# Innovative Machine Learning Approach to Classifying Terrorist-Related Multilingual Communications

**1.G.Krishna Vasudeva Rao,**
**Dept. Of CS&SE, Andhra University, Visakhapatnam**
**2. Prof. P. V. G. D. Prasad reddy,**
**Professor, Dept. of CS&SE,**
**Vice-Chancellor,**
**Andhra University**
**3. Prof. M. James Stephen,**
**Ambedkar chair professor,**
**Andhra University**
**4. Prof. P. Srinivasa Rao,**
**Professor, Dept. of CS&SE,**
**Andhra University**

**Abstract:**
A proposed framework presents a fresh way to addressing global concerns such as terrorism, suspicious activity, and law and order infractions by using multilingual instant messaging services to track short-text messages. There is a considerable barrier to detection and prevention due to the fact that criminals, terrorists, and those involved in unlawful operations typically use other languages to discuss their plans. While multilingual content is commonplace on social media sites, the existing methods available are insufficient to effectively combat online crime. Criminals use a variety of languages to communicate with each other and coordinate their global operations. Prior research into messaging apps has mainly concentrated on detecting suspicious communications inside a single language, ignoring the complexity of predicting suspicious messages across many languages at once.
The suggested approach incorporates a multilingual strategy comprising several elements such as semantic web ontology, a database of suspicious actions with specified judgment rules, machine learning algorithms, and language translators informed by past learning experiences. This framework quickly determines the nature of the crime being discussed in microblogs when a user uses suspicious wording in a cross-lingual setting. Then, the cybercrime division receives timely updates on the criminals, reducing the workload of other security agencies. This cutting-edge system effectively identifies and prevents criminal behaviors through multilingual instant messaging services on a global scale, making it possible to oppose terrorism, suspicious crimes, and law and order breaches.

***Keywords:*** Artificial Intelligence (AI), Multilingual Machine Translation (MMT), Machine Learning (ML); Statistical Natural Language Processing (SNLP); Instant Messenger-based Social Networking; Association Rule Mining (ARM); Suspicious Communication Detection System; SCDs.
DOI Number: 10.48047/NQ.2022.20.3.NQ22975 NeuroQuantology2022;20(3): 1132-1150

## 1. Introduction

There has been a dramatic increase in the frequency and severity of cybercrimes, such as the recent terrorist strikes. The effectiveness of present detection technologies has been called into question [1]. In order to keep tabs on the many channels through which attackers communicate, a new framework has been developed that requires the creation of a multilingual message detection system. Digital communication sometimes involves multiple languages, which makes it more difficult to keep tabs on illicit activity. The framework suggests using multilingual translation techniques like statistical machine translation and the Multilingual Word Net to improve

existing instant messaging systems and better detect suspicious messages written in different languages.

Criminals may be able to discuss illicit acts and avoid detection by using a variety of communication methods [2]. These methods include instant messaging and social networking sites. These illegal acts have been connected to social networking platforms in several instances. Law enforcement agencies therefore require cutting-edge tools for criminal identification and tracking. The suggested system employs multilingual message detection and cutting-edge data recovery techniques to enhance the detection and tracking of digital crimes. Using this method, law enforcement may better anticipate criminal behavior and stop it before it does harm to the community.

The usage of many languages in digital communication is something the framework hopes to fix. Because attackers may utilize more than one language in their communications, e-crime departments may have trouble identifying potentially malicious emails[2]. The approach suggests employing multilingual translation methods like statistical machine translation to remedy this situation. Statistical machine translation makes use of a massive corpus of human-made translations between multiple languages. This method is useful for identifying suspicious texts written in a wide range of languages, even ones with lower frequency of use. The system utilizes statistical machine translation to recognize and translate messages in real time, enhancing the ability of law enforcement to monitor and prevent illegal activity.

The framework also suggests using the Multilingual Word Net as an additional method. The Multilingual Word Net is an extensive dictionary with information on terms from many different languages. This database was designed specifically for sifting through and analyzing IMs saved in text archives. To add to its usefulness, it can be used to classify words taken from unstructured text.

The framework can enhance existing IM systems by offering early detection based on suspicious conversation texting across several systems, all thanks to its ontology-based data recovery mechanism. By using this method, law enforcement organizations will have a better chance of identifying and following up on potentially criminal messages in real time.[3]

There is a pressing need for novel methods to detect and trace illicit activities in light of the proliferation of cybercrime. Using cutting-edge data recovery techniques and multilingual message identification, the suggested framework provides a workable solution. With the help of statistical machine translation and the Multilingual Word Net, law enforcement authorities will be able to monitor and respond to suspicious communications in real time. While the framework is still in its infancy, there is hope that it will help e-crime agencies become more efficient at identifying and avoiding cybercrime.

**2. Literature Survey**

This work investigates the role of automated text categorization (ATC) in knowledge organization and extraction in the Internet age. Word embedding and other rule-based approaches to document representation have become increasingly important thanks to developments in neural language processing. This systematic mapping study surveyed main works on word embedding in rule-based and machine learning approaches for automatic text classification to provide a thorough grasp of the topic. Research is conducted in several fields, including as the social sciences, e-commerce cataloging, digital libraries, and spam detection. This study adds to the growing body of knowledge about TC methods by finding and analyzing key articles in the field of text classification.

It is the goal of this Framework to eliminate the impediment to network connectivity and cyber security posed by untraceable terror and suspicious messages delivered via IM and SNS. We build a surveillance system using OBIE and ARM to detect and anticipate these communications, together with the nature of the cyber threat activity and information about the perpetrator. Decision-making is aided by predefined knowledge-based criteria based on historical experiences with suspicious datasets like GTD, facilitating the rapid eradication of cybercrime.

[6]. In-depth information about NERC (Named Entity Recognition and Classification) is provided in this academic article. Entity extraction and classification from free-form text is what this talks about, so that means people, places, and numbers. Trends in

activities, languages, text types, and entity types are highlighted in this survey that covers the years 1991-2006. The necessity of machine learning and feature selection for efficient NERC systems is emphasized throughout the paper as it delves into algorithmic methodologies, proposed features, and evaluation methods in the field. Researchers and professionals alike can use it to have a deeper understanding of NERC developments.

This research [7] suggests a text-processing strategy and a CNN 2D visualization approach that may be used with datasets in a number of languages, including Korean and English. With an average accuracy of 0.9957, the results show that the CNN 2D image-based detection model performs better than string-based models like RNN, LSTM, and CNN 1D. This demonstrates the efficacy of image-based processing for string data and the potential of multilingual processing using the CNN 2D model.

[8]Images containing text in multiple languages are being used by spammers and criminals to spread their malicious messages via email. Multiple language support has been added to a new framework for filtering suspicious messages sent in text-based and image-text formats. Information about prospective cybercrime and perpetrators is categorized by the system. The experimental results show that the suggested method is more accurate and generates less false positives than the state-of-the-art spam filters.

