Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques

Hissah AL-Saif, Hmood Al-Dossari
College of Computer and Information Sciences,
King Saud University,Riyadh,
Saudi Arabia

Abstract—Crime analysis has become a critical area for helping law enforcement agencies to protect civilians. As a result of a rapidly increasing population, crime rates have increased dramatically, and appropriate analysis has become a time-consuming effort. Text mining is an effective tool that may help to solve this problem to classify crimes in effective manner. The proposed system aims to detect and classify crimes in Twitter posts that written in the Arabic language, one of the most widespread languages today. In this paper, classification techniques are used to detect crimes and identify their nature by different classification algorithms. The experiments evaluate different algorithms, such as SVM, DT, CNB, and KNN, in terms of accuracy and speed in the crime domain. Also, different features extraction techniques are evaluated, including root-based stemming, light stemming, n-gram. The experiments revealed the superiority of n-gram over other techniques. Specifically, the results indicate the superiority of SVM with tri-gram over other classifiers, with a 91.55% accuracy.

Keywords—Crimes; text mining; classification; features extraction techniques; arabic posts; twitter

I. INTRODUCTION

Ssecurity is a very important element of life. Our most important needs cannot be met unless we are secure. Therefore, security is a necessity in human life that allows us to collectively or individually achieve our goals. Nowadays, with an increasing number of Internet users and the ease of accessibility afforded by the expansion of mobile data technology, there is a corresponding increase in the volume of information related crimes to utilize and analyze. Most of this information is unstructured, in the form of "free text." This trend has led to an increased importance upon devising methods to manage unstructured data.

Particularly in the mobile world, social media has become one of the most popular means of communication for private messages, pictures, and video, and various social networking sites have even become sources of global news, both social and political. Currently, there are many social media sites, such as Facebook, Twitter, and Snapchat Twitter is one of the most common social networking sites for casual chats, sharing photos and ideas, and the transfer of information and news through text, limited to 140 characters,called "tweets". The number of users of Twitter— around 500 million— are tweeting approximately 340 million times per day. Many people in Arabic countries are using Twitter regularly, which makes it suitable for this study. Due in part to its small, readable tweets, Twitter has become an important means of communicating news of criminal activity.

In Saudi Arabia the official spokesman for the Ministry of the Interior stated in a 2015 news conference that the crime rate in Saudi Arabia had reached 270 crimes per day, about 100, 000 crimes annually. The volume of this activity places extreme burdens on the State to adequately analyze crime data [1].This requires the careful study and analysis of crime, its escalation, and its geographic spread, in order to objectively develop strategies to slow the crime rate. A major challenge faced by law enforcement is that there is too much information concerning criminal activity as a result of the increase in the number of crimes, technological advances, and increasing population density. The sheer amount of data requires significant time and effort to analyze and utilize.

Rising crime rates coupled with the spread of this news through social networking sites was a major motivation for the current study. In fact, the public was made aware of many of these events through such platforms. The main objective of this research, therefore, is to extract usable, credible information in order to identify the nature of crimes and to assist law enforcement with future crime prevention, thereby contributing to ensuring the security of humanity. Text classification poses a challenging task for the Arabic language, due to its richness and complexity. In this research, we attempt to address this issue by performing an intensive comparison to assess different machine learning algorithms and the impact of various feature extraction techniques on accuracy. In particular, this research involves four classification algorithms, including SVM, DT(C4.5), CNB, and KNN, in terms of accuracy, speed of training, and execution time. Also, four well-known feature extraction techniques are evaluated, including root-based stemming, light stemming, characters-based n-gram, and words-based n-gram for the Arabic language.

The research seeks answers to the following questions:

- How can we detect and classify crimes from Twitter posts?
- What is the best algorithm for classification, from selected machine learning algorithms for the Arabic language in general and in the crime domain in particular?

- What is the best method for features extraction from selected techniques for the Arabic language in general and the crime domain in particular?

The research is organized as follows: section 2 discusses overview of text mining while section 3 presents background information on Arabic language, section 4 descriptions of crime categories, section 5: presents related work section 6 overview of machine learning algorithms, section 7 illustrates of the methodology and data collection, section 8 presents of the experiments and results, and the conclusion and future work section 9.

