ISSN:0975-9646

V.Sreenivasulu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2856-2858

Bi-Variate Gaussian Mixture Model Based Techniques for Effective Analysis of Cyber Crimes

V.Sreenivasulu
Department of Computer Science and Engineering
Gandhiji Institute of Science and Technology,
Krishna District - Andhra Pradesh - India

Dr. R. Satya Prasad
Department of Computer Science and Engineering
Acharya Nagarjuna University - Guntur,
Andhra Pradesh - India

Abstract **- With the wide usage of internet, communication by means of textual messaging has become a trendy. This type of electronic conversation is mostly used in messaging services like chat servers, description forms, emails, news groups and relay chats. These services regularly produced huge quantity of textual data there by, it is obligatory to use remaining techniques for identifying related patterns. The technological developments helped the law breakers to process illegal activities. In this paper a methodology is presented to identify the criminals by analysing the chat messages based on Bi-Variate Gaussian Mixture Model.**

Keywords: Chat messages, mining patterns, Bi-Variate Gaussian Mixture Model , textual messages.

I. INTRODUCTION

The recent development in science and technology has helped towards development of Communication technology. Lot of communication devices and media have been evolved. This communication technology helped to broadcast the messages across the globe with a minimum time stamp. At the same time the technology is equally utilized by the law breakers to execute the illegal activities. Therefore, it is the need of the hour to safe guard the national security. Many models have been addressed towards the issue. Among these communications, textual communications have become easier. Messages like Emails and Chatting are mostly preferred. Since the number of messages across the globe is increasing anonymously, effective methods are to be developed to analyse these messages. To analyse the textual messages called chat transcripts a concept of coding is mostly preferred in this process. Each chat like in the message is analysed and grouped with a specific class label. Chat transcripts are the sequence of words used during the chatting session by a particular actor. In the log file, these sessions are stored with a particular label.

In the chat conversation, the messages communicated between the users during the chat period are called snippets. In general in the chat messages the communication takes place in one to one fashion or one to many fashions. In order to initiate the chatting the user needs to identify the communication to which group the message is to be sent. This messaging helped the law breakers to adopt methodologies for processing law breaking issues. Many models are presented in the literature based on the classification and clustering techniques [1] [2] [3] [4] [5]. By considering the problem a mere classification technique where the classification is based on either the community on the time stamp as a feature. However, some researches considered this stylometric feature of the chatting persons to analyse the user [6] [7] [8]. In order to retrieve the related user, methodologies based on PCA, LDA were also highlighted in the literature [9] [10]. However, in most of the literature, the study is more focused towards an assumption of considering unit features and there by using the features for classification. However, to retrieve more meaningful patterns and assess the criminal investigation, it is necessary to formulate the relationships between the chatting communication and the author. There fore, Bi-variate models will be more advantage than uni-variate model for effective analysis. Hence, in this paper a methodology is based on Bi-Variate Gaussian Mixture Model is presented to analyse the chatting messages and to identify the law breakers. The rest of the paper is organized as follows. In Section–2 of the paper Bi-Variate Gaussian Mixture Model is presented. Section–3 of the paper deals with the dataset considered. In Section–4, the methodology is presented and deals with the experimentation together with the results derived. In Section-6 the conclusion is given.

II. FINITE BI-VARIATE GAUSSIAN MIXTURE MODEL

Every Chatting dataset is a collection of several chat messages. Every message in the database is quantified by group. In order to model the prime investigation process each chatting message group and the participating actor are considered to follow Bi-Variate Gaussian Mixture Model. The probability density of the Bi-Variate Gaussian Mixture Model is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2(\sqrt{1-p^2})} e^{-\left[\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right]}$$

Where x_1, and x_2 are the chatting community and the chatting persons. σ1, σ2, μ1, μ2, denote the variance and the mean of the groups

III. DATASET

In order to present the proposed methodology a chat log dataset is considered from www.perverted-justice.com. This Dataset consists of chat logs from various instants message servers like Yahoo, Google etc. The chatting message conversion is mainly focussed on several issues and the chatting is involved between the youngsters, children and adults. The dataset consists of 250 log files and 500 users. It also contains 2000 lines of chatting messages from 20 different chatting sessions. The manual interpretation highlights 375 topic blocks. The original chatting messages

constitute of 10 classes and after the identification of the class labels. The most focused group identified are movie topics, personal topics, studies, finance, miscellaneous etc.

