TruthSetu: An AI-Based Multi-Agent System for Real-Time Crisis Misinformation Detection in Environmental Disaster Contexts

Pranav Marekar, Aarush Parate, Jeet Ghegade, Arnav Karmankar, Prof. Anuja Gaikwad
Department of School of Computing
MIT-ADT University, Pune, Maharashtra, India

***Abstract*—Environmental crises — including cyclones, floods, dam failures, and seismic events — generate high-volume misinformation on platforms such as WhatsApp, accelerating public panic and impeding coordinated disaster response. Existing fake news detection systems are predominantly content-centric and fail to operate at the speed, scale, and linguistic diversity demanded by such crisis environments. This paper presents TruthSetu ("Truth Bridge" in Sanskrit), a real-time, multi-agent AI system designed specifically for the Indian crisis information ecosystem. TruthSetu deploys a five-stage pipeline — SCOUT, VERIFY, TRANSLATE, DEPLOY, and LEARN — integrating large language model (LLM) reasoning, Retrieval-Augmented Generation (RAG) over a FAISS semantic index, and automated WhatsApp delivery to return verified corrections to citizens within three to ten seconds of query receipt. The system incorporates trust-weighted source hierarchies privileging government and fact-checking outlets, multilingual response generation in Hindi and Marathi via the Groq LLaMA-3.1-8b-instant model, and a continuous self-improvement loop that seeds verified false claims as retrievable templates. Empirical evaluation of the underlying machine learning detection module using nine classifiers on a 44,898-article dataset demonstrates that XGBoost achieves peak performance with an accuracy of 0.9967 and an F1-score of 0.9964. We further situate TruthSetu within the broader sociotechnical literature on fake news detection, drawing connections to the SHAPE framework (Veerasamy and Badenhorst, 2026) and arguing that effective crisis misinformation detection demands the integration of computational precision, social-network awareness, and human-centred design.**

Keywords: **Fake news detection, Environmental crisis misinformation, Multi-agent systems, Retrieval-Augmented Generation, WhatsApp, Natural language processing, Machine learning, FAISS, Sociotechnical systems**

I. INTRODUCTION

The rise of the misinformation ecosystem during environmental disasters and natural calamities has become a significant hazard for people's security. Whether there is an impending cyclone that will hit the coastlines, or if a massive dam faces any problems that might lead to disaster, or if there are floods that have taken over an urban settlement, the resulting vacuum left by these crises will be quickly filled with rumors, fake warnings, and tampered information from various media outlets. This is especially true within a country like India, which has a very high mobile presence rate and depends mostly on WhatsApp for all news updates.

As used in this paper, the term "misinformation" ranges from accidental sharing of erroneous claims to deliberate dissemination of disinformation meant to deceive. As Aïmeur et al. note in their paper published in 2023, "misinformation refers to falsehoods being shared in good faith, disinformation refers to falsehoods being shared intentionally, and malinformation refers to truthful information being shared for malicious purposes." All three types of misinformation can be seen during an environmental crisis: ordinary citizens who are scared of floods pass around unverified reports about water levels (misinformation); political groups spread fake stories about the government's failures to address the situation (disinformation); and true emergency evacuation orders are selectively disseminated to particular populations (malinformation).

Progress in computational methods for detecting fake news is immense within the last decade. Standard methods of machine learning like logistic regression, support vector machines, random forests, and gradient boosting algorithms trained on top of TF-IDF features yield over 0.99 accuracy on conventional benchmark datasets (Al Ibraheemi and Jabardi, 2024). CNNs and LSTM-based deep models

further leverage these methods by modeling sequences and hierarchies within the language. Models based on the transformers paradigm, including BERT, RoBERTa, and their fine-tuned counterparts, are state-of-the-art when it comes to lexical identification of fake news (Veerasamy and Badenhorst, 2026). Large language models provide even more flexibility by facilitating few-shot classification, claims decomposition, and evidence-based reasoning while not relying on task-specific annotated corpora.

However, despite all these developments, there is still one key challenge that exists: technocentric approaches do not take into consideration the social and context-based nature of crisis disinformation. As per Veerasamy and Badenhorst (2026), fake news emerges in contexts where there are complicated social dynamics, specific platform advantages, and human intelligence. In addition, algorithmic models are easily bypassed through new adversarial language and fall apart in situations involving low-resource languages, and are inherently unable to get to the person at the time of decision-making.