II. TEXT CLASSIFICATION

Text mining aims to take advantage of natural language to find new, previously unknown, relationships between a large volume of ambiguous text documents that contain a large number of words and various grammatical structures [2]. The importance of appropriate document categorization has increased dramatically over the last two decades, due to the explosion of web-based contents. According to Hotho [2] and Vandana [3] 85% of information is stored as text in form reports, news, etc. This clearly demonstrates the utility of text mining to classify document to desired category according to its contents. Before beginning of the text mining, it is necessary to transform each document into a more appropriate form for text classification, this is called pre-processing. The pre-processing generally covers the process of structuring the input text in steps, such as tokenization, eliminating stop words, feature extraction and features weighting. It is a critical step in text classification especially and has a large impact positively or negatively on classification accuracy [3] [4].

Features extraction is important step to mitigate language complexities, especially in Arabic language. Features extraction includes different techniques such as stemming that based on morphological analysis. Stemming takes advantage of the fact that most word variations have similar semantic interpretations and can be handled as one root word [5] [6].

In Arabic, stemming is separated into two types based on morphological analysis: root –based stemming, and light stemming. The basic idea of stemming is to reduce a given word to its root, while light stemming only seeks to remove common affixes in order to produce the stem of a given word, rather than producing a root. For example, in stemming, all of these words— المدرسه (school), مدرس (teacher), and الدروس (lessons) — share a stem, though they have different meanings. On the other hand, light stemming reduces ,المدرسين "المدرسون " which means “teachers,” to (مدرس) which means “teacher” [6] [7]. Khoja’s stemmer [8] one of the most widely Arabic root-based stemmer. It is a dictionary-based algorithm that removes affixes and extracts the root word by matching the residual word with patterns. The main drawbacks of this method are its dependence on a dictionary, which must be updated on an ongoing basis, and it replaces the vowel characters with the letter “و” which could lead to mistakes in the extracted root [7]. The Larkey stemmer, light 10, is well-known light stemmer that seeks to remove most frequent suffixes and prefixes [9].

Stemming is useful for extracting features and constructing feature vectors, as well as for enhancing retrieval performance, due to the reduction of variations of a given word to its grammatical root. Stemming also reduces the complexity of the indexing structure, which leads to an improvement in the overall performance [5] [10]. In some cases, however, stemming can reduce the performance of classification, because many conflicting words can have the same root. This reflects the main benefit of light stemming, which focuses on the meaning of the word rather than identifying its root. Therefore, light stemming can improve classification performance, as it maintains word meaning, unlike the stemming approach [6].

Rather than extracting roots, some researchers prefer to utilize statistical methods such as n-gram, which is usually used to classify documents without any stemming. The basic idea is to create a profile for each document by generating all possible continuous n-item slices. This method can be used for a single word, in order to generate all possible continuous n-characters slices, or for a single sentence, in order to generate all possible n-words slices. After generating N-grams for all words in a document, the profile is saved, in order to compare word similarity. The main advantages of this method are that it is language-independent and it works very well with files that contain linguistic errors and noise [11] [12].

After completing pre-processing, text documents transform to vectors by calculating terms weight, which are used in the learning phase for machine learning algorithms. The most common method is term frequency–inverse document frequency (TF–IDF). TF–IDF reflects the importance of a word in a document to collection of documents as in the following equation [13]:

$$w_{dt} = tf_{dt} \times idf_t \tag{1}$$

Where tf_{dt} represents the word occurrences in the document divided by the total number of words, and

$$tf_{dt} = \frac{n_{ij}}{\sum_n n_{kj}} \tag{2}$$

idf_t represents the importance of word in documents

$$idf_t = log \frac{|D|}{\{d:t_i \in d\}} \tag{3}$$

III. ARABIC LANGUAGE

The Arabic language is extremely important, as it is the native language of Arabic countries, and the second language for Islamic countries. Arabic is spoken by more than 310 million people as their native language, and more than 250 million people as a second language, according to the Summer Institute of Linguistics (SIL International) statistics for languages and science professional studies [14] . Unlike Romantic and Germanic languages, Arabic language is an agglutinative language written from right to left. It consists twenty-eight different characters: أ ب ت ث ج ح خ د ذ ر ز س ش

و ه ن م ل ك ق ف غ ع ظ ط ض ص. Arabic contains diacritical marks (ḍammah, fataha and kasra,) that may change the meaning, for example, (مدرَسه) means "school" while (مدرِسه) means "teacher". It has many synonyms that are used frequently, such as "الحزن، الغم، الغمة، الأسى", all of them have the same meaning: "a sense of sadness". Arabic has an extremely complex morphology that relies on more than 11,000 roots and 900 patterns listed in largest Arabic dictionary [15].

