MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

AI-Based Tension Prediction and Human-Agent Interaction Analysis in U.S. Crisis Response Simulators

Md Toushif Pramanik[1], Joy Chakra Bortty[2], Mahuma Akter[3], Tanjina Tuly[4], Md Alal Udden[5], Farhad Uddin Mahmud[6], Saniah Safat[7] and Mahamuda Akter Shati[8]

[1]Master of Embedded Software Engineering, Gannon University, Erie, USA
[2]Department of Computer Science, Westcliff University, Irvine, CA, United States.
[3]Master of Science in Cybersecurity (MSCS), Washington University of Science and Technology (WUST).
[4]MSc in Business Analytics, Trine University
[5]Engineering Management, Trine University
[6]Master of Business Administration in Management Information Systems, International American University
[7]Department of Information Systems and Operations Management, The University of Texas at Arlington
[8]Master's in Business Analytics, Grand Canyon University
Corresponding Author: **Md Toushif Pramanik, Email:** pramanik001@gannon.edu

Abstract

Predicting when a conversation is about to boil over during a crisis is a tough task for any support system. Emotions tend to spike fast, sometimes in just a few back-and-forth exchanges. Because of that, spotting signs of distress early on is a big deal for anyone trying to help. This research looks at a hybrid machine learning setup designed to see these tension spikes coming. It uses a tool called the Composite Tension Index (CTI) to measure emotional weight by looking at how language, timing, and interaction patterns shift over time. To build this index, the framework pulls in various signals like the mood of the words used, how urgent someone sounds, and how much the emotional tone swings. It also tracks things like how long it takes for someone to reply or who is dominating the talk. After gathering this data, the study tests a few different machine learning models, mostly ensemble methods and regularized regressions, to see if they can guess if the next part of the chat will get more heated. The results show that this system actually picks up on these rising patterns pretty well. It turns out that looking at the conversation as a moving, changing process works way better than just analyzing a single message in a vacuum. The data also suggests that the way people interact matters. When responses are reflective and the dialogue stays balanced, tension usually stays low. On the other hand, long silences or bossy tones make things much more unpredictable. Essentially, these escalations work like a tipping point in any complex system. This framework acts as a digital early warning, catching those rising signals before things get out of hand, which helps human responders know exactly when to step in.

Keywords: Artificial intelligence; Conversational analysis; Crisis communication; Machine learning; Tension escalation prediction.

1. Introduction

1.1 Background: Today's crisis response scenes are messy. They are full of unknowns, information that changes by the minute, and high-stakes conversations between the people calling the shots and those on the ground. When things are this tense, spotting the first signs that a situation is about to boil over is huge. If a team can catch those early warnings, they might actually stop a disaster or at least step in before it is too late. People are looking at artificial intelligence more and more as a way to sift through all that noise. Standard math and old-school analysis often trip up when relationships between variables aren't linear or when the "hints" of a coming problem are buried deep. Machine learning is proving it can find those needles in the haystack. Jakir (2025) points out that AI can actually sniff out quiet signals of a crisis even in messy financial or institutional settings. The idea there is that looking at the signal-to-noise ratio can show systemic stress long before the usual metrics start flashing red [22]. It really points to how much AI could help us see tensions growing in these complicated human-tech systems.Whether a crisis is about politics, the economy, or social issues, it usually involves a lot of different people pushing and pulling against each other. Information moves fast, and small problems can feed back into the system and get way bigger, way faster. It is hard to use a stiff, static model to figure this out because the "escalation" usually happens bit by bit, hidden under a mountain of useless or even wrong data. This is why machine learning early warning systems are getting a lot of attention. They can watch how a system is moving and guess when things might blow up. Rahman (2025) showed that these frameworks can pick up on tiny, micro-level shifts that happen right before a whole economy hits a rough patch [26]. Basically, these AI tools see the patterns that humans or basic spreadsheets miss until the house is already on fire. In the world of government and policy, waiting too long to make a move or getting the facts wrong can be a total nightmare. A bad call during a crisis can mess up the economy or national security for years. Mahlik (2026) makes the case that when we are dealing with big macroeconomic shifts, we need tools that can guess how a system will react to a specific policy, especially when info is scarce, and nobody is sure what's going to happen next [24]. This really highlights why we need better ways to help leaders when things get hairy. Plus, public institutions have their own internal baggage that makes predicting a crisis even tougher. Edelberg, Harris, and Sheiner (2025) mention that things like budget pressure and structural economic risks make it really hard to balance trade-offs, which just makes it even more important to have solid tools for planning and figuring out risks [13].

Modern governance is just plain complicated. A single policy environment is tied to a million things like debt, taxes, global markets, and local laws. Dallas and Scavone (2024) show how public finance is this giant web of debt structures and growth paths where everything affects everything else in ways that aren't always obvious [11]. Traditional forecasting usually flops here because it can't handle how all these risks overlap. Because of that, a lot of researchers are turning to simulations and computer models to play out "what if" scenarios. Places like the American Action Forum (2025) use long-term modeling to see how economic shifts might play out over time, emphasizing that we need tough analytical tools to judge these big policy moves [4]. The Congressional Budget Office (2025) also puts out long-term projections that show just how hard it is to guess where an economy is going ten or twenty years out, especially when weird policy shifts and structural surprises start clashing [10]. If you look at all of this together, it is pretty clear that we need AI-led frameworks that can find those early red flags in messy systems. By mixing machine learning with simulated environments, we can actually see how a crisis grows and how the way people talk and act affects whether a situation calms down or gets worse. It is a promising way to get a better handle on how crises work and to give the people in charge better info so they can make a move before it's too late.

1.2 Problem Statement: Even with all the fancy tech we have now, figuring out when a crisis is going to escalate is still a huge pain. These situations move fast, the information is never complete, and people act in ways that are hard to predict. It makes finding a reliable "early warning" really tough. Most of the time, the tension builds up slowly and gets buried under regular daily chatter or conflicting reports. This is why the old ways of analyzing things, using fixed checklists or simple models, don't really work well. Leaders end up missing the moment when a situation turns critical, which leads to late reactions and choices that aren't great. One of the biggest headaches is just telling the difference between a real warning sign and the normal stress of a crisis. Complex systems spit out a lot of data that looks important but ends up being nothing. Jakir (2025) says that a good prediction system has to be able to separate those weak but real signals from all the background noise [22]. He thinks AI is the way to do it. But actually putting that into practice in a real office or command center is another story. You have to worry about whether the data is even available, if the model is actually reliable, and if a human can even understand why the computer is saying what it's saying.

Another problem is that we don't have a lot of good, open data from actual crises to study. Real-world data is usually locked away because of privacy or national security. Researchers end up using "proxy" data, basically, recordings of high-stress meetings or games that feel like a crisis but aren't the real thing. While these help us study how things escalate, they might miss some of the grit and complexity of a real-life disaster.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

Because of that, making a model that works in one situation and also works in a completely different one is still a major goal that hasn't been hit yet. Then there is the issue of trust. If a machine learning model is a "black box" that nobody understands, a leader probably won't want to bet a country's safety on it. People need to be able to justify their choices. Rahman (2025) points out that these early warning systems have to be both accurate and easy to explain if they're going to be useful in a professional setting [26]. It's a balancing act. We need models that don't just say "danger," but actually explain what is causing that danger.

The sheer complexity of policy systems makes everything harder. Government decisions involve so many moving parts and institutional rules that a single choice can cause a domino effect through the whole society. Mahlik (2026) notes that making policy when you aren't sure of the facts can lead to all sorts of unintended consequences [24]. And as Edelberg et al. (2025) mentioned, when you have structural uncertainties in the government itself, it's even harder to see what's coming, especially when risks are slow-rolling and pile up over the years [13]. Simulations might be the best way around some of this. By building a controlled digital world, we can study how humans and AI agents interact and see what specific behaviors or words lead to things getting out of hand. Looking at work from Dallas and Scavone (2024), it's clear that we have to look at the whole system and how variables interact, not just individual red flags [11]. The work done by the CBO and the American Action Forum also shows how these simulation frameworks help us look at long-term risks when we don't have all the answers [4][10]. So, there is a big gap here. We need a way to find those early hints of tension that are actually understandable and work in the real world. New developments in machine learning and language processing mean we can start looking at how people talk and act in simulated crises. By combining predictive models with an actual look at how humans and agents interact, we can build tools that don't just predict trouble but actually show us why it's happening in the first place.

1.3 Research Objectives: The main goal of this work is to build an AI setup that can figure out when a conversation is about to boil over during a crisis simulation. These kinds of high-stress talks usually go back and forth in a sequence, and the emotional energy tends to shift, sometimes things calm down, but other times they get much worse. Figuring out how these moods change as the clock ticks is a big part of helping response systems spot trouble before it happens. This framework looks at the flow of the conversation to find the specific patterns that suggest tension is climbing. A big piece of this involves making a combined measurement that can show how intense a conversation actually is. Since people don't just express stress through one specific word or tone, the analysis has to pull together a bunch of different clues that show a shift in behavior. This "Composite Tension Index" is meant to bundle up linguistic signals—like how fast someone is talking or the specific way they use words, to see how those feelings build up over time. By looking at these chunks of dialogue, the system tries to guess if the next few minutes are going to get heated.Outside of just building a predictive tool, this study also looks at how different types of computer models handle the data. Because crisis talks happen in a specific order, the timing of what is said matters a lot. The research tests out several ways to model this, comparing older, more basic methods against newer ones that are better at tracking changes over time. The idea here is to see if these more complex models are actually better at catching those early warning signs of a blow-up. Another thing this research dives into is how the way people interact affects where the tension goes. In these simulations, there is usually a person talking to a responding agent, and what that agent says can totally change the vibe of the talk. By looking at how these different responses nudge the tension levels up or down, the study tries to figure out how human-agent communication really works in a pinch. This could eventually help make better strategies for how these agents should talk to people in the real world.Lastly, the work checks how easy it is to actually understand why the AI is making certain guesses. In a serious situation, a person making decisions needs to know why a computer thinks things are going south. This study uses specific techniques to peek inside the "black box" of the AI to see which words or behaviors are triggering the alerts. This isn't just about making a model that works well; it is about making sure the people using it can trust and understand the logic behind the results.

