ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(27), DOI: 10.17485/ijst/2019/v12i27/145999, July 2019*

Integrated Cyber Forensic for E-Mail Analysis Framework

Sally Dakheel Hamdi and Abdulkareem Merhej Radhi

Department of Information and Communication Engineering, Al-Nahrain University, Baghdad - 64074, Iraq; sallydakeel@gmail.com, abdulkareemradhi@gmail.com

Abstract

Objectives: This study presents a distinct technique for classifying emails based on data processing and mining, trimming, refinement, and then adapts several algorithms to classify these emails. **Methods/Statistical Analysis:** SWARM algorithm to obtain practical and accurate results. **Findings**: The proposed system is capable of learning in an environment with large and variable data. To test the proposed system, we have to select available data which Enron Data set. A high accuracy rate (95%) was obtained, which is higher than the classification rates mentioned in previous research papers presented in section 2 in this paper. **Application/ Improvements:** In the past two decades, the Internet has become as an open, publicly and widely used as a source of data transmission and exchanging the messages between criminals, terrorists and those who have illegal motivations. Moreover exchanging important data between various military and financial institutions even ordinary citizens. From this view, there is one of the important means of exchanging information widely used on the Internet medium is e-mail. Email messages are digital evidence which have been became one of the important means to adopt by courts in many countries and societies as evidence relied upon in condemnation.

Keywords: Digital Forensic, K-means, Learning, Mining, SWARM Algorithm

1. Introduction

Two decades ago, the world has witnessed a quantum leap in the use of digital data to communicate and share ideas and messages via web and technological media which are easy, familiar and cheap. That the availability of this multimedia and the Internet in its simple form and its cheap price led to the emergence of large groups abused the use of it to transfer data as criminal events. The emergence of this type of non-conventional crime has prompted authorities and governments to support a new type of criminal investigation based on the analysis of digital data by a group of experts specialized in the field of digital information and the analysis of e-mails in order to use it in the courts as an important tool and evidence of the commit such acts. So, in these decades we saw another form of forensic criminal investigation is the digital criminal analysis[1]. "The Internet provides an appropriate platform for cybercriminals to carry out their illegal activities such as anonymization, such as phishing and spam, and as a result, in recent years, the authoring analysis of anonymous e-mails has received some attention in forensic and data mining communities[2]. The preservation of such important evidence in its original form as evidence of condemnation is the primary objective of digital forensic goals. E-mail can be considered as an easy, common and inexpensive way to communicate and exchange messages and data in various formats, textual and digital and through the Web. For these and other

*Author for correspondence

reasons, e-mail has become an easy and attractive way for many people and criminals who have malicious thoughts and bad intentions towards others. They work on sending "spam, threats, spamming emails, spreading malware such as viruses and worms, Child pornography, and other criminal activities, so it is necessary to secure our e-mail system as well as to identify the offender, collect evidence against them and punish them under the law of the Court"[3]. Emails are an easy and important means used by criminals and terrorists to harm others through which forensic workers can obtain the digital evidences that is rigid to use it to be convicted in the courts of justice and criminal. "Forensic analysis of e-mail and other electronically stored data are critical when the evidence becomes digital"[4].

"Digital forensics is the application of investigation and analysis technique to collect and defend evidence from a particular computing device in a way that is proper for presentation in a court of act"[1]. Digital Forensic analysis introduces data processing after collection, analysis, and mining features of digital evidences. Analyzing data for the several and different crimes via computer based means is called as Digital Forensic Analysis (DFA). So to recover forensic analysis process needs text clustering and classification methods.

2. Previous Work

In[5] proposed work "relies on swarm intelligent agents and modification of Voronoi algorithm such that the issues of the messages, including suspicious messages, are divided into communities. Moreover, these communities are divided into categories, each given a specific rank, depending on the quality and size of the threat messages". In[6] focus on Machine Learning-based spam filters and their variants. In[3] "review working and architecture of current email system and the security protocols, further email forensics which is a process to analyze e-mail contents". In[7] present "some feature selection techniques such as Mutual information, Chi-Square, Information gain and TF-IDF. Classification was performed using support vector machine provided by weka9 tool". In[1] "introduce Clustering Technique Cascaded with Support Vector Machine to enhance the expert's job and investigation process". In[8] "employed Naive Bayes (NB) classifier in order to classify the texts to their authors". In[9] "Introduce systematic process for email forensic through which integrate into the normal forensic analysis workflow, and which accommodates the distinct characteristics of email evidence". In[10] "Propose a Hybrid Naive Bayes classifier which is the combination of a machine learning algorithm (Naive Bayes) and a special lexical dictionary (SentiWordNet3.0)". In[11] "perform e-mail Statistical Analysis, e-mail clustering and classification, e-mail authorship identification and social network analysis". In[12] "based on comparing the similarity between a given unknown documents against the known documents using various features so that an unknown document can be classified as having been written by the same author by application of unsupervised techniques for authorship verification problem". In[13] "focus on the problem of mining the writing styles from a collection of e-mails written by multiple anonymous authors". In[14] "Enhanced Document Clustering by means of different algorithms such as K-Means with Support Vector Machine (SVM) for a large data set". The last one of these researches has been compared with our proposed research.

3. Method/Experimental Work

Machine learning can be considered as the most famous techniques having interest of researcher 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 processing, 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 Naïve 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

