Advanced Machine Learning Approach for Detection of Multilinguistic Terror Message to save human Lives

Syed Hussain[1], Dr. Pakkir Mohideen S[2]

[1]Research Scholar, Computer Science and Engineering, B.S Abdur Rahman Crescent institute of science & Technology, Hyderabad, Telangana, India.

[2]Asso.Professor/Computer Application, B.S Abdur Rahman Crescent institute of science & Technology, Hyderabad, Telangana, India.

DOI: 10.47750/pnr.2023.14.02.310

Abstract

Solutions to avoid terrorist attacks, suspicion crimes, and misbehaving of law & order globally. With the use of multilingual through instant messaging applications via short-text messages are traced using the proposed framework. Criminals, terrorists, underworld dons are sitting in one place and implanting their criminal plans globally using different languages. To date, there are no stringent solutions were proposed for mitigating online crimes in Social networking sites, where multilingual words are used while chatting with other users. To date, no proper solution to stop the crimes that are happening through multilingual. Criminals use more than one language to pass activity messages among teammates who may be in living in any corner of the World. Earlier works in messaging applications were based on the prediction of suspicious messages for a unique language at a time (i.e. either English language or china) ignored multiple languages at the same time. The proposed framework developed using a multilingual framework comprising of various components namely Semantic web Ontology, Suspicious Database assisted with Pre-defined decision rules, machine learning technique, language translator guided with past learning experiences. When a user communicates the suspicious terminology using multilingual language this framework expedites in predicting the type of crime from microblogs before it gets executed by criminals. Details of criminals will be alerted to cybercrime the department that reduces the tension for the various security departments.

Keywords: Artificial Intelligence, Multilingual machine translation, Machine Learning; Statistical natural language processing (SNLP); Social networking with instant messengers; Association rule mining (ARM); Suspicious message detection system; SMDs.

1. INTRODUCTION

The Recent terrorist attacks carried on Sri Lankan church, New Zealand Masjid, Indian Cities, and other criminal doings need to be noticed. Past terror activities detecting agencies all over the world have failed to detect such attacks earlier. In this article, the authors seek to propose a framework that would help in the execution of attacks aforetime through the various communication methods among the attackers. We need to develop a multilingual message detection system because the attacker may use multilingual messages while communicating. Our Framework easily detects the accused before any commitment of crime and detected by the crime agencies which would help the society from crime.

Web advancement prompted (evoke) the development of never-ending digital offenses. Culprits deliver suspect text messages by using cellular phones, Instant Messengers, and Social Networking Destinations, which can be very rigid to follow their illegal practices powerfully. We see many cases were registered relating to crime using social networking. The E-crime Agency should be adlibbed with the advancement in innovation to look for culprits. Most of the instant messaging systems (IMS) are created by specifying their very own limit for sending messages, video, and sound conferencing.

Many cases were registered using different languages while communicating at the season of assaults, but E-crime Departments are not all around prepared to recognize the multilingual suspicious messages detection. Digital offense exercises are expanding worldwide. The CIA, FBI, and other government organizations are effectively gathering digital data related to crime, remote knowledge data to anticipate future digital assaults. As of late, the Internet Crime Complaint Center (IC3) released the report in 2012 about digital offenses, with the most recent information furthermore, patterns of online criminal movement. We reviewed the literature on various structures of Mobile Phones, Instant errand people, and Social Systems administration locales using multilingual for a crime. These examinations helped us to figure out a new Framework (Shrestha & Spezzano, 2019).

Multilingual Translation frequently uses a technique termed statistical machine translation. The database is full of trillion of the multilingual translation done by humans. For example, in a book that was written in Arabic translated to English and pattern, the trend is inserted into the machine, an algorithm is designed to find it out. An example of translation is checked as human translated the books the best quality translation is considered as more human translated record quality may vary with

different languages.

Multilingual Word Net is a lexical database that contains tremendous calculates of data comprising of words. In this database, 155288 words classify into 117000 Synsets with a total of 207000 word-sense stands. This database is used for examination for filtering and scrutinizing the instant messages put away in TDB (Text Database). But, multilingual Word Net is utilized even as highlights as characterization of words from unstructured content. Thus, the data extraction method is built upon Multilingual Word Net Ontology.

Our commitment incorporates improving the current IMS utilizing Ontology-based data recovery methods (probability models). This method is managed by pre-defined Knowledge-based standards and ARM. Early recognition based on suspicious conversation texting from frameworks (cell Phone, Instant messenger, and social networking sites) is conceivable. Our framework will detect the E-Crime and track criminal details, IM is lacking such kind of facility doesn't have.

Following English! Hindi pair of sentences: (Johnson et al., 2017)

Thumaamah ibn Athaal killed many Muslim on the battlefield get caught, he begs for his life please don't kill me take ransom how much money you want I will give you,

जुमा मा इब्न अठाल ने युद्ध के मैदान में कई मुस्लिमों को मार डाला, वह अपने जीवन के लिए भीख माँगता है, कृपया मुझे मत मारो फिरौती ले लो कि आप कितना पैसा चाहते हैं मैं आपको दे दूंगा,

Prophet Muhammad (peace and blessings of Allah be upon him) released him just in 3 days with no consideration later he accepted Islam because of kindness to a prisoner.

पैगंबर मुहम्मद (अल्लाह के शांति और आशीर्वाद) ने उन्हें केवल तीन दिनों में रिहा कर दिया, बिना किसी विचार के बाद उन्होंने कैदी के साथ दया के कारण इस्लाम स्वीकार कर लिया।

Mixed Hindi and Tamil: போர் எதிர்ப்பாளருடன் பார்த்ததில்லை विनम्रता युद्ध विरोधी के साथ कभी नहीं देखा गया (politeness with war opponent never seen)(Johnson et al., 2017)