1.4 Research Questions: The questions driving this study center on whether AI can actually keep up with the way tension rises in a crisis. These talks are full of tiny hints that things are getting worse, but those hints are usually scattered across a whole conversation rather than being obvious in one single sentence. The first question asks if a combined index of these language clues can really map out the way a crisis escalates. By watching how this index moves during a talk, the study looks for proof that these gathered signals are a solid way to measure stress. A second question is whether models that pay attention to the order of events do a better job than standard machine learning tools. Since a crisis unfolds step by step, a model that remembers what happened two minutes ago might have a better shot at predicting what happens next. This research puts a few different types of models side-by-side to see if these time-sensitive versions are actually more accurate or more reliable when the pressure is on.Then there is the question of whether mixing old-school rules with modern machine learning makes for a better system. In a field where safety is everything, a model that only learns from data might do something weird if it sees a situation it wasn't trained for. By adding some human-defined rules into the mix, the system might stay more grounded. This study looks at whether these hybrid setups are actually ready for the real world. The research also explores which specific factors matter the most. Tension might come from the words someone chooses, or it might just be the speed of the conversation, or the way the participants are acting. Pinpointing which of these things actually drives the risk of a blow-up helps clarify what is really happening during a crisis and makes future tools even sharper. Finally, the study asks how different ways of talking to people change the path a conversation takes. The style an agent uses when responding to someone in a bad spot can either settle things down or accidentally make them worse. By looking at these interactions in a controlled way, the research aims to get a better handle on how communication styles shape the whole experience and how AI systems can be built to keep things as calm as possible.

2. Literature review

2.1 Affective Computing in Crisis Communication: Affective computing has turned into a pretty vital area for anyone trying to figure out how people show and pick up on emotional signals when they talk. When things hit a breaking point in a crisis, people usually release their stress through the way they use words. Because of that, looking at the tone of a chat and how intense those feelings are has become a big part of building systems that can spot and handle emergencies. Back in the day, most work in this field just stuck to sentiment analysis. It was basic stuff, mostly trying to check if a text was just "good" or "bad" in terms of mood. But lately, things have moved way past that. New research tries to map out a whole range of complicated mental states like being scared, anxious, or just plain fed up. This shift matters a lot for crisis work because emotions aren't static; they change fast and can show when a situation is about to boil over or when someone's mental health is sliding.

Machine learning has been the heavy lifter in making it possible to read these emotional cues in messy conversations. Even in fields like government policy forecasting, people are seeing that these models are great at pulling real meaning out of huge piles of confusing data. For example, Bova and the team in 2023 looked at using machine learning to predict where public debt was heading in tricky economic systems. They found that these data models could see weird, non-linear patterns that old-school math usually misses [9]. Even though they were looking at money and not how people talk when they are upset, it shows that machine learning is really good at finding a signal hidden in a lot of noise. That is exactly what you need for crisis communication, where the signs that someone is spiraling might be buried in a long, rambling conversation. Other researchers are looking at how to plug these machine learning tools into policy frameworks to help people make better choices when things get complicated. Escolano and Shabunina in 2024 showed that these methods can be used in economic modeling to see how tax or spending decisions actually affect people's well-being when the future is uncertain [14]. Their work proves that AI systems can help people make decisions by spotting trends that aren't obvious if you just use traditional methods. This logic applies to crisis chats, too. People making

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

the calls need to understand those emotionally loaded messages so they can step in before things get worse.Beyond just predicting policy, these techniques help find hidden connections in complex systems. Dola and others in 2024 came up with a way to find secret networks of collusion in the corporate world, proving that models can find relationships that stay hidden during a normal audit [12]. Being able to find these "hidden" layers in data is huge for understanding how a conversation flows during a crisis. These dialogues are full of hints, indirect cries for help, and weird back-and-forth patterns between the people talking. By using machine learning that can see these underlying structures, we can build tools that actually track the emotional path of a crisis. All of these points point to the fact that affective computing and machine learning can give us a much deeper look into the raw emotions that define these high-stakes moments.

2.2 NLP for Mental Health and Crisis Detection: Natural language processing, or NLP, is now a go-to tool for digging into text about mental health and personal struggles. Places like online forums, social media, and even transcripts from crisis hotlines are full of people talking about their pain. This makes them a goldmine for building systems that can automatically spot when someone is in trouble. More and more, researchers are using machine learning to find specific word patterns that scream "crisis," like when someone is thinking about suicide or dealing with massive anxiety. These systems usually look at things like "mood scores," word lists related to feelings, and the way words are used in context to figure out how much danger a person is in. One of the biggest headaches in spotting a crisis is that human language is always changing. Someone who is really hurting might not say it straight out. They might be vague, and the signs that they are getting worse can show up slowly over a long time. Because of this, some people have suggested using "adaptive" models that can learn as they go. Bhowmik and a few others talked about these self-adapting frameworks for dealing with financial data that doesn't stay still, showing how algorithms can shift when the environment does [8]. Even though their focus was money, the lesson fits crisis talk perfectly. Feelings in a conversation don't stay in one place; the way someone uses language might change as they move through different stages of a breakdown.Newer "deep learning" setups have made these NLP systems even better in real-world scenarios. Alam and a team in 2026 suggested using hybrid models that mix different types of neural networks to predict when industrial machines might break down [2]. That work shows how mixing different models can help catch both tiny details and big-picture trends in a dataset. You can use that same strategy for crisis communication. A conversation has both quick emotional flares and longer patterns of interaction that tell you if things are escalating. Using AI for big, messy systems has also taught us how to use these models in actual operations. Al Montaser and Bhuiyan in 2025 showed that machine learning can help manage energy in smart cities, where you need to keep a constant eye on things and change your forecast on the fly [3]. Their work is about power grids, sure, but the idea is the same for crisis detection. You need a system that watches conversations constantly so it can flag when a situation starts to go south. Even with all these cool updates, a lot of the crisis tools we have now are still a bit too stiff. They often look at one message at a time and forget to look at the whole conversation. Doing it that way means you miss the "timing" of how feelings grow or fade over a long talk. Because of that, researchers are starting to realize they need to use techniques that respect the order of the chat and how one message leads to the next.

2.3 Temporal Modeling in Dialogue Analysis: Looking at dialogue data means finding a way to track how a conversation actually moves. It isn't like classifying a single piece of text where everything stays still. Crisis communication is a moving target. Every message depends on what came before it, and every new word has the potential to push the whole interaction in a different direction. You need sequential modeling to really see how tension builds up or dies down. If a researcher uses old-school machine learning that treats every sentence like its own separate island, they end up missing the "glue" that connects them. That makes it pretty hard to predict when things are about to boil over. People are starting to see that you get better results when you mix structural models with machine learning. Gauthier and Moyen (2025) did this when they looked at France's public debt. They took these heavy-duty economic models and paired them with forecasting algorithms to handle the messy, nonlinear parts of the data [16]. It showed that you don't have to choose between a solid theoretical framework and a flexible learning model. You can use both to get a clearer picture of a system that is constantly evolving.authier and Moyen (2026) kept pushing this idea, proving that machine learning actually works better for macroeconomics when it's tucked inside a simulation that understands the "rules" of the system [17]. This same logic fits dialogue analysis in a crisis. A conversation is just another dynamic system where emotions, behaviors, and the situation itself all bump into each other. If a model can handle both the rigid structure of a conversation and the way people adapt on the fly, it's going to be way more accurate. There is also some interesting work happening with networks. Islam et al. (n.d.) have been using graph neural networks to look at financial risk [21]. They map out how different banks and markets are tied together, because in a crash, it's those connections that matter most. In a crisis talk, you can think of the people as points on a map and their messages as the roads between them. By modeling it this way, researchers can watch how an emotional spark in one spot travels through the whole network and changes the mood of the entire talk. So, temporal modeling is really the bedrock here. By keeping track of the order of things and how people influence each other, these models let us see how emotional signals pile up. That's how we eventually figure out if a crisis is heading toward a resolution or a total meltdown.

2.4 Human-Agent Interaction in Crisis Response Systems: We are seeing more AI being dropped into places where they have to talk directly to people who are having a really hard time. These crisis systems use chatbots or tools to help humans on the other end of the line. Because the stakes are so high, the way the AI is built matters a lot. A well-timed response from an agent can calm someone down, but a clunky or tone-deaf one might actually make their distress worse. AI is already great at smoothing out big, complicated operations. Islam et al. (2025) looked at how deep learning makes US supply chains run better by spotting patterns in the data and tweaking strategies [20]. Even though shipping packages is different from talking to someone in a crisis, the core idea is the same. You use machine learning to figure out the best move and then adjust. In a crisis, the "optimization" isn't about speed; it's about an AI agent reading the room and picking a response that stabilizes the situation.

But as these systems get more involved in big decisions, we run into the "black box" problem. A lot of these models are so complex that nobody knows why they're saying what they're saying. Bitetto et al. (2025) pointed out that in finance, you need explainable AI because if people don't understand the reasoning, they won't trust the prediction [7]. This is even more vital in a crisis. If a model says a situation is dangerous, the person in charge needs to know why before they take action. Luckily, there are new ways to peek inside these models. Krantz and Richert (2023) came up with something called SHAPoly, which uses Shapley values to explain deep neural networks in economics [23]. It basically highlights which specific pieces of data are doing the heavy lifting for a prediction. If you apply that to a crisis conversation, you can pinpoint exactly which words or behaviors are triggering the "high tension" red flag. At the end of the day, any AI used in a crisis has to be both smart and transparent. It isn't enough to just be right. A system that can spot a problem early and then explain its logic is a huge asset for the people who have to make the tough calls.