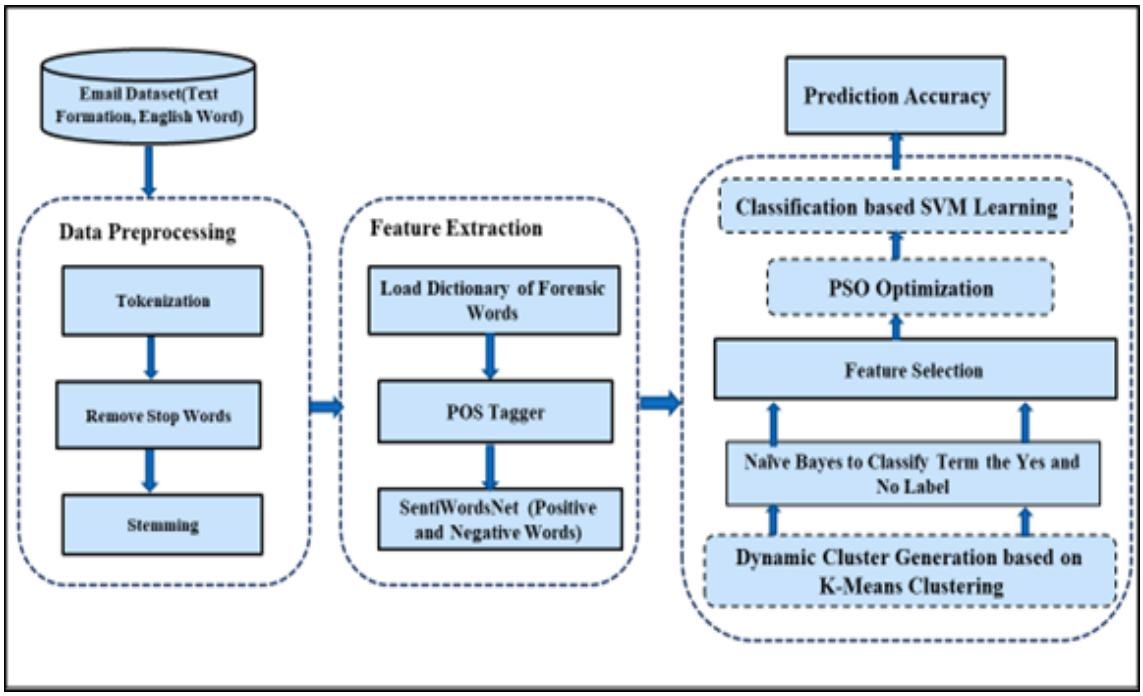


Figure 1. Block diagram of the proposed model.

accuracy. In Figure 1 summarize the frame work phases of our proposed research. The system is composed of several different phases each phase has a specific function.

3.1 E-mail Dataset

The main corpus used in the proposed system is Enron email dataset that contains (619446) text messages in their original form. It downloaded from the predefined corpora on Internet. These data are stored to be used in training and testing phases. Enron email dataset available in this Website: https://www.cs.cmu.edu/~enron/. In this work Load the E-mail Dataset, the sentiment analysis is going to be performed on this dataset. Each file that are presented inside the folder contain the values of contains sender, description, receiver, email id, subject, and file size of the email. Data selection show in Figure 2.

3.2 Data Preprocessing

Data preprocessing is an important step in the data mining process. It is done to represent the data in a form that can be used for next phases. There are many types of representing the documents like, graphical model, Vector-Model, etc. Many measures are also used for weighing the documents. Thus, the representation and quality of data is first and foremost before running a machine learning algorithms. Often, data preprocessing is the main phase of a machine learning project. If there is much irrelevant and redundant information present or confused and unreliable data, then discovering knowledge during the training is more difficult. The following sections present data preprocessing stages:

3.2.1 Tokenization Process

Typically, the term "tokenization" referred to the process of phase sentences are divided into streams of text into its constituent meaning full units (called tokens). From a linguistic perspective, tokenization is accomplished on a word level, hence breaking the stream of text out into words. A predefined set of separators is used for this purpose. In the proposed tokenizer the punctuations mark,

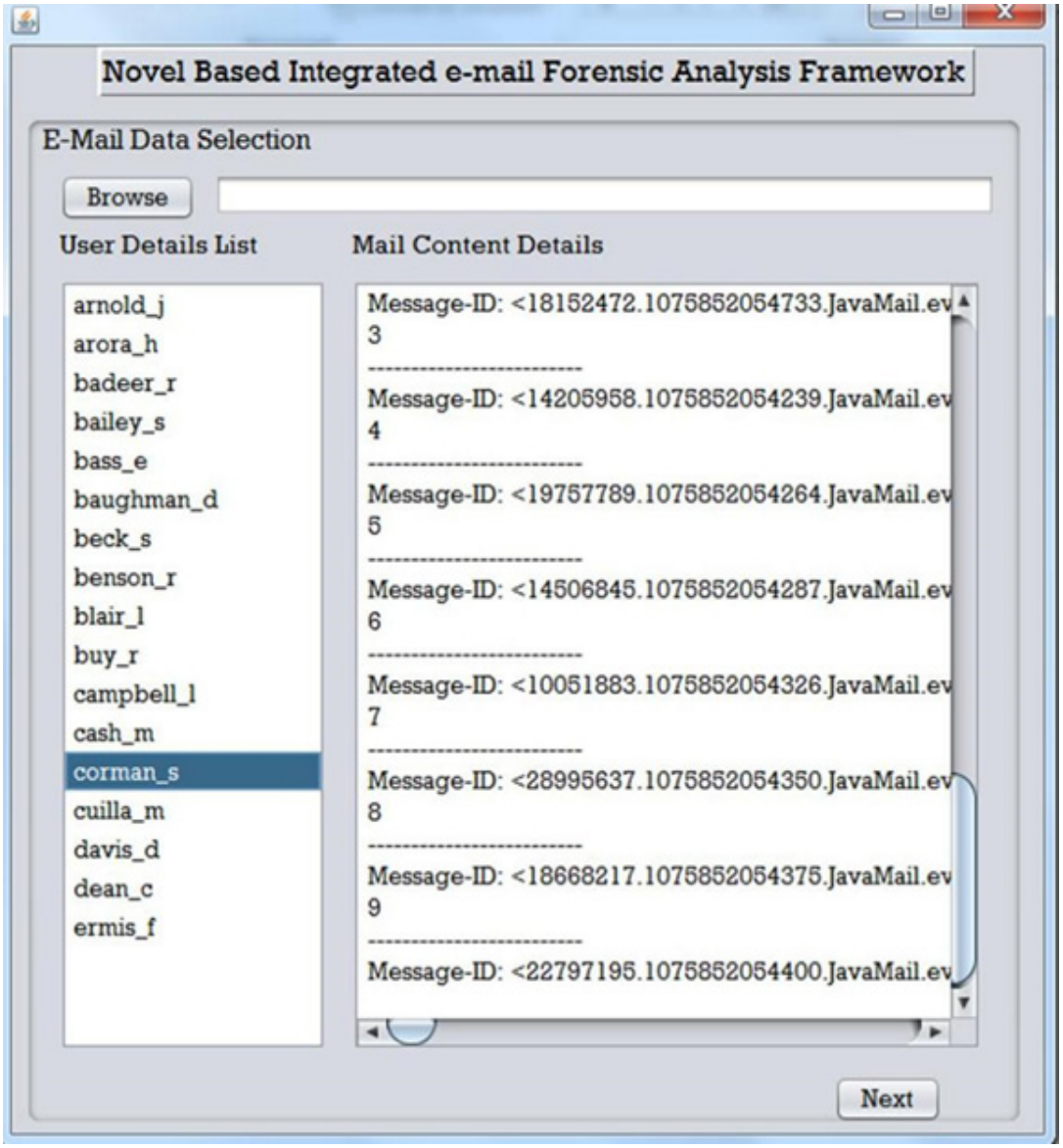


Figure 2. Data selection.