2. Literature Survey (2.0.- 2.7)

In the present situation of the world, we unable to find conversation without instant messenger as the client is dependent on. Plenty of messages are sent and received by Social networking sites and instant messengers like Whatsapp, Facebook, Skype, Zoom, Google talk, Twitter, etc. This is the new trend of communication with a colleague, personal, and business. When constrained with work areas, prevalent texting frameworks are searching toward direction into handled gadgets and cell phones, enabling clients to visit basically from any place. The social way to deal with recognize malignant data on the web for IM with security heretic is constrained toward distinguishing malevolent Uniform resource locator links. Previously, historic conversation have been utilized for detecting culprit but results are misleading correspondingly, engineering for identification of phishing assault from IM for instant message utilizing information mining system was proposed the majority of this strategy is wasteful to anticipate the sort of suspicious digital risk exercises like homicide, psychological oppressor assault, coordinate fixing, tranquilize, carrying, capture, burglary and robbery, debasement accusations and lewd behavior completely (Mohd Mahmood Ali & Rajamani, 2012). The communication that occurs in multilingual cannot be detected by earlier work like కత్తి, కాడు, ஏற்கனவே, तेज हथियार, يوليو منتصف , పదునైన, मारणे, قتلWe proposed framework to detect suspicious words from more than 49 different languages. (Rajamani et al., n.d.), (Fujs et al., 2019)

2.1 Language Detection

A simple method to detect language is matching with the dictionary of all languages, but this method requires a huge database dictionary and we may face problems like inflections. Compound words building corpus is so expensive to require more knowledge of script and language to avoid these issues using of Naïve Bayes and N-gram algorithm (Lui & Baldwin, 2012). With this algorithm, we can calculate probabilities of spelling from features with this language detection library is generated in java which makes language profiles from training collection. It creates the possibilities of every in all language .give back with their probabilities for input text. (Adeel, 2010), (Jauhiainen et al., 2018).

2.1.1 Language Detection Mechanism

Step 1: Classify documents of languages like English, French, Chines, Japanese, etc.

Step 2: Updating back end feasibilities by factor possibilities in every category

- $p(C_k|X)^{(m+1)} \propto p(C_k|X)^{(m)} \cdot p(X_i|C_k)$
  - where C_k:category, X:document, X_i:feature of document

Step 3: terminate detection process if maximum probability (normalized) is 0.99999, early termination for the performance (Lui & Baldwin, 2012).

2.2 How it works with N-Gram

The exact frame is “Unicode’s code point of N-Gram”. The result of a comparison is very less than the size of the word.

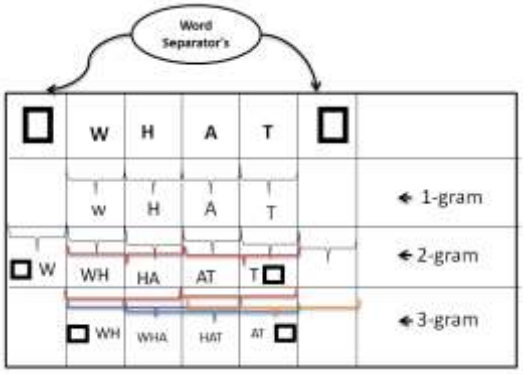


Figure 1: N-Gram, 1, 2, 3 Gram

Every language has its own character and its rule in spelling such as “ é ”. It is utilized in Spanish Language and Italian language etc. This rule cannot be applied in the English Language. Similar to that “Z” is mostly used in the German language as starting of a word, and very little usage in English. With a similar example, the letter “C” and “Th” is more common in English and rarely used in German. With the help of this feature, it is easy to guess language with maximum probability. Above algorithm detect with 90% accuracy except for traditional chines, Arabic, Farsi, Persian, and Japanese, these language show low precision value to improve this noise filter is utilized, K means nearest algorithm is used to gain maximum probability..(Lui & Baldwin, 2012)

	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

Table 1: Maximum probability table for 3 languages

After the detection of language, the English sequence is converted into the vector by Lstm Rnn Mechanism, and output is generated by decode (Dunning, 1994) (Lui & Baldwin, 2012).

2.3 Multilingual Detector and Translation Algorithm:

Algorithm 1: Language Detection (L.D.T.A) followed by Flow Chart1:

Step 1: Initialize

Step 2: Lda. ("action"="detect language") //Language detection algorithm

Step 3: Input Text. ("text"=" Multlingual text.")

Step 4: Detect language ("language"="classify with posterior probability ")

Step 5: if probability value ("probability ">=1)

Step 6: Language Detected ("text"="Name of the Language..")

Step 7: if probability value ("probability "=0)

Step 8: OutPut text. Language not Detected (“text”:=”Unknown”)

// Detected language will be process For language Translation which starts from step 9.

Step 9: Lstm encoder (Source text to vector values) // LSTM RNN encoder

Step 10: If (vector value = 1 match found) // Attention mechanism

Step 11: Lstm decoder (vector to target text) // LSTM RNN decoder

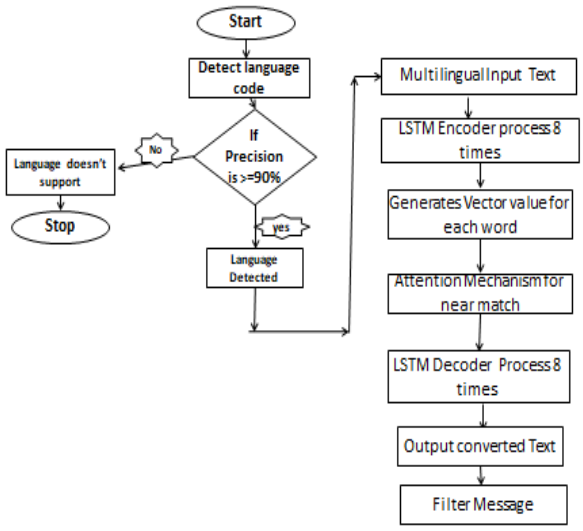
Step 12: If else (vector value # 1 find nearest value upto 0.7) // Attention mechanism

Step 13: Lstm decoder (vector to target text) // LSTM RNN decoder

Step 14: Repeat step 1 to 5 , eight times

//for better understating of syntax and semantics

Step 15: show output text.