2.5 Explainable AI in Safety-Critical Systems: AI is finding its way into more high-stakes spots lately. This creates a big need for models that are accurate. They also need to be easy for a person to understand. Safety-critical areas like medicine, money rules, power grids, or crisis response need systems that show their work. This lets a human operator see the logic behind a computer's guess. When a choice hits a person's health, a bank's stability, or a city's safety, the people in charge have to check the work. They need to trust that the computer is finding real patterns. It cannot be some random fluke. Because of this, explainable AI is a major focus for researchers right now. This is true for any spot where a model talks to a person. Crisis response is a very sensitive area for this technology. It involves deep feelings. It involves danger. Transparency plus accountability are required parts of any tool used there.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

The need for a clear view is obvious when looking at how messy a crisis can be. These spots involve a lot of confusion. Information moves fast. Many different people are involved. AI built for these spots must give answers that a human can verify. Without these clear paths, a leader might be scared to follow a computer's advice. This is common when the model is a deep neural network or something else very thick. Explainable tools help bridge this gap. They let users check for bias. They help find errors in how the model acts. This makes the whole process safer for everyone involved. Research has found several ways to peek inside these black boxes. One common way involves looking at which inputs matter most. Bitetto et al. (2025) proved that this matters in finance. Their work showed that clear models let analysts see which economic signs drive the cost of capital. This builds trust. It shows exactly why a prediction happened. Money is not the same as a crisis. The core logic holds up anyway. In a crisis system, a worker needs to see which words or emotional shifts made the AI flag a rise in tension.

Newer tech has made these explanations even better. Krantz and Richert (2023) built the SHAPoly framework. This tool uses Shapley values to map out complex, non-linear links inside deep neural networks. It shows how specific variables change an outcome even when the math is very messy. This is great for looking at chat data during a crisis. Tiny linguistic parts, context clues, or tone shifts might mix in strange ways to create a blow-up. These attribution tools help researchers see which parts of a talk lead to an escalation. Explainable AI is a key part of making systems people can trust. Adding these features helps ensure tools do more than just flag a risk. They provide a reason. This helps a human understand the situation. It helps them make a better choice. In spots where AI plus humans work together, being able to read the model is the only way to make the partnership work.

2.6 Research Gaps: Even with all the progress in AI or language tech, there are still some massive gaps. We have holes in our knowledge about how tension works in a crisis. A lot of the old work focuses on single messages. It looks for a sign of distress in one isolated spot. This might tell you if someone is sad right now. It does not show how that feeling grew. A crisis talk is a living thing. People react to each other. They change their style. Their feelings move around. Analyzing one text without the rest of the talk gives a broken view of what is happening. Another big limit is the lack of a timeline. Most sentiment tools just slap a label on a message. They rarely map out how the heat in a room rises over time. This stops us from seeing the slow build-up of a problem. In a real crisis, you want to see the small shifts before they turn into a disaster. We do not have many models that look at the path of a conversation. This is a huge missing piece in the field.

Most systems also wait too long. They are reactive. They only see trouble once it is already loud. This helps with spotting a major breakdown. It does nothing to stop one from happening. We need predictive models. We need tools that estimate the chance of a future blow-up. Very few studies have looked at chat patterns to see if things will get unstable later on. We also ignore the bot's role. In many systems, a person talks to an automated tool. The way that tool acts changes the whole vibe. It can make things calm. It can accidentally make things worse. We do not have enough research on how different bot styles change the emotional path of a talk. These gaps show we need a new way of doing things. We need a framework that looks at time, makes guesses about the future, plus watches how the bot plus the human interact. This study aims to fill those holes. It wants to build AI that predicts a crisis, tracks the talk, plus helps people handle the situation better.

3. Methodology

The authors used a mix of old-school rule-based logic and newer machine learning to figure out when a crisis conversation is about to boil over. This hybrid setup looks at text-based chats step-by-step. It lets the system spot early warning signs while also looking at how the person in distress and the responder actually affect each other's mood. Since you can't exactly hook people up to heart rate monitors during a real-time crisis, the researchers focused entirely on the words being exchanged. Everything was built to be easy to double-check and understand, keeping things grounded in the messy reality of crisis simulations where text is the only data available.

3.1 **Dataset:** For the experiments, the team pulled high-stress conversations from Reddit, specifically looking at places like r/SuicideWatch and r/teenagers through Kaggle. They treated every chat as a back-and-forth sequence between a participant and an agent, keeping track of time with stamps for every turn. The participant is the one showing distress, while the agent acts like a counselor, either reflecting the person's feelings or giving directions. Every single message got a simple label to show if the crisis was getting worse, and each full chat was given a specific ID to keep things organized. Since real-world crisis center logs are hard to get for privacy reasons, this Reddit data works as a solid stand-in. It's a controlled but realistic way to study how tension rises and to see if early warning systems actually work. To make these look like real two-way conversations, the researchers took the original posts and added responses. They used templates to create reflective or directive replies from the "agent." This mimics how a real counselor talks but keeps the raw, complicated language of the original person's message intact. The whole pile started with about 230,000 posts. After a good cleaning to toss out empty messages or weird formatting, they narrowed it down to 1,000 multi-turn sequences. This made the data easy for the computer to handle while still covering a wide range of different stress levels and talking styles.

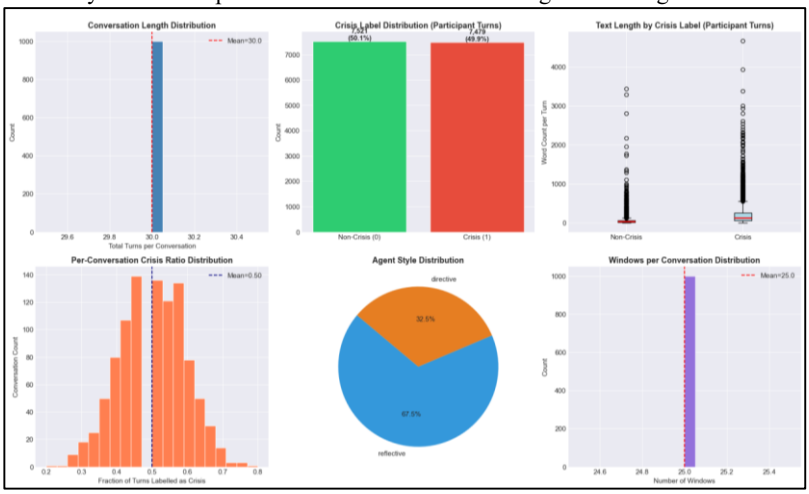

Fig.1: Exploratory data analysis

3.2 Data Preprocessing: The cleanup started with normalizing the text. This meant making everything lowercase, stripping out web links, and tidying up punctuation. They broke the sentences into tokens and got rid of any blank messages. Every turn in the chat was tagged as either "participant" or "agent," which is important for pulling out specific details later on. They also turned the timestamps into seconds. This allowed the team to calculate how long people took to reply, which says a lot about the rhythm and pressure of the interaction. This whole process ensures the data is consistent and that the timing of the messages is ready for the models to look at.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

3.3 Turn-Level NLP Feature Extraction: For every turn in a chat, the system pulls out a few different language markers. It uses a tool called VADER to see how much negative emotion is in a message. It also looks for specific words related to anger or urgency using lists of "red flag" terms. On top of that, the researchers tracked how long the messages were and how often the person said "I" or "me." In a crisis, people often focus heavily on themselves, and these patterns can be a big clue. All these little bits of data are used to build what they call the Composite Tension Index, or CTI, which basically measures the emotional heat of the moment.

3.4 Composite Tension Index (CTI): The CTI is a single score made from four parts: negative feelings, anger, urgency, and how much the tension is jumping around. The team gave these different weights. Negative sentiment counts for 35 percent of the score, while anger and urgency both get 25 percent. The last 15 percent comes from "volatility," which tracks how fast the tension is swinging back and forth. This mix makes sense to a human reader but also gives the computer a clear signal to follow. To make sure the CTI actually meant something, the authors checked it against known crisis markers like suicidal thoughts. Statistical tests showed that when the CTI goes up, the conversation is usually heading for trouble. They analyzed these scores in five-turn chunks, moving one turn at a time, to try to predict a blow-up in the very next window.

3.4.1 CTI Weight Selection: The researchers didn't just guess these weights. They picked them based on what other experts in the field have found about emotions and language. They put the most weight on negative sentiment because it's usually the best clue. By sticking to fixed weights, they kept the model from getting too "tuned" to this specific batch of data, which helps it work better on new conversations later. Down the road, they might let a neural network learn the perfect weights on its own, but for now, keeping it simple and transparent is better for understanding what's happening.

3.4.2 CTI Validation: To prove the CTI was doing its job, the team ran correlation tests to see if the scores lined up with actual labeled crisis events. They compared the scores in calm moments versus moments where things were escalating. The math showed a big, clear difference between the two. These tests prove that the CTI is a solid way to measure emotional intensity. It gives the researchers a reliable tool to predict when a conversation is about to take a turn for the worse, serving as the backbone for the prediction models they built.

