DIGITAL CYBER FORENSIC EMAIL ANALYSIS AND DETECTION BASED ON INTELLIGENT TECHNIQUES

Sally D. Hamdi [1], **Abdulkareem M. Radhi** [2]
[1,2] College of Information Engineering, Al-Nahrain University, Baghdad, Iraq
{sally.dakhel[1], akmurhij} @coie-nahrain.edu.iq [2]
Received:22/10/2019, Accepted:20/1/2020

***Abstract*- The Internet has become open, public and widely used as a source of data transmission and exchanging messages between criminals, terrorists and those who have illegal motivations. Moreover, it can be used for exchanging important data between various military and financial institutions, or even ordinary citizens. One of the important means of exchanging information widely used on the Internet medium is the e- mail. Email messages are digital evidence that has been become one of the important means to adopt by courts in many countries and societies as evidence relied upon in condemnation, that prompts the researchers to work continuously to develop email analysis tool using the latest technologies to find digital evidence from email messages to assist the forensic expertise into to analyze email groups. This work presents a distinct technique for analyzing and classifying emails based on data processing and extraction, trimming, and refinement, clustering, then using the SWARM algorithm to improve the performance and then adapting support vector machine algorithm to classify these emails to obtain practical and accurate results. This framework, also proposes a hybrid English lexical Dictionary (SentiWordNet 3.0) for email forensic analysis, it contains all the sentiwords such as positive and negative and can deal with the Machine Learning algorithm. The proposed system is capable of learning in an environment with large and variable data to test the proposed system will be select available data which is Enron Data set. A high accuracy rate is 92% was obtained in best case. The experiment is conducted the Enron email dataset corpus (May 7, 2015 Version of the dataset).**

keywords: **Digital forensic, Mining, SentiWordNet 3.0, Clustering, SWARM, Classification.**

I. INTRODUCTION

Email appears as a very important application on the Internet for data communication, which is utilized not only by computers but also by numerous electronic devices [1], such that it is a common way to communicate between parties and it transfers information between servers on a specified port number. Typically, email composed using client-side applications with identity sender, then store as a file and delivered to the destination user through one or more servers. Some individuals have found ways to exploit Email for malicious purposes although e- mail connections are designed to make things simple, powerful and effective [2]. The popularity and the low cost of e-mail made it the medium of choice by criminals or persons having mischievous intent [3]. It is one of the most important resources of many criminal behaviors on the Internet. Email analysis is challenging due to not only various fields that can be forged by hackers and the wide range of email applications is used, but also due to imposed law restrictions in the analysis of email [4]. Email messages are digital evidence which have become one of the important means to adopt by courts in many countries and societies as evidence relied upon in condemnation. The problem of gathering significant evidence against adversaries by examining suspected e- mail accounts to identify the most appropriate author from group of potential suspects. Cyber forensics apparatus is enhancing its hardiness and offset these inexorable threats [5]. In the last decade's digital forensics has become a prominent activity in modern investigations. Seized digital devices can provide precious information and evidences about facts and/ or individuals on which the investigational activity is performed. Due to the complexity of this inquiring activity and to the large amount of the data to be analyzed, the choice of appropriate digital tools to support the investigation represents a central concern [6].

Data Mining is an application of algorithms to extract patterns of information and to make the useful information available in management and has a number of applications in Digital cyber forensics. It includes discovering and classification the forensic information in groups based on relationships, identifying relationships in forensic association, detects patterns in information that leads to helpful forecasting and detects groups of hidden facts [7]. So, these machine learning techniques have been widely used to extract evidence from large email groups. This can help the forensic researcher, to perform a multi- stage analysis of email groups.

Aimes of the proposed method:

- Deal with real- life e- mail dataset (Enron corpus) .
- Design and Implement a reliable system that able to classify emails into different categories (malicious or normal) .
- Optimize of the feature selection method.
- Achieve better performance of the proposed system.

II. Related Work

The researchers provide several of literature of similar work to analysis and classify emails or texts, which focused on classifying in to different categories by using different ways. The researchers' work that related to the work such as: Farkhund Iqbal and et. al, in 2010, [8], demonstrated a problem of how to extract writing style from set of email messages written by many anonymous authors. To be to solve this problem by using clustering method. Sobiya R. Khan et. al, in 2012, [9], discussed applying data mining technique to realize many functions for the implementation of the statistical analysis of e- mail, the clustering, and classification of e-mail, the identification of an e- mail author, and the analysis of the social network of email. Sobiya R. Khan, et. al, in 2013, [10], proposed method to analyze and classify emails are made using a decision tree classifier that showed promising results. Emad E. Abdallah, in 2013, [11], presented an analysis and investigation of anonymous mining email content and suspect writing style. In this work used features stylometric and Machine Learning technique. A major contribution of this work is to reduce the training time by extracting some effective features. Harsh Vrajesh Thakkar,2013, [12], combined the English dictionary SentiWordNet3.0 with existing machine learning naive bayes classifier algorithm which that classifies tweets in positive and negative classes respectively. Prachi K. Khairkar et al. ,2015, [13], applied a method for analyzing official documents on a computer. Clustering techniques cascaded with a support vector machine to improve system performance and quality. V. Sreenivasulu et al. , 2015, [14], proposed semantic ontologies and the " Gaussian Mixture Model" of data mining models to investigate cybercrime, and to analyze massive email forensic clusters. The purpose of the semantic tool (WORDNET) is to analyze grammatical rules and determine the term index and abbreviations. This methodology can be very important in investigation from emails. Nirkhi, S, et al. ,2016, [15] explored the application of unsupervised techniques (multidimensional scaling techniques and cluster analysis) to solve the problem of author verification. Sofea Azrina Azizan and et. al. ,2017, [16], proposed improvise of the sentiment analysis by using machine learning to detect the acts of terrorism more accurately. Shahad Fadhil Abbas,2018, [17], proposed statistical methods for displaying " a template matches the person correlation coefficient" the proposed work utilizing " the back propagation neural network" as a tool for analyzing Enron email messages and viber messages to

identify the sender.

The following are the main conclusions from previous works:

1) Depending on Enron email messages dataset for experimental purpose.
2) Using of stylometric features is generally suitable.
3) For E- mail Mining, in most cases, the machine learning algorithm of this technique is employed.