Flow Chart 1: shows detection and multilingual translation

2.4. LSTM Encoding and Decoding Mechanism

Rnn Mechanism for Encoding and decoding:

The recurrent neural network name itself denotes repeating using feedback of previous output to process new input. It consists of input variable sequence with fixed length X=x1,x2...xt is hidden state h and output Z at every step time t the hidden state h<t> RNN is configured as h<t>=f(h<t-1>xt). Where f denotes nonlinear activation function it can be simple as an element or complex as an LSTM unit. Rnn architecture read encoder variable length and decode same variable-length encoder and decoder length are same, this model learning conditional distribution up variable-length sequence conditioned with different length sequence. (Cho et al., 2014)

Eg $p(y1,..,yT'|x1,...xT)$ Eq-1

where t and T' is input and output sequence length it may differ encoder in rnn read input sequence x continuously accordingly hidden state changes by above equation Eq-1. At the end of the sequence flag assigned as an end in the hidden state

arbitrary C complete input sequence utilizes for other Rnn at decoder side which makes to learn and to generate output sequence by predicting the next symbol y<t>given the hidden state h<t> both y<t>and h<t>are also conditioned on y<t-1> C as a summary from the input sequence of state hidden of the decoder at time t is defined as

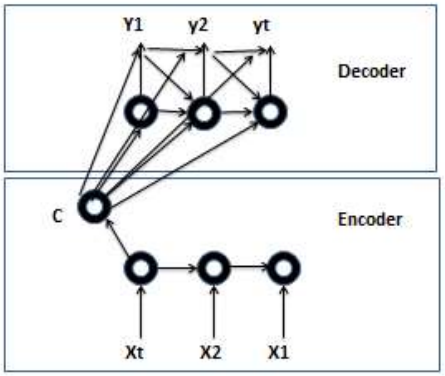
h<t>=f(h<t-1>,y<t-1>,c) condition distribution for the next symbol is

$$P(yt|yt-1, yt-2 \ldots\ldots y1, c) = g(h<t>, y<t-1>, c)$$

For activation function f,g should outcome with valid probabilities eg softmaxEncoder and decoder collectively learned to Magnify conditional log-likelihood(Cho et al., 2014)

$$max_\theta \frac{1}{n}\sum_{n=1}^{n}(\log \ p\theta(y_n|x_n)$$

Where θ is a set of model parameters and x_n is a sequence of input y_n is a sequence of output.Joining learned set use of a gradient-based algorithm to estimate model parameter that is why output decoder is different from the input (Zennaki et al., 2019). It generates input sequence to output fixed-length vector architecture of encoder and decoder is given below (Cho et al., 2014)(Yu et al., 2020).



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

Figure 2: RNN encoder-decoder architecture

2.5 Attention Mechanism

More attention is given to that English word while entering the Telugu word "మీరు" it looks like the only word it would consult is the English word you" (Sherstinsky, 2020). The "you" the same goes for oats the model learns to focus its attention only on the English word "knife" while generating the Telugu word కత్తి. In this way, the model looks at thousands of other English sentences and their corresponding Telugu translations. It learns which English words to focus its attention on while producing the words of the Telugu translation this alignment is learned by an extra unit called an attention mechanism. It sits between the encoder and decoder so during translation an English sentence is fed to the encoder it's encoded into some vector which is just numbers the computer understands it is the same English sentence in the computer's eyes, then we use an attention mechanism asking which Telugu word will be generated by which English words. The decoder will then generate the Telugu translation one word at a time concentrating its attention on the words determined by the attention mechanism so that is sweet this performs better than the original encoder/decoder architecture (Johnson et al., 2017).

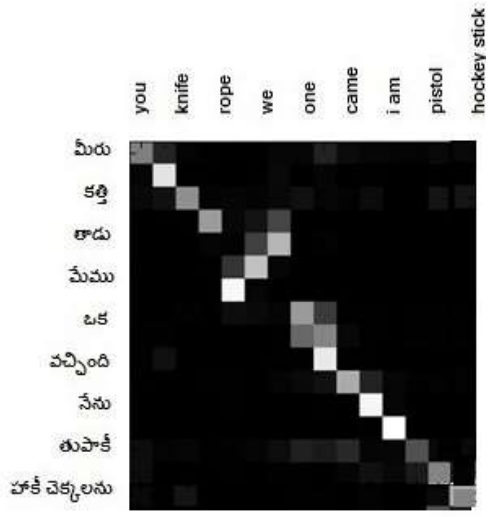


Figure 3: Attention mechanisms

2.6 Encoder and Decoder using attention Mechanism

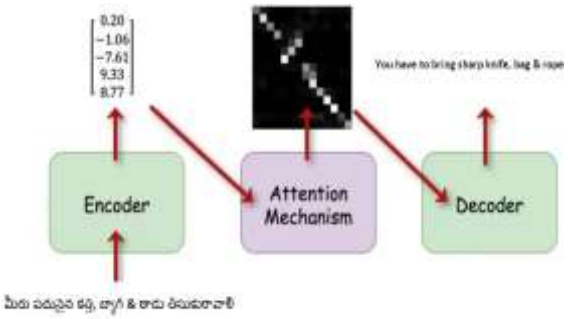


Figure 4: Encoder and Decoder using attention Mechanism

2.6.1 The Architecture Of Encoder And Decoder

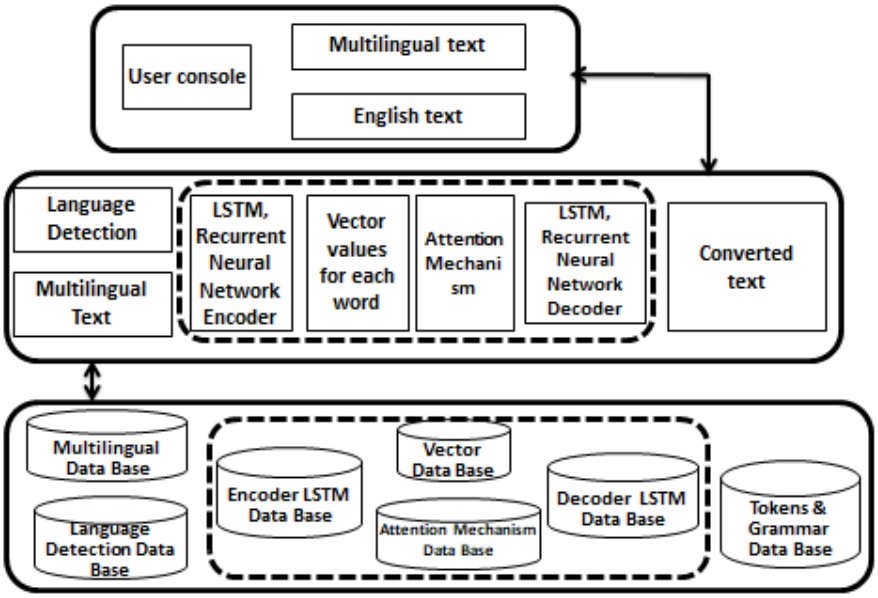


Figure 5: Multilingual Encoder and Decoder Architecture

2.7 The overall scenario of the multilingual translator

The sentence translation is now more narrowly aligned with the original machine translates. AI works exactly like this the only difference is everything is scaled up by this mean instead of using one LS TM for the encoder and decoder we use 8. We do this because deeper networks help better model complex problems so this network is more capable of understanding the semantics of language and grammar just to recap on the final network. You want to translate Hindi to English you pass the Hindi text or word by word to the encoder and it converts these words into several word vectors that are the numbers demonstrating these words these are just numbers that represent the words themselves of the sentence these words are then just passed into an attention mechanism(Johnson et al., 2017). This determines the English words to focus on while generating some Hindi word this data is passed to the decoder which generates the translated Hindi sentence one word at a time to English (Yu et al., 2020), (Johnson et al., 2017)