numbers, white space, and html tags were considered in breaking out the body of the email message into a sequence of features.

3.2.2 Remove Stop Word

In natural language processing, stop words means word that do not have any meaning such as "and", "the", "a", "an", and similar words, and are thus eliminated prior to classification. The stop words are not necessary for an alyzation so we are going to load and remove the stop words from our dataset. Stop words available in this Website: https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords.

3.2.3 Stemming Process

This process reducing words to their original root for instance, finance, financial, and financing may be converted to finance and data preprocessing process shown in Figure 3.

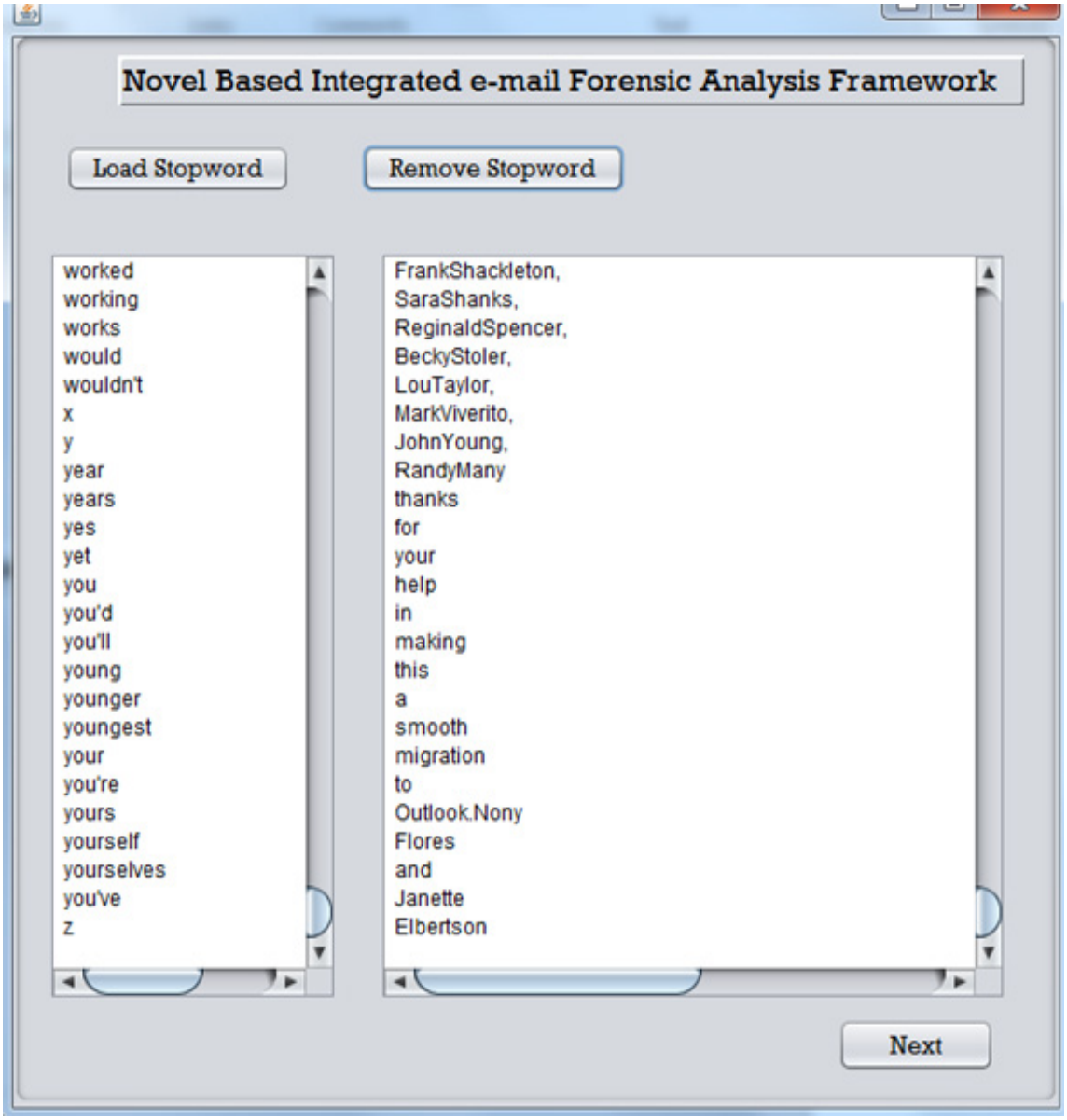


Figure 3. Data preprocessing phase.

3.3 Feature Extraction

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

3.3.1 Forensic Words

In English, there are a lot of specific words for different types of crimes and the criminals who commit them. Unfortunately, the list of crimes and criminals is long! Because the words have specific legal meanings, there are some need-to-know Forensic words vocabulary words. To assist you in learning more about the cyber forensics system, we compiled a list of 647 Forensic vocabulary words. The Forensic words dictionary library datasets are load Then forensic words searched in the email dataset for doing the analyzation. Forensic words available in this Website (https://myvocabulary.com/word-list/crime-vocabulary/).

3.3.2 POS Tagging

The Part-Of-Speech (POS) is a tagging tool it tags each word and assigns parts of speech to each word (and other token). Part-of-speech categories include noun, verb, adjective, preposition, pronoun, adverb, conjunction and interjection. Example word has POS tagging (JJ, JJR, JJS, VB, VBD, VBG, VBN, VBP, and VBZ) of an adjective score and verb score and so as. POS tagger parses a sentence or document and tags each term with its part of speech.

