Automated Cyberbullying Detection using Clustering Appearance Patterns

Walisa Romsaiyud[1], Kodchakorn na Nakornphanom[2] , Pimpaka Prasertsilp[3], Piyaporn Nurarak[4], Pirom Konglerd[5]

Walisa.rom@stou.ac.th[1], Kodchakorn.nan@stou.ac.th[2], Pimpaka.pra@stou.ac.th[3], Piyaporn.nur@stou.ac.th[4], Pirom.kon@stou.ac.th[5]

[1,2,3,4,5] School of Science and Technology,
Sukhothai Thammathirat Open University, Chaengwattana Rd., Bangpood, Pakkret,
Nonthaburi, 11120, Thailand

Abstract— **Cyberbullying is an activity of sending threatening messages to insult person. To prevent cyber victimization from the activity is challenging. This paper enhanced the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering. The algorithm included two main methods: (1) creating partitions by iteratively relocating from entire datasets into clusters using k-mean clustering and (2) capturing any specific partition with the frequency of words with multinomial model feature vector and drawing the probability of words occurring in a document for predicting the eight classes. The proposed method resulted in increasing accuracy and reliability of an experiment.**

Keywords- Cyberbullying Prevention; Big Data Streaming; Naïve Bayes classifier; Pattern Clustering

I. INTRODUCTION

Due to the advances of internet and information technology, Online Social Network (OSN) services, such as Facebook, Twitter, and MySpace are gaining in popularity as a main source of spreading messages to other people. Messaging is widely used and very useful in various purposes, for example, business, education, and socialization. However, it also provides opportunity to create harmful activities. There are numerous evidences showing that messaging can introduce the very concerned problem, namely cyberbullying.

Cyberbullying involves the offensive information such as harassment, insult, and hate in the messages which are sent or post using OSN services for the purpose of intentionally hurting people emotionally, mentally, or physically [1]. It can cause low self-esteem, anxiety, depression, a variety of other emotional problems, and even suicide [2]. Its tragic consequences have continuously reported typically among the school-age children. Since the number of cyberbullying experiences has recently been increasing [3], an intensive study of how to effectively detect and prevent it from happening in real time manner is urgently needed. To prevent victims from the incidents, blocking the message is not an effective way. Instead, texts in the messages should be monitored, processed and analyzed as quickly as possible in order to support real time decisions.

As the problems mentioned, a number of studies are dedicated to explore various techniques to detect cyberbullying efficiently. Manual detection is considered the most accurate detection, but it is hardly employed because it takes too much time and lots of resources. Automatic cyberbullying detection system is therefore emphasized. Text mining technology and technique are mostly used [4]. Karthik [5] applied multi classifiers such as Naïve Bayes, JRip, J48 and SMO with YouTube comments. Vinita [6] used LDA to extract features and employ weighted term frequency-inverse document frequency (TF-IDF) function to improve the classification with datasets from Kongregate, Slashdot, and MySpace. Homa [7] used Support Vector Machines classifier with datasets from Instagram and Romsaiyud [8] applied the Expectation-Maximization (EM) for clustering documents from data streams for threat cyberbullying detection.

Even though cyberbullying detection system has extensively been exploring, cyberbullying remains a growing concern and the existing approaches are still inadequate especially when dealing with a large volume of data. Various kinds of OSN services can represent different forms or patterns of data. Additionally, reduction in computation time becomes very crucial. The detection of cyberbullying is therefore still challenging.

This paper, we developed an automatic cyberbullying detection system to detect, identify, and classify cyberbullying activities from the large volume of streaming texts from OSN services. Texts are fed into cluster and discriminant analysis stage which is able to identify abusive texts. The abusive texts are then clustered by using K-Mean. Naïve Bayes is used as classification algorithms to build a classifier from our training datasets and build a predictive model. Moreover, we also used Naïve Bayes to classify the abusive texts into one of the eight pre-defined categories. The categories include activities approach, communicative, desensitization, compliment, isolation, personal information, reframing, and relationship. Our proposed approach consists of two main methods. The first method aims to clean and pre-process our datasets by removing non-printable and special characters, reducing the duplicate words and clustering the datasets. The second one concerns

978-1-4673-9077-4/17/$31.00 ©2017 IEEE

classification model to predict the text messages for preventing cyberbullying.