IV. METHODOLOGY

In order to use the dataset, the first step to be followed is the pre-processing of the feature sets has been identified. Stemming is applied together with stop words removal. The feature sets correspond to a line of chatting message. In order to extract meaningful information, the features play a vital role. However, considering each line having a chat message of a feature increases the time complexity and minimizing the accuracy of analysis. Therefore, effective methods are to be applied to minimize the features. The various techniques adopted are removal of suffixes and stop word removal. After processing the data each term in the chatting message against the classified groups is considered. The term frequencies are computed to identify the repeated patterns. In order to maximize the efficiency of the investigation process, it is necessary to identify only the most frequent terms. In general the criminals try to navigate by using un related terms and use of frequent words having a specific synonym. Therefore, in this model a methodology is highlighted to correlate the user along with the groups with highest frequent terms.

V. EXPERIMENTS AND RESULTS

Algorithm for investigation analysis.

Step1: Consider the chatting messages.

Step2: Preprocess the data to remove stopwords and stemming is applied for suffix removal.

Step3: Cluster the data to identify different groups.

Step4: Identify the most frequent word in each group.

Step5: Obtain the probability density function (pdf) using equation - 1 of Section-2 by considering the most frequent terms with the user as the variables.

Step6: Obtain the text data and process the step 2 to 5.

Step7: Compare the pdf of text data against the training data using KL Divergence.

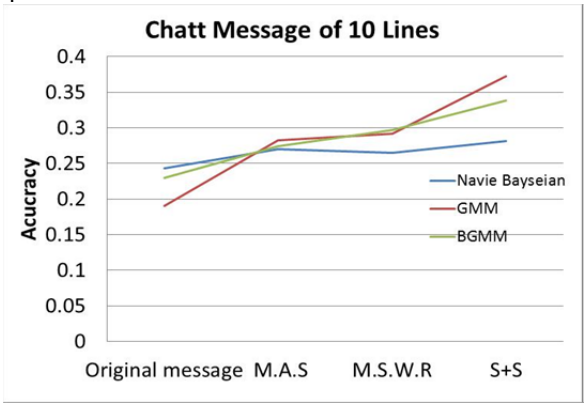
TABLE-1 : THE STYLISTIC FEATURES USED IN THE EXPERIMENTS

Feature category character usage	Features in the category frequency of each character	Possible feature values low, medium, high
message length	average message length	short, average, long
word length	average word length	short, average, long
punctuation usage	frequency of punctuation marks	low, medium, high
punctuation marks	a list of 47 punctuation marks	exists, not exists
stopword usage	frequency of stopwords	low, medium, high
Stopwords	a list of 88 stopwords	exists, not exists
smiley usage	frequency of smileys	low, medium, high
Smileys	a list of 29 smileys	exists, not exists
vocabulary richness	number of distinct words	poor, average, rich

TABLE-2 REPRESENTING THE NUMBER OF CLASSES AND THE NUMBER OF INSTANCES IN EACH CLASS

Text Set	No. Of Classes	No. Of Instances
1	4	22
2	12	17
4	3	20
6	12	22
7	4	12
5	17	24

From the above table2 presents the various classes together with the instances in the considered dataset. The data is processed and stopwords are removed, suffix is removed and the output derived is presented by considering a chat message of ten lines and is presented in the following graph



From the above graph, it can be seen that the accuracy of the recognition by the proposed model is far better than that of the model based on Gaussian Mixture Model and Naive Bayesian.