The solution to this challenge is presented by TruthSetu, which provides a sociotechnical framework that integrates LLM-powered reasoning with reliable information sources from governments and fact-checkers and provides multilingual support through WhatsApp while continually enhancing its knowledge base via a learning agent. This research offers the following contributions:

- A fully operational multi-agent pipeline for real-time crisis misinformation detection, integrating LLM reasoning, semantic vector retrieval, and trust-weighted source hierarchies.
- A comprehensive empirical comparison of nine machine learning classifiers on a large fake news dataset, with XGBoost achieving peak performance (accuracy 0.9967, F1-score 0.9964).
- A multilingual response system supporting Hindi and Marathi via Groq LLaMA-3.1-8b-instant, with a pathway to IndicTrans2 for production deployment.
- A continuous self-improvement mechanism through which verified false claims are stored as retrievable templates, enabling sub-100ms response times on recurring false narratives.
- A theoretical synthesis situating TruthSetu within the SHAPE sociotechnical framework for fake news detection, arguing for the inseparability of computational and human-centred elements.

The rest of the paper is organized as follows. Section 2 provides a review of existing literature regarding fake news detection using cognitive, social network, and technical approaches. Section 3 describes the system architecture and the pipeline used in TruthSetu. Section 4 evaluates the performance of the classifiers using the machine learning model. Section 5 analyzes the sociotechnical aspects of the proposed system and how it fits into the SHAPE framework. Section 6 explores the limitations of the work and future directions.

II. LITERATURE REVIEW

2.1 The Misinformation Problem in Crisis Contexts

It is widely recognized in the academic discourse that there is a clear connection between environmental disasters and fake news. However, social media magnifies this effect because false information spreads more rapidly and extensively than factual information in online social networks, where emotionally laden messages – exactly those created by disasters – receive the most exposure (Vosoughi et al., 2018). The engagement-based dynamics of the platform make matters worse by silencing counterarguments and promoting previously engaging content (Bouchaud, 2024).

There are some distinct problems associated with the WhatsApp system in India. Since the content shared on broadcast media is searchable, it can be tracked by analysts or other platform officials, but in the case of the WhatsApp platform, there is no such thing because of its end-to-end encryption technology. Any forwarded message regarding the risks of dam bursting or the dates for cyclones may reach millions of Indians before any correction.

The cognitive elements further aggravate the problem. The familiarity effect, whereby continuous exposure to falsehoods renders them increasingly believable, becomes even more pronounced when citizens receive the same message again and again from sources that are personally known to them as reliable and trustworthy social actors (Veerasamy and Badenhorst, 2026). The availability effect, which involves giving undue weight to accessible information, explains why emotional claims

regarding impending harm tend to gain more traction than others. Lastly, confirmation bias drives people to believe crisis stories that conform to their existing worldview. All three effects are deliberately manipulated by professional fake news producers.

2.2 Machine Learning and NLP Approaches to Fake News Detection

Classical machine learning methods for fake news detection involve feature extraction based on text followed by supervised classification. One popular approach to feature extraction is the use of TF-IDF (Term Frequency-Inverse Document Frequency), which allows us to represent text as a weighted vector, where the frequency of less relevant terms is reduced, whereas discriminative terms are preferred. Al Ibraheemi & Jabardi (2024) show that a combination of TF-IDF representations and gradient boosting techniques results in high classification accuracy rates higher than 0.99 on the Kaggle fake news dataset.

Out of the traditional models, the best results can be achieved using support vector machines due to their ability to detect the optimal decision hyperplanes in sparse and linearly separable spaces. Logistic regression is comparable to SVMs in terms of accuracy and, at the same time, allows for better interpretation; ensemble techniques such as random forest, gradient boosting, AdaBoost, and XGBoost use several weak models to mitigate the risk of overfitting. It should be mentioned that both Naïve Bayes and k-nearest neighbours demonstrate relatively poor results because of their assumptions.

Data pre-processing is very important for the classifier's success. The conventional steps in NLP data pre-processing, including stop-word removal, removal of punctuation marks, lemmatization, stemming, and case normalization, decrease the number of features and allow the model to concentrate on the semantic value of the information. Al Ibraheemi and Jabardi (2024) use all of the mentioned procedures before applying the TF-IDF vectorization, producing an unbiased data set of 44,898 documents (23,478 false; 21,420 true).

2.3 Deep Learning and Transformer-Based Approaches

Moreover, deep learning frameworks go beyond the bag of words representation, taking into account sequence and context features. Local n-gram features can be extracted using hierarchical convolution from CNN, while RNN and LSTM frameworks utilize sequence dependency across the length of articles. Hybrid frameworks like CNN-LSTM and Bi-GRU-BiLSTM frameworks include both local and sequence aspects and often demonstrate better results compared to single models (Veerasamy and Badenhorst, 2026).