In our experiment, the datasets of streaming texts are collected from two different types of sources: (1) posted messages by members in Perverted-justice used as training datasets, and (2) Twitter datasets from Stanford University as testing datasets. The main goal of our research is to improve accuracy of our automatic cyberbullying detection system and to process and analyze texts in nearby real time. However, the cyberbullying is detected. The faster the action is taken, the more the victims are safe.

The rest of the paper is organized as follows. Section II presents related researches on techniques used in automatic cyberbullying detection system. Section III describes our proposed methodology, and then followed by the detail of our experiment and its results as well as discussion on its potential and limitations in Section IV. Section V presents the conclusion with proposed ideas and how to improve our system as future work.

II. LITERATURE REVIEW

This section reviews the relevant researches related to cyberbullying prevention, big data streaming, Naïve Bayes classifier, and pattern clustering.

A. Cyberbullying Prevention

According to the accumulated literature, cyberbullying can be defined as the use of electronic forms of communication by an individual or group to involve repeatedly in posting or forwarding content about an individual or group that a regular person would consider threatening, embarrassing, or harassing [9-11]. Some of the most common forms of cyberbullying are forwarding someone's private e-mails to others, sending threatening or inappropriate messages via instant messaging, and hacking into another person's account to send unsuitable content to others via social network sites. In addition, cyberbullying is especially deceptive because it affords a degree of anonymity and the opportunity to reach a greater amount of victims without a substantial threat of penalty.

There are a number of empirical studies regarding cyberbullying prevention that disclose the approaches for state and local governments, students, and families to use when addressing cyberbullying [12-13]. These approaches are classified into three main categories: (1) laws and regulations to control the use of the Internet and to establish policies related to cyberbullying and other methods of abuse, (2) curricular programs considered to educate children and adolescence about safe using of the Internet and electronic media and how to avoid and address the consequences for cyberbullying, and (3) technological approaches to prevent the cyberbullying. This paper falls into the last category in which using Naïve Bayes classifier as a technological method, to extract the texts and examine clustering pattern, and to monitor text streams transmitted via the Internet.

B. Big Data Streaming

The concept of a data stream is more suitable than a dataset. Generally, a stored dataset is a proper model when significant portions of the data are queried repeatedly, and updates are small and quite rare. In the opposite, a data stream is a fit model when a huge volume of data is arriving uninterruptedly and it is unreasonable to store the data in some forms of memory. In addition, data streams are applicable as a model of access to big datasets stored in secondary memory in which performance requirements need access through linear scans [13-14].

Past research in the data stream phenomenon has revealed the development of stream mining algorithms [15]. Henzinger et al. [15] reviewed the attraction of data stream to the attention of data mining community. They proposed two types of data stream mining algorithms: data-based and tasked-based techniques. There are a number of clustering, classification, frequency counting and time series analysis have been conducted grounded by these two techniques. They also considered to the significant issues of data stream pre-processing and data stream mining technology. Another group of researchers also indicated that the data pre-processing is the method of designing a light-weight pre-processing techniques that assure the quality of the mining results [16]. The challenge to this research was how a quality process was automated and integrated with the mining techniques.

C. Naïve Bayes Classifier

A Naïve Bayes classifier uses the concept of probability to classify large volume of data by finding models that differentiate classes of data [17-18]. Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other features. The model can be applied to several kinds of problems, such as real time prediction, multi class prediction, and text classification. A Naïve Bayes technique uses machine learning and data mining techniques to filter unseen information [19].

