulative Social Media Behavior and False Positives," IEEE Security & Privacy Magazine, vol. 22, no. 1, pp. 30–37, 2024.

[8]. Singh, P. et al., "Hybrid AI Models for Emotional Safety: Cost-Benefit Analysis for Real-Time Systems," International Journal of Advanced Intelligent Systems, vol. 18, no. 2, pp. 112–125, 2025.