Transformers constitute the latest technology. The BERT model (Bidirectional Encoder Representations from Transformers) and its variations, which include RoBERTa and XLNet, employ self-attention networks to encode bidirectional semantic contexts and can achieve great results when trained extensively with a specific task. When applied to detecting fake news, FakeBERT (an implementation of BERT for analyzing texts in social media) performs well in various benchmark tasks (Kaliyar et al., 2021).

Language models (LMs), including GPT and LLaMA, generalize the power of transformers for emergent reasoning on claims and evidence. Papageorgiou et al. (2024) review the role of LMs in detecting fake news and outline three approaches: fake text classification, fact checking, and contextual analysis. Methods such as step-wise prompting and weak supervision enhance effectiveness while mitigating hallucinations. Importantly, LMs continue to be relevant even when BERT-based approaches surpass their performance in text classification tasks, owing to the fact that LMs are adept at generating instructive rationales from multiple perspectives that assist human verification (Hu et al., 2024).

2.4 Social Network and Graph-Based Approaches

Approaches relying on content alone cannot represent the dynamics involved in the spreading behavior that sets fake news apart from real news. The Graph Neural Network (GNN) approach is one method for representing the unique spreading behaviors of fake and real news within social networks, where structural information such as network centrality, fast diffusion properties, and cluster tendencies are utilized (Sivasankari and Vadivu, 2022). Other methods involving GNNs include temporally extended GNN approaches such as delayed time graph convolutions and dynamic graph snapshots (Song et al., 2022).

Methods based on knowledge extract information from the news articles and validate them with an

external knowledge database, like Wikidata, treating the problem of fake news identification as a fact-checking task (Mayank et al., 2022). Graphs combining multiple modalities of information include visual features alongside the text, allowing for identifying inconsistencies in the meaning of texts and images that single modality graphs cannot identify. Finally, Social Network Analysis (SNA) provides a means to evaluate the reliability of users, the community structure, and even cross-domain cooperation, pointing towards fake behavior (Phan et al., 2023).

2.5 Sociotechnical Frameworks for Fake News Detection

The growing body of literature is becoming increasingly aware of the need for more than just improvements in algorithms to tackle the problem of fake news; rather, sociotechnical approaches that consider issues in cognition, society, and technology at the same time are needed. According to Veerasamy and Badenhorst (2026), the framework called SHAPE – Short, Healthy Habits, Analyse Source, Probe Facts, Evaluate – has been devised for designing the Informed Fake News Advisor (IFNA). The concept behind SHAPE is that people should be encouraged to use System 2 processing, or in other words, analytical thinking, rather than System 1 processing, which involves quick cognition.

SHAPE's relevance becomes most apparent in situations of crises, when emotions serve as a key conduit through which misinformation is disseminated. The use of sensationalist titles and urgent language induces emotional states like fear, anxiety, and anger that interfere with analysis and judgment, reducing the likelihood of individuals verifying information prior to dissemination (Beauvais, 2022). In this sense, it follows that any effective system for detecting misinformation in a crisis scenario needs to act more swiftly than emotions, intervening with verified information before the decision to share can be made.

III. SYSTEM ARCHITECTURE: TRUTHSETU

3.1 Overview and Design Philosophy

The TruthSetu System ("Bridge of Truth" in Sanskrit/Hindi) is a crisis misinformation detection tool built specifically to operate in the Indian information space. It is based on three main principles: responsiveness (a reply comes in 3-10 seconds after receiving the query); linguistic accessibility (the ability to analyze content written in several languages including Hindi, Marathi, and English); and continuous self-learning (the system continuously improves itself).

This process is designed to solve a particular operational problem, which involves a situation where a citizen receives a claim via WhatsApp about a cyclone movement, risk of dam failure, earthquake damage, flooding water levels, or an emergency order by the government, and sends that information to the TruthSetu verification phone number. This process automatically confirms the claim receipt, verifies it using the pipeline of five agents and then gives a decision in the language spoken by the citizen, stating the sources used.