Naïve Bayes method compiled with text mining to detect the emotion classification of Twitter and tested on the validated model [20-21]. Wikarsa and Thahir [21], focused on security bug and presented a bug mining system for the identification of security and non-security bugs. Behl et al. [22] proposed a bug mining tool to identify and analyze security bugs using Naïve Bayes and TF-IDF. Naïve Bayes classifier also used to identify context of text documents by considering the context, which is very useful in information retrieval [21-23]. Nalini and Sheela used LDA with Naïve Bayes classifier to build a sentiment classifier for cyberbullying messages in Twitter [24].

D. Pattern Clustering

Clustering is a data mining technique to group data considered by data similarity. Each group of data is clustered based on the distance measured from data center and the deviation measurement. The popular clustering model is namely K-means. The clustering process is an unsupervised learning technique in which the clustering data is grouped without the target [25]. In addition, clustering has been used in various fields including text mining, pattern recognition, image

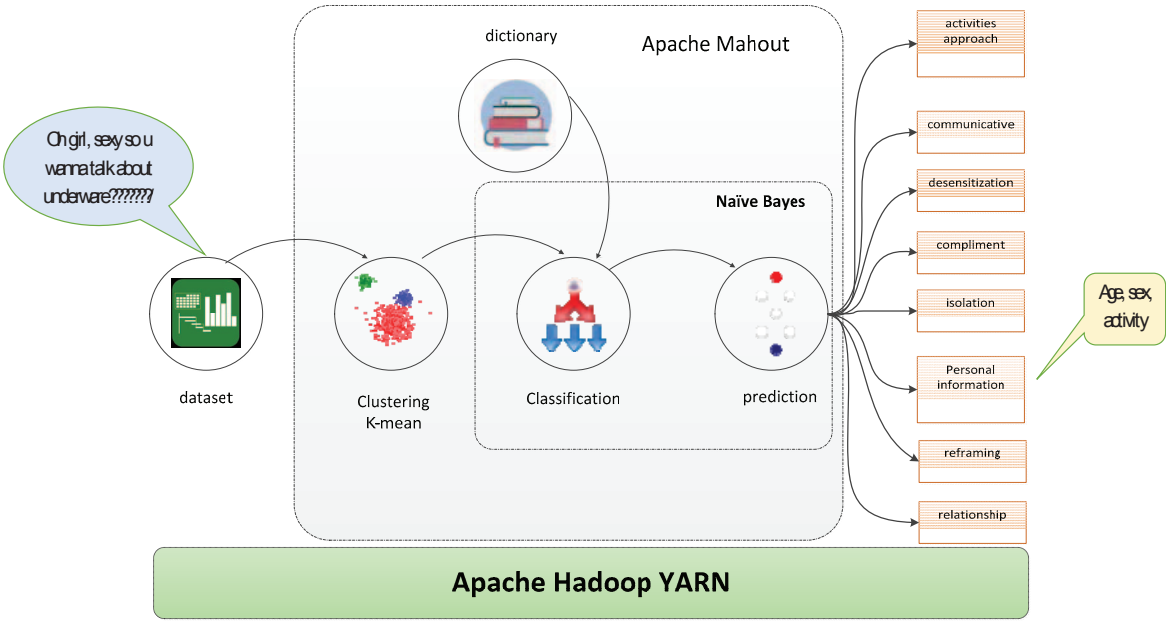


Figure 1. The architecture overview and proposed method

analysis, bioinformatics, machine learning, and image processing. To perform the task of clustering, a variety of data mining tools is freely available. These tools have their own features to efficiently generate clusters automatically for a given set of data [26]. Agnihotri et al. [27] used cosine similarity for measuring the distance among the words to mining the important information from texts. K-means and hierarchical agglomerative clustering algorithm are used to form the clusters. Clustering is also applied with the multi-modal crime patterns on the different modalities, which refer to the different data sources; offender data and weapon data. The GSVM has deployed for a hierarchical clustering approach by identifying patterns that exist in different levels of granularity for different modes of crime data [28].

III. THE ARCHITECTURE OVERVIEW AND PROPOSED METHOD

Figure 1 illustrates the system that consists of four major steps; preparing dataset, generating clusters, training data, and predicting data. The details of each step are as follows.