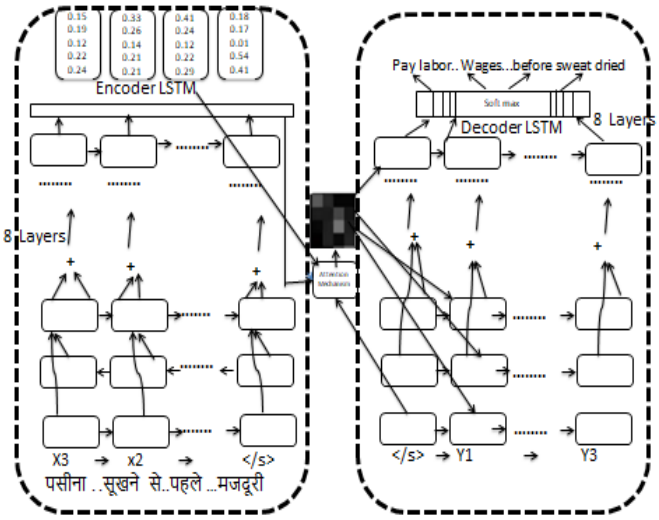


Figure 6: Complete Machine Translation Architecture

3. Proposed Framework Architecture for Multilingual Suspicious Word Detection

In this section, our work on investigating the practical periods on the Framework proposed as appeared in Figure 1. Suspicious Detection of Pattern (SPD) calculation starts with means into catch texts that abide imparted among the clients if the message is multilingual it translates it to the English language and storing in a data basses because distinguishing suspicious message. The SPD calculation will appear in Figure 7, aside from additional effects the E-crimes checking framework programs would follow the guilty party subtleties for the E-crimes departments.

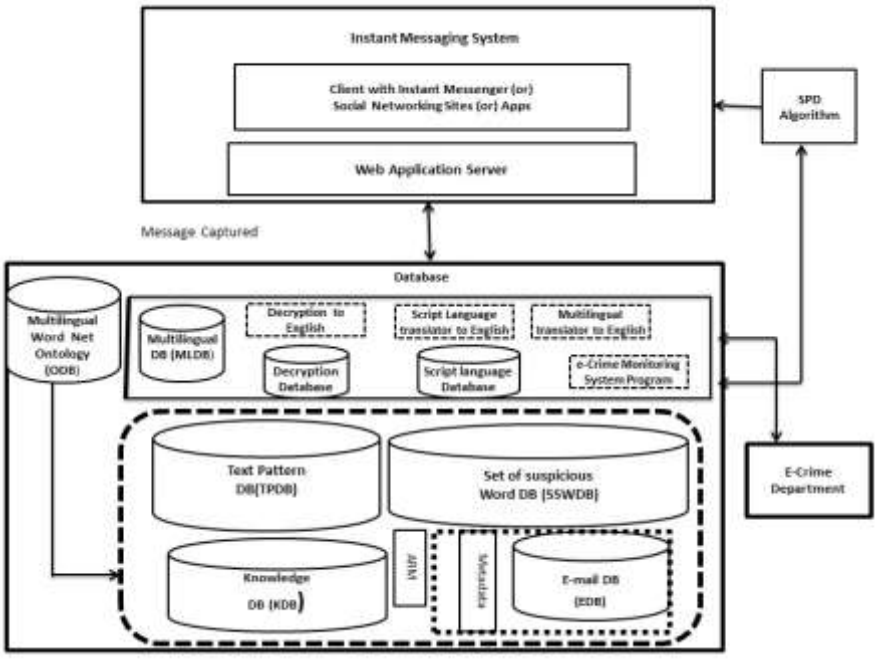


Figure 7: The architecture of language detection, Multilingual translation with Suspicious Messages Detection.

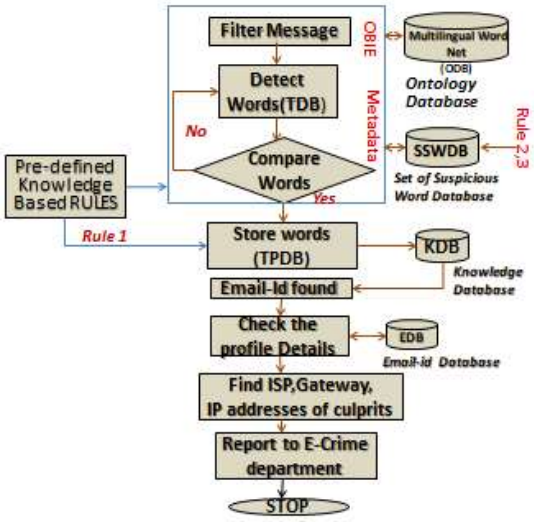
The proposed framework was prepared by utilizing multilingual databases for storing the progressive multilingual message. The extraction of information was done based on ontology for identifying multilingual suspicious words between messages as pre-defined expert system principles compare with ARM. this is real errands performances(Mohammed Mahmood Ali & Rajamani, 2013) (Mohammed Mahmood Ali et al., 2014), (Guo et al., 2020), (Leung, 2019).

1) Extraction of word from disordered content 2) E-crime observing framework 3) SPD calculation (a nonexclusive methodology). pseudo-code working of our Framework Is exhibited utilizing algorithm cum Schematic portrayals is displayed.

3.1 Steps involved in the proposed algorithm

Flow Chart 2: language detection and translation

1) As received input from a sender, it detects the languages by using LDTA algorithm if not detected process stops Sec 2.2 (Lui & Baldwin, 2012), detected languages then translate into English (Johnson et al., 2017) and forwarded to filter the message.