A root is the core form of a word, something one cannot further analyze without losing the word's meaning. Simply put, it is the word without any additions at the beginning (prefix), in the middle (infix), or at the end (suffix). Usually these additions, called affixes, are added in order to create new words and meanings [4] [7].

IV. Crimes

Before discussing the types of crimes and their classifications, crime must be defined first. Crime is a breach of the rules of society by committing and act that is detrimental to community [16]. In reality, criminality varies from one country to another, and is determined by weighing the openness of a society against its adherence to religious and cultural traditions.

For the sake of this study, we have investigated what kinds of crimes occur in Arab countries. To do so, we have utilized several sources such as the Statistical Yearbook of crimes issued by the Ministry of the Interior in Saudi Arabia. It divided crimes into ten major categories, such as crimes related person, crimes related mind, crimes related money, etc [17]. The second source is the Uniform Crime Reporting (UCR) that was issued by the Federal Bureau of Investigation (FBI) in the U.S. Department of Justice that examines all types of crime in American society in detail [18]. The study constructing a tree representing three levels of crimes. The first level includes detecting crimes in "tweets." The second level divides crimes according to types of victims—property, individuals, or society. Crimes against persons include all of those crimes whose victims are individuals, with the aim of hurting a particular person, such as murder, sex offense, or kidnapping. Crimes against property include all of those crimes whose purpose is usually to get money or property, such as a robbery, bribery, or burglary. Crimes against society are crimes that are aimed at hurting the community in general and usually do not have a specific victim. The last level classifies crimes according to the offense, such as murder, theft, etc.

This study accounts for the customs and traditions in Arabic countries in its tree construction— for example, sex crimes in America are limited to coercive sexual crimes, while in Arabic countries they include both coercive and consensual outside the frame of marriage. Some types of minor offenses were excluded, because they are not typically tweeted, such as driving infractions. Also excluded were crimes that rarely occur in Arabic society, based on the statistics issued by the Ministry of the Interior, such as technical crimes.

V. Related Work

Extensive English-language research has been conducted on text mining. However, there is little research in the Arabic language on text mining in general and even less research concerning specific crimes. To the best of our knowledge, this is the first study that has focused on crime detection and classification in Arabic social networking sites. This section will address two fields: classification in Arabic language and classification in crime domain in general.

A. Classification in Arabic language

The study [19] is a model by Support Vector Machine (SVM) to categorize Arabic documents. In order to improve the performance, Inverse Document Frequency (IDF) is applied. IDF represents the importance of a given word by calculating the word's occurrence in a document against a collection of documents. To reduce high dimensionality of features in documents, Chi Square x^2 statistics are applied to measure independence between feature words within each category, which eliminates the least important features. The authors conducted an experiment on more than 1,500 documents across nine categories. The experiment used SVM and then compared it with other classifiers, such as Naïve Bayes and KNN. The results showed better performance for SVM than other classifiers, where it reached 88.11% accuracy, while Naïve Bayes and KNN reached 84.54% and 72.72% respectively.

The study [20] conducted experiments on Arabic documents to classify them into predefined categories according to the subject of each document. The authors used root-based stemming that introduced by [2] to extract roots in the pre-processing phase. The Naïve Bayes classifier was used for the categorizing phase on 300 documents and the results showed an average accuracy of 62.23% for the documents tested.