K.L Divergence

In order to correlate the emails from the text data and the training set, KL Divergence is used. The indexed data is given as input to the Gaussian Mixture Model (gmm) presented in equation (1) and the corresponding probability density function (pdf) is estimated. The process is repeated for all the data in the chat and all the chantings in the dataset. This is called training data. For testing purpose the same procedure is adopted and pdf of the query chat is generated. The pdf is compared with the pdfs in the training data using KL divergence (Khaul Loullie Formulae). It is used to identify the relation between the chats in the database and the email which happens to be false email i.e., the suspected email. MathCAD is used for the calculation purpose where the log of A and B are first calculated and then the integration is carried out. The MathCAD is used to perform the mathematical calculation of integration, differentiation and volume integrals. The alternative to K L Divergence is identifying the likelihood estimate. The formulae for calculating the KL divergence is given by

$$KL\ (A,B)\ = \int A(x) \log\left(\frac{A(x)}{B(x)}\right) dx$$

Where A(x) is the corpus of a person x and B(x) is the chat which is under consideration.

VI. CONCLUSION

In this paper a methodology is presented for analysing the fraud messages by proposing a model based on Bi-Variate Gaussian Mixture Model helps to identify the chat messages and helps to identify the chatting messages together with the actor. The methodology is compared with the models based on Gaussian Mixture Model and Naive Bayesian. From the results it can be clearly seen that the proposed model performs better than the existing models.

REFERENCES

[1] Noora.Al.Mutawa, et al (2012), Digital Investigation (Elsevier), available at http://www.dfrws.org/2012 /proceedings/DFRWS2012-3.pdf

[2] P.H. Adams, C.H. Martell, Topic detection and extraction in chat, in: Proceedings of the 2008 IEEE International Conference on Semantic Computing, IEEE Computer Society, 2008, pp. 581–588.

[3] E.R. Budlong, S.M. Walter, O. Yilmazel, Recognizing connotative meaning in military chat communications, in: Proceedings of Evolutionary and BioInspired Computation: Theory and Applications III, SPIE, 2009

[4] L.G. Boiney, B. Goodman, R. Gaimari, J. Zarrella, C. Berube, J. Hitzeman, Taming multiple chat room collaboration: Real-time visual cues to social networks and emerging threads, in: Proceedings of the Fifth International ISCRAM Conference, ISCRAM, 2008, pp. 660–668.

[5] C.D. Berube, J.M. Hitzeman, R.J. Holland, R.L. Anapol, S.R. Moore, Supporting chat exploitation in DoD enterprises, in: Proceedings of the International Command and Control Research and Technology Symposium, CCRP, 2007.

[6] P.M. Aoki, M.H. Szymanski, L. Plurkowski, J.D. Thornton, A. Woodruff, W. Yi, Where's the "party" in "multi-party"? Analyzing the structure of small-group sociable talk, in: Proceedings of the Conference on Computer Supported Cooperative Work, ACM, 2006, pp. 393–402

[7] J. Bengel, S. Gauch, E. Mittur, R. Vijayaraghavan, ChatTrack: Chat room topic detection using classification, in: H. Chen, R. Moore, D.D. Zeng, J. Leavitt (Eds.), Proceedings of the Second Symposium on Intelligence and Security Informatics, in: Lecture Notes in Computer Science, vol. 3073, Springer, Berlin/Heidelberg, 2004, pp. 266–277.

[8] E. Bingham, A. Kabán, M. Girolami, Topic identification in dynamical text by complexity pursuit, Neural Processing Letters 17 (2003) 69–83

[9] Friedman, N., Geiger, D., and Goldszmidt, M. Bayesian network classifiers. Machine Learning, 29:131–163, 1997.

[10] Lafferty, J. D., McCallum, A., and Pereira, F. C., Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning,. 2001, Pp 282-289

[11] Lee, L.. Measures of distributional similarity. Proceedings of the 37th Annual Meeting of the Association For Computational Linguistics on Computational Linguistics Association for Computational Linguistics, Morristown, NJ, 1999, pp 25-32.

[12] Mccallum, A., and Nigam, K. A comparison of event models for naïve bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, 1998.

[13] Nigam, K., Mccallum, A. K., Thrun, S., Mitchell, T. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39:, 2000, pp 103-134

[14] llan, J. 2002.Topic detection and tracking: event-based information organization Kluwer Academic Pub. 1–16.