The issue of digital advertising funding misinformation and propaganda on the internet is investigated in this study [9]. It presents a machine learning approach to spot and label sites spreading false information in an effort to protect the public. Using multilingual text embeddings, the model predicts the likelihood of dangerous content and produces a list of publications for human assessment. By deploying this approach, internal teams can proactively blacklist harmful content, protecting the advertising provider's name in the process.

False information spread via the internet must be challenged because of its potential to affect many facets of society. In this investigation, a vaccination-related false news dataset was used to evaluate several automated fake news detection algorithms. According to the findings, a convolutional neural network (CNN) model performed well in distinguishing between bogus and reliable news, while a gradient-boosting decision tree technique using feature stacking performed better in identifying satire. A strategy for combining embeddings and redundant data was also developed in the study to enhance satire detection.[10].

According to research (Al-zoubi & Faris, 2021) This research looks into the potentially dangerous phenomenon of spam profiles on Twitter and suggests using freely available language-independent elements to enhance spam detection. Five popular classification algorithms and five filter-based feature selection approaches were tested on four datasets containing text in four distinct languages (Arabic, English, Korean, and Spanish). Taking into account essential traits across multiple languages led to increased classification performance and a deeper knowledge of social spam, both of which led to enhanced detection approaches.

"(Younis and Younis, 2015)"In this study, we explore how people are using social media—and Twitter in particular—to share their thoughts on various topics. Using text mining and sentiment analysis, it provides a free method for gathering and analyzing feedback from customers. The study shows the significance of sentiment analysis for organizations in understanding customer perspectives, informing marketing strategies, and decision-making procedures, and centers on user evaluations about Tesco and Asda stores during the Christmas period of 2014 in the UK.

Table 1: Strategies in online extremism research: Feature Engineering and classification techniques

| Year | Strategies | | Feature Extraction / Reduction Methods | Features Selection | Techniques |
|---|---|---|---|---|---|
| 2020 | | ML | | | |
| 2020 | | DL | | NA | |
| 2020 | | ML | | NA | |

| | | | | | |
|---|---|---|---|---|---|
| 2020 | | ML | TF-IDF | | |
| 2019 | | ML | TF-IDF | NA | |
| 2019 | | DL | | NA | |
| 2019 | | | NA | | |
| 2019 | | NA | NA | | NA |
| 2019 | | NA | NA | | |
| 2019 | | NA | NA | | NA |
| 2019 | | | | | |
| 2019 | | NA | NA | | NA |
| 2019 | | ML | | NA | -- |
| 2019 | | ML | | | |
| 2019 | | ML | | Features, Hate Features | -- |
| 2019 | | | BoW | | |
| 2018 | | DL | NA | NA | |
| 2018 | | ML | NA | | |
| 2018 | | NA | NA | | NA |
| 2018 | | ML | NA | | |
| 2018 | | NA | NA | Topic | |
| 2018 | | ML | NA | | |
| 2017 | | ML | NA | | |
| 2017 | | ML | NA | | |
| 2017 | | | NA | | |
| 2016 | | ML | NA | | |
| 2016 | | | NA | | |
| 2016 | | ML | NA | | |
| 2015 | | ML | | | |

| | | | | | |
|---|---|---|---|---|---|
| 2015 | | ML | NA | | |
| 2015 | | | NA | NA | |

**3. Language Detection**

A common method for identifying a text's language is to check it against a database of words in every known language. This approach, however, requires a large database and may have trouble with inflections and compound words. The Naive Bayes and N-gram algorithms have been presented as solutions to these problems (Lui & Baldwin, 2012). Using these approaches, one can create a language detection library in Java by calculating probabilities based on spelling features. The language probabilities for any given input text are generated by this library, which uses a training collection to build the language profiles. This method overcomes the shortcomings of dictionary-based methods to produce a language detection technique that is both efficient and accurate.[11], [12].

**3.1 Language Detection Mechanism:**

Step 1: involves sorting documents by language (e.g., English, French, Chinese, Japanese, etc.).

Step 2: Calculate the probabilities for each class and implement those changes in the backend.

Step 3: when the greatest normalized probability is more than 0.99999, the detection procedure is aborted prematurely to improve performance. [13].

**3.2 Working with N-Gram**

The Unicode code point of a particular N-Gram frame is used in the N-Gram method, resulting in a smaller comparison size than the actual word size. Different languages have different alphabets and spelling conventions; for example, the accent "é" is used in Spanish and Italian but not in English. This feature can be exploited to estimate the language with a high degree of accuracy, reaching around 90%, similar to how the letter "Z" is commonly used in German but not in English, whereas the letters "C" and "Th" are used frequently in English but rarely in German. It's possible, nevertheless, that classical Chinese, Arabic, Farsi, Persian, and Japanese have less stringent requirements for precision than other languages. A noise filter is used to improve precision, and the K-means closest algorithm is used to get the best possible results. as reported by (Lui & Baldwin, 2012).

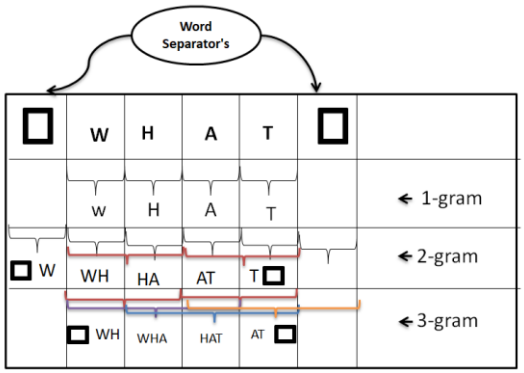



**Figure 1:** N-Gram, 1, 2, 3 Gram

Table 1: Language Maximum Probability

| | C | L | Z | TH |
|---|---|---|---|---|
| English Language | 0.751 | 0.471 | 0.021 | 0.741 |
| Germany Language | 0.101 | 0.371 | 0.531 | 0.03 |
| French Language | 0.381 | 0.691 | 0.01 | 0.01 |

Once the target language has been determined, the LSTM RNN technique is applied to the detected English sequence to convert it into a vector. In the end, the decoding procedure produces the output. This method makes use of LSTM RNN to effectively process and generate relevant results by capturing and analyzing the sequential character of English text. [14] [13].

**3.3 Algorithm for Multilingual Detection and Translation:**

Flowchart-style multilingual language detector and translation algorithm, "Algorithm1," is also known as "Language Identification Mechanism and Translation Algorithm" (L.D.T.A.). Here are the steps of the algorithm:

Step 1: The first step is to start the algorithm.

Step 2: The second step is to use the "detect language" operation of the Language Detection Algorithm (Lda).

Step 3: Input composed of many languages is provided at Step 3.

Step 4: Using posterior probability to determine language classification.

