Indian Journal of Science and Technology, *Vol 10(14), DOI: 10.17485/ijst/2017/v10i14/111364, April 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Mining user Message Pattern for Suspicious Behavior on Terrorism using NLP in Social Networks with Single Sign-On

Mercy Paul Selvan and Renuka Selvaraj

Department of Computer Science, School of Computing, Sathyabama University, Chennai – 600119, Tamil Nadu, India; mercypaulselvan@gmail.com, smilewithrenu@gmail.com

Abstract

Objectives: To find an effective way to find Suspicious Behavior on Terrorism Using NLP in Social Networks with Single Sign-On in advance and prevent the massive destruction of life and property. **Methods/Statistical Analysis:** A survey has been made to understand the behavior of people using the social networking. Social networking has seen a massive growth for more than a decade and popular among people. So, taking this into consideration it aim at developing a monitoring system which continuously monitors user activity in the social network to find any suspicious activity regarding terrorism. Natural Language Processing Technique is used for analyzing the text data and Least Significant Bit algorithm is used for Steganographic images. **Findings:** In our research work, we have considered two social network sites - Gmail and Twitter. We work on real time dataset fetched from Gmail and Twitter. These social networks are continuously monitored for user behavior. The user data from Gmail inbox, sent items, tweets, sent and replied twitter messages are fetched and passed to Natural Language processing to extract the data patterns related to terrorism. Based on the threshold of terrorism related data, the user behavior would be classified as normal, little suspicious and offensive. It considers a single sign on of the user into social network. It has considered the user activity in two social networks to get a precise data. There are existing works analyzing on user sentiments, sarcasm, psychological effects, political, expert advice, and recommendations on user ratings. To my knowledge there is no work done on analyzing user data on more than one social network on text and image data to find their suspicious behavior on terrorism. **Application/Improvements:** In our application, we are using the real data set of the users from Gmail and Twitter account using single sign on. The work is to analyse both the users text information and image shared via Gmail to find the suspicious behavior on terrorism and categorize them to normal, little suspicious and offensive. This information can be shared with the investigation team, where they can take any precautionary action preventing the world from massive destruction of life and property.

Keywords: NLP, Single Sign on, Social Network, Suspicious Behavior, Terrorism

1. Introduction

Growth in internet is tremendous which has impacted in every one of different ages and fields. It has become a part of their life for easy communication, purchase, blogging and reviews. People started to share more emotional part of their life, their desires, opinions, creativity, suggestions and everything on internet.

People spend enormous amount of time in chatting and messaging. They could be on random topics, something specific, personal, and social or sometimes it could be offensive.

In this paper it propose to characterize and detect the suspicious behavior of the internet users' through NLP processing and compute the threshold of the sequence of suspicious data been shared among the users. Based on the computation it conclude that the user behavior is normal, little suspicious and offensive.

For more specific and precise monitoring, it proposes our system to extract the user's activity on real time web

**Author for correspondence*

application data set on Twitter and Gmail. Using our technique it can monitor the user's message pattern based on their session id on multiple applications with single sign on email and twitter. It used the documents of inbox and sent box mail of Gmail contents and twitter's tweet and individual chats to extract the topic and mining the user's activity. It extracts the topic of document stream content using Stanford Natural Language Processing. Using this NLP processing and monitoring dynamic user's different activities can be extracted and monitored effectively.

Rest of the paper is organized as follows. Section II briefs the motivation of the work. In Section III, it explains the framework and methodology proposed for monitoring. Section IV, shows the results and Section V, discussion and Section VI concludes the paper.

2. Motivation

As mentioned above, monitoring the users' behavior for suspicious activity can help detecting the crimes prior to their occurrence. Terrorism has grown much in the world and the terrorist attack occur quiet often causing massive destruction of public and property. The loss caused by the terrorist to the public and property are high creating many emotional and economic problems.

There are many works done in analyzing the tweet data for sentimental analysis through users' positive and negative words in the tweet and sarcastic behavior.

In[1] analyze the tweet data based on the positive and negative emotions.

In[2] works on finding the sarcastic behavior of the users' in twitter for sentimental analysis.

In[3] proposed an approach to find the experts in twitter as the information from the experts are valuable.

In[4] Mine the micro blog text in twitter to find and recommend good restaurants.

In[5] worked on finding a systematic approach to detect social networking users' intensity and intention of feelings.

There are various diagnostics done in twitter to detect the users' behavior. Throughout our work it work on forming a framework of monitoring system for diagnosing the suspicious activity of the users through NLP techniques in twitter and Gmail.