In [21], the authors applied a Neural Network, a simulation of a human brain, to classify "Nine Books," which contains a total of 453 documents across 14 different categories. The network had three layers: a layer for the introduction of the documents, a hidden layer containing an activation function, and an output layer for the classification of the documents. In the learning phase, the initial value of weights for each layer is given randomly. It then updates the weights based on the computing error rate, until the best weight values for accuracy are reached. In the experiment, the input layers contained 739 nodes, which was equal to the number of features, while there were only 10 layers in the hidden layer. The output layer contained 14 layers, equal to the amount of categories in the text. The authors used 20 epochs for training, as this was deemed the best tradeoff between efficiency of the neural network and classification accuracy. The experiment yielded positive results, with 88.33% as the average accuracy.

In [22] a new tool called Arabic Text Classification (ATC) was built, which is used to clean Arabic documents and calculate the importance of each word using a Chi Square. The Chi Square calculates the correlation between the document and each class in order to generate high quality matrices. These matrices are used in weighting features in order to extract the most important features and reduce the volume of

documents. The classification was carried out by C5.0 Decision Tree and Support Vector Machine (SVM) on seven Arabic corpora, each of which contained a number of documents and categories. The results showed a better average accuracy for the C5.0 algorithm, which achieved 78.42% accuracy, than the SVM algorithm, which only reached 68.65% accuracy.

The study [23] conducted an experiment on documents containing Facebook comments in Arabic text, then compared the results to English text. The author argues that the classification of short texts, such as Facebook comments, adds another challenge to classification due to the limited number of words. The sample contains a large number of comments that belong to the classes "food" and "weather" in both Arabic and English. The four types of classification algorithms applied were SVM, Naïve Bayes, K-Nearest Neighbor, and Decision Trees. The experiment yielded interesting results, with the accuracy of the Arabic language 11% higher than the English language in both categories. The authors highlighted that this result was because Arabic contains more discriminative words than English does, especially in the "food" and "weather" categories.

In [24], the authors used three classifiers, including DT,NB, and Equational Minimal Optimization (SMO), to construct a classification model to predict a document category. The dataset was collected from multiple websites and Saheeh AL-Bukhari's book, then divided into four categories: economics, politics, sports, and sayings of the prophet Mohammed. The algorithms were applied to 1,000 documents, and the results reflected a better performance from Naïve Bayes than the other classifiers, as it reached 85.25% accuracy.

The study [11] applied Manhattan distance to measure dissimilarity and Dice's to measure similarity when classifying Arabic documents. The corpus collected from several online newspapers. Next, a characters-based N-gram was applied to documents to generate all possible slices. The authors conducted an experiment on the generated profile across four categories: sports, the economy, technology and weather. The results showed better performance from Dice's measurement than the Manhattan distance; the former reached 89% accuracy, while the latter reached 66% accuracy.

In [7] set of experiments were conducted on seven Arabic corpora in order to classify them into predefined categories according to the subject of each document. The corpora were collected from: *BBC* Arabic, *Aljazeera* corpus, and multiple other websites. The classification was carried out according to DT, *KNN*, *SVMs*, and NB variants. Furthermore, the experiments evaluated two features extraction techniques: root-based stemming and light stemming. The author recommended light stemming as a features extraction technique in order to enhance classification accuracy because it preserves word meanings. The results demonstrated the superiority for SVM over other classifiers, especially as it averages 94.11% accuracy. Another study [25] evaluated two approaches of stemming light 10 and Khoja stemming—on Arabic text. The author used the public dataset "Arabic Articles," which consisted of 2,700 documents classified according to nine categories. The experiments conducted by two tools suit Weka and Rapidminer to compare between them. The results showed a better average accuracy for the light 10, which achieved 98.20% accuracy, than the Khoja stemming, which reached 97.80% accuracy. Adding to work of previous study , the study [26] evaluates these approaches by multiple similarity measures while the study [27] used KNN to evaluate these approaches. The results demonstrated the superiority of light 10 over Khoja stemmer, which supported the study's [25] results.

The study [28] evaluated three preprocessing tasks, including normalization, stop word removal, and light stemming, in terms of accuracy. The author utilized machine learning algorithms NB, KNN, and SVM to categorize Arabic documents. The results showed that light stemming demonstrated the best accuracy among them. It also indicated that, in some cases, removing the stop word can have a negative effect on performance.