Step 5: If the discovered language's probability value is 1 or greater, If the language cannot be determined, "Unknown" should be displayed.

Step 6: The sixth step is to print the identified language's label.

Step 7: Beginning with step 9, translate into the language that was discovered.

Step 8: Converting the source text into a vector representation is the step. The Lstm Rnn encoder is used for this.

Step 9: Analysis of the attention mechanism's vector value yields a 1.

Step 10: The Lstm Rnn decoder is used in Step 10 to change the vector into the desired text format.

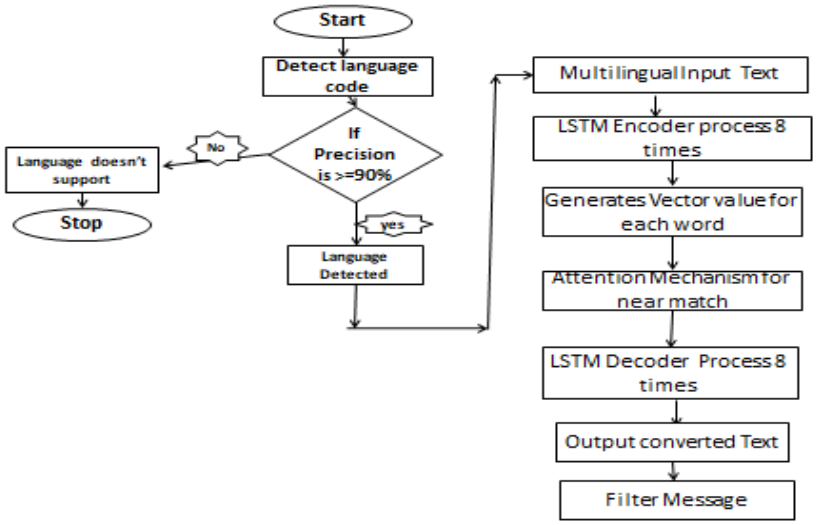
Step 11: If in Step 10 you didn't locate a matching vector value of 1, then in Step 11 you'll use the attention technique to zero in on the closest vector value up to 0.7.

Step 12: Convert the vector to the desired text by using the Lstm Rnn decoder.

Step 13: Perform the process described in Steps 1 through 5 eight times to improve syntactic and semantic understanding.

Step 14: Show the output text that was generated

**Flow Chart 1:** Displays recognition and multilingual translation




**3.4. Mechanism for LSTM Encoding and Decoding**

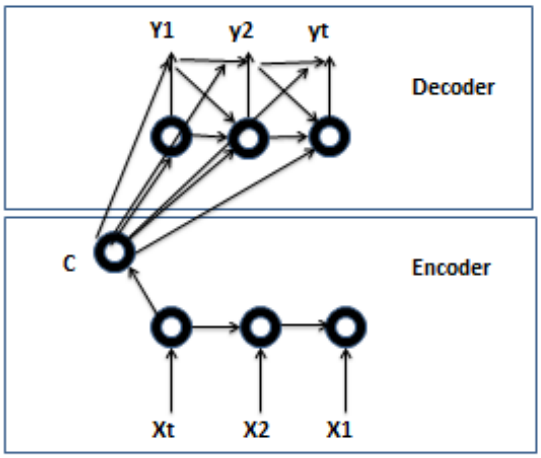
**RNN Encoding and Decoding Mechanism:**

The encoding and decoding operations of a Long Short-Term Memory (LSTM) RNN (Recurrent Neural Network) are informed by previous outputs. Input sequences of fixed length X=x1,x2,...xt are fed into the RNN, together with a hidden state h and an output Z at each step time t. The RNN's hidden state ht is set up as ht=f(ht-1xt), where f might be an element or an LSTM unit. The RNN framework takes an encoder of variable length, reads it, and then uses a decoder of the same length. This model learns the conditional distribution of the up variable-length sequence by using a training sequence of varying length. Cho et al. (2014).

Take the phrase p(y1,..., yT' | x1,..., xT), where t and T' are the lengths of the input and output

sequences, respectively. As the encoder in an RNN continuously scans the input sequence x, the hidden state is modified by the aforementioned equation Eq-1. The sequence concludes with the allocation of a flag as a final state in the covert mode. In the decoder, a second RNN learns to predict the next symbol yt given the hidden state ht from the random C complete input sequence. Furthermore, yt and ht> rely on yt-1 C, where C is a summary of the input sequence of the state hidden from the decoder at time t, and ht=f(ht-1,yt-1,c). The conditional distribution for the following symbol is P(yt|yt-1,yt-2.....y1,c)=g(ht,yt-1,c), where g should yield accurate probabilities, as in the case of softmax. The conditional log-likelihood is improved by training both the encoder and the decoder at the same time. Cho et al. (2014).

The algorithm seeks to optimise itself by increasing the conditional log-likelihood of the sequence given its input as much as possible. Together, the encoder and decoder are trained to achieve the best possible results for the objective function, where is a set of model parameters, x_n is an input sequence, y_n is an output sequence, and is the maximum of the logarithms of the distances between the two sequences (max 1/n, n=1). Since the learning set is used in a gradient-based method to estimate the model parameters, the output decoder is distinct from the input. The following figure depicts the encoder and decoder's architecture for transforming a sequence into a fixed-length vector. The research was published recently (Zennaki et al., 2019). [15][16].




Where **x** is input sequence **C** is hidden state summary **y** is output sequence

**Figure 2:** Architecture for RNN encoders and decoders

### 3.5 Mechanism for Attention

Sherstinsky (2020) claims that while translating from English to Telugu, the model prioritises specific English words, such as "you" when translating the Telugu word "మీరు" and "knife" when translating the Telugu word "కత్తి".. By analysing several English sentences and their Telugu equivalents, the model is taught to zero down on certain English terms. To do this, a third component known as an attention mechanism is used in between the encoder and the decoder. The English text is fed into the encoder, where it is converted into a numerical vector. Word by word, the decoder pays special attention to the words specified by the attention mechanism, translating them into Telugu. The attention mechanism then determines which English words have corresponding Telugu words. This technique outperforms the first encoder/decoder implementation significantly. [17].