Flow Chart 2: Suspicious pattern detection

In the second step, unwanted words are removed by using filters from the messages in this process we will identify suspicious words with the help of algorithm 2. The resultant detected words are stored in TPDB for the next process (Mohammed Mahmood Ali & Rajamani, 2013)(Rajamani et al., n.d.).

2) If suspicious words are found, the message is considered to be suspicious, as given in Table I set of suspicious words (SSWDB) in rule 1. KDB has a record of detected stem words with domain and which kind of activity been performed and metadata help to track the sender and receiver email id of suspicious word belongs(Mohammed Mahmood Ali & Rajamani, 2013) (Rajamani et al., n.d.).

3) While creating an email id the details like contact number name age, sex, and all details are stored in EDB can be accessed with the help of the Relational Wrapper Algorithm(Mohammed Mahmood Ali & Rajamani, 2013) (Rajamani et al., n.d.).

4) The suspicious email id account holder is traced by IP-Address, ISP location by a program (R2D Wrapper) which generates a report by using an algorithm(Mohammed Mahmood Ali & Rajamani, 2013) (Rajamani et al., n.d.).

5) The final report consists of details of criminals with IP addresses, ISP, Email address details according to report crime department can take action under their corresponding act(Mohammed Mahmood Ali & Rajamani, 2013),(Rajamani et al., n.d.)

The OBIE assumes a critical job that predicts and maps the space (subject) to which these suspicious words have a place. For above process we use (,ODB, , KDB, SSWDB Metadata, TPDB and TDB). In this method, TDB (Transactional database)is utilized for the progressive message between the sender and receiver Ontology Database (ODB) is a lingual database that helps to identify synonyms of words and relationships between words as axioms with concept hierarchy(Mohammed Mahmood Ali et al., 2014)

Transactional database (TDB) keeps data about extracted stem words removing the non-suspicious words and OBIE uses a database of multilingual Word Net named (ODB) when it is necessary. The set of the suspicious word database (SSWDB) consists of a database created by a domain expert shown in table 2 comparison of this word with TPDB words under the guidance of ODB. The ODB was utilized two times to avoid ambiguity of each word between SSWDB and TPDB if the word is found suspicious then SPD (suspicious pattern detection algorithm is initiated to e-crime monitoring which appeared in flow chart 2. The EDB (E-mail database), keeps up the email subtleties of the clients that demonstrate the username, father's name, consider subtleties, work and area subtleties, telephone number, and other significant data. (Mohammed Mahmood Ali & Rajamani, 2013) (Rajamani et al., n.d.)(Jadhao & Agrawal, 2016), (Mohd Mahmood Ali & Rajamani, 2012), (Mohammed Mahmood Ali et al., 2014).

Metadata is the fundamental part, that keeps up data of all databases utilized, clients data to whom the message has a location and other important data relating to Framework (time, date, recipients and senders subtleties, and so on.). Much the same as a log of history kept up by the vast majority of IMS. The pre-characterized guidelines of Table 2, explicitly rule 1 is given to OBIE, utilizing acquainted principles are surrounded cautiously by breaking down meeting to generate new ideas of a constant

set of data is collected from FBI and CBI examinations of settled cases and GTD global terrorism database (Access the GTD | GTD, n.d.)(Mohammed Mahmood Ali & Rajamani, 2013),

3.2 Proposed Framework Detection Table For Multilingual Suspicious Messages Detection

Table:2 Proposed Framework For Multilingual Suspicious Messages Detection.

knowledge-based rules (pre-defined) 1 Rule	
Category of threat	**Example detected words (Stem words)**
(Domain)	**context**
Murder →	slaughter, strike, kill, wipe out, firearm knife, blade, cut, area, cash, threat, fear death, bullet, Supari, assassinate, plan, murder, a pistol of 6mm, sharp weapon.
Kidnap →	Commandeer, catch, seize, snatch, usurp, get, firearm, take hostage, area, sum, kill, land, plot
Terror Outbreak →	granite, human bomb, area, attack, ak47, machine gun, hostage, tear gas, blast timer, fake sim, sack, holy place, workstation, obliterate, installment, money
Trafficking & Drug supplier →	RDX, heroin, charas, narcotic, money, M.tabs, Methoquoline, opium, charas, area, injection, Morphine, LSD STR/ECA, dibucaine
Fraud →	Fake Paper, dupe, misrepresentation, mislead peoples, misstatement, looting people Scam, fake advertisement, fake certificate, Area, money, official corruption, virgin girl, wines bank corrupt, installment
Bribery →	The new complex, cash, account number, check, store, jewel, remove duty, PC, Swiss account.
Extortion →	a sharp knife, bag, rope, extortion, kidnap, ransom amount, sharp weapons, delivery of the child, send location. Expensive Jewelry, Location, secrete account, weapon, blade, area, night planning, Master key.
Sexual abuse →	Harassment calls, abusive messages, wonderful, installment, Night offer, area, offering amount, parking, vehicle gift, nude photos, blackmail, flat number, midnight, sexy dress.
In 2 RULE (initiation value)	
Checking the user-defined initiation value for every stem word which may exist in many domains with help of precision formula in OBIE (Mohammed Mahmood Ali et al., 2014), rule 1,2 is applied for (TPDB) for information retrieval. After that, it sends it to the ontology editor (SSWDB) accordingly then to the Knowledge database KDB(Mohammed Mahmood Ali & Rajamani, 2013)(Mohammed Mahmood Ali et al., 2014)	
In 3 RULE (not detected words)	
not detected words tracked and rectified automatically with the help of nearest stem word by ontology taxonomy construction(Mohammed Mahmood Ali et al., 2014) by 1 rule. (This rule 3 is used in OBIE it is applied for (TPDB) for information retrieval. After that, it sends to the KDB Knowledge database). (Mohammed Mahmood Ali & Rajamani, 2013)(Mohammed Mahmood Ali et al., 2014)	