Most of the recent researches are given to the information retrieval and analysis to understand the users' behavior, intention, interest in order to recommend, determine the mental health of the users' and so on.

In[6] describes different methodology and implementation details of question answering system for general language and also proposes method to retrieve more precise answers using NLP techniques.

The basic idea behind the QA system is that the users just have to ask the question and the system will retrieve the most appropriate and correct answer for that question and it will give to the user. In order to solve this they are working in various domain such as Web Mining, NLP, Information retrieval and information extraction.

In[7] proposes the helpfulness of online physician reviews. It uses review ratings, psychological, linguistic and

Semantic features as input to classify these reviews into helpful or unhelpful categories. The results demonstrate a significant impact of review ratings on the helpfulness of online physician reviews.

They extract the linguistic and psychological features using the NLP tool.

A large number of pages are added to the web every day. This causes the problem of data overload.

In[8] proposes a system for recommendations to suggest users the newly added data based on their interest and search behavior.

NLP is used when sites are dealing with matching of items based on data, text entered and for integrating the information regarding content and hyperlinks.

In[9] proposed an automatic framework for document validation instead of human manual verification. They define set of rules in the framework and the rules are automatically extracted using the NLP techniques.

In[10] proposed a work that compares the patients and normal users in the social network for psychiatric disorder determination. NLP is used in semantic based text classification of the users in social media.

In[11] are focused on identifying the mood of the users in twitter regarding some social issues or topic from tweeter posts.

NLP is used to enhance the sentiment classification by adding semantics in feature vectors and thereby improving the classification.

In[12] uses a Document Object Modelling (DOM) tree modelling to remove the irrelevant data from the newspaper like ads and user comments. They also use wordnet processing to extract the content that semantically match the web page. The semantically gathered information are grouped based on the users' interest and preferences.

In[13] proposes a rule based approach to query the data from the database by a common man instead of a structured Query Language.

In[14] proposed an approach to extract the keywords from a video instead of providing the entire text data. This could help the user to understand how relevant the video is to their search. They use a regular expression grammar rule to detect the keywords from the video transcript.

In[15] with the wide spread use of social networking and exchange of news and opinions, political learning has become far easier through it. Different discussions, opinions and realities are wide spread over the network. The tweet and retweet of the political information are collected with similar types of users.

In[16] several mining algorithms exists, here we refer to the sequential mining pattern algorithm to identify the sequential messaging pattern of the users in social networks.

In[17] this site is referred to collect the terrorism related data to map with the user behavior and identify the matching patterns.

3. Proposed Approach

Our proposed approach is to monitor the user behavior in both Gmail and Twitter dataset for a particular user using single sign on. Our approach deals real data sets.

3.1 DataSet

As mentioned above, real time datasets Gmail and Twitter are considered. For experimental purpose, user registration with the Gmail and Twitter ID are done. These two IDs should be same. IMAP (Internet Message Access Protocol) is used for retrieving the Gmail message.

3.1.1 Gmail Data Set

a. Inbox: Text data and images in the Gmail inbox of the users are retrieved and analyzed. This contains the data sent by other users to the current user. This could be a plain text or a HTML data.

b. Sent box: Text data and images in Gmail Sent items are retrieved and analyzed. This contains the mail messages sent by the current user to any other user.

3.1.2 Twitter Data Set

a. Timeline Messages: This includes the messages posted by us or any other user on the current users' timeline.

b. Sent Messages: This contains the messages sent by the current user to others in twitter.

c. Received Messages: This contains the messaged received from the other users on the current users twitter account.

3.2 Tools

To retrieve the messages IMAP (Internet Message Access Protocol) a standard email protocol is used to store the messages on a server and allow the users to access them as if they are able to retrieve from their local server.

To perform different Natural Language Processing (NLP) like POS tagging, chunking, WordNet processing and spell checking Apache OpenNLP is used.

Steganography, LSB algorithm is used to find the hidden text data in the images sent via mails.

3.3 System Framework

Our framework contains various functionalities to extract the users' suspicious behavior. Figure 1 shows the architectural diagram of our proposed system.

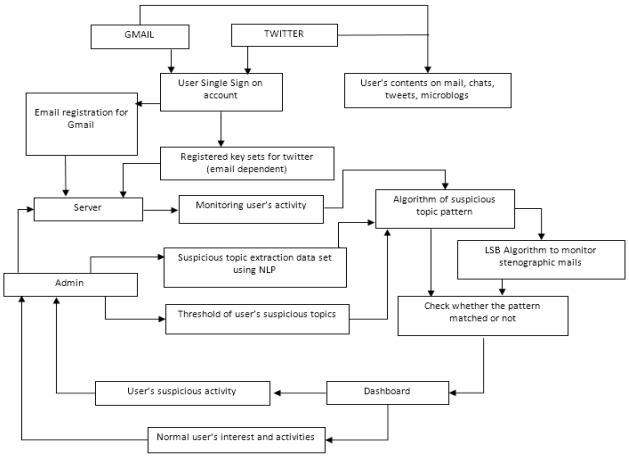

Figure 1. System Architectural Diagram.