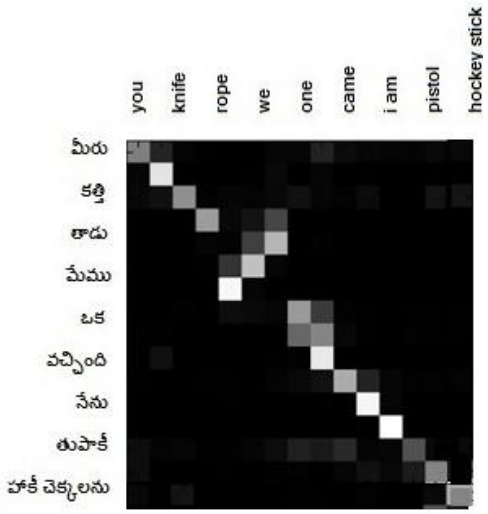



**Figure 3:** Attention mechanisms

### 3.6 Encoder and Decoder using Attention Mechanism

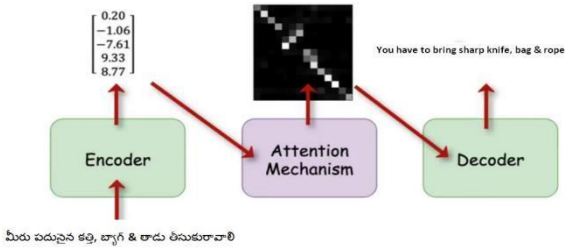



**Figure 4: Attention-based Encoder and Decoder Mechanism**

### 3.6.1 The Encoder And Decoder Architecture

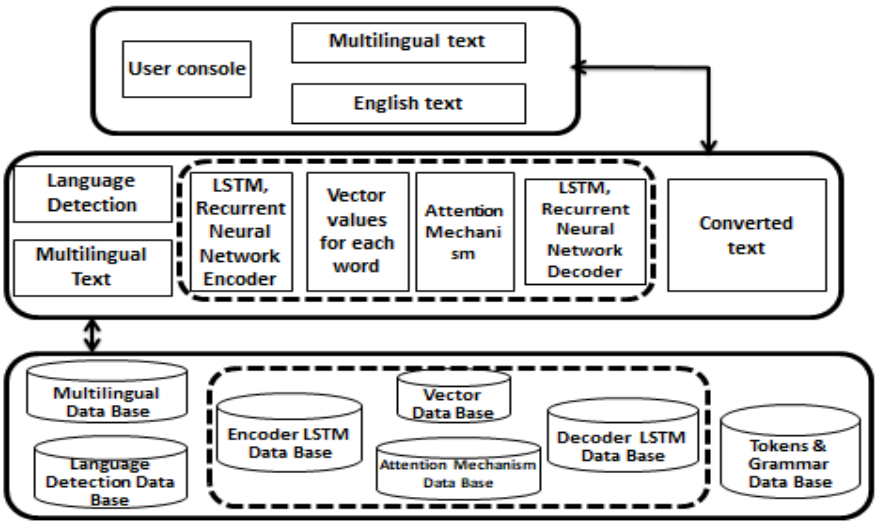



**Figure 5:** Architecture for Multilingual Encoder and Decoder

### 3.7 The overall scenario of the multilingual translator

Machine translation processes are now more closely aligned with the original translation processes, which improves the accuracy of sentence translation. This is accomplished by using eight LSTMs rather of just one in the encoding and decoding processes. The AI model benefits from the deeper network's enhanced comprehension of linguistic context and syntax. In order to use this final network for Hindi to English translation, the encoder receives the Hindi text word by word and transforms it into a series of numerical representations of words called "word vectors." Then, the Hindi word that corresponds to the English word is generated with the use of an attention mechanism that prioritises certain English terms. The decoder receives this information and then produces

the English translation of the Hindi sentence word by word. The efficiency of this method greatly exceeds that of the first encoder/decoder design. This has been demonstrated by multiple groups (Johnson et al., 2017; Yu et al., 2020).

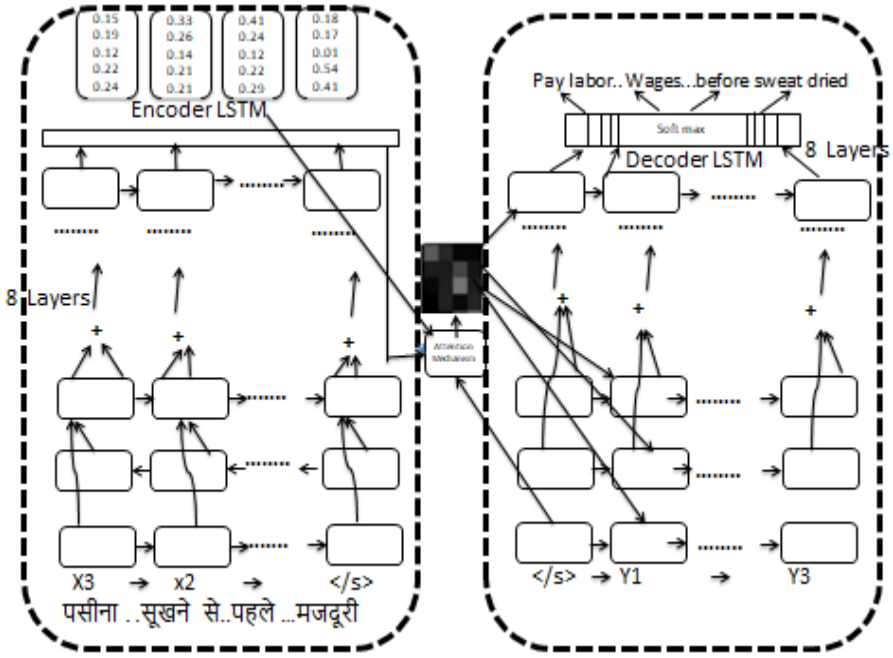



**Figure 6:**Detailed architecture for machine translation

**4. Suspicious Word Detection Framework Architecture Proposed for Multilingual Text**

Studying useful time intervals is central to the suggested architecture depicted in Figure 1. Sharing texts between users triggers the Suspicious Pattern Detection (SPD) algorithm. The communication is translated into English and then kept in a database for later examination if it is written in a language other than English. The SPD algorithm was developed to spot potentially harmful communications. Figure 7 provides a comprehensive breakdown of the SPD method and its related impacts. Information about cybercriminals can be tracked through the E-crimes division's monitoring system.

In contrast to ARM, the suggested architecture makes use of databases that can contain more than one language in order to track evolving multilingual communications and ontology-based information extraction to spot cross-language red flags. It has been shown that this paradigm can be put to good use in a variety of contexts. (Guo et al., 2020), (Leung, 2019), and (Mohammed Mahmood Ali & Rajamani, 2013) are all examples of scholarly works on the topic. There are three primary components to the framework: Word recognition from jumbled text, monitoring for cybercrime, and a non-standard method of computing SPD are all discussed. Algorithmic and graphical representations of the framework's pseudo-code functionality are presented

.