3.3.3 SentiWordNet3.0

It is an opinion lexicon mining from the Word Net

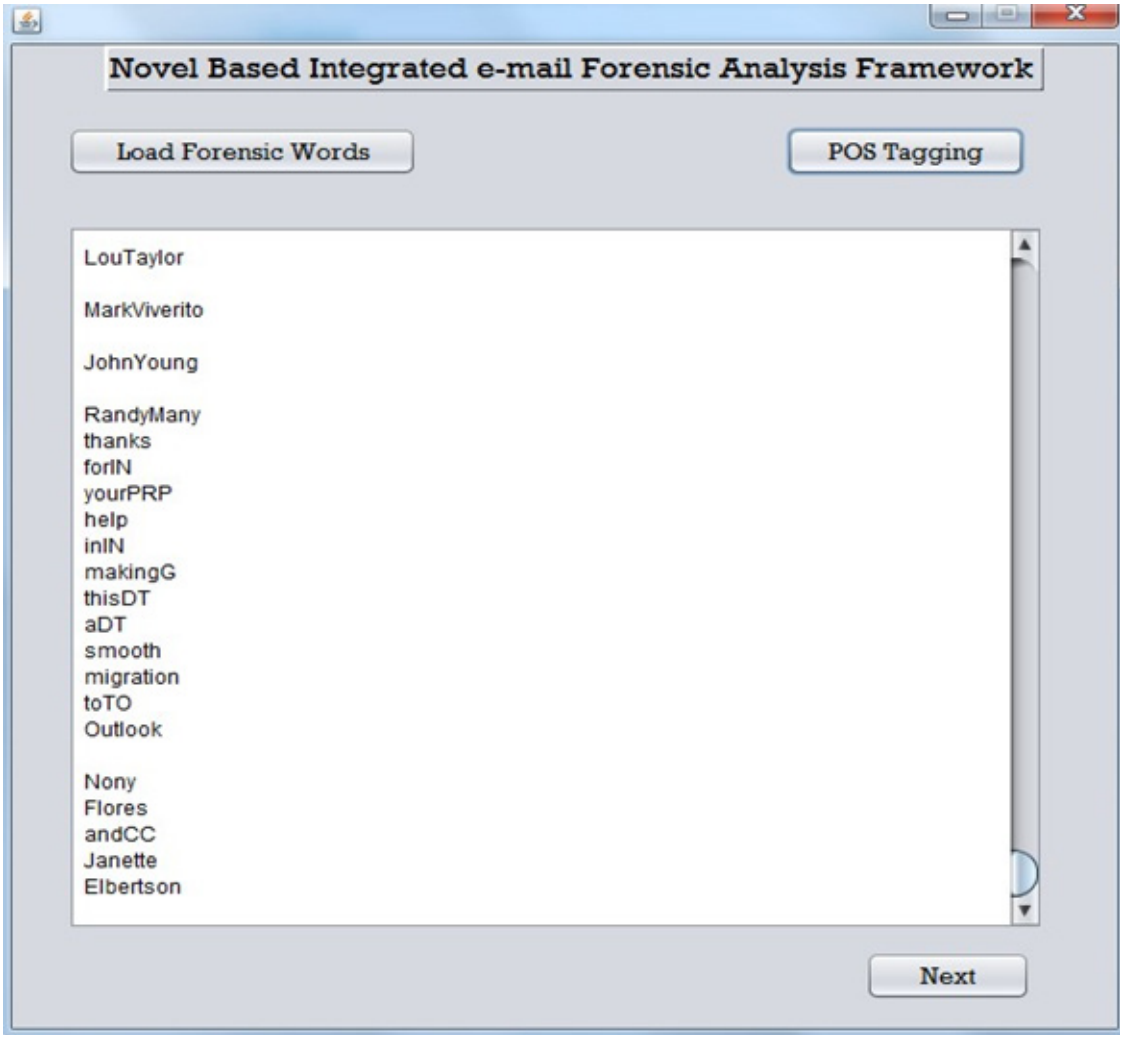


Figure 4. Feature extraction phase.

database. Each token is related to numerical scores representing positive and negative sentiment information SentiWordNet3.0. Score calculated using SentiWordNet3.0. SentiWordNet3.0 provides positivity and negativity scores for Part Of Speech (POS)-tagged synsets (synonym sets). If the score is greater than zero, this feature is classified as positive, whereas if the score is less than zero, it is classified as negative available in this Website: http://sentiwordnet.isti.cnr.it/. Feature Extraction process as shown in Figure 4.

3.4 Clustering

In K-means Clustering the center point is defined. It is not dynamically generated in our process we create the