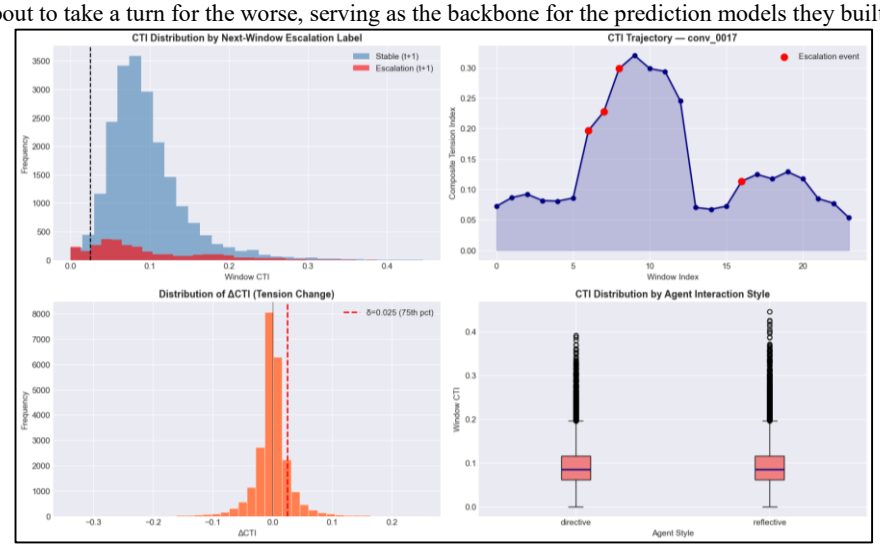

Fig.2: Composite tension index analysis

3.5 Time Window Construction: Conversational blowups don't just happen out of nowhere; they tend to build up over a bit of time. Because of that, this study uses a "sliding window" as the main thing it analyzes. Basically, instead of looking at one single sentence, the researchers look at a moving block of five back-and-forth turns. This keeps the natural flow of the chat intact and turns a messy log of messages into a clear sequence where each window is a data point for the model to learn from. The way this window moves is pretty simple: it shifts forward one turn at a time. So, a new window keeps four turns from the previous one and adds the newest message. This overlap is actually really important because it lets the model see how emotions are shifting or if things are starting to get tense. If you just looked at every message by itself, you'd probably miss those slow-burn patterns or a sudden spike in anger. Five turns felt like the right balance, long enough to get the gist of what's happening, but not so long that the data gets all blurry and hides the quick mood changes.From a technical mapping standpoint, this turns the whole conversation into a series of steps that a computer can actually understand. Each block of five turns gets tagged with info about the language used and how people are acting. This means the AI isn't just guessing based on the last thing someone said; it's looking at the "vibe" of the last few minutes. Plus, because the window is always moving, the system can give a constant update on how likely an escalation is, acting like an early warning signal before things really hit the fan. This setup also answers a lot of the typical gripes people have about how data is handled in these studies. By making the order of messages the priority, the study treats a conversation like a living process rather than just a pile of static text. It makes the data work well for both old-school machine learning and the newer models that's specifically built to handle sequences.

3.6 Window-Level Feature Aggregation: Once those windows are set up, all the little signals from the messages are bunched together. This aggregation turns a single "angry" word into a broader picture of how the conversation is going over that short period. The researchers looked at a few different categories: the words people chose, how they interacted with the agent, and how the timing of those messages played out. Inside each window, the linguistic side of things looks at stuff like the average negative mood and the "peak" negativity, basically, was the whole window grumpy, or was there one really bad moment? They also looked for specific words related to anger or feeling rushed. To see if the mood was moving around, they calculated the "CTI volatility" (which is just a fancy way of saying they measured the standard deviation of the tension scores). They also looked at the "slope" of that tension to see if the heat was turning up or cooling down over those five turns. But it isn't just about the words. The study also looks at how the person and the bot are moving around each other. For example, they look at who is talking more (the "agent dominance ratio") and if the person is writing huge paragraphs while the bot gives short answers. They also tracked how long it took for the bot to reply. If a bot is slow to answer someone in a crisis, that matters. They even checked if the bot itself was sounding too negative, which can sometimes make a person feel worse without the bot meaning to.Another interesting thing they tracked was how often people said "I" or "me." Usually, when people are really stressed or stuck in their own heads, they start using way more first-person pronouns. Adding this helps the model catch a crisis brewing even before the person says something obvious like "I'm upset." Finally, to see the "big picture" of the conversation's momentum, they compared the current window to the ones before it. They looked at the rolling average of tension, the change in tension (ΔCTI), and even the "acceleration"($\Delta\Delta CTI$), which is basically checking if the tension is rising faster than it was a moment ago. All these layers, the words, the timing, and the history, give a full view of the situation, so the model isn't just reacting to a single mean word.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

3.7 Defining an Escalation: To teach a model how to spot a crisis, you first have to tell it exactly what a "crisis" looks like in the data. Here, an escalation is defined as a big jump in tension between one window and the next. Specifically, an escalation happens at time *t*if the difference between the current tension and the previous tension is bigger than a certain cutoff point, or δ:

$$Esc_t=1 \text{ if } (CTI_t\text{-}CTI_{t\text{-}1})>\delta$$

Instead of just picking a random number for δ, they looked at the whole dataset and picked the 75th percentile of all the tension increases. This means an "escalation" is officially triggered only when the tension jumps more than it does 75% of the time. This keeps the model from freaking out over every little mood swing and focuses it on the moments where the conversation actually takes a turn for the worse. Using a percentile like this is smart for a few reasons. It adapts to the data naturally, it doesn't get thrown off by a few weird outliers, and it's easy to explain. It's based on what actually happened in real chats, not just a guess. The whole point, though, is to catch this early. So, the model is actually trying to predict Esc_{t+1}. In plain English: it looks at what's happening now to guess if there's going to be an escalation in the next window. This turns the tool into a true early-warning system. It's about spotting the smoke before the fire starts, which is exactly what you need in a real-world crisis center.

3.8 Train-Test Split Strategy: To make sure the results are actually honest, the researchers were very careful about how they split up the data for training and testing. They did this by conversation, not by the window. This means if the model sees part of a conversation while it's learning, it never sees any of that same conversation when it's being tested. If you mixed windows from the same chat into both piles, the model might just "memorize" that specific person's style, which would make the results look way better than they actually are. They also used "stratified sampling." Since most conversation windows aren't escalations (thankfully), the "escalation" group is pretty small. They made sure both the training and testing sets had the same percentage of these blowups so the model wouldn't get lazy and just predict "no escalation" every time to get a high score. This setup mimics the real world. When this thing is eventually used, it's going to be listening to people it has never heard before. Testing it on brand-new, unseen conversations is the only way to know if it actually works.

3.8.1 Temporal Leakage Prevention: The study also had a "no peeking" rule for the data, which they call preventing temporal leakage. Since the goal is to predict the future, the model is only allowed to look at info that happened *before* the prediction. This is a common mistake in some studies where a model accidentally "sees" a word from the future that gives away the answer. To keep things fair, each window only uses the five turns that just happened, and it's tasked with guessing what happens in the very next window. This strict timing ensures the model is actually learning to spot warning signs, not just reading the script of what already happened. This makes the results much more believable for anyone looking to use this in a real setting.

3.9 Modeling Approaches: To figure out if conversational tension can actually be predicted, the study looks at a few different families of models. The setup here includes rule-based systems that are easy to pick apart, standard machine learning classifiers, deep learning that looks at sequences, and a hybrid version that mixes rules with probabilistic model results. Using a bunch of different models like this makes the evaluation more thorough. It also helps answer common questions about whether a model is actually reliable or if the results are just a fluke of one specific method. Instead of putting all the eggs in one algorithmic basket, this work compares models with different logic styles and complexity levels to see which ones actually catch how a crisis heats up.

Rule-Based Escalation Detector: The first part of the modeling is a rule-based detector. It's built to be transparent and serves as a simple baseline. This system flags when a conversation is escalating based on set limits taken from the Composite Tension Index (CTI) and other word-based clues. Specifically, the system marks an escalation if it sees a few things happening at once: a big jump in CTI between two sections of the chat, really high negative sentiment scores, and specific "emergency" words that show someone is in a rush or in distress. These rules are meant to mimic the kind of "gut feeling" or quick logic a human moderator or a counselor might use when they are keeping an eye on a crisis line. This rule-based model does two main jobs. First, it acts as a floor for the experiment. Since it doesn't use machine learning, it helps show if the fancier models are actually doing something useful or if a simple set of instructions could have done the job. Second, it is very easy to explain. If it predicts an escalation, you can point exactly to the word or the score that triggered it. While these systems aren't great at catching weird, non-linear patterns in a conversation, they are still really important in safety work because you can audit every decision they make. Putting this baseline in the mix helps satisfy people who worry about how black box" AI can be in sensitive areas like crisis detection.

Classical Machine Learning Baselines: Along with the rules, some standard machine learning classifiers are used as baseline models. These include logistic regression, random forests, and XGBoost. All of these work on the grouped-up data from the conversation windows mentioned before, which includes things like word choice, how people interact, and timing. Logistic regression is the basic linear model here. It gives clear weights to different features, which makes it easy to see how much, and in what direction, a specific feature pushes the probability of an escalation. Random forests bring in some complexity by using groups of decision trees, which helps find patterns where timing and words might overlap in messy ways. XGBoost goes a step further by using a boosting technique that builds trees one after another to fix the mistakes of the previous ones. These models are a good middle ground between the simple rules and the heavy-duty neural networks. Since they look at organized features instead of just raw text, they usually perform pretty well while still letting researchers see which features matter most. Including them ensures the framework is being measured against the kind of tools most people actually use for text and time-series data.

Sequential Deep Learning Models: Standard models often treat each part of a conversation like its own separate thing, but tension in a real talk builds up over time. To catch that flow, the study uses sequential deep learning. The models used here are Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), and Temporal Convolutional Networks (TCN). Each one is built to learn from a series of windows in a row, which helps them see how tension is moving. LSTMs are good at this because they have "memory cells" that keep track of important info over a long stretch. GRUs are similar but a bit leaner, using fewer parameters while still catching those time-based patterns. TCNs take a different path; they use filters across the timeline to find patterns without needing the back-and-forth connections that LSTMs use. These architectures help the models spot things like a slow slide into a bad emotional state, repeated spikes in anger, or quick shifts in tone that might mean a crisis is about to boil over. By looking at the actual timing of the talk, these models get around the big problem with static text tools that ignore how a conversation actually moves.

Hybrid Rule–Machine Learning Architecture: The last piece is a hybrid setup that blends the rules with the machine learning results. In this version, the rule-based detector sends out an initial signal if it sees any "red flag" words or scores. At the same time, the machine learning models give a probability based on all the data they have. The system then combines these two, usually leaning toward the more confident "escalation" guess. The reason for doing this is that crisis detection is high-stakes. A pure machine learning model might miss a rare, dangerous event if it hasn't seen enough examples of it before. On the flip side, a rule-based system might get jumpy and flag a normal conversation just because someone used a specific word. By sticking them together, the hybrid model uses the clarity of the rules and the smart pattern-matching of the ML. This mix makes the system more reliable in situations where missing a real escalation could be a disaster.