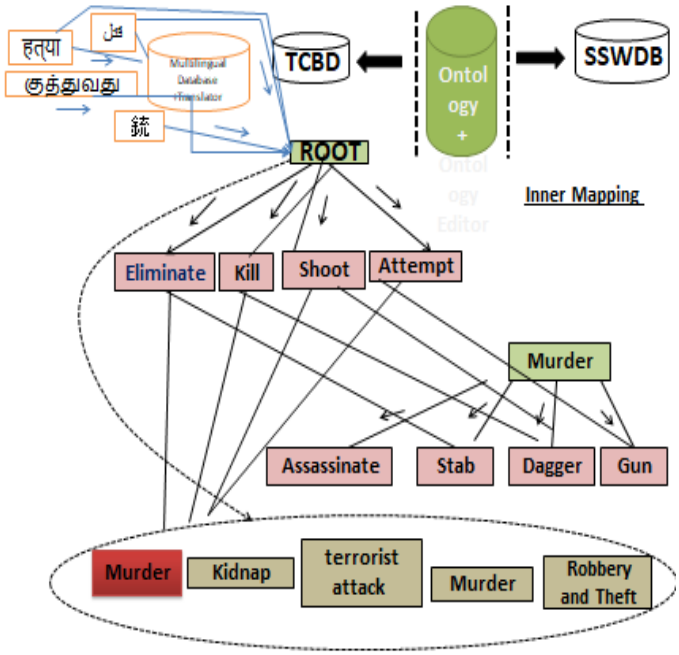


Figure 8: Internal Mapping Of Multilingual Suspicious Messages Detection

3.3 Complete Algorithm of the proposed system

1. Create instant messenger input

// create an account, perform user interaction, maintain database, and store conversation

2. Call L.D.T.A Algorithm

// detect Multilingual then translate Multilingual to the English language.

// page number 3

3. Call SDP algorithm (Mohammed Mahmood Ali et al., 2014)

4. // filter unwanted words detect the suspicious word and forward to TPDB its map with a pre-defined knowledge-based rule which is available in SSWDB.

// page number 8,9.

5. // Conversation of the user are stored in TDB

6. //Check for the suspicious word if detected store in TPDB and map with SSWDB with help of OBIE

7. //at first stage complete stem words will be mapping to empty root with the help of OBIE (TPDB) termed as tree alignment algorithm

8. //perform GSHL algorithm it will check stem word threshold value of TPDB compare to the SSWDB it also finds out which kind of threat it is like come under Rule1, Rule2, Rule 3

9. //scan TDB and forward it to TPDB as stem word

10. //comparing TPDB with inbuilt SSWDB if found similar forward to KDB stem words with domains are stored in KDB

11. //if TPDB equal to KDB then domain considered as suspicions, as it consists of a suspicious word

12. //check with KDB checking which kind of domain word come under with the help of R2D wrapper algorithm and using E-crime database

13. //if KDB equal to True, then Suspicious word Match

14. //check with EDB which will track the user name, phone number, location, Email id, IP address

15. Details of a user should be reported to the E-Crime Department.

16. The report consists of suspicious conversation, details of users

//Output

4. Summary of proposed System:

In this article, authors seek to propose a framework that would help in the execution of attacks a foretime though the various communication methods among the attacker process start with the first conversation between criminals will detect language, secondly translate multilingual conversation to the English language, thirdly words are filtered, fourthly suspicious words are compared with pre-defined rules and multilingual wordnet synonyms, fifthly criminal are tracked by Metadata.

5. Experimental Analysis

Table 3: Shows output obtained from Extortion

The domain of threat detected.	User – 1	User - 2
Extortion Extortion	We have planned to kidnap the son of a famous businessman "మీరు పదునైన కత్తి, బ్యాగ్ & తాడు తీసుకురావాలి "(You have to bring sharp knife, bag & rope) " நீங்கள் பல குழந்தைகளை கடத்திச் சென்றீர்கள் "(You already kidnaped many children's) " मैं तुम्हें अपनी कार दे दूँगा 2 मजबूत अपराधी के साथ तेज हथियार लाना "(I will give you my car be prepared with 2 strong criminals with	"ठीक है, इस जबरन वसूली मामले में मेरी भूमिका क्या होगी"(Ok what will be my role in this extortion case) "Pehele aap mere ko training diyo iske "(Foremost give me training for the same) " मुझे कैसे शुरू करना चाहिए स्थान भेजें "(how should I start)

	sharp weapons) Take 10 lakh advanced and take 30 lakh on delivery of child	" ముందస్తు విమోచన మొత్తాన్ని ఇరవై లక్షలు ఇవ్వండి " (Give advance ransom amount of 20 lakhs) Ok send the location on my mail
	Output	
Suspicious multilingual words Detected in communication	పదునైన కత్తి, బ్యాగ్ & తాడు, குழந்தைகளை , విమోచన , तेज हथियार.	
Multilingual suspicious identified by Underlined green as Telugu, yellow as Hindi, violet as Tamil, and brown as English	Above multilingual word translated to English words respectively. (Sharp knife, rope, bag, kidnaped, ransom, sharp weapon)	
Undetected words	जबरन वसूली, अपराधी स्थान भेजें(Extortion,criminal,send location)	
Suspicious English word	kidnap, delivery of child	
The precision for multilingual translated words + English words divided by no. of pre-defined words in SSWD in a particular domain	6 = Multilingual Word 2 = English Word 6+2=8 8/21 =38.09	

Table:4 Shows output obtained from Fraud

The domain of threat detected.	User – 1	User - 2
Fraud	" मैंने भूमि के नकली कागजात पाया "(i found fake papers of land) " కాగితాన్ని తప్పుగా చూపించడం ద్వారా మనం చాలా మందిని మోసం చేయవచ్చు"We can dupe many people by misrepresentation of paper Around 50 Acers worth will be 25 Cores in a good location The good idea many greedy people and land grabbers come forward to take land at a low price	" तो हम इसके साथ क्या कर सकते हैं डुप्लीकेट पेपर "(So what we can do with that duplicate paper) "Oh! For how many acres we can mislead peoples Will give misstatement to people that we are selling this land at half rate After looting people will move to another country.
	Output	
Suspicious multilingual words Detected in communication	नकली कागजात, కాగితాన్ని తప్పుగా, మోసం,	
Multilingual suspicious identified by Underlined green as Telugu, yellow as Hindi, brown as English	Above multilingual word translated to English words respectively. (Fake Paper, misrepresentation, dupe)	
Undetected words	डुप्लीकेट पेपर (duplicate paper) land grabbers, location	
Suspicious English word	Mislead, Misstatement, looting	
The precision for multilingual translated words + English words divided by no.of pre-defined words in SSWD in a particular domain	3 = Multilingual word 3= English word 3+3=6 6/14=42.85	