Sentiment analysis is special type of text classification that aims to extract subjective information, such as emotions and opinions, in order to classify them into positive or negative categories. The study [29] investigated different representation models, as well as features reduction techniques and their impacts on sentiment analysis.

TABLE I. COMPARISON OF THE REVIEWED STUDIES

Study	Dataset	Study goal	LS	RS	SN	WN	Machine learning algorithms
Mesleh [19]	Newspaper	Text Classification	_	_	_	_	SVM, NB, KNN
Hatem et al.[20]	Arabic Documents	Text Classification	_	✓	_	_	NB
Fouzi et al.[21]	Books	Text Classification	_	_	_	_	Neural Network
Harbi et al.[22]	Online Newspapers	Text Classification	_	_	_	_	SVM, DT
Nawaf et al.[23]	Facebook Comments	Text Classification	✓	_	_	_	SVM, NB, KNN, DT
Khreisat [11]	Online Newspapers	Text Classification	_	_	✓	_	_
Saad et al.[7]	Online Newspapers	Text Classification	✓	✓	_	_	SVM, NB, KNN, DT
Hmeidi et al.[25]	Arabic Articles	Text Classification	✓	✓	_	_	SVM, NB, KNN, DT
Froud et al. [26]	Online Newspapers	Text Classification	✓	✓	_	_	_
Duwairi at al.[27]	Arabic Documents	Text Classification	✓	✓	_	_	KNN
Ayedh et al. [28]	Arabic Documents	Text Classification	✓	_	_	_	SVM, NB, KNN
Duwairi et al.[29]	Online Newspapers	Opinions Classification	✓	✓	✓	✓	SVM, NB, KNN
Brahimi et al.[30]	Twitter Posts	Opinions Classification	✓	✓	✓	_	SVM, NB, KNN
The Proposed Study	Twitter Posts	Text Classification	✓	✓	✓	✓	SVM, NB, KNN, DT

The experiments were performed on two different datasets: the first dataset consisted of 322 reviews of political articles collected manually from the Aljazeera2 website, while the second dataset was public and contained 500 reviews of movies. The results showed that the accuracy of opinion classification was affected directly by dataset type and preprocessing techniques. Also, the study [30] investigated the impact of the preprocessing stage on the opinion classification of two datasets collected from Twitter. The results were comparable to those of the study [29]. Table 1 summarized the reviewed studies where LS: light stemming,

RS: root-based stemming, SN: character-based n-gram and word-based n-gram.

B. Classification in Crime Domain

The authors in [31] applied sentiment analysis to English tweets related to crime in order to identify the users' behavior and attitude regarding crimes. They aimed to identify cities in the USA where the most crimes occur and where the least crimes occur. This has been tested by a geographical analysis of crimes that have occurred in the ten most dangers cities and ten safest cities, as determined by Forbes magazine. These results were similar to those of the study provided by Forbes magazine.

In [32] a web-based system was built for crime analysis and detection in Sri Lanka. The system collected articles from different newspapers by using the Crawler4j web crawler. The study covered the crimes occurred between 2012 and 2014. The articles were classified into two categories: crime and non-crime, using SVM. The results were extremely positive, with an accuracy of 95.71%. The authors in [33] used classification techniques to construct new software called "Z-crime" based on the Decision Tree (ID3) algorithm. The software was used to detect any terrorist attacks that might occur via email analysis.

The study [34] set up a system to efficiently detect the patterns of crimes that have occurred in India. Data concerning the crimes was gathered from several sources, including specialized news sites, blogs, and social sites. The collected data was classified by type of crime using the Naïve Bayes classifier and the results were extremely positive, with an accuracy of 90%.

VI. Machine Learning Algorithms

Recently, there has been a trend toward using machine learning to classify documents by building classifiers through learning instead of the old methods. In general, text classification in machine learning consists three phases. The first phase is data pre-processing to make the text convenient to train a chosen classifier. The second phase uses the classifier to construct a model based on a set of labeled examples, or training set. The last phase is evaluating the constructed model by various performance measures to gauge the model's success [19][35]. In this study different text-classification algorithms are used, such as a Naïve Bayes (*NB*), decision tree (*DT C4.5*), support vector machines (*SVMs*), and *K*-nearest neighbors (*KNN*).