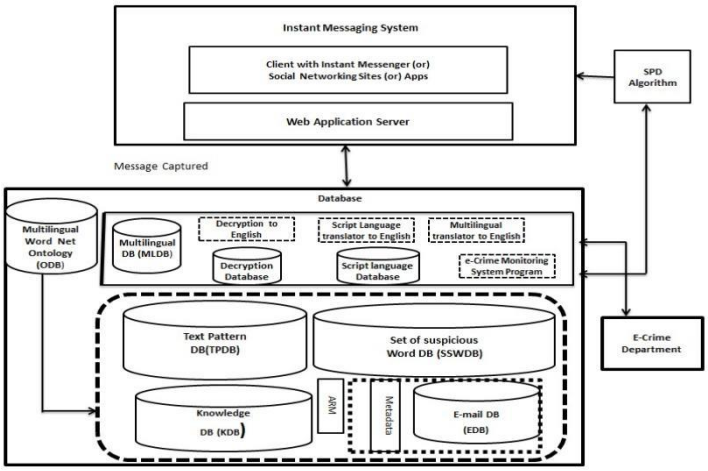
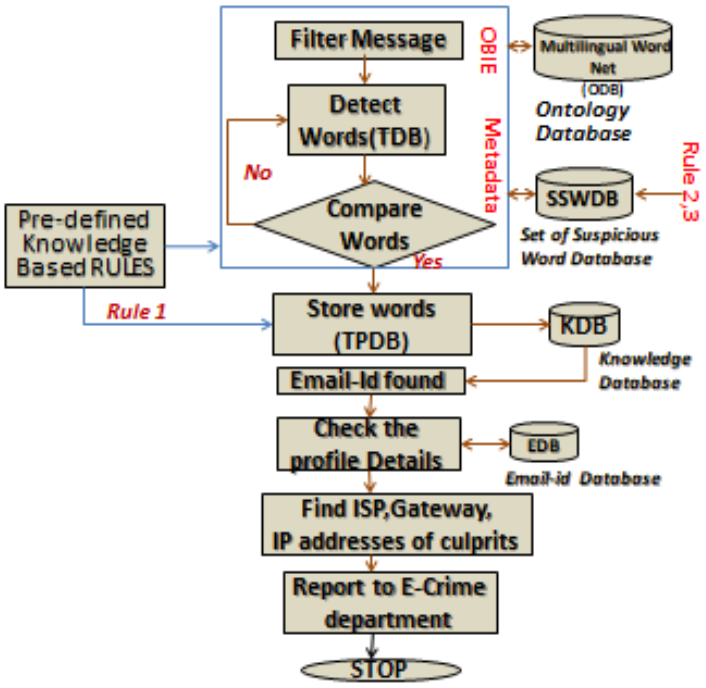


**Figure 7:** Language detection, multilingual translation, and suspicious message detection are all part of the system's framework.




Flow Chart 2: Suspicious message detection

**5.1 The steps in the proposed algorithm**

**Flow Chart 2:**

1. Detection of Languages and The LDTA method is initially applied to the input message in order to determine the language. If language detection fails, the procedure ends at Section 2.Second, if the language is recognized, it is translated into English before being submitted through the screening process.
2. If any words from the list of suspicious words (SSWDB) provided in Rule 1 of Table I are found in the filtered message, the message is flagged as suspicious. To aid in tracing the origin and destination email addresses of suspicious phrases, the KDB keeps a record of detected stem words with domain and activity data.
3. The EDB stores data provided when an email address is created, including a phone number, name, birth date, and gender. This information can be retrieved with the help of the Relational Wrapper Algorithm.
4. The suspect's Internet service provider (ISP) and IP address are used by the R2D Wrapper application to locate their email account. The software use an algorithm to produce the report.
5. The information in the final report, which includes the criminal's IP addresses, ISPs, and email addresses, allows the crime department to take action in compliance with the applicable laws. There have been numerous studies on this topic (Mohammed Mahmood Ali & Rajamani, 2013; Rajamani et al., n.d.).

In order to determine what domain suspicious terms belong to, Ontology-based Information Extraction (OBIE) is an essential tool. Ontology databases (ODB), knowledge databases (KDB), sets of suspicious words (SSWDB), metadata, transaction processing databases (TPDB), and transactional databases (TDB) are only some of the databases utilized to get the job done. While ODB is a multinational Word Net database used to find synonyms and associations between words, TDB is a record of extracted stem words after questionable ones have been removed. Using ODB, we compare and contrast the terms used in the TPDB with those used in the SSWDB, which was compiled by a domain expert and is shown in Table 2. If a suspicious phrase is discovered, the Suspicious Pattern Detection (SPD) algorithm is activated for cybercrime monitoring (as shown in Flow Chart 2). Email data is stored in EDB, along with other user data including username, father's name, education, employment, and contact details. Metadata is a crucial part of the system

since it records details like the databases used, who received and sent what, and when. akin to a record of past events, which many IMS systems keep. Table 2 contains the pre-defined rules that OBIE can use to do a full analysis of the data set obtained from the FBI and CBI investigations of settled cases and the GTD global terrorist database (Access the GTD | GTD, n.d.)(Mohammed Mahmood Ali & Rajamani, 2013).

**5.2 The proposed process for identifying multilingual questionable texts includes a table.**

| **(Pre-defined) knowledge-based rules Rule 1** | |
|---|---|
| **Category of threat** | **Examples of detected words (Stem words)** |
| **(Domain)** | **Factors** |
| Murder → | homicide, slaughter, butchery, manslaughter, axing, elimination, termination, removal, destruction, eradication, wiping out; synonyms include Supari, an assassin armed with a 6mm handgun, two hockey clubs, and a pair of gloves. |
| Kidnap → | Synonyms, nab, remove, entangle, imprison, shanghai, lure away, hold captive, abduct, snatch, spirit away, carry off, run off with, nab, remove, abduct, snatch, seize, capture, hijack, kidnapping, take hostage, hold for ransom. |
| Terror Outbreak → | Alarm, dismay, consternation, anxiety, fear, apprehension, panic, confusion, unrest, commotion, upheaval, turbulence, agitation, disturbance, disorder, upheaval, crisis, havoc, mayhem, pandemonium |
| Trafficking & Drug supplier → | synonyms include: smuggling, peddling, distribution, dealing, running, trade, contraband, transportation, illicit trade, drug trafficking, drug dealing, drug running, drug distribution, drug smuggling, drug trade, narcotic supply, peddling, transporting, dealing, running, and distributing drugs; dealing; running; distributing; smuggling; and peddling |
| Fraud → | Words like "deception,""scam,""hoax,""cheat,""dupe,""dishonest,""cheating," and "Fake Paper" come to mind. |
| Bribery → | Subornation, inducement, payola, hush money, backhander, sweetener, buy off, grease, bung, sop, palm-greasing, graft, unlawful gratuity, under-the-table payment, filthy money, favoritism, and nepotism are all forms of corruption. |
| Extortion → | Extraction, pressure, ransom, force, compulsion, demand, tyranny, coercion, blackmail, intimidation, shakedown, Kidnapped for a ransom, with a sharp knife, rope, and a bag in tow. moving, sugar, breaking the law |
| Sexual abuse → | Sexually-motivated aggression, rape, molestation, harassment, violation, misconduct, exploitation, perversion, indecent exposure, assault, harassment, violence, exploitation, perversion, indecent exposure, lewdness, obscene behavior, and indecent exposure. |
| **Rule II:** This rule may apply to several different fields, and it requires checking the user-specified initiation value for each stem word. The TPDB is queried by the OBIE using rules 1 and 2 and a precision formula. Knowledge database (KDB) and ontology editor (SSWDB) are the next repositories to receive the data. | |
| **Rule III:** Unrecognized words can be found and fixed with the use of an ontology taxonomy construction that looks for the nearest stem word. The OBIE system uses this rule to access the TPDB for data analysis and retrieval. The generated information is subsequently uploaded to the KDB. | |