The technology stack that powers TruthSetu consists of carefully selected elements at each level. FastAPI running on Python 3.12 delivers the asynchronous backend, which supports simultaneous verifications. The Groq API with LLaMA-3.1-8b-instant allows us to make LLM predictions with a latency of about 0.5 seconds, while the free tier limits our capacity to 14,400 requests per day, which is enough for proof-of-concept and pilot deployments. The MiniLM-L6-v2 model creates 384 dimensional semantic embeddings offline with zero costs and without any networking required. The Facebook AI Similarity Search (FAISS) library finds similar vectors within indexed documents in a sub-millisecond timeframe. The MongoDB Atlas stores facts, verdicts, templates, translations, and training data asynchronously. The Twilio WhatsApp Sandbox facilitates the two-way messaging experience. The APScheduler library executes three scheduled tasks: fact-checker monitoring every 15 minutes, RSS feed updates every 30 minutes, and index rebuilding every 6 hours.

3.2 The Five-Agent Pipeline

TruthSetu orchestrates five specialised agents in a sequential asynchronous chain, with graceful degradation at each step. Table 1 summarises the role, core technology, and key functions of each agent.

Agent	Role	Core Technology	Key Functions
SCOUT	Claim extraction &	LLM (Groq LLaMA-3.1)	Virality gate, 24-hour

Agent	Role	Core Technology	Key Functions
	deduplication		semantic dedup, question-to-statement conversion
VERIFY	Evidence retrieval & verdict	FAISS + Groq tool calling	Template cache, RAG over 1500+ docs, DuckDuckGo fallback, trust-weighted scoring
TRANSLATE	Multilingual response	Groq LLM (IndicTrans 2 planned)	Language detection (en/hi/mr/10 + languages), Devanagari-script output
DEPLOY	Message delivery	Twilio WhatsApp API	Formatted verdict delivery, simulation fallback, deployment logging
LEARN	Continuous improvement	MongoDB + Sentence Transformers	FALSE template seeding, training data storage, source reputation tracking

Table 1: TruthSetu Multi-Agent Pipeline — Agent Summary

3.2.1 Agent SCOUT: Claim Extraction and Deduplication

SCOUT receives raw citizen text — which may be a WhatsApp forward, a question, a news snippet, or conversational text — and produces a clean, deduplicated factual claim ready for verification. The agent applies a virality gate (WhatsApp messages always pass, reflecting the deliberate tipline design), then uses a Groq LLM prompt to convert natural language inputs into verifiable statements: the question "Is it true that the Mullaperiyar dam has broken?" becomes the statement "The Mullaperiyar dam has broken." This conversion is essential because semantic search against a FAISS corpus of news documents requires declarative, noun-verb claim structures to produce meaningful similarity matches.

Then, SCOUT encodes the claimed statement by utilizing the all-MiniLM-L6-v2 model and searches in the claims database using the MongoDB claims database for semantically similar statements within 24 hours before the search (similarity threshold cosine = 0.85). Duplicate claims are sent to VERIFY along with their original claim text, which results in almost instantaneous cache hit. Inputs without any verifiable claims, not related to a crisis event, are labeled NO_CLAIM and ignored.

3.2.2 Agent VERIFY: Evidence Retrieval and Verdict

The VERIFY agent is the most resource-intensive one, as it has to go through multiple phases of evidence extraction and reasoning. The first phase involves checking whether the statement exists in the cache of templates, which contains all previously proven false claims together with their embeddings generated by the LEARN agent. If there is a match, the result is returned in around 50 milliseconds without any LLM interaction.

When there is a cache miss, VERIFY encodes the statement and searches for the relevant information using the FAISS database of over 1,500 indexed articles obtained from eleven sources, including four government RSS channels (Press Information Bureau, India Meteorological Department, National Disaster Management Authority, and WHO), three fact-checking websites in India (Alt News, BOOM Live, and Vishvas News), two Tier 1 Indian national newspapers (The Hindu and Indian Express), and two Tier 2 publications (Times of India and Hindustan Times). Articles are ranked based on their source credibility and recency, with more recent or ephemeral articles having lesser scores as they become older.

The five most highly rated retrieved sources, which are tagged as government, free content, and news, are fed to the Groq LLM, along with its zero-hallucination prompt. The LLM is responsible for determining whether there is enough evidence to provide a judgment or whether it needs to continue with web searching for more evidence. If the LLM invokes the search_web tool, VERIFY uses the cached RSS entries from the past six hours, and only if it does not find any relevant information does it perform a DuckDuckGo web search, thus protecting itself from being flagged by the rate limiter on the

primary evidence route. There is an invocation limit of up to three searches per query to prevent the previously observed infinite loops in the system.