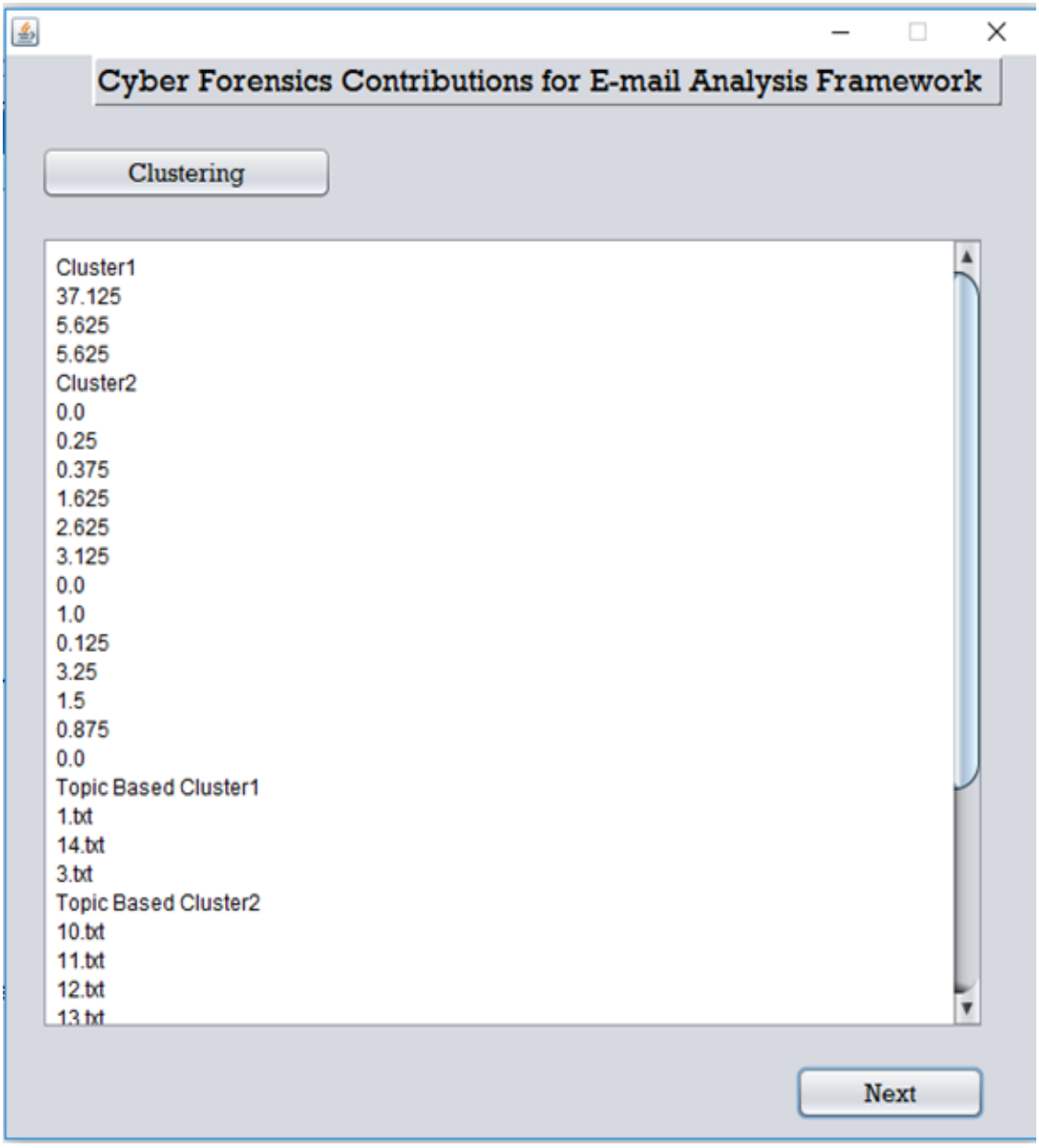
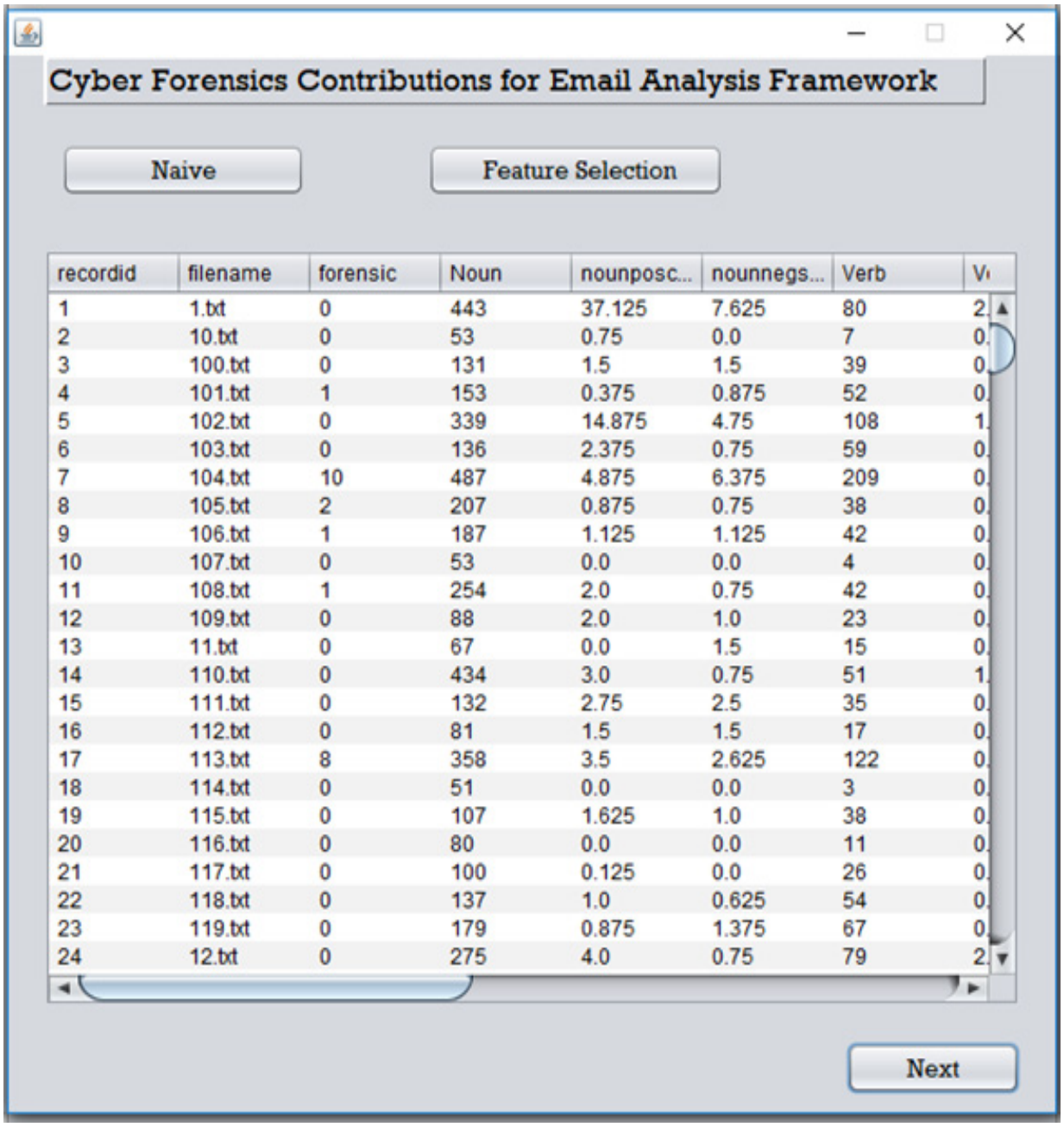


Figure 5. Clustering.

center point node in k means at dynamically as depicted in Figure 5.