**Table 2 Proposed Framework For Detecting Suspicious Messages in Multiple Languages.**

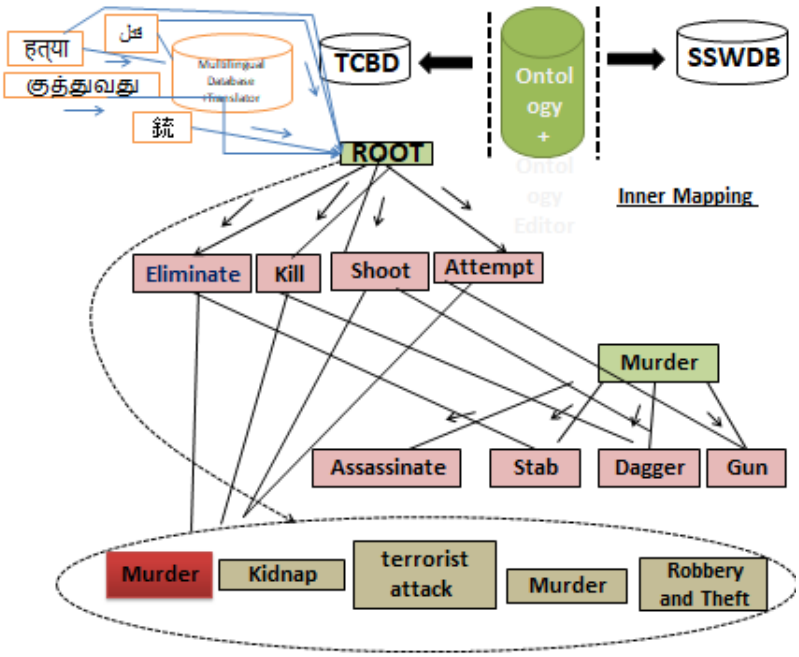



**Figure 8:** Multilingual Suspicious Messages Detection Internal Mapping

**5.3 Complete System Proposed Algorithm**

1. You'll need to sign up for an IM service, connect with users, manage your database, and save your chats.
2. Call the LDT.A program that can translate foreign words into English.
3. SDP algorithm(s) should be called[5].
4. Clean up the SSWDB by removing superfluous keywords, flagging suspicious words, and sending them on to the TPDB to be mapped with the help of previously defined knowledge-based criteria (pages 8-9).
5. Have the user talk into TDB.
6. Utilize OBIE to map questionable words from TPDB to SSWDB after scanning the conversation for them.
7. Map whole stem words to a blank root using OBIE (TPDB) and the tree alignment technique.
8. Determine the severity of the threat by comparing the TPDB and SSWDB stem word threshold values (Rules 1, 2, and 3).
9. Send the stems from TDB to TPDB after scanning it.
10. Stem words with domains are sent to the KDB if the SSWDB, TPDB, and TPDB already installed on your machine are equivalent.
11. If both the TPDB and the KDB are the same, then the domain is suspicious since it contains a suspicious phrase.
12. Check KDB with the R2D wrapper algorithm and the E-crime database to determine the type of domain word.
13. If KDB is correct, then there is a questionable word match.
14. Confirm EDB, which stores details such the user's name, email, IP address, and geographic location.
15. Report the user's details to the E-Crime Division.
16. The output is a report containing data on users and potentially suspicious conversations.

This article analyzes the various forms of communication employed by attackers in order to propose a paradigm for stopping attacks before they occur. Language detection, translation of multilingual conversation to English, filtering of words, monitoring criminals using metadata, and evaluation of suspicious terms with predefined rules and multilingual wordnet synonyms are just a few of the steps in this process, which begins with the initial conversation between criminals.

**6. Groundbreaking Observation**

**Table 3:** Illustrates the output obtained through Extortion

| The realm of identified threat. | Subscriber – 1 | Subscriber - 2 |
|---|---|---|
| Extortion<br><br>Extortion | I need kidnap son of minister<br><br>నాకుకట్టడానికిపదునైనకత్తి, బ్యాగ్&తాడుఇవ్వండి"(give me sharp knife, bag & rope to tie )<br><br>"நீங்கள்பலகுழந்தைகளைகடத்தியுள்ளீர்கள்என்றுநம்புகிறேன்"( hope you kidnapped many childern )<br><br>"आपइसमौकेकोकभीनचूकेंअपनेसाथधारदारहथियारलेकरचलें"(you should never miss this chance take sharp weapons along with you )<br><br>15 lakhs are advance and 15 on delivery of child | इसजबरनवसूलीमामलेमेंमैंआपकीकैसेमददकरसकताहूं(how can i help you in this extortion case )<br><br>"main aapakee madad karoonga kyonki mere paas achchha anubhav hai "(I will help you as i have good experience)<br><br>"मुझेकार्ययोजनादो"(give me action plan )<br><br>"30 లక్షలవిమోచనమొత్తాన్నిచెల్లించండి"<br>(pay ransom amount of 30 lakhs )<br><br>Done , give me location |
| | **Output** | |
| **The presence of suspicious language in multilingual communication has been detected.** | పదునైనకత్తి, బ్యాగ్&తాడు,குழந்தைகளை ,விமோచன ,तेजहथियार. | |
| **The multilingual suspicious are highlighted with green for Telugu, yellow for Hindi, violet for** | The multilingual word mentioned above converted into English terms respectively.<br>(Sharp knife, rope, bag, kidnaped, ransom, sharp weapon ) | |

| | |
|---|---|
| **Tamil, and brown for English.** | |
| Undetected words | जबरनवसूली,अपराधीस्थानभेजें(Extortion,criminal,send location ) |
| **suspicious term in the English language** | kidnap, delivery of child |
| **precision is dividing the sum of multilingual and English words translated by the number of pre-defined words in a specific domain.** | 6 = Multilinguistic word<br>2 = English Word<br>6+2=8<br>8/21 =38.09 |