Table:5 Shows output obtained from Murder

The domain of threat detected.	User – 1	User - 2
Murder	" ఒక వ్యాపారవేత్తను హత్య చేయడానికి మాకు విమోచన విమోచన లభించింది "(We got ransom to assassinate a businessman) "25 lakhs advance 25 lakh after the assassination	I already killed many businessmen for money

	الموقع في تاريخ ٢٥ يوليو منتصف الليل في باندرا (location and date On 25 july midnight at bandra) " Bring pistol of 6mm, sharp weapon and 2 hockey sticks and don't forget gloves	"当我们必须计划谋谋杀"(When we have to plan murder) "సరే, నేను సిద్ధం కావాలి స్థానాన్ని పంపండి "(Ok, what I have to prepare for) " త్వరలో తెస్తాను "I will bring soon)

Output	
Suspicious multilingual words Detected in communication	హత్య , విమోచన , 谋杀 الموقع في تاريخ
Multilingual suspicious identified by Underlined green as Telugu, Orange as Arabic, blue as Chinese, brown as English	Above multilingual word translated to English words respectively as assassinate, ransom, murder, location, and date,
Undetected words	Money, స్థానాన్ని పంపండి (send Location)
Suspicious English word	(Assassination, the killed pistol of 6mm, sharp weapon hockey sticks, gloves)
The precision for multilingual translated words + English words divided by no.of pre-defined words in SSWD in a particular domain	4 = Multilingual 6 = English word 4+6=10 10/18=55.55

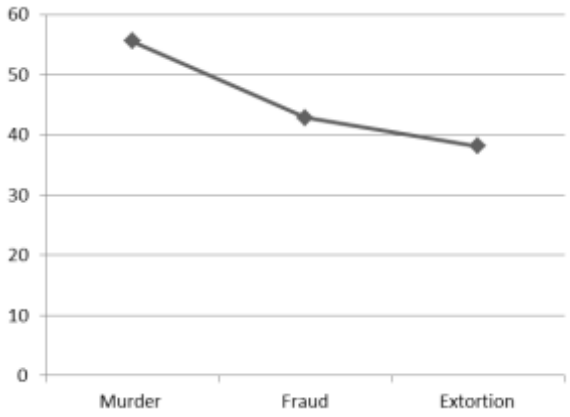


Figure 9: Comparative analysis of three domain

Murder, Fraud, Extortion the domain murder have the highest precision value of 55.5 of domain murder > Hence, it maps to suspicious word database.

6. Experimental Results

To evaluate our proposed system we use precision(Shiri, 2004) Matrics efficiency of suspicious word extracted only on two factors one is precision and another is a recall

$$\text{Precision} = \frac{\text{extarcted corectly words}}{\text{total correctly extarcted words}}$$

$$\text{Recall} = \frac{\text{extarcted corectly words}}{\text{total number of words possible}}$$

For the domain murder we received database from GTD(Global Terrorisms Database)(Access the GTD | GTD, n.d.) the most famous database records of terrorists and criminal all over the world with latest to 2018 this is the updated version till date it consists of 191465 rows, size 90 MB its codebook can be downloaded with this below link. we can get a complete picture of a terrorist attack or any criminal activity by using the following database we analyze our system and compare our work with previous system Comparison table is shown below. https://www.start.umd.edu/gtd/downloads/Codebook.pdf

Table:6 Output Obtained from domain expert data GTD

Term	Proposed system output
Total correctly extracted	1880
Extracted correctly	1804

Total number of word possible	1833
Precision	**95.95%**
Recall	**98.41%**

Here dataset was utilized by generous brainstorming sessions and for domain experts using GTD.

7. Observation

It is observed that naïve basin algorithm unable to detect properly for few languages such as traditional Chinese, Japanese, Persian, and a few Arabic characters cause of this is bias, noise. And it is improved by noise filter and character normalization. In multilingual translation, while matching word translation with attention mechanism few words unable to match exactly but near meaning help to sort out this issue, this is more seen because of few similar multilingual words have a different meaning or and different word have the same meaning with the help of pre-defined rule error rate is reduced .it is also identified when multilingual few words not exactly matched for such kind of case Rule 3(for undetected words) is applied, It is observed that few words are common in many domains such as location, money for this minimum threshold value is calculated using GHSL algorithm(Mohammed Mahmood Ali et al., 2014) for such cases expert interface is necessary.

Table:7 Comparison Analysis with the Previous system

System	d **Advanced Machine Learning Approach for Detection of Multilinguistic Terror Message to save human Lives** (Proposed system)	Framework for surveillance of instant messages (Mohammed Mahmood Ali et al., 2014)(Mohammed Mahmood Ali & Rajamani, 2013)
Language	Detect multilingually	Can't detect multilingually
Encoder & Decoder	Accurate	No such system
Languages can be translated	Of 49 languages	No
Attention Mechanism	Relate corresponding different languages	No such mechanism
Total no of languages support	49 language's	Only **1** language
Lexical Database	Multilingual WordNet	Only WordNet
Precision for similar test (Murder)	**55.55%**	**31.5%**
Precision for similar test (Extortion)	**38.09%**	**19.0%**
Precision for similar test (Fraud)	**42.85%**	**25.0%**
Accuracy	**64% more accurate than the previous system**	**Less Accurate**
GTD	**191465 rows, size 90 MB**	**59787 rows, size 30 MB,**

8. Future Work and Conclusion

We present a simple solution to the entire world facing problems with terrorist attacks, extortion, fraud, murder, and all criminal activity. There was no proper solution for detecting suspicious multilingual messages till our proposed system. The drawback in our system is to improve precision and decrease recall percentage, the multilingual translation should be improved for more accurate results our system can detect up to 49 languages. This system should be improved for the detection of all possible languages. The multilingual translator translates into English then mapping suspicious word by using our pre-defined database and GTD comparing with suspicious word SSWD if found checking the profile of sender and reporting to the e-crime department.