III. THEORETICAL BACKGROUND

A. K- means clustering algorithm

Clustering is a method of organizing a group of data into clusters and classes where the data reside inside a cluster are more similarity and data from two clusters will be different from each other. Here the two clusters can be considered disjointed. The main goal of clustering is to divide the entire data into multiple clusters. The tools mostly used in the clustering technique are k- mean, k- medoids, hierarchical, density- based and several other techniques. One of the most popular methods in clustering is the k- mean algorithm. The procedure follows a simple and easy way to classify a given set of data by a certain number of clusters (k clusters) constant. The core idea is to determine the centers of k, one for each cluster (group). These centers should be placed in a subtle way due to the different locations causing a different result. So the best option is to put them as far as possible away from each other. The next step is to take each point belonging to a particular dataset and connect it to the nearest center. When there is no hanging point, the first step is completed and an early group age is completed. Finally, this algorithm aims at minimizing an objective function know as a squared error function given by the following equation below [18].

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (\|X_i - Vj\|)^2 \tag{1}$$

Where X_i is data points, V_j is cluster centers, $\|X_i - V_j\|$ is the Euclidean distance between X_i and V_j, C_i is the number of data points in the ith cluster. 'c' is the number of cluster centers.

B. Naive bayesian algorithm

In classification techniques, current feature selection techniques choose features that are appropriate for the 0-1 classification. They do not take into account the accuracy of the class probability estimates provided by the classifier, which is important for numerous applications. Current methods of selecting features are naive Bayesian classifiers with the aim of obtaining accurate class probability estimates [19]. Naive Bayesian classifiers for each example X specify a score between 0 and 1 that can be interpreted, in principle, as an estimate of the probability of class membership. However, it is well known that these estimates are not accurate when the naive Bayesian assumption of conditional independence of features is violated (given the C- class category). Naive Bayesian classifiers the influence of predictor- X on given class- C is estimated, assuming that the predictors are independent with each other. The Bayes theorem is being applied to predict a class for any given text as in the following equation [20].

$$P(C|X) = P(X|C)P(C)/P(X) \tag{2}$$

Where $P(C|X)$ is the probability of class- c given predictor- x, which is called the posterior probability, $P(X|C)$ is the probability of predictor- X given class- c, P(C) is the probability of class- C whether it is positive or negative while P(X) is the prior probability of predictor- X. With n predictors, $P(X|C)$ is defined in the following equation [20].

$$P(X|C) = \prod_{k=1}^{n} P(X_k|C) \tag{3}$$

whereX_k is one feature value of X.

C. Particle swarm optimization

The particle swarm optimization (PSO) is a simple, effective and computationally efficient optimization algorithm. PSO is one of the swarm-based intelligence methods devised to find an optimum solution by imitating the behavior of flocks of birds and fish. It works through initializing a population of random solutions and searching for the optima by updating generations. All particles have a fitness value that is evaluated with the help of the fitness function and have a velocity that directs the movement of the particles [21]. Particle swarm optimization is initialized through a set of random solutions (particles), then searches for optimal by the generational update. In each iteration, each particle is updated by following two "best" values as was shown in the following equation [22].

$$X_{id}^{t+1} = X_{id}^{t} + V_{id}^{t+1} \tag{4}$$

The velocity of the particle i is represented as $V_i = (V_{i1}, V_{i2}, ..., V_{id})$ which is limited by a predefined maximum velocity, $V_{id}^{t} \in [-V_{max}, V_{max}]$. The equation for velocity is given by equation 5[23].

$$V_{id}^{t+1} = W * V_{id}^{t} + C_1 * r_{1i} * (P_{id} - X_{id}^{t}) + C2 * r_{2i} * (P_{gd} - X_{id}^{t}) \tag{5}$$

where t denotes the tth iteration, $d \in D$ denotes the dth dimension in the search space, W is the inertia weight, C_1 and C_2 are acceleration constants or sometimes called the learning rate, r_{1i} and r_{2i} are the random numbers uniformly distributed within the range of $[0, 1]$, P_{id} represent the element of the solution for the particle's individual best (pbest), and pgd is the element of the solution for the particle's global best(gbest) in the dth dimension[23].

D. Support vector machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that is utilized to classify and analyze data and estimate the relationship between variables. It is a supervised algorithm where there is an initial training stage where it feeds the algorithm data that has already been classified (labeled) [24]. Support Vector Machine is one of the most common classifiers that are based on a function of the linear discriminant. Essentially, it is suitable for binary classification. Support Vector Machine can be used two types of data: linearly separable and non-linearly separable. In the non- linear Support Vector Machine classifier, data is generally scattered. It is not appropriate to draw a straight linear plane to separate this data. For this problem, non-linear classification can be fulfilled by applying a kernel to hyperplanes maximum margin. Fig. 1 shows support vectors with a maximum margin hyperplane [25].

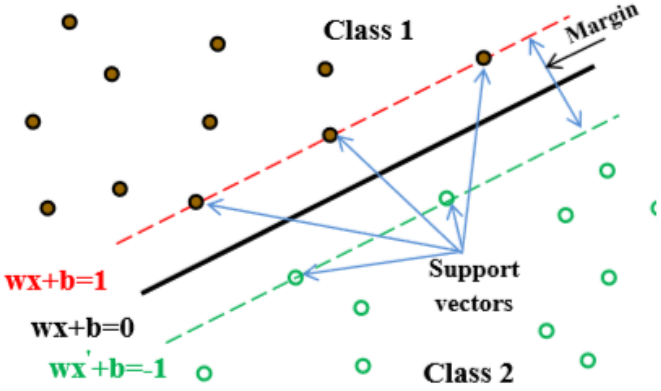


Figure 1: Block diagram of support vector machine [25]

Fig. 1 shows support vectors help creating separate parallel hyperplanes $(H1 : W \cdot X_i + b = 1)$ and $(H2 : W \cdot X_i + b = -1)$ that maximize the margin between the two classes. Intuitively, larger margins reduce the generalization error in the Support Vector Machine classifier. To construct an optimal hyperplane, it must first compute the weight vector by the following equation 6 which is a linear combination of support vectors [25]:

$$W = \sum_{i=1}^{n} a_i Y_i X_i \tag{6}$$

Then, the best hyperplane in the feature space can be defined as [25]:

$$W \cdot X_i + b = 0 \tag{7}$$

Where $(\cdot)$ represents the dot product, X is the row vector of the corresponding sample, w is the weight vector perpendicular to the hyperplane, and b is the bias. The system evaluation is an integral part of system development. The most straight forward and still widely used evaluation measure is the accuracy. Accuracy is calculated by " the percentage of correctly classified emails in the testing set" in the following equation [26].

$$Classification\ Accuracy = \frac{Number\ of\ Correctly\ Classified\ Identities}{Total\ Number\ of\ Identities} * 100\% \tag{8}$$

IV. Proposed Of The Email Forensic Analysis System

Machine learning can be considered as the most famous techniques having interest because of its accuracy and adaptability. For E -mail Mining, in most cases, the learning algorithm of this technique is employed. The system is composed of several different phases each phase has a specific function: Data preprocessing, Feature extraction, Clustering, Feature selection, Optimization, Classification, and then Prediction results. Four Machine learning algorithm used in this work are K- Means for clustering and Naive Bayes for feature extraction, Particle Swarm Optimization and Support Vector Machine for Classification. We optimize the selected features of the results which were improved for enhancing accuracy. Fig. 2 explained the frame work phases of our proposed research.