First step, preparing dataset, data sources comes from the CyberCrime Data and Twitter across cluster networks from data streams.

Second step, generating clusters, data sources are clustered the features two categories of the messages as polite messages and abusive messages, which the contents of the messages are identified, based on a crime pattern and the normalized documents using K-means clustering technique. The K-means clustering is a method of vector quantization, originally from signal processing, that is widespread for cluster analysis in data mining. It aims to do a partition of n observations into k clusters in which each observation belongs to the cluster with the nearest mean for finding the word pattern frequency with N-Gram in each sentence.

Third step, training data, the abusive partition is used to transform each into the feature extractor for classification technique using on Naïve Bayes. We assigned the feature extractor into eight sub categories including; activities approach, communicative, desensitization, compliment, isolation, personal information, reframing, and relationship. The pairs of feature sets and labeled data are fed into the Naïve Bayes classifier to produce a model.

Final step, predicting data, the feature sets are fetched into the model to produce predicted labels in eight categories.

For example, the message “Oh girl, sexy so u wanna talk about underware??????//” is considered containing abusive content by K-means and it is classify by Naïve Bayes into Personal Information category; age, sex, and activity.

IV. EXPERIMENTAL EVALUATION

For these experiments, we used the CyberCrime Data, which is a manually labeled dataset, for 170,019 posts from the perverted-justice dataset and Stanford University dataset that collect from Twitter web site from July 2012 to Jan 2013. There are 467 million Twitter posts from 20 million users that consist of author, time and content. The datasets can be divided into eight clusters as follows; (1) Activities approach, (2) communicative, (3) desensitization, (4) compliment, (5) isolation, (6) personal information, (7) reframing, and (8) relationship. In this section, the evaluation approach and the evaluation results are elaborated and discussed in detail.

A. DataSet and Expermental Settings

We used the CyberCrime Data and standard collection of Twitter dataset for training and testing the split data. Twitter include four direction networks that extracted from user activities including; re-tweeting, replying to existing tweets, mentioning other users, and being friends or followers as social relationships among users involved in the above activities and twitter information during the data collection.

TABLE I. TWITTER TEXT CATEGORIZATION COLLECTION DATASET

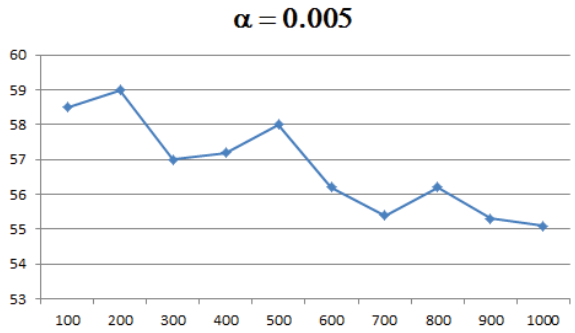
Data	Statistics
Number of users	17,069,982
Number of tweets	476,553,560
Number of URLs	181,611,080
Number of Hashtags	49,293,684
Number of re-tweets	71,835,017

Table I illustrates documents of the Twitter categorization collection dataset in which they consisted of number users, number of tweets, number of URLs, number of hashtags, and number of re-tweets during the seven months.

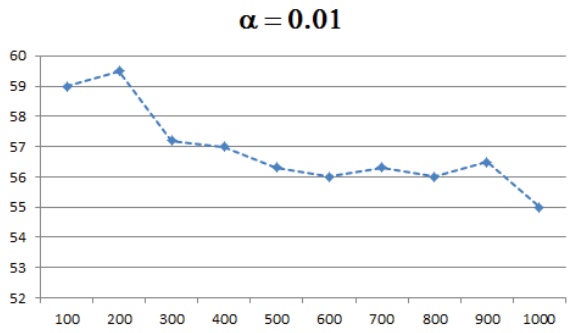
The experiment was tested on HP - Compaq HP z620 Workstation (12-DIMM slots) for master node. Our machine specification was as follows: the Intel C602 Chipset 8-channel ECC DDR3-1866 (Transfer rates up to 1866 MT/s), 4 channels per CPU, and hard disk 12TB. We run on Hadoop flume Version 1.5.0 for streaming data flows based on Apache YARN (for MR2) 2.3.0 framework. Hadoop Mahout 0.10.1 for executes the K-means and Naïve Bayes classifier algorithms.

