Abstract: In the past two decades, the Internet has become as 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 that has been become one of the important means to adopt by courts in many countries and societies as evidence relied upon in condemnation. This paper presents a distinct technique for classifying emails based on data processing and mining, trimming, refinement, and then adapts several algorithms to classify these emails and then using SWARM algorithm to obtain practical and accurate results also using hybrid English lexical dictionary SentiWordNet3.0 for email forensic analysis then deal with a machine learning algorithm. The proposed system is capable of learning in an environment with large and variable data. To test the proposed system, 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.

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 that 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 they 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 like cyberbullying, anonymization, phishing, email forensic and spam. As a result,
in recent years, the authoring analysis of anonymous emails has received some attention in forensic and data mining communities [2]. The preservation of such important evidence in its original form is evidence of condemnation is the primary objective of digital forensic goals. Email 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 reasons, email 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 threats, spamming emails, spreading malware like viruses and worms, Child pornography, and other criminal activities, so it is necessary to secure our email 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 evidence that is rigid to use it to be convicted in the courts of justice and criminal. Forensic analysis of email and other electronically stored data are critical when the evidence becomes digital [4].

2. Previous Work

Radhi [5] proposed work that relies on swarm intelligent agents and modification of the 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. B. Alexey and H.M. Shyamanta [6] Focus on Machine Learning-based spam filters and their variants. Chhabra and Bajwa [3] present review working and architecture of the current email system and the security protocols, further email forensics which is a process to analyze email contents. P.H. Shahana and O. Bini [7] present some feature selection techniques such as Mutual information, Chi-Square, Information gain, and TF-IDF. The classification was performed using the support vector machine provided by weka data mining tool. Priyanka and et.al. [1], introduce the Clustering Technique Cascaded with Support Vector Machine to enhance the expert's job and investigation process. Fatima H., Masnizah M.[8], employed Naive Bayes (NB) classifier in order to classify the texts to their authors. P. Justin, M. Mike, and A. Gail-Joon,[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. Harsh Vrajesh T., [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). Sobiya K.R., Smita M.N., and et. al., [11], perform email Statistical Analysis, email clustering & classification, email authorship identification, and social network analysis. Nirkhi, S., and et. al. [12], Focus on comparing the similarity between 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. Farkhund I., and et. al. [13], focus on the problem of mining the writing styles from a collection of emails written by multiple anonymous authors. K.K. Prachi and P.D.A, [14], Enhanced Document Clustering by means of different algorithms like 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. Digital Cyber Forensic Analysis

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

4. Proposed Work

Machine learning can be considered as the most famous technique having an interest of researchers because of its accuracy and adaptability. For email mining, in most cases, the learning algorithm of this technique is employed. It consists of several phases: Data Preprocessing, Clustering, Feature extraction, Optimization, Classification, and then Prediction results. Four Machine learning algorithms used in this work are k-means for clustering and naïve bayes for class probability estimation, particle swarm optimization for optimizing feature and support vector machine for Classification. optimize the selected features of the results which were improved for enhancing accuracy. Figure [1], summarize the framework of the proposed model as follows:

5. Dataset and Preprocessing

Due to their privacy issues and their secrecy, few e-mail data are available publicly for experiments. The exception to the above statement is the Enron Corpus [12]. It has been followed the concepts and principles mentioned in this section to preprocessing and evaluation metric a reader may be fixed

5.1 Token-Frequency

By equation [1], measure the parameter which depicts in what way appropriate token belonging to a specific email in the Enron dataset. This significance score reflects the number of times a token appears in the email.

\[ TF - idf_i = t_{i,j} \times \log \left( \frac{N}{df_i} \right) \]  

Such that \( TF - idf_i \) is the weight of a term \( i \). \( t_{i,j} \) is the frequency of term \( i \) in sample \( j \). \( N \) is the total number of samples in the corpus. \( df_i \) is the number of samples containing term \( i \).