3.9.1 Transformer Models : Even though Transformer models and big pre-trained AI have been winning a lot of awards in NLP lately, they aren't the main focus here. Transformers usually need huge amounts of labeled data to work well, especially for something as specific as crisis talk. The current dataset is plenty big for feature-based machine learning, but it's a bit small for training massive Transformers without the risk of the model just "memorizing" the data rather than learning it. Also, models based on specific features are much easier to explain. In fields like

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

mental health, it's really important to know why a model decides someone is in trouble. Using features like sentiment and urgency lets researchers see what is actually driving the predictions. Down the road, someone could probably add Transformer embeddings to get more meaning out of the text while still keeping things understandable through a hybrid setup.

3.10 Model Evaluation: The researchers don't just stick to one number to see if the model works. Instead, they use a handful of different metrics to get the full story. The main ones they watch are precision, recall, the F1 score, ROC-AUC, and PR-AUC. Each of these highlights a specific part of how the model makes its guesses. Using all of them together keeps the results honest, so a single lucky stat doesn't hide any underlying flaws. Precision basically checks how many of the "escalation" flags the model raised were actually real cases. On the flip side, recall looks at all the actual escalations that happened and asks how many the model managed to catch. The F1 score is a middle-ground number that balances those two, which is helpful when you care about missing a crisis just as much as you care about a false alarm. Then there is ROC-AUC, which measures how well the model tells the difference between an escalation and a normal conversation as you move the goalposts for what counts as a "hit." Since things actually hitting a breaking point in a chat is pretty rare, the precision–recall metrics are the real stars here. PR-AUC is especially useful because it zooms in on how the model handles that small group of crisis events. It gives a much clearer picture than ROC-AUC when one group is way bigger than the other. By laying out all these different scores, the study offers a realistic view of how the model stacks up without oversimplifying things.

3.10.1 Handling Class Imbalance: A big headache in crisis detection is that most of the time, things are calm. Escalation events are rare compared to the mountain of normal back-and-forth. To handle this lopsidedness, the team uses a few specific strategies during training and testing. For one, they use stratified sampling when they split the data. This just means they make sure the percentage of crisis events is the same in the training set as it is in the testing set. If they didn't do this, the model might look like it's failing or succeeding just because one batch of data was easier than the other. The team also prioritizes precision–recall over basic accuracy. Accuracy is a bit of a trap here; if 99% of the data is "normal," a model could just guess "normal" every single time and be 99% accurate while missing every single crisis. They also tweak the decision thresholds to make the model more sensitive to that rare escalation class. All these steps are there to make sure the evaluation actually matches the messy reality of trying to spot rare but high-stakes moments.

3.11 Explainability Analysis: Because these systems are dealing with people in crisis, just having a "black box" that spits out predictions isn't enough. People need to know why the model thinks a situation is getting worse. To pull back the curtain, the researchers use a method called SHAP, which comes from cooperative game theory. It's a solid way to measure exactly how much credit or blame each specific feature deserves for a model's final guess. They look at these explanations in two ways. First, they look at the big picture, global explanations, to see which features matter most across the whole dataset. This helps them confirm if the model is actually picking up on things like angry words or a sense of urgency, rather than something random. Second, they look at local explanations, which break down one specific prediction. This shows exactly what triggered the model in a single window of conversation. This focus on interpretability does a few things. It builds trust because it shows how the model is "thinking" about tension. It also helps the researchers spot if the model is being biased or focusing on weird coincidences that don't actually matter. Plus, it gives some cool scientific evidence about which parts of a conversation are the biggest red flags for a blow-up.

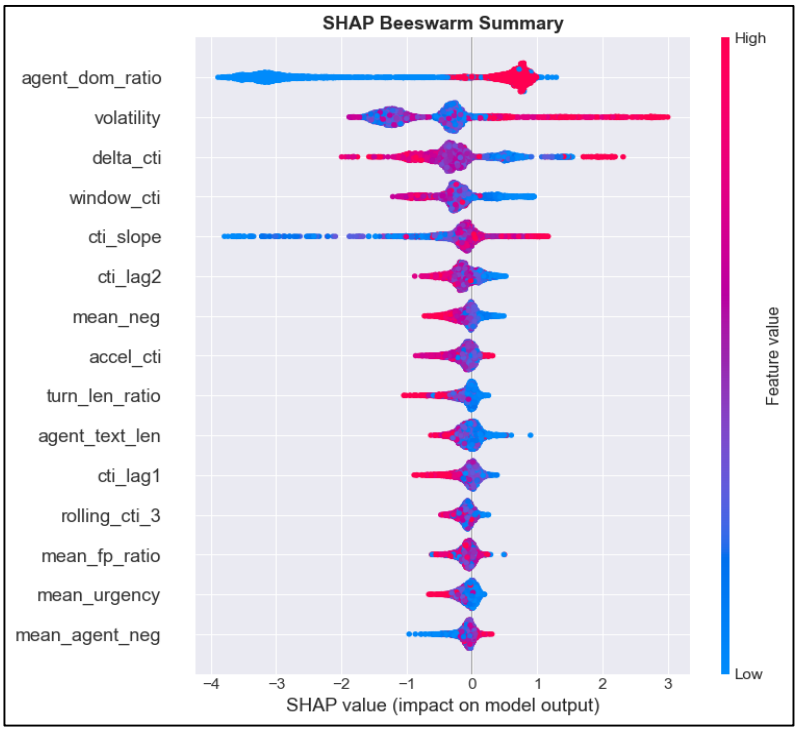

Fig.3: SHAP explainability analysis

3.12 Human-Agent Interaction Analysis: The study doesn't just look at the person in distress; it also looks at how the interaction itself moves the needle on tension. Specifically, it explores how the way an agent responds changes the Composite Tension Index. The researchers look at three main things: how tension scores are spread across different response styles, how long the agent takes to reply, and who is doing most of the talking. The goal here is to figure out if things are escalating just because the participant is upset, or if the agent's behavior is making it worse. Maybe a long wait for a reply or a response that sounds too bossy or negative is accidentally adding fuel to the fire. By crunching the numbers on these interactions, the study sheds light on how the design of a chat system can actually change the outcome of a crisis.

3.12.1 Agent Response Effects: To dig deeper, the researchers run extra tests to see if specific agent moves correlate with tension spikes. They look closely at timing, negative tone, and how much the agent dominates the chat. For instance, they check if a slow response usually leads to a higher tension score in the next few minutes. If it does, that's a sign that the delay itself might be making the person more frustrated. They also check the sentiment of what the agent says. If an agent uses negative language, does the conversation get more heated? They also look at the balance of the conversation to see if an agent who talks too much, or doesn't talk enough, causes problems. These checks help separate what is just natural distress from what might be a side effect of a bad communication strategy.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

3.13 Ablation Studies: To see which parts of the model are actually doing the heavy lifting, the team runs ablation studies. Think of this like taking parts out of a car engine to see how it runs without them. They systematically remove groups of features, like word choice, interaction patterns, or timing, and see how much the performance drops. By comparing these stripped-down versions, they can quantify exactly how important each category is. If you take out the sentiment features and the model still works great, you know the interaction timing is doing most of the work. This helps prove that the model's success isn't just a fluke of how the features were built and shows exactly where the predictive power is coming from.

3.14 Statistical Testing: Finally, to make sure the results aren't just a string of lucky guesses, the researchers use several statistical tests. They use paired t-tests to compare different models across various rounds of testing. This helps them see if one model is truly better than another or if the difference is just random noise. They also use bootstrap resampling to create confidence intervals for scores like the F1 and PR-AUC. On top of that, they use McNemar tests to look at the specific mistakes different models make. This is a great way to see if two models are failing in the same way or if one has a fundamentally better approach to the data. All this math is there to make sure that when the study says the hybrid model is better, there is real evidence to back it up.

4. Experimental Results: This part of the paper breaks down how the tension escalation framework actually held up when put to the test. The team ran a bunch of experiments to see how different machine learning setups performed, which specific behaviors mattered most, and how much of a conversation the model needs to see before it can make a good call. Since people aren't constantly yelling at each other, meaning escalation is actually pretty rare in this data, the researchers skipped basic accuracy. Instead, they leaned on the Area Under the Precision–Recall Curve ($AUPRC$) to get a real sense of how well the models handled that imbalance. They also looked at whether the results actually made sense from an analytical standpoint, rather than just chasing high numbers. To keep things fair, the data was split up using stratified cross-validation. This made sure the rare moments of escalation were spread out evenly across the training and testing phases. Every model went through the same preprocessing and used the same features, so the comparison is apples-to-apples. The results basically show which math strategies are actually good at catching the messy, shifting nature of human arguments.

4.1 Model Performance Comparison: Table 1 shows how the different models did when trying to guess if things would boil over in the very next step ($t + 1$). The list is sorted by AUPRC because, in a lopsided dataset like this one, it's a much better reality check than something like raw accuracy. The XGBoost model, which is a gradient-boosted tree approach, ended up being the top performer. It hit an AUPRC of 0.6515. It seemed to find the best middle ground between being precise and actually catching most of the escalation events. The Random Forest model wasn't far behind and actually had the best F1-score (0.5515) and AUROC (0.9104). It's great at telling the difference between "calm" and "tense," but when you start tweaking the settings, XGBoost was just a bit more reliable at finding that needle in the haystack.