3.4 User Data Extraction (UDE)

The data from Gmail and twitter account are extracted. The Gmail and Twitter accounts are logged in using single sign on id. Since there are abundant data in each account a threshold is maintained to know the amount of data to be retrieved each time to monitor. The type of data set can be categorized like inbox, sent items, mail chats, user's tweets, twitter chats and micro blogs maintained in the database. JavaMail API and Twitter4j API are used to retrieve the data from Gmail and Twitter. For our

experimental purpose, IMAP (Internet Message Access Protocol) is used which is a Mail access protocol which enables the user to access the Gmail accounts to read the inbox and sent mails of the users. Mails might contain stenographic images. LSB algorithm is used to extract the hidden data from the images. These dataset are passed to NLP for processing and the data pattern returned by the Natural Language Processor are stored for further processing.

UDE Algorithm

- ***Procedure*** *CheckSingleSignOn()*
- *Is single sign on ID for Gmail and Twitter*
- *Return true*
- *Return false*
- ***End procedure***
-
- ***Procedure*** *ExtractUserContent()*
- ***If*** *CheckSingleSignOn == true*
- *ExtractGmailContent()*
- *ExtractTwitterContent()*
- ***End if***
- ***End Procedure***
-
- ***Procedure*** *ExtractGmailContent()*
- *Get the threshold level for Gmail content*
- *Use JavaMail API to retrieve the content*
- *Categorize the content as inbox, sent mail*
- and mail chats
- *Store the content in the database*
- ***End procedure***

Steganography is the process of hiding secret data within an image and sending it to the receiver. Least Significant bit algorithm is used for extracting the data from the image sent via mail and the content is sent to NLP processing to find any information related to terrorism.

LSB Algorithm

- ***Procedure*** *DecodeText()*
- Get the bytes of the image
- *Grab each LSB from the image*
- *Shift by 1 and grab the last bit of next image byte*
- *Convert the result to string*
- *Return the message*
- ***End Procedure***

3.5 Natural Language Processing (NLP)

This module is responsible for processing the dataset got as input and return the data pattern. Apache OpenNLP is used for our processing. A sequence of techniques are applied on the data set to retrieve the data pattern.

- *POS tagging* - Given a sentence, POS tagging helps to identify the part of speech of each word in a sentence.

Figure 2 Shows the tags used for generating POS tagging.

Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential "there"	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, preterite (past tense)	*ate*
IN	preposition or subordinating conjunction	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama, snow*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Figure 2. Part of Speech Tags.

Maxent Tagger, an OpenNLP Java API is used, which takes the mail content as input and provides the data with Part Of Speech tagging

The 8th report on terrorism in India published in 2008 defined terrorism as the peacetime equivalent of war crime.

POS tagging output

The/DT 8th/JJ report/NN on/IN terrorism/NN in/IN India/NNP published/VBN in/IN 2008/CD defined/VBN terrorism/NN as/IN the/DT peacetime/NN equivalent/NN of/IN war/NN crime/NN ./.

- *Chunking* - POS tagging just says whether the word is verb, noun or adjective etc., but it doesn't give any idea about the structure of the sentence. Chunking helps to get a structure or phrase of a sentence.

Chunking takes the POS tagged data as input and provides the output. Chunker Model (ChunkerME) API of OpenNLP is used for creating chunks of data.

Chunk output for above *sample*

[The_DT 8th_JJ report_NN] on_IN [terrorism_NN] in_IN [India_NNP] published_VBN in_IN [2008_CD

defined_VBN terrorism_NN] as_IN [the_DT peace-time_NN equivalent_NN] of_IN [war_NN crime_NN] ._.

WordNet Processing – Synsets in wordset is used for identifying the semantic similarities in the words of the dataset. It groups the semantically identical words to the word of interest. The Terrorism related dataset is taken from onelook.com. Each word has a key which is mapped to the value. The chunker output is passed to check the matching key from which the domain is mapped. When the abnormal pattern is more than two, it is considered as a rare pattern.

- *Spell Checking* - This is used to correct the words that are misspelled. It scans the text and extracts the words contained in it. It then compares the words against the correctly spelled words and corrects them when needed.