5.2 Chi-Squared {Selection method}

To measures, the deviation from the estimated distribution was expecting that feature occurrence is independent of class value, by use equation (2) [7].
\[ X^2(f, c) = \frac{N(WZ-YX)^2}{(W+Y)(X+Z)(W+X)(Y+Z)} \]  

(2)

Such that \( W, X, Y, Z \) denotes the frequencies, indicates the presence or absence of a feature in the sample, \( W \) is the count of samples in which feature \( f \) and \( c \) occurred together, \( f \) is the feature, and \( c \) is the class.

5.3 Information Gain {Features Reduction}

The entropy reduction for a specific feature offers a ranking of the features depending on their IG score. as equation (3)

\[ IG(f, c) = - \sum P(c) \log P(c) + \sum P(c|f) \log P(c | f) \]  

(3)

Such that \( P(c|f) \) is the joint probability where class \( C \) and feature \( f \) is co-occurred, \( P(c) \) denotes the marginal probability.

5.4 Evaluation

Finally, to evaluate the extent of resultant clusters and validate experimental results, the frequently used formulation is F-Measure. It is consequent from precision and recall, which are the accuracy procedures employed in the area of Information Retrieval (IR)” [13] as follows:

\[ Recall(N_p, C_q) = \frac{O_{pq}}{|N_p|} \]  

(4)

\[ Precision(N_p, C_q) = \frac{O_{pq}}{|C_q|} \]  

(5)

\[ F(N_p, C_q) = \frac{2 \times recall(N_p, C_q) \times precision(N_p, C_q)}{recall(N_p, C_q) + precision(N_p, C_q)} \]  

(6)

Such that \( O_{pq} \) is the number of members of a natural class, \( N_p \) in cluster \( C_q \), \( N_p \) is the natural class of a data object \( O_{pq} \) and \( C_q \) is the assigned cluster of \( O_{pq} \).

6. Proposed Work Implementation

The major task of the proposed work is to distinguish a forensic e-mail from a normal e-mail. In the proposed work e-mails passes through several phases each one has a specific function to reach the target.

6.1 Email Dataset

Enron's corpus is used for the purpose of experimentation. Enron's 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 and has a thousand samples and categories for the collection that the data is considered to be composed of real messages. 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/). Figures (2) present email data selection, user details list and details content for each user.

6.2 Data Preprocessing

Figure 2. Data Selection.
Data preprocessing is an important phase in the data mining process. It is a data mining technique that involves transforming raw data into an understandable format. In this phase, we remove the unwanted null values and special char symbols and remove the stop words. The apply part-of-speech tagging to assigns parts of speech to each word. The following sections present data preprocessing stages:

6.2.1 Tokenization Process

The text will tokenize into tokens or words to treat with each token separately, the text tokenized depending on the spaces between tokens.

6.2.2 Remove Stop Word

In natural language processing, stop words means a word that does not have any meaning such as "and", "the", "a", "an", and similar words, and is thus eliminated prior to classification. The stop words are not necessary for analyzation so we are going to load and remove the stop words from the Enron dataset as shown in Figure (3). Stop words available on this Website(https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords).

6.2.3 Stemming Process

This process reducing words to their original root. For instance, finance, financial, and financing may be converted to finance.

6.2.4 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 are available on this Website(https://myvocabulary.com/word-list/crime-vocabulary/).
6.2.5 Part-Of-Speech (POS) Tagging

The POS tagger is a tagging tool that tags each word and assigns parts of speech to each word (and another token). Part-of-speech categories include noun, verb, adverb, and adjective. The example word has Part-Of-Speech tags (JJ, JJS, JJR, VB, VBD, VBG, VBP, VBN, and VBZ) of an adjective and verb scores and so on.

6.2.6 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. A 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 categorized as positive, whereas if the score is less than zero. The purpose of this step analyzing the information presented in the email dataset and find a score each term. The term frequency is calculated each term. Then forensic Terms frequency is also calculated and each noun, verb, adverb, and adjective frequency. SentiWordNet3.0 dictionary is available on this Website (http://sentiwordnet.isti.cnr.it/) This process as shown in Figure (4).