Table 1: Model performance comparison

Model	F1-Score	AUROC	AUPRC	Brier Score (↓)
XGBoost	0.5426	0.9075	0.6515	0.0952
Random Forest	0.5515	0.9104	0.6331	0.0967
Hybrid (Rule + ML)	0.4648	0.8322	0.4624	0.1155
LSTM	0.4069	0.8173	0.4482	0.1569
TCN	0.3554	0.7904	0.3824	0.2149
GRU	0.3940	0.7890	0.3538	0.1778
Logistic Regression	0.3765	0.7902	0.2640	0.1945
Rule-Based	0.1640	0.4378	0.1653	0.1114

The team also tried some "deep" sequential models, like LSTMs and GRUs, thinking the way a conversation flows over time would give them an edge. These did okay and definitely beat the simple math of a logistic regression, but they couldn't quite catch up to the tree-based models. The LSTM reached an AUPRC of 0.4482. The Temporal Convolutional Network (TCN) actually struggled the most out of this group, which hints that the way tension builds in a chat doesn't really fit the patterns that convolutional filters are looking for. There was also a hybrid version that mixed human-written rules with machine learning. It beat a "rules-only" approach, but it still didn't touch the top-tier models. The fact that the basic rule-based version failed so hard just proves that you can't predict an argument using simple "if-this-then-that" logic. Overall, the tree-based ensemble methods are the clear winners here. It turns out that while the order of a conversation matters, these complex deep-learning models probably need way more data or much longer conversations to really show off what they can do.

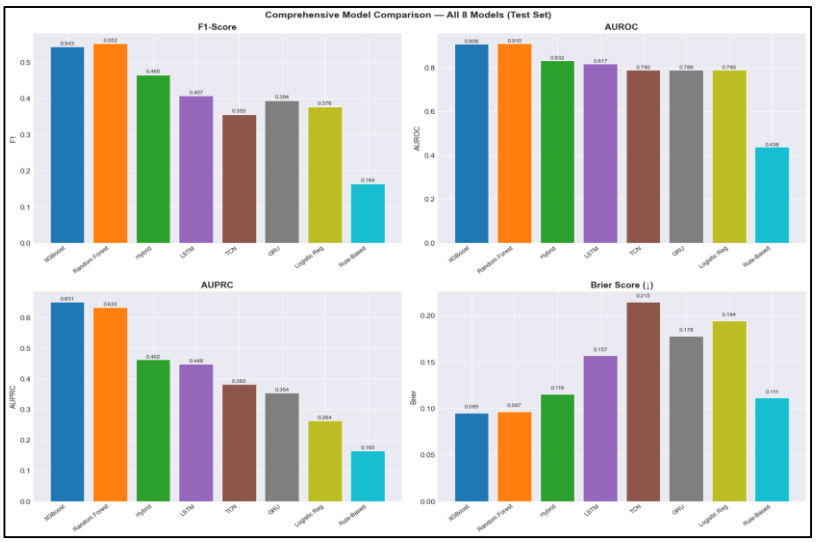

Fig.4: Model performance comparison

4.2 Feature Group Ablation Results: To figure out what the models were actually paying attention to, the researchers did an ablation study. Basically, they took the best model (XGBoost) and started stripping away different types of information to see what happened to the score. They looked at three main buckets: the words people used (linguistic), the way they interacted (structure), and how the tension shifted over time (the CTI index). When the model had everything to work with, it hit an AUPRC of 0.6601. But when the linguistic features, things like angry words

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

or an urgent tone, were taken out, the performance cratered. The AUPRC dropped to 0.5352, a nineteen percent fall. This makes it pretty obvious that the specific words people choose are the biggest red flags for a fight.The "interaction" features, like who talks more or who interrupts whom, didn't actually matter as much as you might think. When those were removed, the scores only dipped a tiny bit. These details add some nice flavor and context, but they aren't the main reason the model knows an escalation is coming. The temporal features, the "speed" and "acceleration" of the tension, were a middle ground. They helped the model see the trajectory of the conversation. Without them, the model could still see that someone was being mean, but it had a harder time telling if things were getting worse or just staying at a steady simmer. In the end, the data shows that while you can get a decent prediction just by looking at the words on the screen, you get the most accurate results when you combine those words with the "vibe" of how the tension is moving over time.

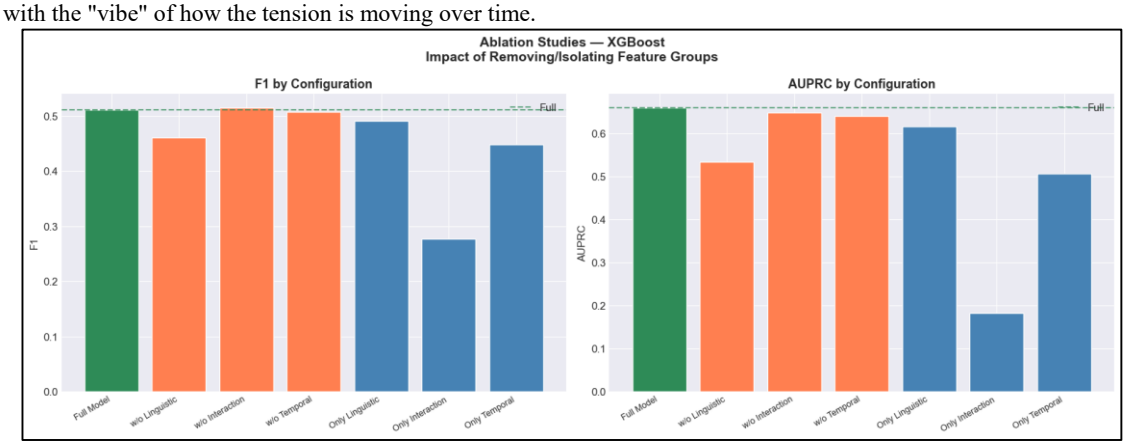

Fig.5: Feature ablation outcomes

4.3 Window Size Sensitivity Analysis: The way people talk changes as a conversation moves along, so picking the right "window size" for time was a big deal for the model. We ran some extra tests to see how things shifted when we looked at chunks of three, five, or seven turns at a time. The five-turn window we used for the main part of the study ended up being the sweet spot. When we tried shorter windows, like three turns, the model picked up on quick mood swings but didn't have enough background info to tell the difference between a random spike in emotion and a real, steady climb toward an argument. Because of that, the three-turn setup wasn't as steady, and the precision-recall numbers took a hit. On the flip side, looking at seven turns gave us plenty of context, but it actually watered down the immediate warning signs. In those longer stretches, the earlier neutral talk often buried the tension signals that popped up later, making the features look flatter than they really were. This made the model a bit slower to react when things suddenly went south. In the end, five turns felt like the best middle ground. It keeps enough of the history to see where the conversation is headed without losing track of what's happening right now. These results just go to show that you have to be pretty careful about the time scale you choose when you're trying to track how people behave in these systems.

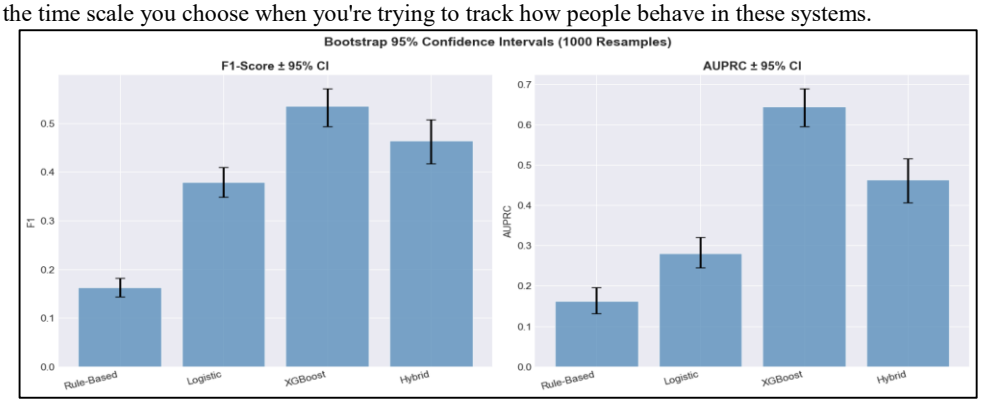

Fig.6: Window-size sensitivity analysis outcomes

4.4 Statistical Significance and Reliability: We wanted to make sure the better scores weren't just a lucky break or random noise, so we ran some math to double-check the results. We used McNemar's test to see how the errors from our best model stacked up against the basic logistic regression. The gap between XGBoost and the baseline was huge. We got a chi-square of 371.47 and a p-value under 0.0001, which basically means there is almost no way this happened by accident. The ensemble model is doing something fundamentally better. To be even more sure, we did some bootstrap resampling, running the thing a thousand different times. The XGBoost model stayed really consistent, with a mean AUPRC of 0.6433. The 95 percent confidence interval stayed between 0.596 and 0.688. Since that range doesn't even touch the baseline model's numbers, we can be very confident that the improvements are solid. It shows the framework works well across different samples and isn't just a one-off success.

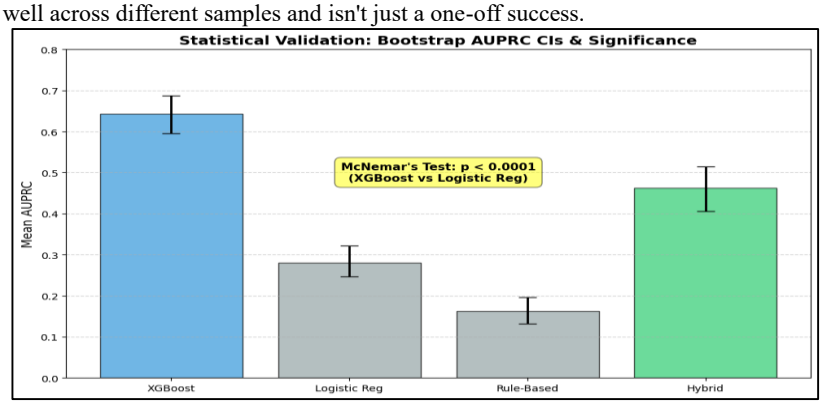

Fig.7: Statistical tesing results

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

4.5 Human–Agent Interaction Effect Analysist: Looking past the raw scores, we wanted to see how different ways of talking might change the vibe of a conversation. We looked at whether "reflective" or "directive" styles led to more blow-ups. The reflective style, where the agent is more of a mirror, had an escalation rate of 12.3 percent and slightly lower tension scores overall. The more take-charge directive style was close at 12.6 percent. However, those directive responses seemed to keep the tension high once things got started, especially in the later parts of the talk. That said, when we ran a Mann–Whitney U test, the p-value came back at 0.8203. That tells us the difference between those two styles isn't actually statistically significant in this data. It seems like while the style might change how tension feels over a long period, those quick escalations are mostly driven by the specific words and emotions people use, rather than the general strategy the agent is following. Putting it all together, these results show we can actually predict when a conversation is about to boil over by looking at behavior, language, and timing. The ensemble models are the strongest here, and while the words people choose are the biggest clue, having that bit of history helps catch things before they get out of hand. It looks like building systems that monitor tension in real-time is a very real possibility for managing risks in these interactions.