3.2.3 Agent TRANSLATE: Multilingual Response Generation

The TRANSLATE function identifies the user's language based on the first 200 characters of the user's original input by using a deterministic Groq model (temperature 0.0, max_tokens 5) that yields a two-letter ISO 639-1 language code. In the case of the Hindi and Marathi languages, the English verdict is translated into Devanagari-script output using an exclusive Groq prompt that keeps emojis, URLs, and the TruthSetu trademark intact. Inputs written in English or an unrecognized language yield plain English outputs. The translation logic is purposely decoupled into a single function to facilitate a future changeover to IndicTrans2 (AI4Bharat) without altering the rest of the pipeline – better grammar in IndicTrans2's output for Indian languages was deliberately not implemented in the current version because of its 4 GB RAM constraint.

3.2.4 Agent DEPLOY: WhatsApp Message Delivery

The output of the verification is delivered by DEPLOY in the form of a WhatsApp message, in accordance with the emoji traffic light guidelines, i.e., red circle if FALSE verdict, green circle for CONFIRMED, and amber circle for UNVERIFIED. These messages have a standard format, containing: the claim text, its description from the sources' perspective, source name, and the date of verification. DEPLOY failure cases are not critical; hence, the pipeline will log what would have been sent as a message but will go on processing the verdict so that it gets fed into the LEARN process. If not running with Twilio credentials, DEPLOY will log messages on the server.

3.2.5 Agent LEARN: Continuous Self-Improvement

Once the verification is completed, LEARN performs its operations in parallel through three streams. First, all verified claims are stored along with metadata within the training_data collection in MongoDB, thereby creating a corpus that can be used for offline classifier training. Second, for FALSE claims, LEARN converts the claim into an encoded version, determines if there is a template with high enough similarity (a cosine similarity of at least 0.92, higher than SCOUT's duplicate threshold), and creates new templates in case of new FALSE claims while increasing the occurrence counter if there are near duplicates. This is what allows VERIFY to achieve sub-100 ms cache results against recurring false narratives, hence allowing for faster processing of those claims. Third, LEARN updates the source score within MongoDB using a citation count for each domain.

The RSS Monitor complements LEARN's responsive learning with a proactive seed-template generation process: every fifteen minutes, it pulls the entire text of the article from Alt News, BOOM Live, and Vishvas News, processes it via Groq LLM to isolate the specific false claim that needs to be countered, and stores this information as a pre-seeded FALSE template. Any queries made by citizens about issues that fact-checkers have already evaluated will lead to an immediate cache hit, even if TruthSetu hasn't processed the specific query via the WhatsApp route.

3.3 Knowledge Base and Source Architecture

TruthSetu's retrieval engine contains the FAISS semantic index that is built and updated through the RSS Monitor. There are 11 chosen sources that are relevant to Indian crises and reliable in their editorship that are included into the index. The sources related to government – PIB, IMD, NDMA, WHO – have the maximum trust weight of 0.90-0.95. Fact-checking organizations – Alt News and BOOM Live – also have high trust weight of 0.95, representing the main authorities for debunking. Tier 1 sources have trust weight of 0.88-0.90; Tier 2 sources have trust weight of 0.80. DuckDuckGo web searches have minimum trust weight of 0.70.

A total of fifteen static facts manually curated are embedded in the index upon start-up and never become outdated. The static facts include descriptions of IMD alert levels, NDMA guidelines for evacuating in case of flooding, how earthquake predictions cannot be made due to technological constraints, EVM machines being air-gapped and thus disconnected from the internet, and RBI being the sole entity authorized to make any currency-related announcements.

Rumour phrase filter for RSS indexing - rejecting any articles that have the following terms in the body text: "rumour goes viral," "forward on WhatsApp," "fake news is spreading," etc. - avoids one of the key mistakes in indexing: without this filter, articles from Alt News debunking cyclone rumours would be selected by VERIFY as actual proof that there is an

approaching cyclone, since they include the text of the rumour that is being debunked.

IV. EXPERIMENTAL EVALUATION

4.1 Dataset

Evaluation with Machine Learning is done with Kaggle Fake News Dataset by BHAVIK JIKADARA containing 44,898 news stories with 4 attributes (title, text, subject, date), along with the labels 1 (for real news) and 0 (for fake news). The dataset comprises 21,417 real news stories and 23,481 fake news stories; hence it has almost an equal ratio of classes, namely 47.7% real, 52.3% fake, thus making it less prone to any class imbalance biases while training. For classification purposes, only the text attribute of the news story is considered.

4.2 Preprocessing and Feature Extraction

NLP pre-processing was carried out for all articles' texts: stripping HTML tags (artefacts of web-scraping procedure), stripping common words (such as "a," "an," "the," "is," "of," etc.), stripping special characters, lemmatisation/stemming for words' root-form extraction, and lowering letters to lower case. Vectorisation of preprocessed texts was done with help of TF-IDF technique, which measures words' weights according to their frequency within document normalised against their scarcity throughout the whole corpus, yielding high dimensional and sparse vectors that emphasise discriminating vocabulary at the expense of common "stop-phrase" artefacts.