A. The Naïve Bayes (NB)

The Naïve Bayes classifier is one of the simplest classification methods. It is part of the probabilistic family based on Bayes' Theorem[36] [37]. The Naïve Bayes classifier assumes that there is no relationship between features in dataset. It calculates the conditional probability of each new feature belonging to each class, then chooses the class that promises the highest probability. In general, there are two models used in text classification based on the Naïve Bayes conditional assumption: Multivariate Bernoulli Model and the Multinomial Model. In the Multivariate Bernoulli Model, a feature vector is represented as a binary vector that takes two values (1, 0) according to the appearance of a word at least one time in the document. Conversely, the Multinomial Model takes into account the frequency of a word, where feature vectors are represented by an integer indicating the repetition of a word, not just the word's presence [38].

Complement Naive Bayes (CNB), an improved version of the Multinomial Naïve Bayes (MNB) that overcomes its original weaknesses. The MNB is affected by the number of examples in a class, giving a high weight for a class that has more examples. In addition, it assumes there is no relationship between features. The CNB overcomes these problems and works very well with text data and has the highest accuracy among these variations [39].

B. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that represents a data set as points in space separated by lines, constructing a "hyperplane" model for prediction and classification [40]. This hyperplane is used as a boundary to make decisions separately to place each new tuple in its optimal class.

In text classification, SVM is an effective method in the classification of high dimensionality feature space, because the complexity of hypotheses is measured by size of the margin rather than the number of features in the document. SVM needs a large memory capacity to execute properly, and therefore the size of the training set affects the speed of execution, which can be very slow. This is contrary to NB, which is simpler and has a significantly lower memory requirement. Further, larger training sets significantly improve the accuracy of NB, unlike SVM, which already has a high benchmark [41][42].

C. Decision Tree (DT)

Decision Tree is one of the most popular classifiers based on statistics, contains a set of nodes and edges that help in decision-making. Many algorithms are used to construct decision trees such as ID3, C4.5. These algorithms aim to construct the appropriate decision tree for a data set that reduces the error rate. The superiority of C4.5 is clear; in general, it deals with numbers and missing values unlike in ID3. Also, it can deal with over-fitting by a burning procedure to stop the growing of the tree [40].

D. K-Nearest Neighbors (KNN)

KNN is one of the most widely used classifiers based on instance-based learning [40]. The basic idea behind KNN is to assign new objects to the class that receives the majority of votes from its neighbors in an n-dimensional space.

VII. Study Methodology

In general, our system consists of four stages, as presented in Fig. 1.

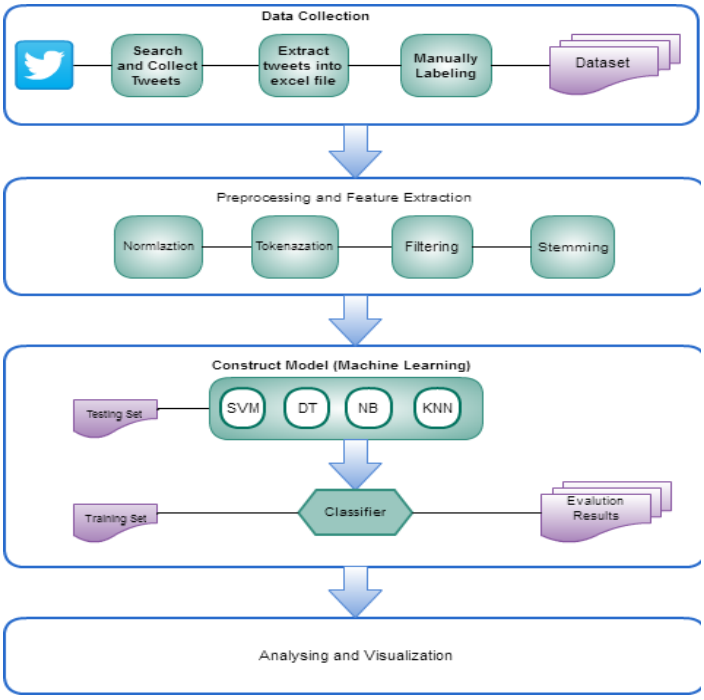

Fig. 1. The Four Stages of The Construct Model

A. Data Collection

In data collection stage, crime-related tweets are collected from Twitter using different tools, such as Twitter API and Topsy, and then extracted to an Excel file with UTF-8 Unicode to support Arabic text. Each tweet is labeled with its corresponding category based on the contents used to create the training set. The collected data set consists of more than 8,000 Arabic tweets from 2013 to 2016. The search targeted specialized news accounts on Twitter. The dataset contains approximately 4,000 different news items about crimes and 4,000 news items about political, health, and new technology issues. The data set contains approximately 187,748 words, with 23,994 distinct words