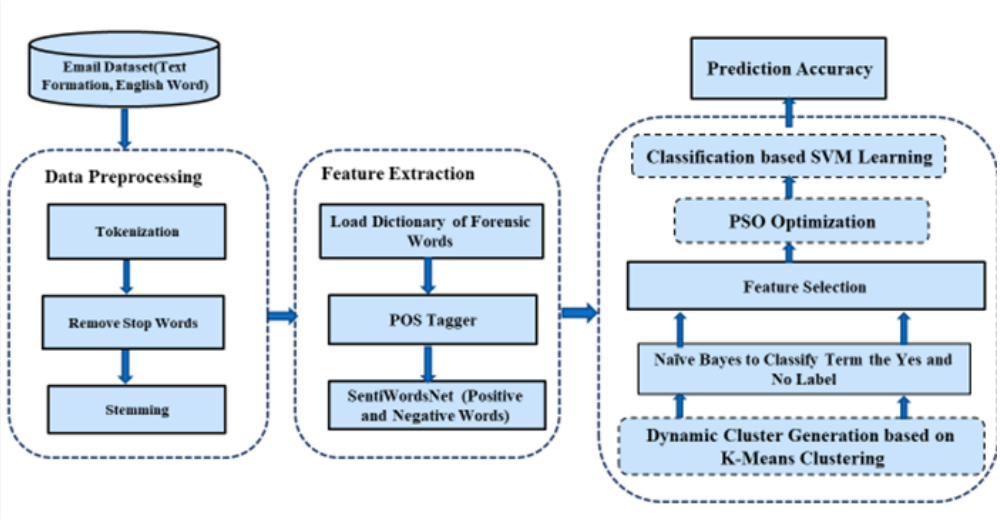


Figure 2: Block diagram of the proposed model

The system is composed of several different phases each phase has a specific function as shown in this section:

A. E- mail dataset for cyber forensic

Enron's corpus is used for the purpose of experimentation. Enron corpus was published during Enron's Corporation's legal investigation and turned out to have a number of integrity problems. This data is valuable. To my knowledge, it's the only large group of public " real" emails. The reason that other datasets are not public is because of the privacy concerns. The fact that data is considered to be real messages is another reason that the Enron dataset owns thousands of categories and different samples. The current version contains 619446 text messages in their original form belonging to 150 employees, mostly senior management of Enron Corporation, organized into folders. Enron email dataset is available on this website (https://www.cs.cmu.edu/enron/) . Fig. 3 presented email data selection, user details list and details content for each user.

B. Data preprocessing

Data preprocessing is a significant phase in the data mining technique. This is done to prepare data in a form that can be used in the next stages. There are many types of document representation, such as vector- model, graphical mode and so on. Many measurements are also be used for document weighting. Often, data preprocessing is the main phase in a project using machine learning. If there is much redundant and irrelevant information or data confusion and unreliable data, then discovering knowledge during the training stage will be more difficult. Data preprocessing stages are described in Fig. 4. The following section presents data preprocessing stages:

- Tokenization process

  The fundamental goal of the tokenization stage is separating the text of message into smaller components. The term " tokenization" is referred to the process of phase sentences to be divided into text streams to its constituent meaning,

as units called " tokens" or " words" . The proposed system is using java tokenizer such that each email message is converted into distinct words or tokens. The list of tokens becomes input for the next steps.

- Removal of stop words

  The next stage after the tokenization process is the removal of stop words. Stop words are common words that can be found in almost all text scripts. There is a need to remove these words because these words do not hold any useful information to help determine whether the e-mail message belongs to a specific classification or not. Elimination of stop words from the email message will reduce the feature space dimensions. The list contains about 488 stop words that are used in this work. Stop words available in this Website (https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords).

C. Feature extraction

Feature extraction is extracting all features which are given in the dataset. Feature extraction is the selection of those data attributes that best characterized a predicted variable. The forensic words are searched in the email dataset and POS were Tagging categories include noun, verb, adverb, and adjective then the score is calculated for each term by using SentiWordNet 3.0. The forward feature extraction method removes irrelevant features of the text and reduces the original feature set. Moreover, classification accuracy is increased while decreasing the time of the learning algorithm. Feature extraction performed on the email dataset after preprocessing stage and extract subjective features identified. This process term frequency for an individual feature is calculated, will calculate the forensic word frequency for each noun, verb, adverb, and adjective frequency. The following sections presents features extraction stages:

- Forensic words

  In forensics Analysis, emails are classified as malicious if their contents match up a specific cybercriminal classification. The list of criminals and crimes is long. Because these words have specific legal meanings, there is a need to know forensic vocabulary words. There are many specific words for different types of crimes and the criminals who commit them. To assist in knowledge more about the email forensic, a list of 647 forensic vocabulary words has been compiled. The Forensic words dictionary datasets are loaded then forensic words are searched in the email dataset for doing the analysis. Forensic words available in this Website (https://myvocabulary.com/word-list/crime-vocabulary/).

- POS tagging

  The Part- Of- Speech tagger is a tagging tool used to tag each word and assigns parts- of- speech to each " word" and another for the " token" . It distributes document or sentence and tags each term with a part-of-speech. The part-of-speech tagging uses the Stanford Part-Of-Speech tag. These tags are used by dividing text scripts into sentences and giving a part-of-speech tag for each token whether it is a " name" , " verb" , " adverb" or " adjective" . SentiWordNet3.0 modeled Part- Of- Speech. Example word has POS tagging (JJ, JJR, JJS, VB, VBD, VBG, VBN, VBP, and VBZ) represent of an adjective score and verb score and so as.

