



A Study on Mining Approach under Cyber Crime Analysis

Priyanka Maan¹, Meghna Sharma²

¹Department of CSE, ITM University, Gurgaon, Haryana, India

²Department of CSE and IT, ITM University, Gurgaon, Haryana, India

Abstract: Today, the use of social network is been increased in daily life of a user. But this increased use of social media has also increased the associated threats. There are number of criminal activities possible over the web or social network conversation. This kind of conversation includes spamming, blackmailing, cyber threatening etc. The identification of these kind of messages or text segments is always a challenge because of vast dictionary and data for social message. In this work, a basic fuzzy effective framework is presented to identify the criminal activities and to perform the message classification. Here a text mining model is presented under fuzzy logic to provide effective solution.

Keywords: Cyber Crime, Message Classification, Opinion Mining, Sentiment Analysis, Neural network

I. Introduction

Opining Mining or Sentiment Analysis is having its significance to analyze the textual information and extract the valuable decisions. This text mining includes the document analysis or message review analysis so that the effective features will be obtained from the text message. This kind of message classification or categorization comes under prediction algorithm. The prediction algorithms are having the importance to provide the some conclusion driven results by applying a series of operations on input set. Here the input will be taken in the form of message from the specific domain or the application on which mining will be applied. One of such problem domain is social network. Social networks such as facebook, twitter etc provides the way to perform message conversion in the form of chats or the emails. These networks also enable the unknown person to interact or communicate some message to user. These kind of messages are having the changes of some interact with unknown person. This kind of message from unknown sources has the maximum changes of inclusion of some criminal activity. These criminal activities can be spam messages, threats, blackmailing etc. To avoid the misuse of the social network, it is required to identify such kind of messages at the earlier stage. To perform this some classification algorithm is required. Some of these classification algorithms are given here under.

The detection of the criminal activity depends on the activity prediction or the message content analysis. A dictionary based adjective specification is done to identify such kind of messages. More explored the word dictionary will be defined, more accurate results are expected from the system. This kind of analysis requires to collect the statistical information as well as the intelligent system. The work can be applied in real time system as an automated system that can be embedded on message communication to social network. It means, as the server will retrieve some message, the message will be check under the possible threats and the earlier decision can be taken for safe delivery of message. This kind of message processing is performed under different tests and constraints specific to the application and the domain. Once the rules are specified, the statistical features are obtained from the messages. Now this feature set work as the input method for classification algorithm. Here some of the most effective classification algorithms are described.

A) DECISION TREE

Decision Tree is one of the most accurate classification approach that is based on the statistical decision. The algorithmic construction of this method is more simple. But it is based on multiple conditions and constraint. This statistical information based approach is not effective when the size of dataset is large. While working with social message classification, it generates a topic set respective to the criminal activities. Now the distance based approach is defined to generate the associated features. Now the attribute prioritization is done based on which classification will be done. With each layer of decision tree, an attribute value is checked and based on this, the classification of message is done. The attribute value is defined in the form of range specification based on which the acceptance or resentence of a message is done. The method comes under supervised learning method that uses the traditional tree based approach. Each nod here represents the conditional constraints or activity to perform the message classification. When a message passes through all the levels of tree and comes to leaf, it itself identify the associated class. In this method, at each level M number of decisions is taken and each

decision generates the respective dataset partition. As the number of attributes or the classes increases, the method is not effective.

B) BAYESIAN CLASSIFICATION

Bayesian Network is considered as the intelligent approach that itself includes the statistical constraint to improve relevancy of network. It uses the conditional probability analysis approach to establish the relationship between different features and then identify the associated dependency. This approach is considered as a graph based approach represented in the form of acyclic graph. The dependency analysis is here based on the weighted probabilistic selection approach. In case of text mining, this kind of weightage can be obtained by analyzing the word frequency, relevancy etc. Once the statistical measures will be obtained, the intelligent classification will be performed under conditional probability approach. The work also includes the decision generation and class identification so that the attribute specification to the network is effective. This network based approach is defined as an acyclic graph based approach that is effective for class identification and to process the discrete predictive analysis. This classification approach also uses the ranking based on the work significance analysis respective to the topic or the criminal activity. The messages with highest ranked can be kept in the category of safe messages. In same way, the structural analysis can be performed to generate the feature vector and to take the decisions regarding the feature exploration.