**Table:4** Illustrates the output obtained through Fraud

| | | |
|---|---|---|
| The realm of identified threat. | Subscriber – 1 | Subscriber – 2 |
| Fraud | "मुझे 200 एकड़जमीनकेनकलीकागजातबनानेहैं"(I need to make fake papers of 200 Acers land)<br><br>"నేనుభూమికాగితాలనుతప్పుగాచూపించినాబందువులు మరియుస్నేహితులనుమోసంచేస్తాను"I will dupe my relatives and friends by misrepresentation land papers<br><br>200 Acers worth will be around 100 crores<br><br>Very good idea many people are greedy | "मैंएकव्यक्तिकीव्यवस्थाकरूंगावह डुप्लीकेटपेपरबनाएगा"( I will arrange one person he will make duplicate paper)<br><br>"people will get easily mislead by us"<br><br>Will advertise as misstatement 80% less offer<br><br>We will be looting people, run away to London |
| | **Output** | |

| | |
|---|---|
| **The presence of suspicious language in multilingual communication has been detected.** | नकलीकागजात,కాగితాన్నితప్పుగా, మోసం, |
| **The multilingual suspicious messages are recognized and distinguished through color coding, with underlined green for Telugu, yellow for Hindi, and brown for English.** | The multilingual word mentioned above converted into English terms respectively.<br>( Fake Paper, misrepresentation, dupe) |
| **Undetected words** | डुप्लीकेटपेपर ( duplicate paper) land grabbers, location |
| **suspicious term in the English language** | Mislead, Misstatement, looting |
| **precision is dividing the sum of multilingual and English words translated by the number of pre-defined words in a specific domain.** | 3 = Multilinguistic word<br>3= English word<br>3+3=6<br>6/14=42.85 |

**Table: 5** Illustrates the output obtained through Murder

| The realm of identified threat. | Subscriber – 1 | Subscriber – 2 |
|---|---|---|
| Murder | "మంత్రినిహత్యచేయడానికినాదగ్గరవిమోచనక్రయధనంఉంది) "మొదటివేతనం 20 లక్షలుమరియుహత్యజరిగినతర్వాత 20 లక్షలు"(I have a ransom to assassinate a minister) " 20 lakhs first pay and 20 lakh after the assassination<br><br>وزيركا مقام بهيجناور تاريخدينا( sending location of minister and giving date )<br><br>" you should carry pistol of 6mm, sharp weapon and 2 hockey sticks , gloves you should never miss thischance | I have killed many ministers<br><br>"我最擅长计划谋杀"( I am best at Plan murder)<br>"నేనుశిక్షణపొందుతున్నాను “అదిఖచ్చితంగాఉంటుందిచింతించకండి"(I am doing training it will be perfect don't worry ) |

| | |
|---|---|
| **Output** | |
| **The presence of suspicious language in multilingual communication has been detected.** | హత్య , విమోచన ,谋杀 الموقع في تاريخ |
| **The multilingual texts that are considered suspicious are identified through the use of different colored underlines. Telugu is marked with green, Arabic with orange, Chinese with blue, and English with brown.** | The multilingual word mentioned above converted into English terms respectively.<br>as assassinate, ransom, murder, location, and date, |
| **Undetected words** | Money, స్థానాన్నిపంపండి (send Location) |
| **suspicious term in the English language** | (Assassination, the killedpistol of 6mm, sharp weaponhockey sticks, gloves) |
| **precision is dividing the sum of multilingual and English words translated by the number of pre-** | 4 = Multilinguistic word<br>6 = English word<br>4+6=10 |

| | |
|---|---|
| **defined words in a specific domain.** | 10/18=55.55 |

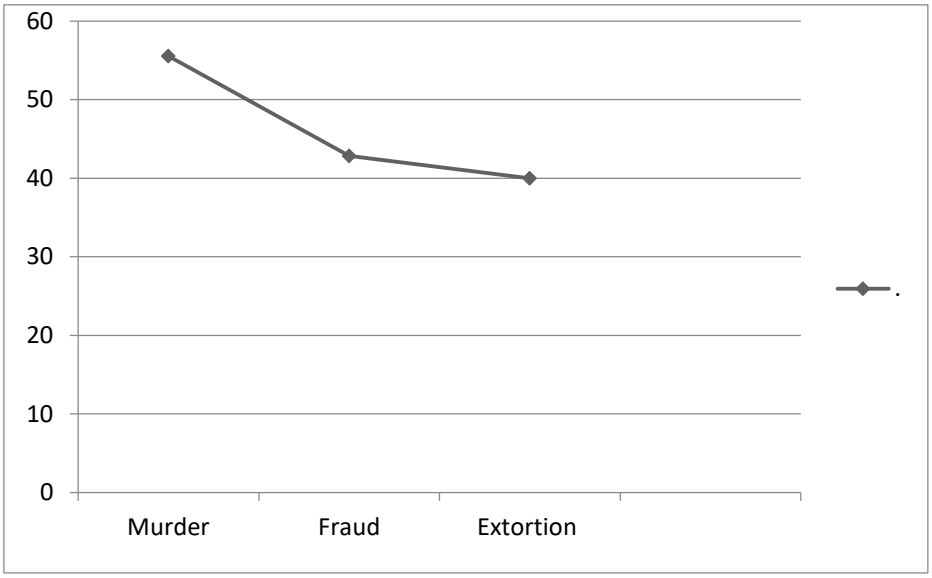



**Figure 9:**a comparison of three domains

The domain associated with murder was found to have the greatest precision value (55.5) after being compared to the fraud and extortion domains. Therefore, it was connected to the database of potentially malicious terms for future investigation.

**7. Experimental Results**

We used two metrics—precision and recall—to measure the quality of the extracted suspicious words and assess the efficacy of our suggested system.

$$\text{Precision} = \frac{\text{Extracted words accurately}}{\text{Total number of accuretly extracted words}}$$

$$\text{Recall} = \frac{\text{Extracted words accurately}}{\text{The maximum possible number of words}}$$

The Global Terrorism Database (GTD), which includes information on terrorists and criminals around the world up until 2018, provided us with a database for the "murder" domain. This 92 MB database has 191,465 rows and is the most current one available. You can get its codebooks by clicking the link below. We can learn a great deal about terrorist attacks and criminal actions by poring over this database. We compared our system to others and provided a table showing the results. Visit https://www.start.umd.edu/gtd/downloads/Codebook.pdf to get the GTD and the codebook.

| **Term** | **Proposed system output** |
|---|---|
| Total number of accuretly extracted words | 1880 |
| Extracted words accurately | 1796 |
| The maximum possible number of words | 1896 |
| **Precision** | **95.25%** |
| **Recall** | **94.72%** |

**Table: 6**Results obtained through the analysis of data provided by domain experts in GTD.

Domain experts participated in group brainstorming sessions using GTD to build the dataset.