- SentiWordNet3.0

  SentiWordNet3.0 dictionary is an opinion lexicon mining from the WordNet database. Each token is related to numerical

scores representing positive and negative sentiment information SentiWordNet3.0. The purpose of this step analyzing the information presented in the email dataset and finds a score each term. The term frequency is calculated each term. SentiWordNet3.0 dictionary is available on this Website (http://sentiwordnet.isti.cnr.it/). Feature Extraction process shown in Fig. 5.

D. Clustering

The scores of each term were achieved. Based on scores clustering performed by using the k-means clustering algorithm. It will cluster (group) the information into two different clusters. In k-means clustering, the center point is defined. It is not dynamically generated in the process such that creates the center point node in k- means dynamically as depicted in Fig. 6.

E. Naive bayes algorithm

In this process e- mail messages are analysis either it is a positive or negative sense by using the naive algorithm for class probability estimation with feature extraction. The naive bayes classification algorithm is used for classifying the yes and no label. Yes, they represent positive scores. No, represent a negative score. Fig. 7 shows Naive Bayes and Extracted Features.

F. Optimization

The obtained result will be optimized to select the best feature by using a particle swarm optimization algorithm. The particle swarm optimization used to have the best prediction optimization for the selected features. The particle swarm optimization begins by randomly initializing the particle population (data attributes that best characterize a predicted variable). A whole swarm moves in the search space to find the best solution (fitness)by updating the position then calculate the velocity of each particle. The output from these phase best features (attribute) is forensic, noun, verb, adverb, and adjective attributes as shown in Fig. 8.

G. Classification

For the email dataset, the given set of emails is divided by randomly selecting into a training 70% of total emails and testing set 30% of total emails. That the ideal case from statistical view in clustering and training techniques is splitting data sets to 70% for training samples, 15% for validation and 15% for Testing. To check the effect of class labels on the accuracy of classifiers, that performed classification experiments for class labels. In this work implementation of SVM by using LIBSVM involves two steps: first, training a data set to obtain a model and second, using the model to predict information of a testing dataset. Fig. 9 depicts classification process. Training a given set of data affects the separator hyperplane produced by the classifier, which may lead to the problem of over- fitting or a biased decision towards the data samples involved in the training phase. To overcome this problem, the evaluation based process affects the performance of the classifier. The testing phase includes all these processes were carried out in the same way as the training phase of the model performance.