C) Neural Network

Neural network comes under soft computing approach used an intelligent neuron based system. This is the approach comes under expert system that divides the complete classification in a layered model. This layered model includes the input layer, hidden layer and output layer. The input layer will accept the feature vectors obtained from the input message set. The statistical values will be taken in this layer as input. Once the input set is defined, the hidden layer will work as the weightage layer. There can be more than one hidden layer to improve the effectiveness of work. Here the transformation rules are defined in this layer to transform the input feature vector to final results. The transformation is applied on the mapping between the data and the neurons. This classification process is defined along with decision so that the intelligent results will be derived from the system, This layered model is shown in figure 1.

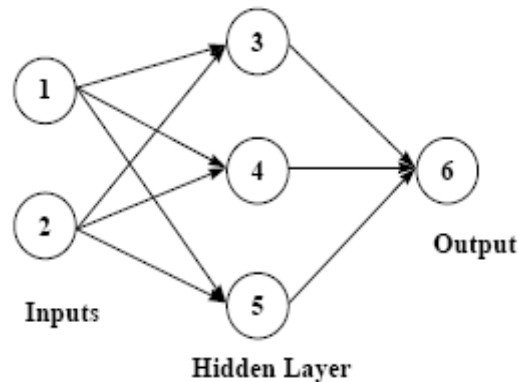


Figure 1: Neural Network System

Here figure is showing the three stage model, with the specification of input, weighted modeling and the output layer. The input layer can be described by one or more feature sets. These feature sets can be defined in the form of word frequency and relevancy to the crime activities. The hidden layer provides the transformation of this input set to the output. Finally the output set is generated as the classified dataset.

D) SUPPORT VECTOR MACHINE

Support Vector Machine is more evolutionary form of neural network. In this form of classification, the analytical system is composed in the form of data points and these data points are analyzed at different dimensional of the dataset. The dimensions are based on the analysis features such as margin analysis, distance analysis or interaction analysis. Once the data points are defined and explored, the conditional analysis is applied over it to filter the dataset and to improve the significant data feature. The vector formation relative to the vector space is defined. The intelligent decisions are taken regarding the neighbor point specification so that more accurate decision will be taken. The vector specific decisions are more accurate as well as intelligent.

In this, a study to the social network conversion message is defined under different criminal activities. As this kind of conversion includes the textual message communication so that the text mining is explored in this paper. In this section, a review to the opinion mining, text mining and review mining is defined. In section II, the work defined by earlier researchers is defined. Here the research in the area of opinion mining and social conversion

mining are defined. In section III, a fuzzy based model is presented for identification of criminal activities over the text. In section IV, the conclusion obtained from the work is presented.

II LITERATURE SURVEY

Cyber threat detection during the web communication or online conversion is one of the major challenge for the researchers. For this, there is the requirement to analyze different kind of communications performed by different users. These communications includes emails, chat, SMS, reviews etc. In this section, the work defined by earlier researchers on cyber threat analysis are given. Chris Clifton[1] defined a work to study the textual passage and to identify the topic of the document. This kind of analysis is helpful to identify the relevant text respective to the specified title or header. While working with web application, different users can provide irrelevant inputs or reviews or feedbacks. This approach is helpful to perform the filtration under topic identification. The work is based on the cluster generation and classification so that the separation of different reviews will be done. Jian Hu[2] has presented a predictive model analyze the browsing behaviour of a web use. Author defined the work to utilize the demographic information so that the usage identification over the web will be done. Author has defined the intelligent learning approach based on the events performed by user. This approach is helpful to track the invalid activity in terms of site visit by the user. Author has presented a Bayesian network based approach for the classification of user visit to track the abnormal behaviour. Faliang Huang[3] has defined a mining approach to analyze the chat messages to identify cyber threat. Author presented the message content analysis under threat structure identification so that message categorization will be done. Author presented the rule based algorithm for message categorization that can improve the social networks under observations. Masafumi Hamamoto[4] has presented a study based work for topic identification schemes. Author used the comparison between the SVD, Clustering, ICA methods so that the effective message classification will be done. Author performed the message analysis under different approaches. Reihaneh Rabbany[5] has presented an assessment behaviour to analyze the message over social network and to perform the actual partitioning. Author defined the study on the student message and provided the significant analysis over the message text. The information structure is identified to exploit the message so that more relevant decision will be taken. Author also explored the summarization under different aspects. Koosha Golmohammadi[6] has defined a data mining approach to identify the fraud from security market data. Author presented the work to analyze and study the associated message and generate the relevant patterns so that the identification of feature from the message will be done. Author presented a detection mechanism so that the more relevant decision regarding the recommendation to the research will be done.