4.3 Classifiers Evaluated

Nine algorithms were assessed: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Gradient Boosting (GB), K-Nearest Neighbors (KNN), Decision Trees (DT), AdaBoost, and XGBoost. The aforementioned algorithms cover linear, instance, tree, and ensemble learning approaches, offering a broad understanding of how different inductive bias algorithms operate with TF-IDF representation of news articles.

4.4 Results and Discussion

Table 2 presents the classification results for all nine classifiers across accuracy, loss, recall, precision, and F1-score metrics. All classifiers except Naive Bayes and KNN achieve accuracy above 0.98.

Classifier	Accuracy	Loss	Recall	Precision	F1-Score
Logistic Regression (LR)	0.9846	0.0154	0.9856	0.9815	0.9835
SVM	0.9920	0.0079	0.9926	0.9904	0.9915
Random Forest (RF)	0.9964	0.0036	0.9978	0.9946	0.9961
Naive Bayes	0.9279	0.0720	0.9117	0.9323	0.9218
Gradient Boosting	0.9953	0.0046	0.9984	0.9917	0.9950
KNN	0.7127	0.2873	0.4200	0.9208	0.5768
Decision Tree (DT)	0.9956	0.0043	0.9955	0.9952	0.9953
AdaBoost	0.9947	0.0052	0.9958	0.9930	0.9944
XGBoost	0.9967	0.0033	0.9984	0.9946	0.9964

Table 2: Classification Results for Machine Learning Methods on the Fake News Dataset

XGBoost provides the best results, with an accuracy score of 0.9967, a recall score of 0.9984, precision score of 0.9946, and an F1-score of 0.9964. XGBoost is a gradient-boosting algorithm that uses decision trees with regularization to model the complicated interactions between TF-IDF variables. The Random Forest and Gradient Boosting models have similar scores, with accuracy scores of 0.9964 and 0.9953 respectively.

The SVM model reaches an accuracy of 0.9920, which illustrates the high efficiency of the SVM classifier in dealing with high-dimensional sparse spaces by identifying optimal hyperplanes. The ability of the SVM model to handle multidimensional data and find an effective decision boundary is one of the key aspects of detecting fake news articles.

The Naive Bayes classifier is less effective (accuracy = 0.9279) since it does not account for the relationships between the variables through the use of conditional independence in the generation of the vocabulary used in fake news articles. The KNN algorithm proves to be the least effective among those used (accuracy = 0.7127), as it is inappropriate for high dimensional data.

The above findings about empirical evidence guide the design of TruthSetu in two ways. First, the success of using XGBoost and ensemble models for offline classification proves the utility of supervised learning as a means of efficiently and quickly filtering out claims, using cheaper methods such as the application of LLMs in reasoning only when necessary, something the LEARN agent aims to achieve through acquired training data. Second, the failure points associated with simple classifier models guide the design of using both machine learning and LLMs in retrieving and reasoning about information semantically.

V. SOCIOTECHNICAL ANALYSIS AND POSITIONING

5.1 TruthSetu within the SHAPE Framework

The SHAPE model presented by Veerasamy and Badenhorst (2026) provides five principles to direct users towards a more thoughtful and reality-based approach to information management: Stop (break the automatism of System 1 processing), Healthy Habits (establish critical examination practices), Analyse Source (confirm authenticity of sources), Probe Facts (employ fact-checking services), and Evaluate (base decisions on facts). TruthSetu may be regarded as a practical application of SHAPE principles through automation of activities which the latter requires from individual users.

The confirmation notice from the SCOUT agent, "We're verifying this..." functions as a Stop similar to that suggested by SHAPE, in that it informs the citizen that their story has been verified thus mitigating the risk of them sharing content before it can be confirmed. The use of the VERIFY agent to retrieve source information with trust weighting constitutes a proper implementation of the Analyse Source method since this relies on a hierarchy of sources that most individuals would find hard to implement themselves. The correction notice provided by the DEPLOY agent is an accurate representation of the Probe Facts and Evaluate methods of information analysis, and helps to provide citizens with the necessary evidence.

An important distinction from the approach taken by SHAPE in its use of an individualistic perspective lies in the action taken at the platform level by TruthSetu. SHAPE, according to Veerasamy and Badenhorst (2026), takes a very individualistic perspective on the issue; a deeper engagement that considers the platform architecture level of interventions has been neglected. TruthSetu, however, works at the platform level by stopping the WhatsApp forward from spreading through secondary sharing cascades. There is no way for user training in critical analysis to keep up with how fast messages travel on WhatsApp.