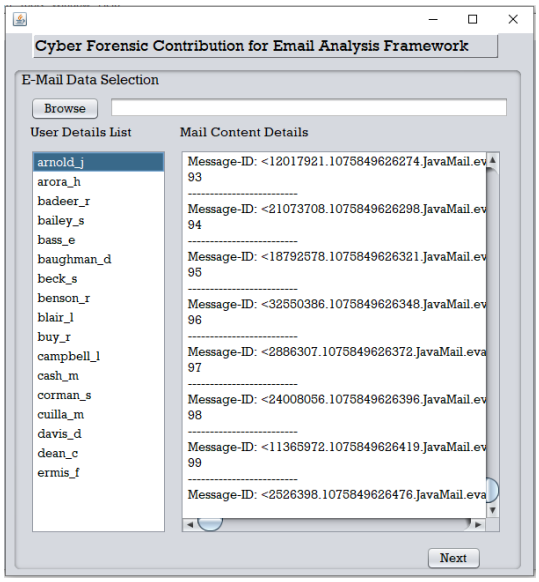


Figure 3: Data selection

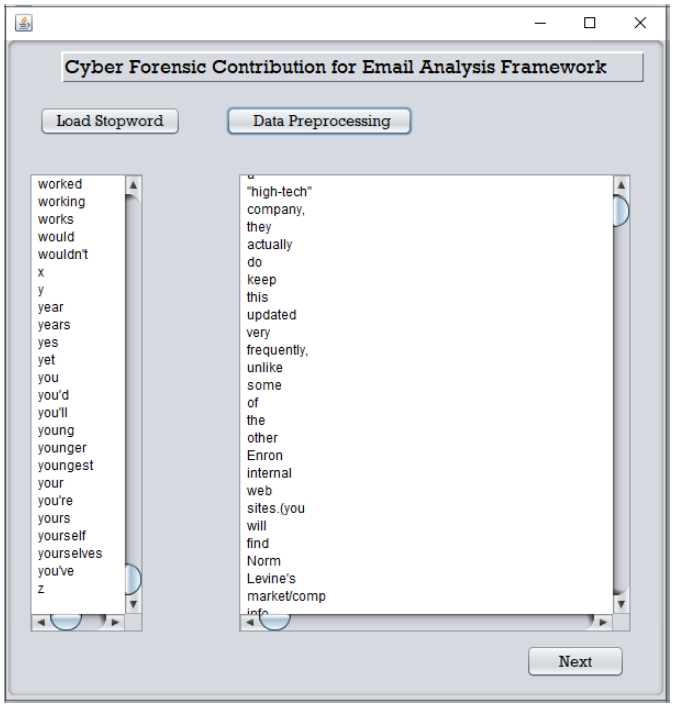


Figure 4: Data preprocessing phase

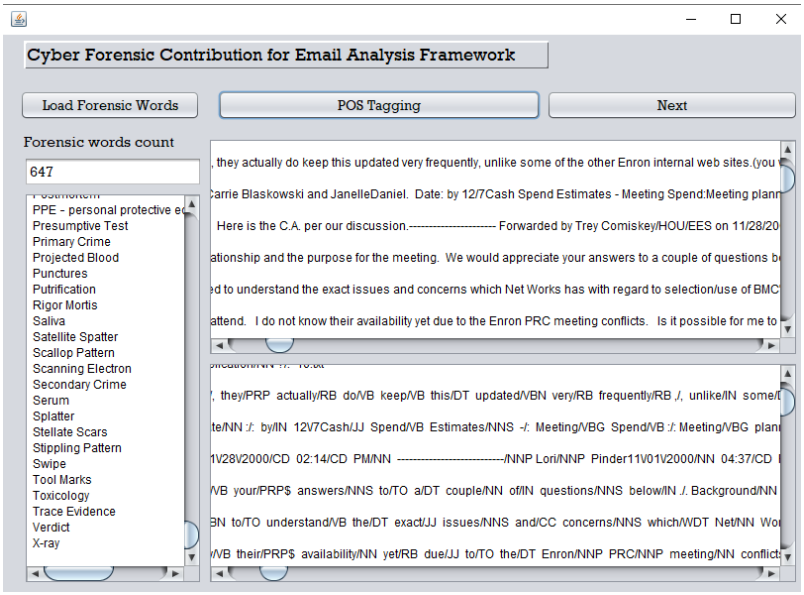


Figure 5: Feature extraction phase

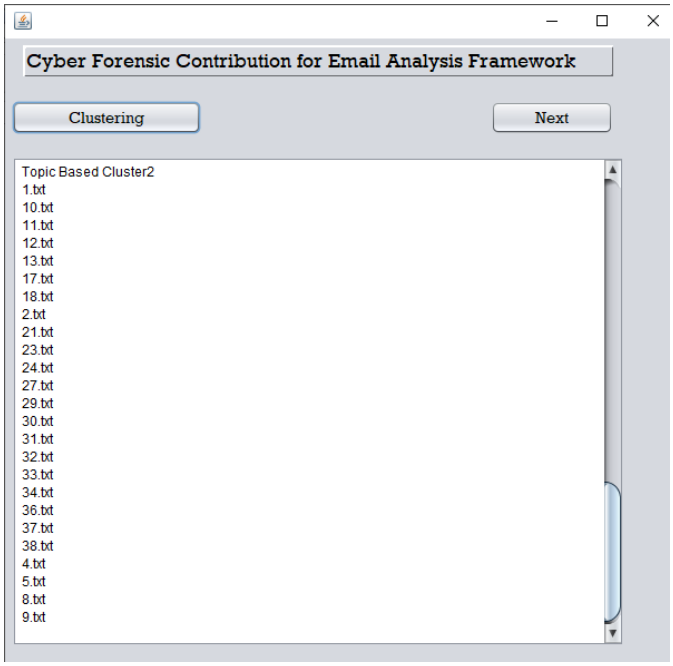
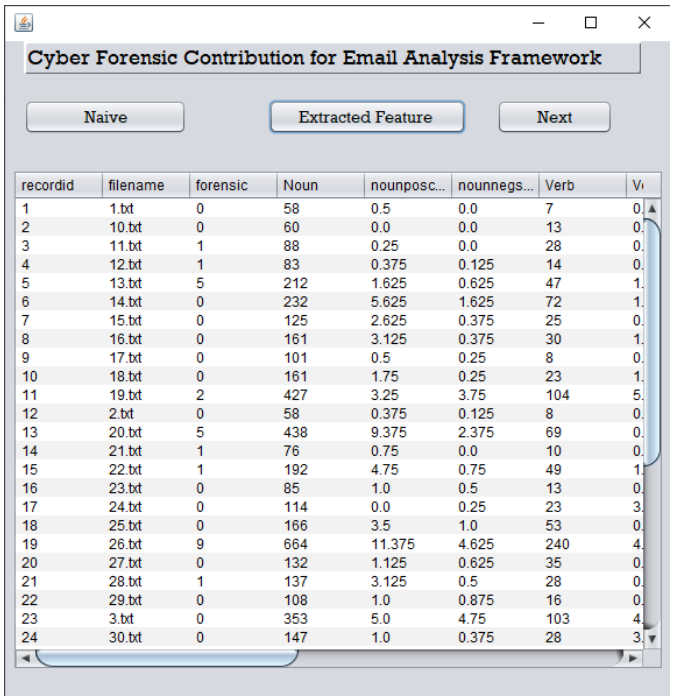
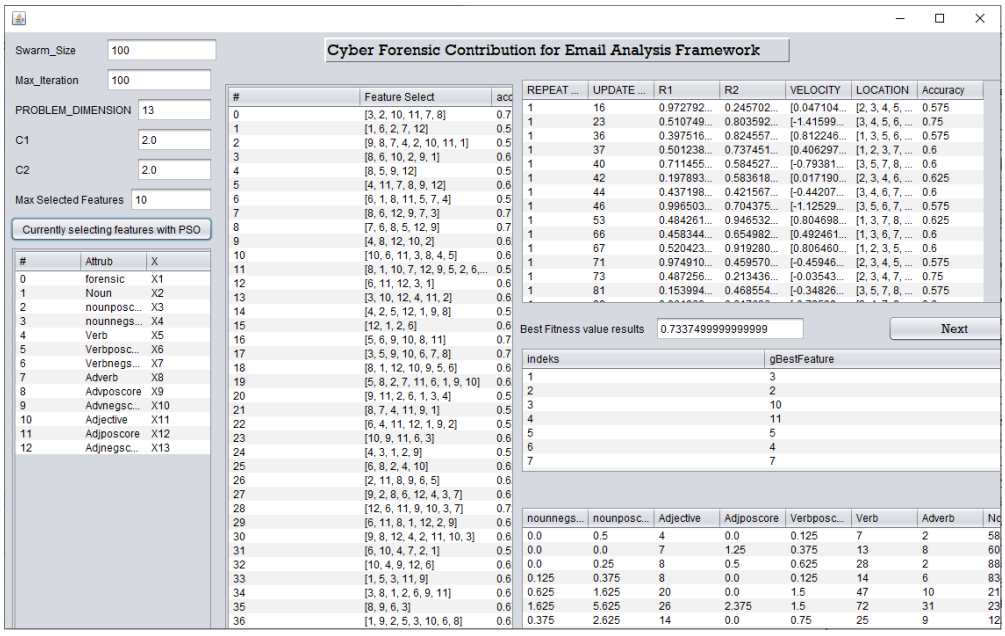


Figure 6: Clustering phase



recordid	filename	forensic	Noun	nounposc...	nounnegs...	Verb	V
1	1.txt	0	58	0.5	0.0	7	0
2	10.txt	0	60	0.0	0.0	13	0
3	11.txt	1	88	0.25	0.0	28	0
4	12.txt	1	83	0.375	0.125	14	0
5	13.txt	5	212	1.625	0.625	47	1
6	14.txt	0	232	5.625	1.625	72	1
7	15.txt	0	125	2.625	0.375	25	0
8	16.txt	0	161	3.125	0.375	30	1
9	17.txt	0	101	0.5	0.25	8	0
10	18.txt	0	161	1.75	0.25	23	1
11	19.txt	2	427	3.25	3.75	104	5
12	2.txt	0	58	0.375	0.125	8	0
13	20.txt	5	438	9.375	2.375	69	0
14	21.txt	1	76	0.75	0.0	10	0
15	22.txt	1	192	4.75	0.75	49	1
16	23.txt	0	85	1.0	0.5	13	0
17	24.txt	0	114	0.0	0.25	23	3
18	25.txt	0	166	3.5	1.0	53	0
19	26.txt	9	664	11.375	4.625	240	4
20	27.txt	0	132	1.125	0.625	35	0
21	28.txt	1	137	3.125	0.5	28	0
22	29.txt	0	108	1.0	0.875	16	0
23	3.txt	0	353	5.0	4.75	103	4
24	30.txt	0	147	1.0	0.375	28	3