3.5 Feature Selection

As shown in Figure 6 Feature Extraction is extracting the feature which is given in the dataset, we use naïve Bayes. Feature extraction is the selection of those data attributes that best characterize a predicted variable. In Naive Bayes, the probabilities for each attribute are calculated independently from the dataset. In these process e-mail



recordid	filename	forensic	Noun	nounposc...	nounnegs...	Verb	V
1	1.txt	0	443	37.125	7.625	80	2
2	10.txt	0	53	0.75	0.0	7	0
3	100.txt	0	131	1.5	1.5	39	0
4	101.txt	1	153	0.375	0.875	52	0
5	102.txt	0	339	14.875	4.75	108	1
6	103.txt	0	136	2.375	0.75	59	0
7	104.txt	10	487	4.875	6.375	209	0
8	105.txt	2	207	0.875	0.75	38	0
9	106.txt	1	187	1.125	1.125	42	0
10	107.txt	0	53	0.0	0.0	4	0
11	108.txt	1	254	2.0	0.75	42	0
12	109.txt	0	88	2.0	1.0	23	0
13	11.txt	0	67	0.0	1.5	15	0
14	110.txt	0	434	3.0	0.75	51	1
15	111.txt	0	132	2.75	2.5	35	0
16	112.txt	0	81	1.5	1.5	17	0
17	113.txt	8	358	3.5	2.625	122	0
18	114.txt	0	51	0.0	0.0	3	0
19	115.txt	0	107	1.625	1.0	38	0
20	116.txt	0	80	0.0	0.0	11	0
21	117.txt	0	100	0.125	0.0	26	0
22	118.txt	0	137	1.0	0.625	54	0
23	119.txt	0	179	0.875	1.375	67	0
24	12.txt	0	275	4.0	0.75	79	2

Figure 6. Features selection.

words are analysis either it is positive or negative sense by using Naïve Bayes algorithm with feature extraction. After the extracted features, the obtained result will be optimized.

3.6 Optimization

Particle Swarm Optimization (PSO) is Optimization algorithm so we optimize the selected features and the result was improved as shown in Figure 7.

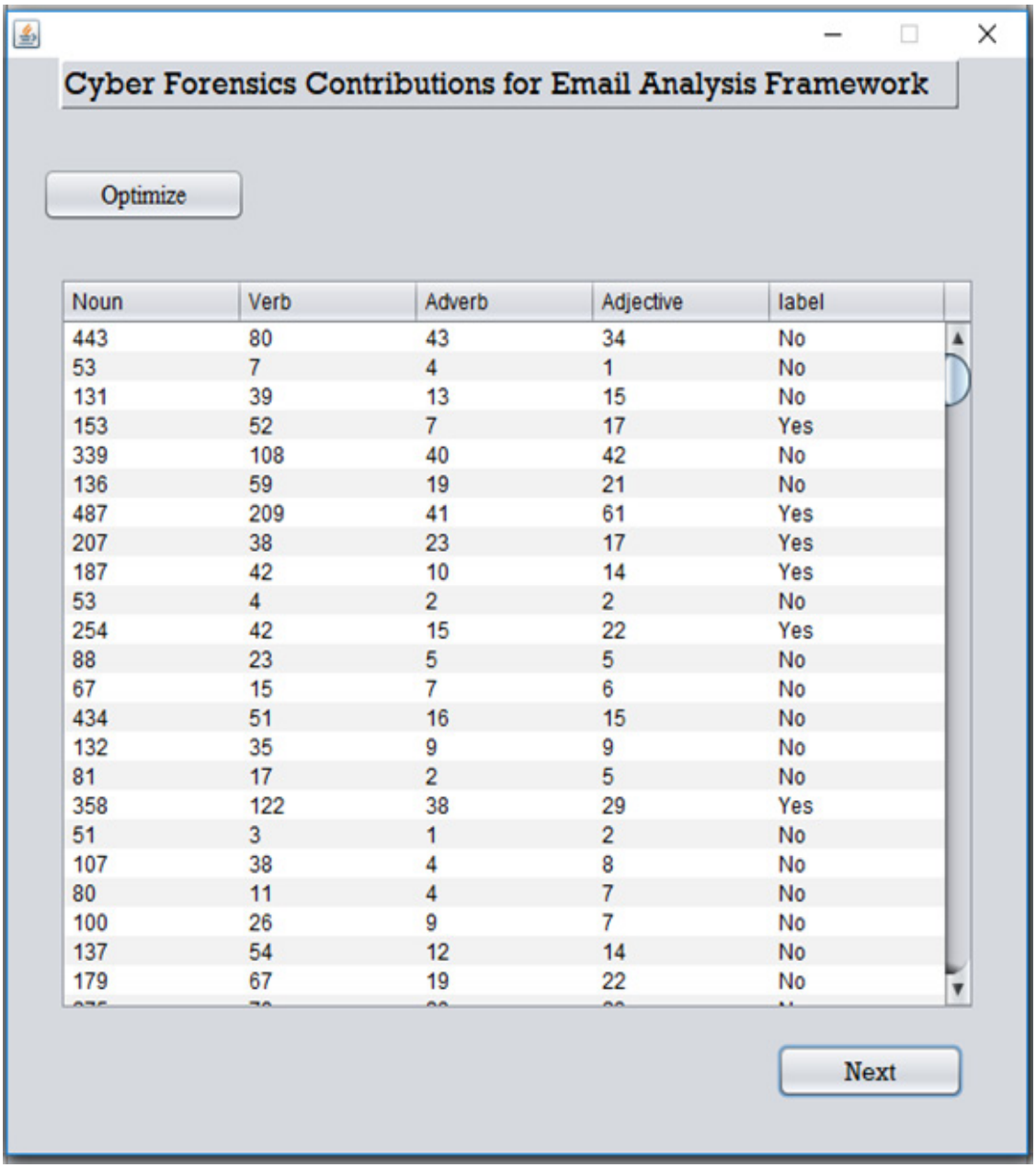


Figure 7. Optimization.