5.2 Crisis-Specific Misinformation Dynamics

Environmental crises create information conditions that differ qualitatively from routine misinformation environments. Three crisis-specific dynamics are particularly relevant to TruthSetu's design.

The first aspect that arises from the nature of crisis information is the time-critical decision-making period in which one must act. If a citizen were to receive news that a dam upstream had burst, they would have only seconds to make a decision about evacuation – either an expensive mistake, since there may be no need for evacuation, or a potentially life-threatening one if they had waited to check the information. The 3-10 seconds in which TruthSetu aims to return results is intended for this time constraint.

Furthermore, this is exacerbated by the fact that crisis disinformation is more likely to be aimed at people who do not have access to verification methods. Urban English-speaking people can verify information on government sites or English fact-checking platforms; however, rural people who receive information about the crisis in Hindi or Marathi are less fortunate because they have much fewer sources for verifying information.

Third, crisis misinformation's time-based nature calls for a time-sensitivity approach in constructing its knowledge base. News at flood level issued six hours ago can be outdated during an ongoing weather crisis. The use of freshness weighting by TruthSetu through EPHEMERAL documents having exponential decay starting from weight 1.0 after two hours and weight 0.1 after three days ensures that the information search process favors more recent and relevant data and does not exclude static data that is always valid irrespective of the course of events.

5.3 Limitations of Purely Technical Approaches

Human judgment is purposely designed within TruthSetu's boundaries. The VERIFY algorithm does not show the credibility score to citizens but simply

gives an English summary of what each source says, leaving the evaluation up to the citizens themselves. This design element is based on one of the key lessons learned in the context of misinformation research: over-reliant behavior is induced in cases where algorithms deliver highly confident decisions, even when they are correct.

In a similar manner, TruthSetu does not assert itself as providing the ultimate resolution of disputed claims. The UNVERIFIED and NO_INFO decisions clearly define the limits of confidence within the system and invite citizens to consult other sources, instead of viewing a lack of FALSE verdicts as an endorsement of the claim's truthfulness. Such epistemic modesty is essential in the realm of environmental crisis situations, which often come with their own inherent uncertainties.

The shortcomings of the machine learning detection component are also directly pertinent to TruthSetu's architecture. As pointed out by Veerasamy and Badenhorst (2026), algorithmic solutions encounter obstacles posed by new forms of misinformation attacks, lack of high-quality training data, and inherent inability to factor in the emotive and social aspects of the situation at hand. While TruthSetu's LLM verification procedure solves some of these problems through flexible reasoning about new phrasings and incorporation of different evidence types, it also suffers from its own shortcomings in terms of hallucination possibilities and restrictions on the rate at which tools can be used.

VI. LIMITATIONS AND FUTURE WORK

Current Limitations in TruthSetu's Framework

The current implementation of TruthSetu comes with certain limitations which need to be overcome before moving forward on the platform's development. Primarily, the knowledge corpus of the TruthSetu platform is restricted to eleven handpicked RSS feeds in English; regional language news channels, district-level government websites, and specialized RSS feeds for environmental monitoring are currently not included in the database.

Furthermore, the LLM translation pipeline, which works fine for Hindi and Marathi, has not yet been evaluated for its accuracy in translating crisis-related vocabulary. Some technical terminologies like cyclone strength classification, flood depth, and seismic magnitude scales may not be translated effectively via the generic Groq prompting system. The integration of IndicTrans2 (AI4Bharat), an advanced neural machine translation system dedicated to Indian languages, is intended to be the main improvement in translation accuracy.

Third, the present system was tested in the demonstration mode with the number of simultaneous users ranging between 1 to 50. For deployment in production mode on a national level for millions of citizens involved in a critical incident, there will be a need to switch from the present single asynchronous chain model to the Celery/Redis Distributed Task Queue for scalability among many worker nodes. The Meta Business WhatsApp API will be used in place of the Twilio Sandbox.

Fourthly, TruthSetu does not currently support multimodal verification. There is a large number of environmental crisis disinformation that comes in the form of image and video manipulation such as satellite images presented out of context, photos of floods that were taken in other locations or different time frames, and fake news reports imposed onto legitimate document formats.

The fifth issue with the LEARN agent lies in the lack of use made of its stored training data, which constitutes a database consisting of verified claims with their verdicts, sources, and embeddings. Using this data to train a classifier such as XGBoost or fine-tuned BERT would result in a substantial drop in LLM API calls for unambiguous claims.