Figure 7: Naive bayes and extracted features phase



#	Attrub	X
0	forensic	X1
1	Noun	X2
2	nounposc...	X3
3	nounnegs...	X4
4	Verb	X5
5	Verbposc...	X6
6	Verbnegs...	X7
7	Adverb	X8
8	Advposcore	X9
9	Advnegsc...	X10
10	Adjective	X11
11	Adjposcore	X12
12	Adjnegsc...	X13

#	Feature Select	acc
0	[3, 2, 10, 11, 7, 8]	0.7
1	[1, 6, 2, 7, 12]	0.5
2	[9, 8, 7, 4, 2, 10, 11, 1]	0.5
3	[8, 6, 10, 2, 9, 1]	0.6
4	[8, 5, 9, 12]	0.5
5	[4, 11, 7, 8, 9, 12]	0.6
6	[6, 1, 8, 11, 5, 7, 4]	0.5
7	[8, 6, 12, 9, 7, 3]	0.7
8	[7, 6, 8, 5, 12, 9]	0.7
9	[4, 8, 12, 10, 2]	0.6
10	[10, 6, 11, 3, 8, 4, 5]	0.6
11	[8, 1, 10, 7, 12, 9, 5, 2, 6,...	0.5
12	[6, 11, 12, 3, 1]	0.6
13	[3, 10, 12, 4, 11, 2]	0.6
14	[4, 2, 5, 12, 1, 9, 8]	0.5
15	[12, 1, 2, 6]	0.6
16	[5, 6, 9, 10, 8, 11]	0.7
17	[3, 5, 9, 10, 6, 7, 8]	0.7
18	[8, 1, 12, 10, 9, 5, 6]	0.6
19	[5, 8, 2, 7, 11, 6, 1, 9, 10]	0.6
20	[9, 11, 2, 6, 1, 3, 4]	0.5
21	[8, 7, 4, 11, 9, 1]	0.5
22	[6, 4, 11, 12, 1, 9, 2]	0.5
23	[10, 9, 11, 6, 3]	0.6
24	[4, 3, 1, 2, 9]	0.5
25	[6, 8, 2, 4, 10]	0.6
26	[2, 11, 8, 9, 6, 5]	0.6
27	[9, 2, 8, 6, 12, 4, 3, 7]	0.6
28	[12, 6, 11, 9, 10, 3, 7]	0.7
29	[6, 11, 8, 1, 12, 2, 9]	0.6
30	[9, 8, 12, 4, 2, 11, 10, 3]	0.6
31	[6, 10, 4, 7, 2, 1]	0.5
32	[10, 4, 9, 12, 6]	0.6
33	[1, 5, 3, 11, 9]	0.6
34	[3, 8, 1, 2, 6, 9, 11]	0.6
35	[8, 9, 6, 3]	0.6
36	[1, 9, 2, 5, 3, 10, 6, 8]	0.6

REPEAT ...	UPDATE ...	R1	R2	VELOCITY	LOCATION	Accuracy
1	16	0.972792...	0.245702...	[0.047104...	[2, 3, 4, 5, ...	0.575
1	23	0.510749...	0.803592...	[-1.41599...	[3, 4, 5, 6, ...	0.75
1	36	0.397516...	0.824557...	[0.812246...	[1, 3, 5, 6, ...	0.575
1	37	0.501238...	0.737451...	[0.406297...	[1, 2, 3, 7, ...	0.6
1	40	0.711455...	0.584527...	[-0.79381...	[3, 5, 7, 8, ...	0.6
1	42	0.197893...	0.583618...	[0.017190...	[2, 3, 4, 6, ...	0.625
1	44	0.437198...	0.421567...	[-0.44207...	[3, 4, 6, 7, ...	0.6
1	46	0.996503...	0.704375...	[-1.12529...	[3, 5, 6, 7, ...	0.575
1	53	0.484261...	0.946532...	[0.804698...	[1, 3, 7, 8, ...	0.625
1	66	0.458344...	0.654982...	[0.492461...	[1, 3, 6, 7, ...	0.6
1	67	0.520423...	0.919280...	[0.806460...	[1, 2, 3, 5, ...	0.6
1	71	0.974910...	0.459570...	[-0.45946...	[2, 3, 4, 5, ...	0.575
1	73	0.487256...	0.213436...	[-0.03543...	[2, 3, 4, 7, ...	0.75
1	81	0.153994...	0.468554...	[-0.34826...	[3, 5, 7, 8, ...	0.575

indeks	gBestFeature
1	3
2	2
3	10
4	11
5	5
6	4
7	7

nounnegs...	nounposc...	Adjective	Adjposcore	Verbposc...	Verb	Adverb	Nc
0.0	0.5	4	0.0	0.125	7	2	58
0.0	0.0	7	1.25	0.375	13	8	60
0.0	0.25	8	0.5	0.625	28	2	88
0.125	0.375	8	0.0	0.125	14	6	83
0.625	1.625	20	0.0	1.5	47	10	21
1.625	5.625	26	2.375	1.5	72	31	23
0.375	2.625	14	0.0	0.75	25	9	12