The generated data pattern is returned to the modules which provided the dataset input

3.6 Mining Suspicious Behavior (MSB)

This module performs the mining of the user's data pattern and pre-defined suspicious data pattern in the server to check the behavior of the users. If any match is found, then the user is suspicious. A threshold is maintained to see whether the suspicion is little suspicious or more. The behavior of the users are analyzed in both Gmail and twitter account to get a strong result. It can be an important clue of illegal activity and trigger targeted investigations.

MSB Algorithm

- ***Procedure*** *GetUserDataPattern()*
- *Get the user data Pattern from the*
- userdataextractor
- ***Return*** *UserDataPattern*
- ***End Procedure***
-
- ***Procedure*** *GetSuspiciousDataPattern()*
- *Get the pre-defined suspicious data pattern*
- *from the suspicious topic pattern extractor*
- ***Return*** *PredefinedSuspiciousDataPattern*
- ***End Procedure***
-
- ***Procedure*** *CheckSuspiciousBehavior()*
- ***If*** *UserDataPattern matches Pre-*
- SuspiciousDataPattern
- ***If*** *threshold level > N*
- *Users is Offensive*
- ***Else***
- . *User is little suspicious*
- ***End if***
- ***Else***
- *User behavior is normal*
- ***End if***
- ***End Procedure***

4. Results

In order to achieve a more precise monitoring and information retrieval two networking sites are Gmail and Twitter which are commonly used among the users. Figure 3 Shows the monitoring system home page view. The access is restricted to user with administrative privileges. So admin login tab is placed to provide the credentials. User registration is required here for experimental purpose.

Figure 3. Monitoring System Home Page.

This page has the threshold settings for each category like inbox, sent box, timeline, sent message, received message to set as how many data needs to be retrieved. This is needed as huge amount of data is been present in the Gmail and twitter accounts.

The user behavior is constantly monitored to find any suspicious activity and it is categorized based on our analysis as shown in the Figure4.

5. Discussion

Two email IDs are registered to monitor their activities for experimental purpose. Each mail ID contains 25 mails and the threshold is set as 5.

The monitoring undergoes 5 iterations to check 25 mails for each user. The Table 1 shows the number of related mails found in each iterations.

USER ID	TIMELINE STATUS	SENT MESSGAE STATUS	RECEIVE MESSAGE STATUS	INBOX MAIL STATUS	SENT MAIL STATUS	OVER ALL STATUS
vishwa021985@gmail.com	littlesupecious	normal	normal	normal	normal	3.2857144
aakash011985@gmail.com	normal	normal	normal	normal	normal	1.0

Figure 4. Dashboard.

The precision and recall are computed based on the table above and the results are shown in Figure 5 and Figure 6.

Table 1. Email Iterations - relevant suspicious mail for user 1 and user 2

Iterations	User 1	User 2
Iteration 1	3	2
Iteration 2	2	1
Iteration 3	1	3
Iteration 4	4	1
Iteration 5	1	1

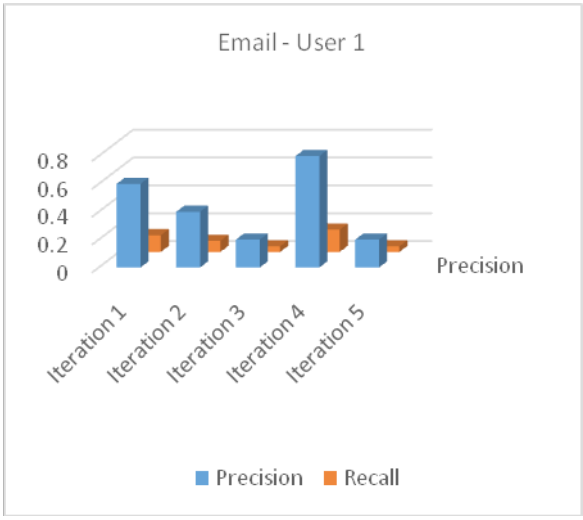

Figure 5. Email Precision and Recall for user1.

6. Conclusion

In this paper, an idea to continuously monitor user behavior in two accounts using single sign on is proposed to detect the suspicious activity of the user and report to the investigation team in order to prevent major disaster. Both the text data and the imagesbeen sent via mail are monitored. This could act as a preventive action from a devastation that might occur.