B. Experimental Results

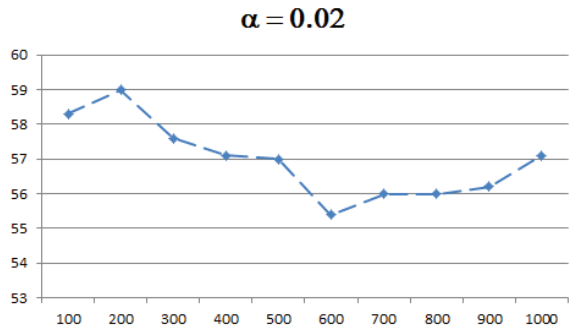
We conducted a maximum number of iterations of 20 (to make a fair comparison) for this entire algorithm. Each experiment is running ten times. We set the threshold of relevant strength between two words to 0.4. For each sentence, we did training the data to optimize the sentiment word strengths that covered spelling correction, negating word list, and repeating letters or punctuation and emotions. The K-means clustering was used to split data sources into an abusive cluster.



(a) α = 0.005



(b) α = 0.01



(c) α = 0.02

Figure 2. (a), (b) and (c) Naïve Bayes classification accuracy against feature set size for various α values

From Figure 2 (a), (b) and (c), the maximum number of iterations is 100 and the feature set is 200 features. There are three α values being tested in this works, which are α = 0.005, α = 0.01 and α = 0.02. Naïve Bayes was used for outperforming without the extended feature set by a lower margin of 0.7%. It performs less well than N-gram *("I love you", "I love", "love you" and "I", "love", you")* with adding the pattern terms. Naïve Bayes performs the classification accuracy = 95.79% is already very high when using all features. All except accuracy are statistically significant difference. In terms of α value, 0.02 tends to perform almost uniformly better than other values for this data set. When α = 0.02, relevant measure will contribute more than redundant measure, which means that the extracting method will focus more on improving classification accuracy than reducing the number of features.

We guaranteed the abusive messages by considering the recurring patterns within the whole datasets. *K-mean* clustering patterns applied for assigning the partitions of a set of objects into the *k* subclasses.

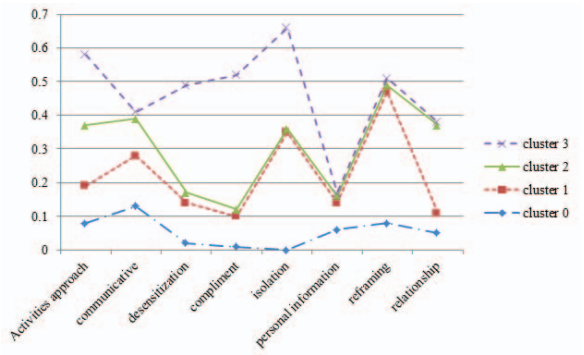


Figure 3. The eight subclasses from four-category clusters

Figure 3 shows the results of a dataset that was collected during 2012-1013 and categorized into eight classes based on four clusters. We found that *k=4* produced the best results that follow from the Cyberbullying type (Child Exploitation, Money Laundering, Internet-based Drug Trafficking, and Mass Marketing Fraud). Based on the results in Figure 3, the isolation, communicative, and reframing were the three highest score on normalized activity count. For example, both of the contributions "are you alone" and "do you have many friends" are considered under the category of "isolation."

TABLE II. EIGHT SUB CATEGORIES

Categories	Sentences
activities approach	"Are you safe to meet"
communicative	"I just want to meet and mess around"
desensitization	"if I don't cum right back"
compliment	"You are a really cute girl"
isolation	"Do you have many friends"
personal information	"Oh girl, sexy so u wanna talk about underware??????//"
reframing	"Let's have fun together"
relationship	"are you married yet"

Table II illustrates the sample coding dictionary [29] used in these experiments contained 475 unique phrases. For example, we assigned the sentences as "Let's have fun together", "Let's play a make believe game", "there is nothing wrong with doing that" for reframing category.