Figure 8: Optimization phase

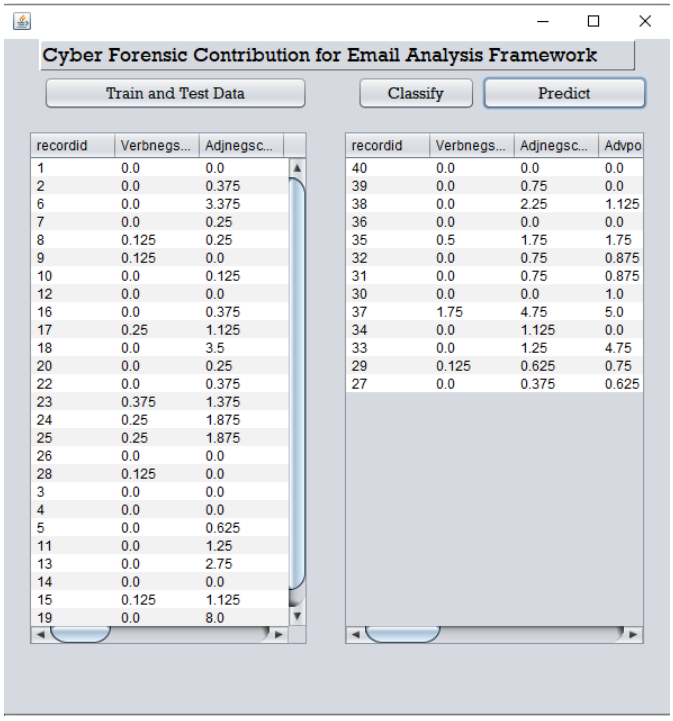


Figure 9: Classification frame by using SVM algorithm

V. RESULTS AND DISCUSSIONS

To evaluate our approach, we used e- mails from the Enron e- mail corpus. For case study are viewing the analysis and classification of seventeen employee (arnold- j, arora- h, badeer- r, bailey- s, bass- e, baughman- d, beck-s, benson- r, blair-l, bu- r,campbell-l,cash- m, corman- s, cuilla- m, davis- d, dean- c, and ermis- f) selected randomly. All documents folder was selected for each employee so that each all document folder contains a certain number of e- mails. The raw e- mail message text is processed into a form that can be tokenized. Firstly, the phase contains a number of methods designed to remove noise from the e- mail (in the form of obfuscation). The output of this phase is a string that contains the cleaned text of the e- mail along with some non- token features. The proposed system was implemented on different subsets of Enron email dataset and accuracy was calculated in each case and the results as shown in the table I. Fig. 10 shows the relation between number of email and accuracy result. The accuracy of classification is calculated by the percentage of the correctly classified emails in the testing set. The best- case of classification accuracy obtained by using the proposed algorithm is 92%. The proposed algorithm was provide a better prediction results. The experiments of this work have been implemented using the environment with the following specifications: Windows 10, Intel(R) Core(TM) i5- 4200U CPU@1.60GHz 2.29 GHz, RAM 8GB and 64- bit system type, the proposed system is programmed in Java Language platform on NetBeans IDE 8.2, Tool: Wampserver to handle MySQL database and used SentiWordNet 3.0 and Stanford (tagger and parser).

TABLE I
RESULT ACCURACY OF CLASSIFICATION

Different subset from enron email dataset	Number of samples	Accuracy %
Subset 1	450	80.7%
Subset 2	950	85.5%
Subset 3	1425	76.9%
Subset 4	1900	87.1%
Subset 5	2345	76.02%
Subset 6	2850	79.02%
Subset 7	3325	92.12%
Subset 8	5000	91.79%
Subset 9	10000	87.7%
Subset 10	20000	85.5%
Subset 11	30000	88.9%
Subset 12	40000	87.1%
Subset 13	50000	89.02%
Subset 14	60000	92.97%
Subset 15	70000	90.12%
Subset 16	80000	92.92%
Subset 17	90000	92.52%
Subset 18	100000	92.72%
Subset 19	200000	90.12%
Subset 20	300000	91.12%

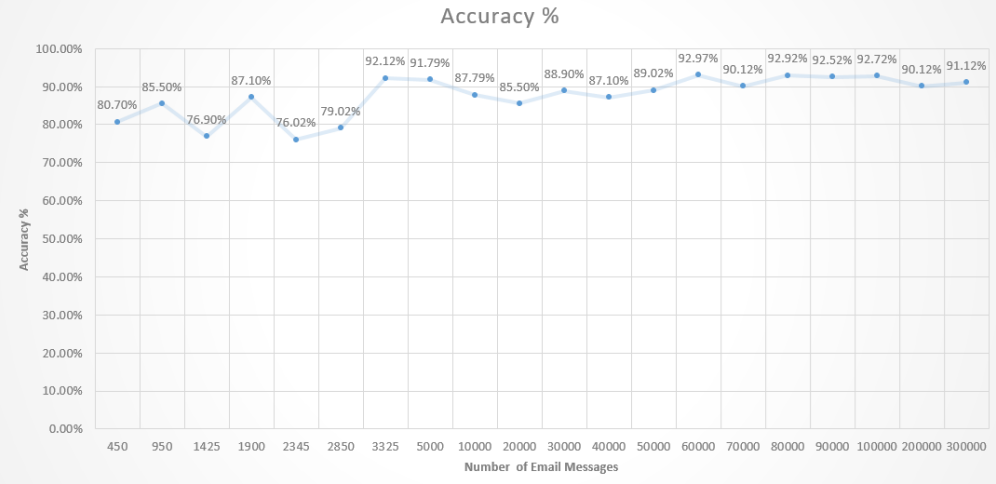


Figure 10: Different subset from enron email dataset and accuracy

VI. COMPARISON

The novelty of the proposed technique is the design and implementation of hybrid machine learning classifications and optimization methods which are offer enhancement of the statistical and other known intelligent techniques. The proposed hybrid technique presents a valuable result for classification email forensic as shown in Table II

TABLE II
COMPARISON THE PROPOSED WORK WITH THE RELATED WORKS

Related work	Method	Corpus dataset	Evaluation
Mining Writeprints from Anonymous E- Mails for Forensic Investigation [9]. 2010	clustering	Enron e- mail dataset	90%
Mining Email Content for Cyber Forensic Investigation [10]. 2012	clustering & classification	Enron	83%
E- mail data analysis for application to cyber forensic investigation using data mining [11]. 2013	decision tree	Enron	82%
Simplified features for email authorship identification [12]. 2013	the Simple Logistic and AdaBoost	Enron e- mail dataset	80% to 90%
Twitter sentiment analysis using hybrid naive bayes [13]. 2013	SentiWordNet and Naive Bayes classifier	Twitter API datasets	90%
Enhanced document clustering using k-means with support vector machine approach [14]. 2015	Clustering cascaded with Support Vector Machine	set of documents collected from the computer	70- 90%
A methodology for cybercrime identification using e- mail corpus based on GMM [15]. 2015	Gaussian Mixture Model (GMM)	Enron e- mail dataset	85%
Authorship verification of online messages for forensic investigation [16].2016	clustering and multidimensional scaling	Enron corpus, considered only 4 authors	70- 90%
Cybercrime. system to identify author of IM, using viber[17]. 2018	template matching Pearson Correlation coefficient	Enron	88%
Proposed work	combination special lexical SentiWordNet dictionary with Clustering, Particle Swarm Optimization and Support Vector Machine algorithms	Enron	92%

VII. CONCLUSIONS

Emails are one of the important means for exchanging information and widely used on the Internet which is a weak secure medium. Emails messages are digital evidence that has been become one of the important means to adopt by courts in many countries and societies as evidence relied upon in condemnation. Due to the huge number of these emails besides its rapid growth, this requires categorizing them to specific classes. The most important of these classes are legitimate emails and illegal emails that are issued from criminal persons whose intents are blackmail, murder, kidnapping, and intimidation of others, threats, rape, and disgraceful sexual acts. Therefore, it is necessary to find a successful and practical way to accommodate and classify these messages. Experiments of the proposed approach were aimed to test the effectiveness of the anonymous e- mails to collect evidence to prosecute criminals in a court of law. This paper presents a distinct technique for classifying emails based on data processing, trimming, refinement, and then adapt several algorithms to classify these emails and then using the SWARM algorithm to obtain practical and accurate results. The proposed system is capable of learning in an environment with large and variable data. To test the proposed system selected available data which Enron Data set. A high classification rate (92%) was obtained, which is higher than the classification rates mentioned in previous research papers presented in section II in this paper.