3.7 Classification

Classification was performed for the training and testing data, predicting the results and providing the better prediction results. Scalar Matrix Based Support Vector Machine (SVM) Learning is classifying the normal (not forensic) and abnormal (forensic) messages depicted in Figure 8.

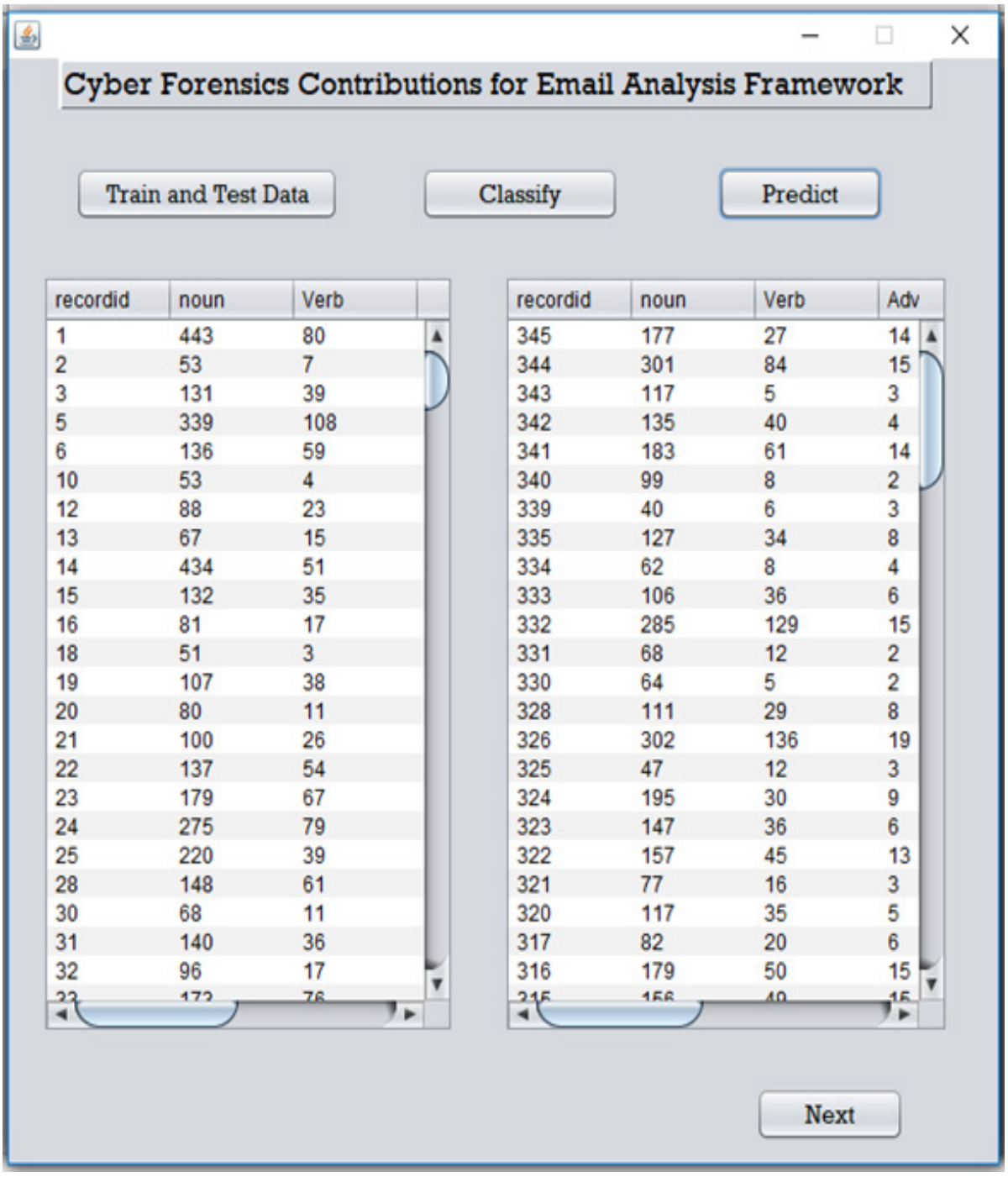


Figure 8. Classification.

3.8 Prediction

Finally, we predict the accuracy of our dataset. It will predict accuracy and classification accuracy of our process as shown in Figure 9.

4. Results and Discussion

As mentioned in section 2 of this paper, we saw that there are different researches trying to analyze data sets or emails by different clustering means. But we saw that[8] was the much nearest approach to our proposed research, so we would like to compare philosophy and results between them in this section. As mentioned in section 2 of this paper, we saw that there are previous researches was trying to analyze data sets or emails by different clustering means. But we saw[8] was the much nearest approach to our proposed research, so, in this section we would like to compare philosophy and results of each between them as follows:

1. The proposed research was process a huge email data set achieved by different means (statistical, textual, and using machine learning).
2. The obtained results and accuracy rate of our proposed technique was 95%, while we saw that the previous researches satisfied the accuracy rate less than 85%.
3. Our proposed research used a textual mean to help for scoring and ranking tokens, sentences, and phrases which are a sentiment lexicon and a specific stem technique. These means have been help for having efficient and high accuracy rate.

To evaluate our approach, we used e-mails from the Enron e-mail corpus. For case study we are viewing the analysis and classification of seventeen employees (arnold_j, arora_h, badeer_r, bailey_s, bass_e, baughman_d, beck_s, benson_r, blair_l, buy_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 employ so that each all docu-