Alexandra Balahur[7] has presented a feature based opinion mining approach for document feature exploration and classification. Author has performed the work to extract these features from the document review and performed a stage level work. At first the sentence level and then feature level approach is defined to perform the document summarization and sentiment analysis. Author has defined a machine learning approach for the categorization of these reviews and to perform the message classification. Author defined a name entity based syntactic parsing approach classification. Chun-hung Li[8] has presented the knowledge extraction approach using machine learning approach. Author has defined a grained analysis approach with specification of threshold for specific features. These features are then evaluated for message classification. Peter Koncz[9] has presented feature selection approach to identify the embedded sentiments in the messages. Author defined a subjective approach for feature classification. Author used the review dataset as an application of work. Amir A. Sheibani[10] has provided the spam detection over the message obtained from social networks. Author applied work for blogs to estimate the brand reputation. Author has presented a complete framework for spam recognition so that the review filtration will be done. Ms.K.Mouthami[11] has defined a work to analyze the review so that the sentiment extraction from the reviews will be done. Author has presented the valuable information analysis with effective of pattern identification. Author has analyze the social media to monitor the customer mood and negative message will be identified. Author used the SVM approach applied on feature set so that the review categorization will be obtained.

Mostafa Karamibekr[12] has performed a work on structure analysis applied in social domain. Author applied the text mining approach to analyze the opinion in the form of a topic or the issue and obtained the statistical measures. These measures are able to obtain the structure from subjective documents and obtained the more accurate results for classification. Rabeah Al-Zaidy[13] has defined a mining work un unstructured documents. Author collected the valuable information to perform forensic analysis so that the suspect identification and criminal activity identification will be done. Author has presented an investigation for criminal network analysis so that the suspect identification will be done. Author has presented a cyber crime dataset for the analysis. Farkhund Iqbal[14] presented a work on classification of chat messages obtained from social networks. The classification is here done under spamming, cyber predation, blackmailing etc. Author has defined an online message system so that the crime investigation will be done. Author has defined chat log effective framework so that the message enforcement will be done. Ching-Yung Lin[15] has defined intelligent work to analyze the

social network to perform the expertise search and to identify the message spams. Author has defined work under community analysis and knowledge specification so that the spam messages will be identified.

III Fuzzy based Analytical Model

In this section, a fuzzy effective predictive model is presented to identify the criminal activities over the social network. Social network is the most common and popular way for communication between web users. This kind of criminal activity detection can be done online or offline. In case of online message filtration and classification, the model can be combined with mail server or chat server to classify the message before delivery. This kind of analysis can generate alert messages if some criminal activity is identified. In case of offline message classification, the particular message can be defined in the form of dataset. These messages will be defined along with the specification of sender and receiver. Now the messages will be extracted from dataset for analysis.

While working with such kind of message classification, there is the requirement of some preprocessing approach to explore the message features. At the earlier stage of work, there is the requirement of identification of the terms involved in different kind of criminal activities and to generate the dataset respective to the activity class. These activities can be in the form of spam messages, hacking, blackmailing, threatening, advertisements. Once the classes will be identified, the criticality of each class as well as the relevant word set will be defined. The work also requires a stop list dataset to filter the input message at earlier stage so that the only relevant message words will be processed. After this all, the processing environment will be defined.

While working with the model, the first stage is to extract the messages under the specification of the user or the message itself. To extract this message, some sql query can be defined. Once the message is extracted, the next work is to filter the message by removing the stop list words. Now it will return the main message words as the input set. The next stage of work is to explore the message features. To explore the message features and represent this features as the main dataset. These features can be in terms of words frequency or the relevancy to the message words. This model is the proposed model and shows all the stages clearly. The model presented here is shown in figure 2.