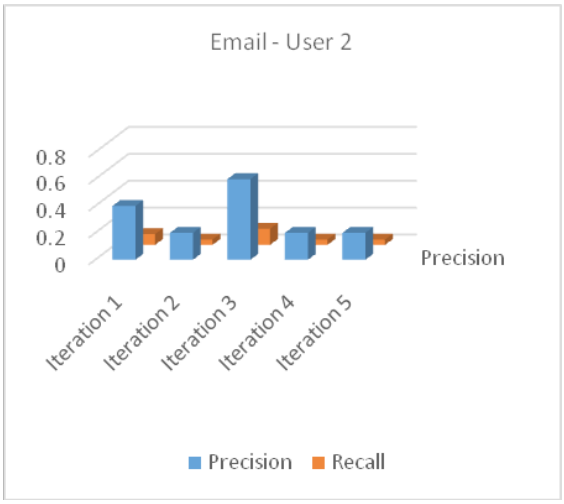

Figure 6. Email Precision and Recall for user2.

Acknowledgement

The authors thank Sathyabama University, School of Computing, Chennai, India for providing the opportunity and the facility to carry out this experiment.

8. References

1. Larsen M, Boonstra TW, Batterham PJ, O'Dea B, Paris C, Christensen H. We Feel: Mapping emotion on Twitter. Proceedings in IEEE Journal of Biomedical and Health Informatics. 2015 Feb; 19:1246-52. Crossref. PMid:25700477.
2. Bouazizi M, Ohtsuki T. A pattern-Based approach for Sarcasm Detection on Twitter. Proceedings in IEEE Access. 2016 Aug; 4:5477-88. Crossref.
3. Cong G, Miao C. Learning to Find Topic Experts in Twitter via Different Relations. Proceedings in IEEE Transactions on Knowledge and Data Engineering. 2016 Mar; 28:1764-78. Crossref.
4. Xiang ZL, Yu XR, Kang DK. Mining Restaurants Information by Micro Blog Text Analysis. Proceedings in 17th International Conference on Advanced Communication Technology (ICACT). 2015 Aug; p. 634-8. Crossref.

5. Tai CH, Tan ZH and Chang YS. Systematical Approach for Detecting the Intention and Intensity of Feelings on Social Network. Proceedings in IEEE Journal of Biomedical and Health Informatics. 2016 Feb; 20:987-95. Crossref. PMid:26955055.
6. Lende Sweta P, Raghuwanshi MM. Question Answering System on Education Acts Using NLP Techniques. Proceedings in IEEE Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCFTR'16). 2016 Oct; p. 1-6.
7. Alodadi N, Zhou L. Predicting the Helpfulness of Online Physician Reviews. Proceedings in International Conference on Healthcare Informatics. 2016 Dec; p. 1-6. Crossref.
8. Tugaonkar PS, Chitre V. Practical Approach for Recommender Systems. Proceedings in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). 2016 Nov; p. 4464-7. Crossref.
9. Roychoudhury S, Bellarykar N, Kulkarni V. A NLP based Framework to support Document Verification-as-a-Service. Proceedings in 20th International Enterprise Distributed Object Computing Conference (EDOC). 2016 Sep; p. 1-10. Crossref.
10. Krishnamurthy M, Mahmood K, Marcinek P. A Hybrid Statistical and Semantic Model for Identification of Mental Health and Behavioral Disorders using Social Network Analysis. Proceeding in ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2016 Nov; p. 1019-26. Crossref.
11. Kanakaraj M, Guddeti RMR. NLP based sentiment analysis on Twitter data using ensemble classifiers. Proceedings in 3rd International Conference Signal Processing, Communication and Networking (ICSCN). 2015 Aug; p. 1-5. Crossref.
12. Kanakaraj M, Kamath SS. NLP based Intelligent News Search Engine using Information Extraction from e-Newspapers. Proceedings of International Conference in Computational Intelligence and Computing Research (ICCIC). 2015 Sep; p. 1-5.
13. Mahmud T, Hasan KMA, Ahmed M, Chak HC. A Rule Based Approach for NLP Based Query Processing. Proceedings of International Conference on Electrical Information and Communication Technology (EICT). 2016 Jan; p. 78-82.
14. Shukla H, Kakkar M. Keyword Extraction from Educational Video Transcripts Using NLP techniques. Proceedings in 6th International Conference - Cloud System and Big Data Engineering (Confluence). 2016 Jul; p. 105-8.Crossref.
15. Wong FMF, Tan CW, Sen S. Quantifying Political Leaning from Tweets, Retweets, and Retweeters. Proceedings in IEEE Transactions on Knowledge and Data Engineering. 2016 Apr; 28:2158-72.
16. Mabroukeh NR and Ezeife CI. A taxonomy of sequential pattern mining algorithms. ACM Computational Survey. 2010; 43(1):1-41. Crossref
17. OneLook: Available from: http://onelook.com/. Date accessed: 01/11/2016.