REFERENCES

[1] Charalambou, E, Bratskas, R, Karkas, G, & Anastasiades A, " Email forensic tools: A roadmap to email header analysis through a cybercrime use case" , Journal of Polish Safety and Reliability Association Summer Safety and Reliability Seminars, Vol. 7, No. 1, 2016.
[2] Meghanathan, N, Allam, S. R, & Moore, L. A, " Tools and techniques for network forensics" , arXiv preprint arXiv:1004.0570, 2010.
[3] Tsochataridou, C, Arampatzis, A, & Katos, V, " Improving Digital Forensics Through Data Mining" , In IMMM 2014, The Fourth International Conference on Advances in Information Mining and Management, September, 2016.
[4] Korasidi Andriana Maria , " Authorship Attribution Forensics: Feature selection methods in authorship identification using a small e- mail dataset" , M. Sc. Thesis, University of Athens, 2016.
[5] Sridhar N, Lalitha Bhaskari D, & Avadhani P. S, " Survey paper on cyber forensics" , International Journal of Computer Science, Systems Engineering and Information Technology,pp. 113- 118, 2011.
[6] Haggerty, J, Karran, A. J, Lamb, D. J, & Taylor, M, " A framework for the forensic investigation of unstructured email relationship data" , International Journal of Digital Crime and Forensics (IJDCF), 3(3) , 1- 18 , 2011.
[7] Bhardwaj, A. K, & Singh, M, " Data mining- based integrated network traffic visualization framework for threat detection" , Neural Computing and Applications, 26(1), 117- 130, 2015.
[8] Iqbal, F, Binsalleeh, H, Fung, B. C, & Debbabi, M, " Mining writeprints from anonymous e- mails for forensic investigation" , digital investigation, 7(1- 2), 56- 64, 2010.
[9] Khan, S. R, Nirkhi, S. M, & Dharaskar, R. V, " Mining e-mail for cyber forensic investigation" , International Conference on Advances in Computer, Electronics and Electrical Engineering, ICACEEE, International Journal of Computer Science and its Applications (UACEE), Vol. 2(2), pp. 112- 116, 2012.
[10] Khan, S. R, Nirkishi, S. M, & Dharaskar, R. V , " E- mail data analysis for application to cyber forensic investigation using data mining" , In Proceedings of the 2nd National Conference on Innovative Paradigms in Engineering and Technology (NCIPET), New York, USA, 2013.
[11] Abdallah, E. E, et al. , " Simplified features for email authorship identification" , International Journal of Security and networks 8.2 :72- 81, 2013.
[12] Thakkar, M. H. V, " Twitter sentiment analysis using hybrid naive bayes" , M. Sc. Thesis, Department of Computer Engineering, sardar vallabhbhai national institute of technology, surat, 2013.
[13] Khairkar P. K, Phalke D. A, " Enhanced Document Clustering using K- Means with Support Vector Machine(SVM) Approach" , International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), Volume: 3 Issue: 6, ISSN: 2321- 8169 ,4112 - 4116, June 2015.
[14] Sreenivasulu, V, and Prasad R. S, " A methodology for cybercrime identification using email corpus based on gaussian mixture model" , International Journal of Computer Applications117. 13 , 2015.
[15] Nirkhi, S, Dharaskar, R. V, & Thakare, V. M, " Authorship Verification of Online Messages for Forensic Investigation" , Procedia Computer Science 78 640 - 645, 2016.
[16] Azizan S. A, " Terrorism Detection Based on Sentiment Analysis Using Machine Learning" , Journal of Engineering and Applied Sciences, Universiti Teknologi Petronas, Bandar Seri Iskandar, Peerak,Malaysia, 12(3) :691- 698, 2017.
[17] Shahad Fadhil Abbas, " Cybercrime System to Identification the Author of Instance Message Using Chat Messages" , M. Sc. Thesis, the University of Technology, Department of Computer Science, Baghdad, Iraq, 2018.
[18] Sailaja, D, Kishore, M, Jyothi, B, & Prasad, N. R. G. K, " An overview of pre- processing text clustering methods" , Int. J. Comput. Sci. Inform. Tech 6, 3119- 24, 2015.
[19] Gidofalvi, G, & Zadrozny, B, " Feature selection for class probability estimation" ,2002.
[20] Wang, Y, Kim, K, Lee, B, & Youn, H. Y, " Word clustering based on POS feature for efficient twitter sentiment analysis" , Human- centric Computing and Information Sciences 8.1 ,17, 2018.
[21] Sharma, T, Tomar, G. S, Kumar, B, & Berry, I, " Particle swarm optimization based cluster head election approach for wireless sensor network" , International Journal of Smart Device and Appliance 2.2 ,2014.
[22] Oludare, O, Stephen, O, Ayodele, O, & Temitayo, F, " An Optimized Feature Selection Technique For Email Classification" , International Journal of Scientific & Technology Research 3.10 ,286- 293,2014.
[23] Anuar, S, Selamat, A, & Sallehuddin, R, " Hybrid particle swarm optimization feature selection for crime classification" , New Trends in Intelligent Information and Database Systems. Springer, Cham,101- 110, 2015.
[24] Patil, M. S, Bewoor, M. S, & Patil, S. H, " A hybrid approach for extractive document summarization using machine learning and clustering technique" , International Journal of Computer Science and Information Technologies, pp.1584- 1586, 2014.
[25] Sally D. Hamdi, " Cyber forensic email analysis and detection based on intelligent techniques", M.Sc. Thesis, Al- Nahrain University, College of Information Engineering, Baghdad, Iraq, 2019.
[26] Howedi, F, & Mohd, M, " Text classification for authorship attribution using Naive Bayes classifier with limited training data" , Computer Engineering and Intelligent Systems, 5(4), 48- 56 ,2014.