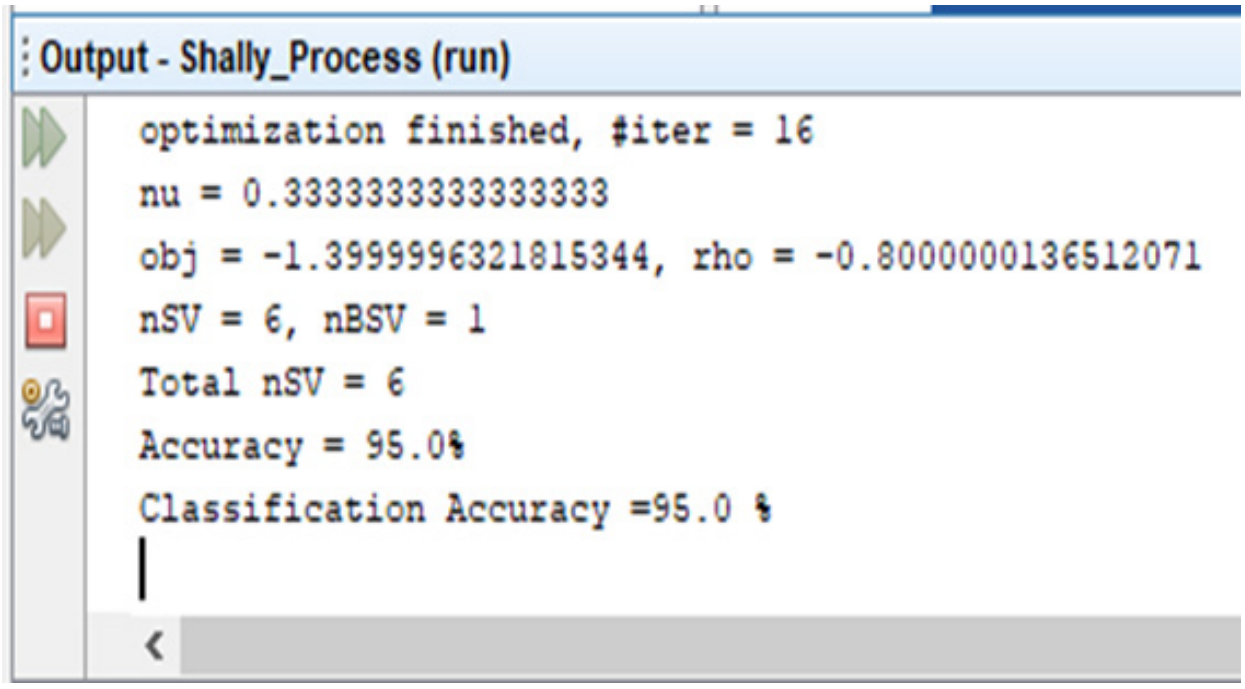


Figure 9. Accuracy of classification an optimization.

ments folder contains a certain number of e-mails. The message in the dataset includes: email addresses of the sender and receiver, date and time, subject, and the body text. In this work, only the "body" sections of the emails were used, and sender/receiver, date/time information were discarded.

Raw e-mail body text is processed into a form that can be tokenized. Firstly, the phase contains a number of methods designed to remove noise from the body of the e-mail (in the form of obfuscation). The output of this phase is a string which contains the cleaned body of the e-mail along with some non-token features.

The e-mail set derived by randomly selecting (2/3) of the e-mails as the training set and considering the remaining e-mails as the testing set. In our experiments we used SVM classifiers. Classification is performed on the training and test data. To check the effect of class labels on the accuracy of classifiers, we performed classification experiments for class labels. The best classification score that obtained in using support vector machine, 95% of accuracy.

5. Conclusion

Emails are one of the important means for exchanging information and widely used on the Internet which is a weak secure medium. Email messages are digital evidence which have been became 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 which 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. This study 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 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, we have to select available data which Enron Data set. A high classification rate (95%) was obtained, which is higher than the classification rates mentioned in previous researches presented in section 2 in this study.

6. References

1. Priyanka K, Prashant N, Annand M. Mining frequent sequences for emails in cyber forensics investigation, International Journal of Computer Applications. 2014; 85(17):1–7. https://doi.org/10.5120/14930-3332.
2. Farkhund I, Liquate KA, Benjamin FCM, Murad D. E-mail authorship verification for forensic investigation, Computer Security Laboratory CIISES. 2010; 1591–98.
3. Gurpal C, Dilpreet B. Review of e-mail system, security protocols and email forensics, International Journal of Computer Science and Communication Networks. 2016; 5(3):201–11.
4. Sobiya KR, Smita NM. E-mail data analysis for application to cyber forensic investigation using data mining, International Journal of Applied Information Systems (IJAIS). ISSN: 2249-0868, Foundation of Computer Science FCS, New York:USA; 2016. p. 1–4.
5. Swarm intelligent agents of e-mail classification. Date accessed: 2017. https://www.semanticscholar.org/paper/Swarm-Intelligent-Agents-of-E-mail-Classification-Radhi/4c6226c2ba51f76c1c03ca7895235ca36322d7ae.
6. Alexey B, Shyamanta HM. Machine learning for e-mail spam filtering: review, Techniques and Trends. 2016; 1–26.
7. Shahana PH, Bini O. Evaluation of features on sentimental analysis, International Conference on Information and Communication Technologies. 2015; 46:1585–92. https://doi.org/10.1016/j.procs.2015.02.088.
8. Fatima H, Masnizah M. Text classification for authorship attribution using naive bayes classifier with limited training data, Computer Engineering and Intelligent Systems. 2014; 5(4):48–56.
9. Towards comprehensive and collaborative forensics on email evidence. Date accessed: 20/10/2013. https://ieeexplore.ieee.org/document/6679965.
10. Twitter sentiment analysis using hybrid naïve bayes. Date accessed: 06/2013. https://www.researchgate.net/publication/318663470_Twitter_Sentiment_Analysis_using_Hybrid_Naive_Bayes.
11. Sobiya KR, Smita MN. Mining e-mail content for cyber forensic investigation. Proceedings of International

Conference on Advances in Computing, Electronics and Electrical Engineering; 2012. p. 415–19.

12. Nirkhi S, Dharaskar RV. Thakare VM. Authorship verification of online messages for forensic investigation, Procedia Computer Science. 2016; 78:640–45. https://doi.org/10.1016/j.procs.2016.02.111.
13. Farkhund I, Hamad B, Benjamin CM, Mourad D. Mining write prints from anonymous e- mails for forensic investigation, Digital Investigation. 2010; 56–64. https://doi.org/10.1016/j.diin.2010.03.003.
14. Prachi KK. Enhanced document clustering using k-means with Support Vector Machine (SVM) approach, International Journal on Recent and Innovation Trends in Computing and Communication. 2015; 3(6):4112–16.