**8. Observation**

Some Arabic, Persian, old Chinese, and Japanese characters are difficult for the naive eye to recognise because to biases and noise in the approach. Noise filters and character normalisation methods, however, can help overcome these restrictions. It is not uncommon for words to be translated into another language with the same meaning but a different spelling. This can cause inaccuracies in attention-based word translation. Using

previously established guidelines is one solution to this problem. When a multilingual term cannot be recognised, Rule 3 is used. In addition, many concepts, such as "geography" and "money," are utilised in many distinct contexts. To solve this problem, the GHSL method is applied to obtain a minimal threshold value, and a specialist user interface is required. The system's ability to recognise and translate items written in several languages has been improved with the use of these improvements and methods.

**Table:7** Analysis of Comparisons with the Former System

| Platform | Cutting- Edge Machine Learning Solution for Identifying Multilingual Terrorism-related Communications (Proposed system) | Framework for surveillance of instant messages [18][19] |
|---|---|---|
| Linguistics | Multilingual detection capability | Lacks multilingual detection capability |
| Encoding & Decoding Module | Highly precise | Absence of a system for such detection |
| Translation capability across languages | Comprising 49 distinct languages | Nonexistent |
| Attention Mechanism | Establishing relationships between corresponding languages | Unavailability of such mechanism |
| Number of supported languages in total | Involving 49 language variations | Limited to a single language |
| Lexical Repository | Multilingual WordNet resource | Restricted to WordNet alone |
| Precision score for a corresponding examination (Murder) | Achieving a precision rate of 55.55% | Achieving a precision rate of 31.5% |
| Precision scores for corresponding examinations (Extortion) | Attaining a precision rate of 38.09% | Attaining a precision rate of 19.0% |
| Precision scores for corresponding examinations (Fraud) | Securing a precision rate of 42.85% | Securing a precision rate of 25.0% |
| Accuracy | 64% improvement in accuracy compared to the previous system | Lower level of accuracy |
| Global Terrorism Database (GTD) | Containing 191,465 rows with a file size of 90 MB | Comprising 59,787 rows with a file size of 30 MB |

## 9. Conclusion

This research presents a revolutionary approach for identifying suspicious signals written in a variety of languages, with the goal of combating the worldwide problem of criminal activities such as terrorism, extortion, fraud, and murder. Our suggested technology fills a gap in the market for detecting potentially malicious texts written in many languages. Though our algorithm can identify up to 49 different languages, increasing accuracy would require better multilingual translation and decreasing the recall rate. This requires improvements to our system so that it can identify any and all possible languages. The suspicious words are translated into English and then checked against our own custom database and the Ground Truth Database (GTD). If the message contains any words from the Suspicious Word Database (SSWD), the

system checks the sender's identity and notifies the appropriate authorities immediately.

Future plans include expanding the system's language support and enhancing the accuracy of the existing multilingual translation. In addition, incorporating sophisticated machine learning and NLP algorithms into our system can significantly improve its performance. To provide a more all-encompassing response to illegal actions, our technology can be expanded to analyse other forms of communication, such as phone and video communications. By identifying potentially fraudulent information across many languages, the suggested method offers a practical answer to the worldwide problem of criminal activity. The system's accuracy, linguistic capacity, and integration of cutting-edge algorithms all have room for improvement. Our method has the potential to significantly contribute to the international fight against crime.

**References**

[1] "IC3 Releases 2020 Internet Crime Report — FBI." https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-the-internet-crime-complaint-center-2020-internet-crime-report-including-covid-19-scam-statistics (accessed Dec. 04, 2021).

[2] "FBI — All About FBI Linguists." https://archives.fbi.gov/archives/news/stories/2008/july/linguists_072908 (accessed May 11, 2023).

[3] Vineet Pande, Viraj Samant, and Sindhu Nair, "Crime Detection using Data Mining," *Int. J. Eng. Res.*, vol. V5, no. 01, pp. 891–896, 2016, doi: 10.17577/ijertv5is010610.

[4] A. M. Aubaid and A. Mishra, "Text classification using word embedding in Rule-based methodologies: A systematic mapping," *TEM J.*, vol. 7, no. 4, pp. 902–914, 2018, doi: 10.18421/TEM74-31.

[5] M. M. Ali, K. M. Mohammed, and L. Rajamani, "Framework for surveillance of instant messages in instant messengers and social neworking sites using data mining and ontology," *IEEE TechSym 2014 - 2014 IEEE Students' Technol. Symp.*, no. February 2014, pp. 297–302, 2014, doi: 10.1109/TechSym.2014.6808064.

[6] K. Shaalan, "A Survey of Arabic Named Entity Recognition and Classification," *Comput. Linguist.*, vol. 40, no. 2, pp. 469–510, 2014, doi: 10.1162/COLI_a_00178.

[7] H. Lee, S. Jeong, S. Cho, and E. Choi, "Visualization Technology and Deep-Learning for Multilingual Spam Message Detection," *Electron.*, vol. 12, no. 3, 2023, doi: 10.3390/electronics12030582.

[8] B. Hills and S. Arabia, "Framework for Surveillance of Multilingual Contents From Emails Problem Statement & Related Work," vol. 1, no. 2, pp. 7–12, 2013.

[9] N. El Manouzi, M. Chen, and S. Ivanov, "Multilingual Disinformation Detection for Digital Advertising," 2022.

[10] G. Guibon *et al.*, "Multilingual Fake News Detection with Satire To cite this version : HAL Id : halshs-02391141," 2019.

[11] M. Adeel, "Soundex Algorithm," 2010.

[12] T. Jauhiainen, M. Zampieri, and T. Baldwin, "Automatic Language Identification in Texts : A Survey Automatic Language Identification in Texts : A Survey," no. April, 2018, doi: 10.1613/jair.1.11675.

[13] M. Lui and T. Baldwin, "langid.py: An Off-the-shelf Language Identification Tool," *Aclweb.Org*, no. July, pp. 25–30, 2012, [Online]. Available: http://www.aclweb.org/anthology-new/P/P12/P12-3005.pdf.

[14] T. Dunning, "Statistical Identi cation of Language Ted Dunning New Mexico State University," no. January 1996, 1994.

[15] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.

[16] Z. Yu, Z. Yu, J. Guo, Y. Huang, and Y. Wen, "Efficient Low-Resource Neural Machine Translation with," vol. 19, no. 3, pp. 1–13, 2020.

[17] M. Johnson *et al.*, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 339–351, 2017, doi: 10.1162/tacl_a_00065.

[18] M. M. Ali, K. M. Mohammed, and L. Rajamani, "Framework for surveillance

of instant messages in instant messengers and social neworking sites using data mining and ontology," *IEEE TechSym 2014 - 2014 IEEE Students' Technol. Symp.*, pp. 297–302, 2014, doi: 10.1109/TechSym.2014.6808064.

[19] M. M. Ali and L. Rajamani, "Framework for surveillance of instant messages," *Int. J. Internet Technol. Secur. Trans.*, vol. 5, no. 1, pp. 18–41, 2013, doi: 10.1504/IJITST.2013.058292.