Lastly, an objective assessment of TruthSetu's effectiveness needs to be carried out through user testing. The main factors that will be analyzed here are whether TruthSetu has any effect on the users' tendency to share the claim (i.e., do citizens receiving a FALSE judgment post the claim less often?), the level of trust citizens have in the platform (by demographic group), and whether the judgments made by VERIFY are accurate, based on expert fact-checkers' judgments on the dataset.

VII. CONCLUSION

TruthSetu has been described in this paper as a multi-agent real-time AI system designed for detecting crisis misinformation. The system has been embedded into the existing body of literature on fake

news detection using cognitive, social networks, and computational methods. In fact, TruthSetu uses a five-agent pipeline consisting of SCOUT, VERIFY, TRANSLATE, DEPLOY, and LEARN agents to implement the SHAPE sociotechnical framework by providing language-specific verified corrections to citizens in real-time.

The empirical evaluation of nine machine learning classifiers on a large-scale fake news dataset confirms that ensemble methods, particularly XGBoost with TF-IDF feature extraction, achieve near-ceiling accuracy (0.9967) on standard benchmark tasks. These results validate the use of supervised learning for fast claim pre-screening and inform the future development of the LEARN agent's local classifier. The documented failure of simpler classifiers — Naive Bayes and KNN — underscores the importance of architectural choices in fake news detection and motivates TruthSetu's hybrid computational-LLM approach.

Crisis misinformation detection represents a domain where the stakes of both false positives and false negatives are exceptionally high, and where the speed, linguistic diversity, and emotional intensity of information flow exceed the capacity of any purely technical or purely human response. TruthSetu demonstrates that a carefully engineered sociotechnical system — one that couples machine speed with source-grounded reasoning, multilingual accessibility, and continuous self-improvement — can operate meaningfully within these constraints. Effective crisis misinformation detection is not a solved problem, but it is an increasingly tractable one.

REFERENCES

[1] Aïmeur, E., Amri, S., Brassard, G., 2023. Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining 13.

[2] Al Ibraheemi, H.A.H., Jabardi, M., 2024. Detecting Fake News Using Machine Learning: A Comparative Study of Techniques. Journal of Kufa for Mathematics and Computer 11(2), 113–120.

[3] Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. Journal of Economic Perspectives.

[4] Beauvais, C., 2022. Fake news: Why do we believe it? Joint Bone Spine.

[5] Bouchaud, P., 2024. Skewed perspectives: examining the influence of engagement maximization on content diversity in social media feeds. Journal of Computational Social Science 7.

[6] Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., Qi, P., 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. Proceedings of the AAAI Conference on Artificial Intelligence.

[7] Kaliyar, R.K., Goswami, A., Narang, P., 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications 80.

[8] Kahnemann, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux.

[9] Lakzaei, B., Haghir Chehreghani, M., Bagheri, A., 2024. Disinformation detection using graph neural networks: a survey. Artificial Intelligence Review 57.

[10] Mayank, M., Sharma, S., Sharma, R., 2022. DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection. Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[11] Papageorgiou, E., Chronis, C., Varlamis, I., Himeur, Y., 2024. A Survey on the Use of Large Language Models (LLMs) in Fake News. Future Internet.

[12] Phan, H.T., Nguyen, N.T., Hwang, D., 2023. Fake news detection: A survey of graph neural network methods. Applied Soft Computing.

[13] Raza, S., Ding, C., 2022. Fake news detection based on news content and social contexts: a transformer-based approach. International Journal of Data Science and Analytics 13.

[14] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter 19, 22–36.

[15] Sivasankari, S., Vadivu, G., 2022. Tracing the fake news propagation path using social network analysis. Soft Computing 26.

[16] Song, C., Teng, Y., Zhu, Y., Wei, S., Wu, B., 2022. Dynamic graph neural network for fake news detection. Neurocomputing 505.

[17] Veerasamy, N., Badenhorst, D., 2026. The Informed Fake News Advisor (IFNA): Toward Sociotechnical Solutions for Fake News Detection. Proceedings of the 21st International Conference on Cyber Warfare and Security (ICCWS 2026), pp. 520–530.

[18] Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. Science 359.

[19] Zhang, X., Ghorbani, A.A., 2020. An overview of online fake news: Characterization, detection, and discussion. Information Processing and Management 57.

[20] Zhou, X., Zafarani, R., 2019. Network-based Fake News Detection. ACM SIGKDD Explorations Newsletter 21.