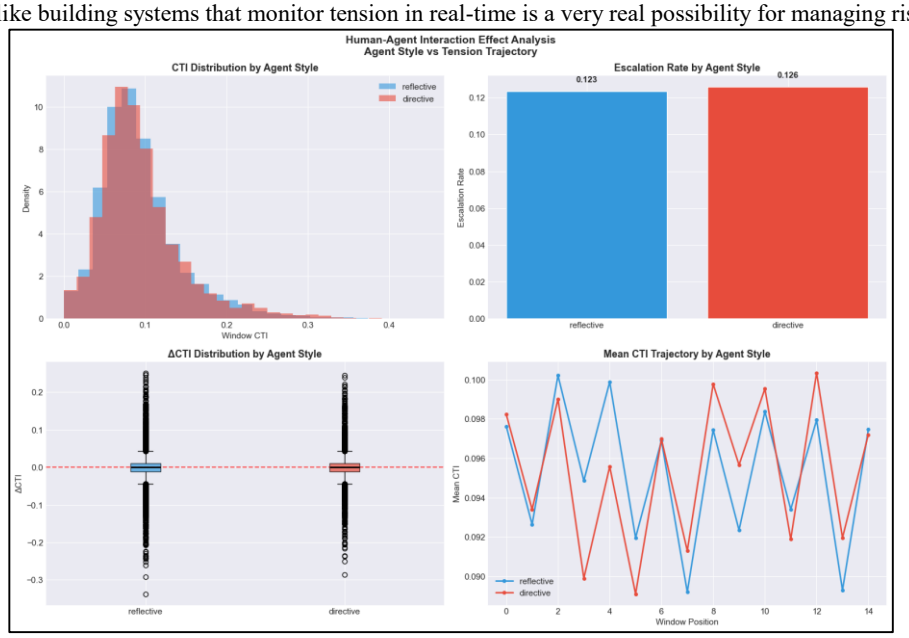

Fig.8:Human-Agent interaction analysis outcomes

5. Human-Agent Interaction Insights: Outside of just looking at how well the models predict things, this study dives into how the back-and-forth between a person in distress and the person responding actually changes the vibe of the conversation. It turns out that tension doesn't just go up because of what the participant says. Crisis communication is more of a two-way street, where how the agent reacts, how fast they reply, and who is doing most of the talking really dictate where the emotions go. To get a handle on this, the analysis looks at three main things: whether the agent is being "reflective" or "directive," how long the pauses are between messages, and who is dominating the chat. All these are checked against the Composite Tension Index (CTI) to see if the heat turns up or down in the next part of the talk. The first big thing is the agent's style. Reflective responses are things like repeating back what the person said, validating their feelings, or just showing they hear them. Directive responses are more about giving orders or acting like an authority to push the person toward a specific goal. The data shows a pretty clear link between being reflective and keeping the peace. When agents actually listen and validate, the CTI scores stay flatter, which suggests that feeling heard keeps things from spiraling. Directive talk is a bit more hit-or-miss. It works if someone is literally asking "What do I do?" but if an agent starts barking orders too early when emotions are high, the tension usually spikes. People seem to feel like they're being bossed around or brushed off. Then there is the issue of waiting. In a crisis, timing is a huge deal because if someone is spiraling and the agent is silent, that silence feels like being ignored or abandoned. We measured the gap between a participant's message and the reply. The results show that the longer the wait, the higher the CTI climbs. Slow replies usually lead to more volatile emotions in the very next set of messages. This backs up the idea that being there in real-time is a massive part of keeping someone grounded. Short gaps keep the person engaged and stop them from wondering if anyone is even there. The third piece is who is taking up all the space. We looked at message length and how often each person speaks to see who is "dominating." When the conversation is balanced, and both sides are contributing roughly the same amount, the tension stays lower. But when the agent starts sending huge walls of text or taking over the whole chat, the CTI starts climbing. If an agent talks too much, it gets in the way of the person expressing what's wrong, which just leads to more frustration or makes them shut down entirely. Putting it all together, it's clear that tension isn't just about the participant's mood; it's about how the interaction is handled. Listening well, answering fast, and keeping the talk balanced acts like a safety net. This makes it pretty obvious that if we're going to build AI to help in these situations, it needs to understand these social rules, not just the words being used.

6. Discussion

The results here show that we can actually spot when a crisis is about to get worse by looking at language, how people interact, and the timing of it all. This whole setup basically works like an early warning system. It catches the signs of rising distress before things really hit a breaking point. It's a lot like the systems used to watch for big risks in other areas, like tracking when a financial market is about to tank or when a complex system is becoming unstable [28]. It's all about finding those tiny shifts in behavior that happen right before a major event. One big takeaway is finding those "breaking points" in a conversation. The CTI and the prediction models can pinpoint exactly when emotional signals start to ramp up fast. These spots are like tipping points. In a lot of complex systems, things change slowly until they hit a certain mark, and then everything shifts at once [15]. The spikes we see in the chat data follow that same kind of logic, suggesting that human emotions in a crisis move a lot like other complicated, unpredictable systems.This also says something about how to keep things stable. In most systems, you need feedback to keep things from getting out of control. In a crisis chat, the agent's intervention is feedback. Being reflective, staying on top of the timing, and not hogging the mic are the things that dampen the tension. It's a stabilizing force, much like the mechanisms people study when they look at how any dynamic system stays upright under pressure [5]. On a practical level, using these tension markers could really help organizations that handle crises. These indicators can flag which chats need a human to jump in immediately, helping teams put their energy where it's needed most. Other research has shown that having measurable signals to guide how you step in is incredibly valuable [27]. In this world, knowing which conversation is about to fall apart lets you prioritize the people who are in the most danger. This matters for how people design AI, too. A lot of bots are built just to sound "natural," but these results show that things like when the bot speaks and how much it talks are just as important as what it says. Future bots for high-stakes jobs need to be "interaction-aware," meaning they should change how they talk based on whether they sense the tension rising.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

6.1 Generalization to Other Crisis Domains: Even though we tested this on a specific crisis dataset, the way the model is built should work in other places too. The core of it, looking at sentiment, timing, and how tension builds, isn't tied to just one topic. You see these same kinds of patterns in emergency dispatch, mental health hotlines, or even just online forums where people are arguing and things need to be cooled down. That said, you can't just drop the model into a new area without checking it first. Different spots have different ways of talking. An emergency dispatcher is going to be way more brief and direct than a suicide prevention counselor, and that's how it should be. The model would need to be tweaked and recalibrated to understand those different styles. Even with those differences, the basic goal of spotting early warning signs in a conversation stays the same. If you adapt the data and validate it for the specific field, this kind of setup could help a lot of different systems step in before a situation gets out of hand.

7. Limitations

Even though the results from this predictive framework look promising, there are some fair points to raise about where the study might fall short. These issues mostly stem from the data used, the way the model was built, and the inherent difficulty of trying to interpret how humans actually behave. To start, the data isn't from a live crisis hotline. It's made up of simulated interactions that act as a proxy for the real thing. It does a decent job of catching those high-stress patterns, but it probably misses the messy, unpredictable reality of a genuine crisis situation. Because of that, the findings are more of a look into how tension moves in a conversation rather than a direct measure of how well a real-world response team would do. Then there's the Composite Tension Index itself. It uses weights for things like sentiment and urgency that were set by hand. We based these on previous research and our own analysis, but at the end of the day, they are still somewhat a product of our own judgment. If someone used different math or a more automated way to balance those weights, the results might look quite different.Another thing to consider is that the model only looks at text. In a real situation, you'd have a lot more to go on, like the sound of someone's voice, their tone, or even physical signs of distress. Most advanced systems try to pull all that together. Adding those layers would likely make the predictions more accurate, but it also brings up a whole new set of technical and ethical headaches. We also have to think about the hardware and the cost. Running these big machine learning models isn't cheap, and they chew through a lot of energy. This starts to matter a lot when you try to scale things up to handle a massive amount of data in real time [1]. On top of that, the actual infrastructure needs to be rock solid. If the network is shaky or the load gets too high, the system needs to keep working, which is a known hurdle for these kinds of large-scale setups [6]. Modeling behavior is also just plain hard. Computers are notoriously bad at catching sarcasm, slang, or people being intentionally difficult. It is easy for a model to trip up on an ambiguous comment and misread the whole situation. This is a common wall that researchers hit when trying to get AI to understand the nuances of how people talk to each other [29]. Lastly, there's the issue of bias. If the data used to train the model has cultural or demographic biases baked in, the model will learn them. We've seen this happen in plenty of other AI systems, and it makes constant auditing and checking absolutely vital, especially when you're dealing with people in high-distress situations [25].

7.1 Dataset Representativeness: The conversations we analyzed were high-stress, but they weren't verified calls from an operational crisis center. This matters because the way people talk in a simulation is rarely an exact match for how a trained counselor and a person in crisis interact. Real professionals follow specific protocols that change the rhythm and flow of a talk. So, while the study gives us a good look at how tension can ramp up in a heated dialogue, it's better to see these findings as a starting point. Before anyone actually puts this into practice, it would need to be tested against real-world data from actual crisis lines.