Figure 4. Loading Forensic Word and Part-Of-Speech Tagging

6.3 Clustering

After the data preprocessing phase, 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 create the center point node in k-means dynamically as depicted in Figure. 5.
6.4 Naïve Bayes Algorithm

In this process e-mail messages are analyzed either it is a positive or negative sense by using the naïve bayes algorithm for class probability estimation with feature extraction. The naïve bayes classification algorithm is used for classifying the yes and no label. Yes, it represents positive scores. No, represent a negative score.

6.5 Feature Extraction

As shown in Figure 4, feature extraction is extracting the feature which is given in the dataset. The dataset is processed to get all the counters of (forensic, nouns, nounposscore, nounnegscore, verbs, Verbscore, Verbnegscore, adverbs, Advposcore, Advnegscore, adjectives, Adjposcore, and Adjnegscore) features, were consists of 13 columns. Each column represents the features and each row represents a feature of extracted from messages. The term frequency is calculated for each word. The columns represent the term frequency. Forensic word frequency is also being calculated in this process. As shown in figure (6).

6.6 Optimization

After the extracted features, 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) are forensic, noun, verb, adverb, and adjective attributes as shown in Figure (7).
The supervised classification relies on training the classifier using a set of labeled samples and evaluating model performance with another independent set. The goal of training a classifier is to create separations between groups of different class categories. Classification is performed on the training and testing data, predict the result and then it will provide a better prediction result. In experiments, using a support vector machine (SVM) algorithm for classification. The support vector machine algorithm learning in classifying normal and forensic email messages.

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. 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.

The frame for the classification process as shown in figure (8).

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.

As mentioned in section II 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 Ref. [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

![Figure 7. Optimization](image1)

**Figure 7. Optimization**

6.7 Classification

The supervised classification relies on training the classifier using a set of labeled samples and evaluating model performance with another independent set. The goal of training a classifier is to create separations between groups of different class categories. Classification is performed on the training and testing data, predict the result and then it will provide a better prediction result. In experiments, using a support vector machine (SVM) algorithm for classification. The support vector machine algorithm learning in classifying normal and forensic email messages.

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. 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.

The frame for the classification process as shown in figure (8).

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 8. Classification Frame by Using SVM Algorithm.](image2)

**Figure 8. Classification Frame by Using SVM Algorithm.**

7. Results and Discussion

As mentioned in section II 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 Ref. [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 Ref. [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 processed 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 of 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 helping for having efficient and high accuracy rates.

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, buy_r, campbell_l, cash_m, corman_s, cuillas_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 sets of data and accuracy was calculated in each case and the results as shown in the table (1). 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 95%. The proposed algorithm will provide a better prediction result. 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: Wamp server to handle MySQL database and used SentiWordNet3.0 and Stanford (tagger and parser).

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Number of Samples</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>450</td>
<td>80.7%</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>950</td>
<td>85.5%</td>
</tr>
<tr>
<td>Dataset 3</td>
<td>1425</td>
<td>76.9%</td>
</tr>
<tr>
<td>Dataset 4</td>
<td>1900</td>
<td>87.1%</td>
</tr>
<tr>
<td>Dataset 5</td>
<td>2345</td>
<td>76.02%</td>
</tr>
<tr>
<td>Dataset 6</td>
<td>2850</td>
<td>79.02%</td>
</tr>
<tr>
<td>Dataset 7</td>
<td>3325</td>
<td>95.12%</td>
</tr>
<tr>
<td>Dataset 8</td>
<td>3894</td>
<td>76.79%</td>
</tr>
</tbody>
</table>

8. Conclusions

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 that has 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, 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 research papers presented in section II in this paper.

9. References
2. Farkhund I, Liquate KA, Benjamin FCM, Murad D., (2010) "E-mail authorship verification for forensic investigation", Computer Security Laboratory CIISES, Concordia University Montreal, Quebec, Canada., pp.1591-1598.