As illustrated in Figure 4, the codebook used as the baseline for defining the conversation dialogs based on the eight sub categories that were previously tested with PV1 and SafeChat. In order to improve the accuracy, we executed ten times and repeated on the labeled datasets (i.e., Twitter and CyberCrime) and on the unlabeled datasets based on eight subclasses. Considering the results of the Figure 4, we realized that when using PV1, the accuracy for clustering the conversation dialog (in each sub category) for the labeled dataset was approximate to SafeChat and for the unlabeled dataset was increased 33.48%, respectively.

Figure 4. Based line result for verification of PV1

Knowing that many other studies (previously conducted in this field) merely used labeled datasets collected from SafeChat, we aimed to compare our results based on both labeled and unlabeled datasets through PV1. Consequently, the results of PV1 and the Codebook baseline approaches from SafeChat were less accurate dealing with the unlabeled datasets, compared to the labeled datasets.

V. CONCLUSION AND FUTURE WORK

The main objective of the paper is to partition abusive messages from big data streaming with K-means clustering method on top of word probabilistic on each document for determining the similarity sentences score based on the improving accuracy and computation time. The paper proposes a novel method which can generate a predictive model from large volume of data sets for supporting the analysis services on business. The primarily study is to enhance the features of a Naïve Bayes classifier for extracting the words and generating model on text streaming. Moreover, a local optimum is guaranteed by our proposed method. The method was executed on CyberCrime Data, which is a manually labeled dataset, for 170,019 posts and Twitter web site for 467 million Twitter posts.

The empirical results demonstrate that our method; (1) is able to classify abusive messages from sentences frequency by using statistics score and partition data sources, and (2) is capable of classification and prediction model from the feature sets into eight sub categories as; activities approach, communicative, desensitization, compliment, isolation, personal information, reframing, and relationship. However, in future we will study more on the streaming K-mean clustering using Apache Spark for increasing a performance of computation time and cost on the different data types from many data sets.

ACKNOWLEDGMENT

The author would like to acknowledge the financial support of this work by grants from Institute for Research and Development, Sukhothai Thammathirat Open University, under the IRD 0522.25/58 program on 2015.

REFERENCES

[1] Cyberbullying Research Center, 'What is Cyberbullying?', 2016. [Online]. Available: http://cyberbullying.org/. [Accessed: 10-Jul-2016].

[2] R.M. Kowalski and S.P. Limber, “Psychological, Physical, and Academic Correlates of Cyberbullying and Traditional bullying,” J. Adolescent Health, 2013, vol. 53, no. 1, pp.513-520.

[3] Cyberbullying Research Center, 'Summary of Our Cyberbullying Research (2004-2016)', 2016. [Online]. Available: http://cyberbullying.org/summary-of-our-cyberbullying-research. [Accessed: 10-Jul-2016].

[4] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, “Improving cyberbullying detection with user context,” Advances in Information Retrieval, Springer, 2013, pp.693-696.

[5] D. Karthik, R. Roi, and L. Henry, “Modeling the detection of textual cyberbullying,” International Conference on Weblog and Social Media - Social Mobile Web Workshop, 2011.

[6] N. Vinita, L. Xue, and P. Chaoyi, “An Effective Approach for Cyberbullying Detection,” Communications in Information Science and Management Engineering, 2013, vol. 3, no. 5, pp.238-247.

[7] H. Homa, A. M. Sabrina, I. R. Rahat, H. Richard, L. Qin, and M. Shivakant, “Detection of Cyberbullying Incidents on the Instagram Social Network,” 2015.

[8] W. Romsaiyud, “Expectation-maximization algorithm for topic modeling on big data streams,” IEEE 7th International Conference on Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, USA, 2016, pp. 1-7.