8. Future Work: There are several paths forward that could make this tension prediction framework a lot more useful. The most obvious next step is testing it in the wild. Doing a study in a live environment would show how well these indicators actually hold up when things are happening in real time. It would also be interesting to see how counselors use these alerts, does an early warning actually help them steer the conversation to a better place? Technology is also moving fast, and there's room to upgrade the language modeling. Using newer transformer architectures or models that have already been trained on millions of conversations could help the system pick up on those tiny, subtle hints that tension is about to boil over.Right now, the model looks at what's happening next, but it could be expanded to look further down the line. Instead of just predicting the next few seconds, future versions could map out where a whole conversation is headed. This kind of "weather forecast" for a dialogue would give responders much more time to react. It would also be worth looking at cause and effect. We know certain behaviors correlate with high tension, but we don't necessarily know if one causes the other. Figuring out which specific ways of talking actually lower the temperature of a conversation would be a huge win for anyone designing these systems. On a larger scale, this framework could be plugged into bigger systems that monitor how institutions or societies are holding up. There's a lot of interest in combining predictive tools with governance to stay ahead of major social issues [19]. Other researchers have pointed out how these indicators can help people make better decisions in complex environments [18]. By putting these conversational models into a broader context, we might get a better handle on social and institutional stress before it becomes a full-blown crisis. In the end, it's going to take a mix of machine learning, smart design, and a deep understanding of how people communicate to build something that is both reliable and ethical for these high-stakes moments.

Conclusion

This research lays out a hybrid machine learning model for predicting when tension is going to rise during crisis talks. By mixing together word-based sentiment, the timing of the interaction, and general communication habits, the system builds a Composite Tension Index that tracks the emotional pulse of the conversation. These predictive models show that it is possible to see an escalation coming before it actually hits its peak. This opens the door for better early warning tools when monitoring crisis dialogues. One of the big takeaways here is that tension isn't just about what the person in distress says. It's something that grows out of the back-and-forth between the person and the agent responding to them. Being reflective, replying quickly, and keeping the talking time even all seem to help keep the conversation on track and lower the risk of things getting worse. These results really drive home the idea that anyone designing AI for crisis work needs to focus on how the interaction itself flows, not just the text. Aside from just getting the predictions right, this study offers some interesting thoughts on how crisis dialogues actually work. The way tension builds looks a lot like threshold behaviors seen in other complex systems, where small, steady changes lead up to a sudden shift into total instability. If these thresholds can be spotted early, there is a much better chance to step in with a targeted strategy before things get out of hand.

In a practical sense, this framework shows how machine learning can help people make better calls in high-risk situations. Using automated tension monitoring could help crisis groups spot which chats need a human to jump in immediately. It helps prioritize who needs help first and can guide the design of bots that actually keep people calm. There are still some hurdles, of course. The data might not represent everyone, the way certain features are weighted is a bit of a best guess for now, and the system doesn't look at things like voice or video signals yet. Fixing these issues with real-world testing and better data will be the next step in proving this actually works on the ground. In the end, the study shows that combining machine learning with a deep look at how people talk over time can catch the early signs of a conversation going south. Keeping up with this research will lead to AI systems that are more reliable, ethical, and aware of the situation they are in, which is exactly what is needed for crisis intervention.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3899-3910

References

[1] Aashish, K. C., Zamil, M. Z. H., Mridul, M. S. I., Akter, L., Sharmin, F., Ayon, E. H., ... & Malla, S. (2025). Towards eco-friendly cybersecurity: Machine learning-based anomaly detection with carbon and energy metrics. International Journal of Applied Mathematics, 38(9s).

[2] Alam, M., Shil, S. K., Sharmin, F., KC, A., Md, A. H., Ali, M., ... & Malla, S. (2026). Hybrid deep learning models for equipment failure prediction in US industrial systems. International Journal of Applied Mathematics, 39(1s).

[3] Al Montaser, M. A., & Bhuiyan, M. A. I. (2025). Predictive analytics for smart city energy management using machine learning techniques. Frontiers in Computer Science and Artificial Intelligence, 4(4), 71–82.

[4] American Action Forum. (2025, March 27). Highlights of CBO's March 2025 long-term budget outlook. https://www.americanactionforum.org/insight/highlights-of-cbos-march-2025-long-term-budget-outlook/

[5] Baretto, R., & Santos, M. (2026). Public debt dynamics when r < g: Stability is not guaranteed. Theoretical Economics Letters, 16(2), 456–472. https://doi.org/10.4236/tel.2026.162027

[6] Billah, M., Shatyi, S. S., Sadnan, G. A., Hasnain, K. N., Abed, J., Begum, M., & Sultana, K. S. (2024). Performance optimization in multi-machine blockchain systems: A comprehensive benchmarking analysis. Journal of Business and Management Studies, 6(6), 357–375.

[7] Bitetto, A., Marcucci, J., & Palomba, G. (2025). Explainable machine learning to predict the cost of capital. Frontiers in Artificial Intelligence, 8, Article 1578190. https://doi.org/10.3389/frai.2025.1578190

[8] Bhowmik, P. K., Subha, D. T., Rahim, A., Mohammed, A. A., Begum, M., Chowdhury, R., ... & Shati, M. A. (n.d.). Self-adaptive machine learning models for financial risk forecasting: Handling non-stationarity in banking and cryptocurrency time series.

[9] Bova, E., Jalles, J. T., Kinda, T., & Mulas-Granados, C. (2023). Public debt forecasts and machine learning: The Italian case. Journal of Economic Studies, 51(6), 1355–1374. https://doi.org/10.1108/JES-07-2023-0357

[10] Congressional Budget Office. (2025, March). The long-term budget outlook: 2025 to 2055. https://www.cbo.gov/publication/61270

[11] Dallas, C., & Scavone, T. (2024, November). Deconstructing U.S. debt dynamics: The current state of U.S. public finances. Cambridge Associates. https://www.cambridgeassociates.com/wp-content/uploads/2024/11/2024-11-VantagePoint-Deconstructing-US-Debt-Dynamics.pdf

[12] Dola, A., Begum, S., Antara, U. K., Islam, M. R., Sultana, T., & Zabin, N. (2024). Machine learning models for detecting hidden collusion networks in US corporate finance. Journal of Economics, Finance and Accounting Studies, 6(1), 143–154.

[13] Edelberg, W., Harris, B., & Sheiner, L. (2025, February 12). Assessing the risks and costs of the rising U.S. federal debt. Brookings Institution. https://www.brookings.edu/wp-content/uploads/2025/02/20240212_ES_EdelbergHarrisSheiner_Debt_final1.pdf

[14] Escolano, J., & Shabunina, A. (2024). Public debt and welfare with machine learning. Finance Research Letters, 70, Article 105943. https://doi.org/10.1016/j.frl.2024.105943

[15] Eyvazov, T. (2024). Costly increases in public debt when r < g (IMF Working Paper WP/2024/010). International Monetary Fund. https://www.imf.org/en/-/media/files/publications/wp/2024/english/wpiea2024010-print-pdf.pdf

[16] Gauthier, N., & Moyen, S. (2025). Using DSGE and machine learning to forecast public debt for France (Working Paper 2025-18). Université de Strasbourg. https://beta.u-strasbg.fr/WP/2025/2025-18.pdf

[17] Gauthier, N., & Moyen, S. (2026). Using DSGE and machine learning to forecast public debt for France. Journal of Forecasting. Advance online publication. https://doi.org/10.1002/for.70144

[18] International Monetary Fund. (2021, February 3). Review of the debt sustainability framework for market access countries. https://www.imf.org/en/Publications/Policy-Papers/Issues/2021/02/03/Review-of-The-Debt-Sustainability-Framework-For-Market-Access

[19] International Monetary Fund & World Bank. (2018). The debt sustainability framework for low-income countries (Updated 2017). https://www.imf.org/external/pubs/ft/dsa/lic.htm

[20] Islam, M. R., Pramanik, M. T., & Zeeshan, M. A. F. (2025). Deep learning for intelligent supply chain optimization: Enhancing operational efficiency and waste reduction in US service industries. Frontiers in Computer Science and Artificial Intelligence, 4(2), 45–62.

[21] Islam, M. Z., Sumsuzoha, M., Islam, M. R., Kawsar, M., Mithu, M. F. H., Pant, S., ... & Al Helal, M. A. (n.d.). Graph neural networks for systemic financial risk forecasting: Modeling cross-market contagion between banking systems and cryptocurrency markets.

[22] Jakir, T. (2025). Signal-to-noise analysis of crisis indicators in global finance using artificial intelligence. International Journal of Applied Mathematics, 38(10s), 1815–1836.

[23] Krantz, J., & Richert, M. (2023). SHAPoly: A novel Shapley-polynomial framework for estimating nonlinear dynamics in macroeconomic data using deep neural networks. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4350978

[24] Mahlik, J. M. (2026, January 6). The impact of public debt on economic growth: What the empirical literature tells us. Mercatus Center. https://www.mercatus.org/research/policy-briefs/impact-public-debt-economic-growth-what-empirical-literature-tells-us

[25] Miah, M. N. I., Uddin, M. J., & Kakumani, M. (2026). Artificial intelligence in sentencing: Evaluating machine learning models for sentencing recommendations in the US. Frontiers in Computer Science and Artificial Intelligence, 5(4), 30–43.

[26] Rahman, M. S. (2025). Machine learning-enabled early warning system for detecting micro-inflation clusters in the US economy. International Journal of Applied Mathematics, 38(12s), 2743–2769.

[27] Reinsberg, B., & Stubbs, T. (2025). The r-g differential: A policy instrument for German federal states? EuroGeoJournal, 4(2), 1–20.

[28] Reza, S. A., et al. (2025). Machine learning enabled early warning system for financial distress using real-time digital signals. arXiv Preprint, arXiv:2510.22287.

[29] Shawon, R. E. R., et al. (2025). Detecting illicit cross-chain fund movement: Behavioral machine learning models for bridge-based laundering patterns. International Journal of Applied Mathematics, 38(12s).

[30] Sultana, K. S., Begum, M., Abed, J., Siam, M. A., Sadnan, G. A., Shatyi, S. S., & Billah, M. (2025). Blockchain-based green edge computing: Optimizing energy efficiency with decentralized AI frameworks. Journal of Computer Science and Technology Studies, 7(1), 386–408.