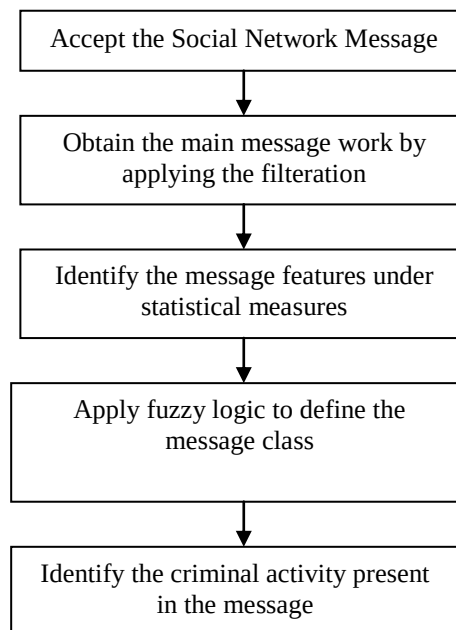


Figure 2: Proposed Model

Once the features will be explored, the next work is to define some rule so that the relevancy between the criminal activity and the input message can be identified. Here some distance analysis based approach can be applied along with fuzzy logic. Fuzzy logic is considered as the soft computing approach used to taken the decision based on current statistics and the integration of the message attributes. The inter relation analysis is here performed under fuzzy restriction. These restrictions or constraints are defined along with fuzzy operators. These fuzzy operators will be applied on all features individually as well as collectively. Based on these all rules, the message class will be identified. Once the message class will be identified, the criminal activity can be recognized from the message along with message criticality. The use of fuzzy logic improves the effectiveness of work in decision making.

IV. CONCLUSION

In this paper, a study based work is presented to perform the social network mining and to identify the criminal activities of web users. The paper has explored different classification algorithms respective to the social conversation. The paper also presented a fuzzy based model for message classification.

References

- [1] Chris Clifton, "TopCat: Data Mining for Topic Identification in a Text Corpus", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING @2003
- [2] Jian Hu, "Demographic Prediction Based on User's Browsing Behavior", Predictive Modeling of Web Users, 2007, ACM 978-1-59593-654-7/07/0005
- [3] Faliang HUANG, "An Approach to Mining Social Networks in Chat Room", Journal of Computational Information Systems 2011
- [4] Masafumi Hamamoto, "A Comparative Study of Feature Vector-Based Topic Detection Schemes for Text Streams", Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05) 0-7695-2414-1/05 © 2005 IEEE
- [5] Reihaneh Rabbany k., "Social Network Analysis and Mining to Support the Assessment of Online Student Participation".
- [6] Koosha Golmohammadi, "Data Mining Applications for Fraud Detection in Securities Market".
- [7] Alexandra BALAHUR, "A Feature Dependent Method for Opinion Mining and Classification", 978-1-4244-2780-2/08 ©2008 IEEE
- [8] Chun-hung Li, "Sentence Factorization for Opinion Feature Mining", 2009 International Conference on Computational Aspects of Social Networks 978-0-7695-3740-5/09 © 2009 IEEE
- [9] Peter Koncz, "An approach to feature selection for sentiment analysis", INES 2011, 978-1-4244-8956-5/11 ©2011 IEEE
- [10] Amir A. Sheibani, "Opinion Mining and Opinion Spam-A Literature Review Focusing on Product Reviews", 6th International Symposium on Telecommunications (IST'2012), 978-1-4673-2073-3/12 ©2012 IEEE
- [11] Ms.K.Mouthami, "Sentiment Analysis and Classification Based On Textual Reviews".
- [12] Mostafa Karamibekr, "A Structure for Opinion in Social Domains", SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013, 978-0-7695-5137-1/13 © 2013 IEEE
- [13] Rabeah Al-Zaidy, "Mining Criminal Networks from Unstructured Text Documents", 2012
- [14] Farkhund Iqbal, "Mining Criminal Networks from Chat Log", 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 978-0-7695-4880-7/12 © 2012 IEEE
- [15] Ching-Yung Lin, "SmallBlue: Social Network Analysis for Expertise Search and Collective Intelligence", IEEE International Conference on Data Engineering, 1084-4627/09 © 2009 IEEE