References

1. Ali, Mohammed Mahmood, & Rajamani, L. (2013). Framework for surveillance of instant messages. International Journal of Internet Technology and Secured Transactions, 5(1), 18–41. https://doi.org/10.1504/IJITST.2013.058292
2. Ali, Mohammed Mahmood, Mohammed, K. M., & Rajamani, L. (2014). Framework for surveillance of instant messages in instant messengers and social neworking sites using data mining and ontology. IEEE TechSym 2014 - 2014 IEEE Students' Technology Symposium, 297–302.

https://doi.org/10.1109/TechSym.2014.6808064
3. Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena, 404(March), 1–43. https://doi.org/10.1016/j.physd.2019.132306
4. Yu, Z., Yu, Z., Guo, J., Huang, Y., & Wen, Y. (2020). Efficient Low-Resource Neural Machine Translation with. 19(3), 1–13.
5. Zennaki, O., Semmar, N., & Besacier, L. (2019). A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages. Natural Language Engineering, 25(1), 43–67. https://doi.org/10.1017/S1351324918000293
6. Shrestha, A., & Spezzano, F. (2019). Online misinformation: From the deceiver to the victim. Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019, 847–850. https://doi.org/10.1145/3341161.3343536
7. Fujs, D., Mihelič, A., & Vrhovec, S. L. R. (2019). The power of interpretation: Qualitative methods in cybersecurity research. ACM International Conference Proceeding Series. https://doi.org/10.1145/3339252.3341479
8. Guo, S., Li, X., & Ma, Z. (2020). Association rule mining of anaphora based on parcorfull corpus. ACM International Conference Proceeding Series, 91–98. https://doi.org/10.1145/3379247.3379277
9. Jauhiainen, T., Zampieri, M., & Baldwin, T. (2018). Automatic Language Identification in Texts : A Survey Automatic Language Identification in Texts : A Survey. April. https://doi.org/10.1613/jair.1.11675
10. Access the GTD | GTD. (n.d.). Retrieved April 27, 2020, from https://www.start.umd.edu/gtd/access/
11. Adeel, M. (2010). Soundex Algorithm.
12. Ali, Mohd Mahmood, & Rajamani, L. (2012). APD: ARM deceptive phishing detector system phishing detection in instant messengers using data mining approach. Communications in Computer and Information Science, 269 CCIS(PART I), 490–502. https://doi.org/10.1007/978-3-642-29219-4_56
13. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Transactions of the Association for Computational Linguistics, 5, 339–351. https://doi.org/10.1162/tacl_a_00065
14. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1724–1734. https://doi.org/10.3115/v1/d14-1179
15. Dunning, T. (1994). Statistical Identi cation of Language Ted Dunning New Mexico State University. January 1996.
16. Jadhao, A. R., & Agrawal, A. J. (2016). A digital forensics investigation model for social networking site. ACM International Conference Proceeding Series, 04-05-Marc, 3–6. https://doi.org/10.1145/2905055.2905346Access the GTD | GTD. (n.d.). Retrieved April 27, 2020, from https://www.start.umd.edu/gtd/access/
17. Adeel, M. (2010). Soundex Algorithm.
18. Ali, Mohammed Mahmood, Mohammed, K. M., & Rajamani, L. (2014). Framework for surveillance of instant messages in instant messengers and social neworking sites using data mining and ontology. IEEE TechSym 2014 - 2014 IEEE Students' Technology Symposium, 297–302. https://doi.org/10.1109/TechSym.2014.6808064
19. Ali, Mohammed Mahmood, & Rajamani, L. (2013). Framework for surveillance of instant messages. International Journal of Internet Technology and Secured Transactions, 5(1), 18–41. https://doi.org/10.1504/IJITST.2013.058292
20. Ali, Mohd Mahmood, & Rajamani, L. (2012). APD: ARM deceptive phishing detector system phishing detection in instant messengers using data mining approach. Communications in Computer and Information Science, 269 CCIS(PART I), 490–502. https://doi.org/10.1007/978-3-642-29219-4_56
21. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1724–1734. https://doi.org/10.3115/v1/d14-1179
22. Dunning, T. (1994). Statistical Identi cation of Language Ted Dunning New Mexico State University. January 1996.
23. Fujs, D., Mihelič, A., & Vrhovec, S. L. R. (2019). The power of interpretation: Qualitative methods in cybersecurity research. ACM International Conference Proceeding Series. https://doi.org/10.1145/3339252.3341479
24. Guo, S., Li, X., & Ma, Z. (2020). Association rule mining of anaphora based on parcorfull corpus. ACM International Conference Proceeding Series, 91–98. https://doi.org/10.1145/3379247.3379277
25. Jadhao, A. R., & Agrawal, A. J. (2016). A digital forensics investigation model for social networking site. ACM International Conference Proceeding Series, 04-05-Marc, 3–6. https://doi.org/10.1145/2905055.2905346
26. Jauhiainen, T., Zampieri, M., & Baldwin, T. (2018). Automatic Language Identification in Texts : A Survey Automatic Language Identification in Texts : A Survey. April. https://doi.org/10.1613/jair.1.11675
27. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Transactions of the Association for Computational Linguistics, 5, 339–351. https://doi.org/10.1162/tacl_a_00065
28. Leung, C. K. (2019). Pattern mining for knowledge discovery. ACM International Conference Proceeding Series. https://doi.org/10.1145/3331076.3331099
29. Lui, M., & Baldwin, T. (2012). langid.py: An Off-the-shelf Language Identification Tool. Aclweb.Org, July, 25–30. http://www.aclweb.org/anthology-new/P/P12/P12-3005.pdf
30. Rajamani, L., Ali, M. M., & Rasheed, M. A. (n.d.). OSMD : Online Suspicious Message Detection Framework for Instant Messaging Systems. 380–385.
31. Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena, 404(March), 1–43. https://doi.org/10.1016/j.physd.2019.132306
32. Shiri, A. (2004). Introduction to Modern Information Retrieval (2nd edition). Library Review, 53(9), 462–463. https://doi.org/10.1108/00242530410565256
33. Shrestha, A., & Spezzano, F. (2019). Online misinformation: From the deceiver to the victim. Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019, 847–850. https://doi.org/10.1145/3341161.3343536
34. Yu, Z., Yu, Z., Guo, J., Huang, Y., & Wen, Y. (2020). Efficient Low-Resource Neural Machine Translation with. 19(3), 1–13.
35. Zennaki, O., Semmar, N., & Besacier, L. (2019). A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages. Natural Language Engineering, 25(1), 43–67. https://doi.org/10.1017/S1351324918000293