[9] K. L. Mason, “Cyberbullying: A preliminary assessment for school personnel,” Psychology in the Schools, 2008, vol. 45, no. 4, pp.323-348.

[10] J. W. Patchin and S. Hinduja, Eds, Cyberbullying prevention and response: Expert perspectives. Routledge, 2012.

[11] H. Vandebosch and K. Van Cleemput, “Defining cyberbullying: A qualitative research into the perceptions of youngsters,” CyberPsychology & Behavior, 2008, vol. 11, no. 4, pp.499-503.

[12] T. Diamanduros, E. Downs, and S. J. Jenkins, “The role of school psychologists in the assessment, prevention, and intervention of cyberbullying,” Psychology in the Schools, 2008, vol. 45, no. 8, pp.693-704.

[13] R. Ortega-Ruiz, R. Del Rey, and J. A. Casas, “Knowing, building and living together on internet and social networks: The ConRed cyberbullying prevention program,” International Journal of Conflict and Violence (IJCV), 2012, vol. 6, no. 2, pp.302-312.

[14] S. Stauffer, M. A. Heath, S. M. Coyne, and S. Ferrin, “High school teachers' perceptions of cyberbullying prevention and intervention strategies,” Psychology in the Schools, 2012, vol. 49, no. 4, pp.352-367.

[15] M. R. Henzinger, P. Raghavan, and S. Rajagopalan, “Computing on data streams,” vol. 198. Technical Note 1998-011, Digital Systems Research Center, Palo Alto, CA, 1998.

[16] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Data stream mining." Data Mining and Knowledge Discovery Handbook. Springer US, 2009, pp.759-787.

[17] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, “Streaming-data algorithms for high-quality clustering,” In ICDE vol. 2, pp. 685, March 2002.

[18] D. T. Larose, Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, 2014.

[19] M. Bramer, “Principles of data mining,” London: Springer, 2007, vol. 131.

[20] Analytics Vidhya, ‘6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python)’, 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/.[Accessed: 10-Jul-2016].

[21] Wikarsa and S. N. Thahir, "A text mining application of emotion classifications of Twitter's users using Naive Bayes method," 1st International Conference on Wireless and Telematics (ICWT), Manado, 2015, pp. 1-6.

[22] D. Behl, S. Handa, and A. Arora, "A bug Mining tool to identify and analyze security bugs using Naive Bayes and TF-IDF," International Conference on Reliability Optimization and Information Technology (ICROIT), Faridabad, 2014, pp. 294-299.

[23] A.R. Kulkarni, V. Tokekar and P. Kulkarni, "Identifying context of text documents using Naïve Bayes classification and Apriori association rule mining," CSI Sixth International Conference on Software Engineering (CONSEG), Indore, 2012, pp. 1-4.

[24] K. Nalini and L. J. Sheela, Classification using Latent Dirichlet Allocation with Naive Bayes Classifier to detect Cyber Bullying in Twitter, Indian Journal of Science and Technology, 2016, vol. 9, no.28, pp. 1-5.

[25] K. M. A. Patel and P. Thakral, "The best clustering algorithms in data mining," International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, Tamilnadu, India, 2016, pp. 2042-2046.

[26] P. Aalam and T. Siddiqui, "Comparative study of data mining tools used for clustering," 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 3971-3975.

[27] D. Agnihotri, K. Verma and P. Tripathi, "Pattern and Cluster Mining on Text Data," Fourth International Conference on Communication Systems and Network Technologies, Bhopal, 2014, pp. 428-432.

[28] Y. L. Boo and D. Alahakoon, "Mining Multi-modal Crime Patterns at Different Levels of Granularity Using Hierarchical Clustering," International Conference on Computational Intelligence for Modelling Control & Automation, Vienna, 2008, pp. 1268-1273.

[29] B. Michael and K. Jacob, “Text Mining Applications and Theory”, in “Text mining and cybercrime” (Ed. K. April, E. Lynne and L. Amanda), John Wiley & Sons, Ltd., UK, 2010, Ch. 8.