B. Preprocessing

In our system, preprocessing contains four steps, outlined as follows:

- Normalization is applied to make data more consistent. For example, some of the same words can be written differently in Arabic. For example, "ابتزاز ," which means "extortion," can be written as "إبتزاز." These spelling variations can negatively affect classification accuracy. Thus, normalization is used to overcome this problem. In Arabic, three letters must be normalized:
  - "أ," "آ," and "إ" will be normalized as "ا";
  - "ه" will be normalized as "ة";
  - "ي" will be normalized as "ى."
- Tokenization splits the textual data into a sequence of tokens and removes spaces so that each word is separated by only one white space. This step is very important, especially to text data, which is in an unstructured form that needs to be transformed into a suitable form for processing.
- Filtering reduces the size of the file and improves the efficiency of the text classification process. It is used to remove all non-alphabetic characters, especially signs frequently used in Twitter, such as (#) for hashtags and (@) used for usernames. Moreover, stop word filtering is used to remove all the frequent words that do not affect the meaning of a sentence and carry no value, such as prepositions sentence by comparing them to a list of Arabic built-in stop words. In addition, Length filtering is used to remove any words that exceed a specific length or that are less than specific length. We set the shortest length at 3 because there are many words in our data set with a length of 3 characters, such as "سرق" and "قتل."
- Features extraction, three methods have been used to extract features, including the light stem, root-based stem, N-gram, as well as original features as "bags-of-words" to establish the most suitable approach for our dataset. We have adopted TF–IDF to create vectors, since the initial experiments indicated that its results were significantly better than the binary results.

C. Construct model and Analysis

Machine learning algorithms use training dataset to construct a model that used to classify new tweets while the testing set is used to evaluate the constructed model. In this study four algorithms were used, including *NB, DT C4.5, SVM and KNN*. The classification accuracy is influenced by many of the characteristics of the volume of data in the training phase, diversity, the right selection of features, the type of classifier, and targeted language all have an impact on accuracy [13] [43]. Several measurements are used to evaluate the model for accuracy, recall and precision as in the following equations:

- Accuracy represents the ratio of tuples that are correctly classified by the model. It is calculated using the following equation: $\frac{(TP+TN)}{N}$. (4)
- Recall represents the number of items correctly classified as positive divided by the total number of positive. It is calculated using the following equation: $\frac{TP}{(TP + FN)}$. (5)
- Precision represents the number of items correctly classified as positive divided by the total number of positive predictions. It is calculated using the following equation: $\frac{TP}{(TP + FP)}$ (6)

where True positives (TPs) represent all tuples that were correctly predicted as crimes. False Negatives (FNs) represent all tuples that were incorrectly predicted as non-crimes. False positives (FPs) represent all tuples that were incorrectly

predicted as crimes. True negatives (TNs) represent all tuples that were correctly predicted as non-crimes. In this study, we measure the accuracy of the model by applying k-fold cross validation. The dataset is split into K groups, each of which has its own training set and test set. In our experiment, we selected K as 10, and then the original data set was divided into 10 folds, each one containing the same amount of data.

VIII. Results and Evaluation

The experiments were conducted by Rapid Miner. Rapid Miner is a type of open source software developed in 2006 to provide a complete environment for machine learning, text mining, and analytical prediction which make it suitable to our study. The experiments were run on a laptop device with a 64-bit machine and 16 GB memory.

A. Classifiers' Performance

The performance of the model, evaluated by different measures, including accuracy, recall, and precision. Accuracy is the most important measure for evaluating the model. Fig. 2 represents the accuracy of different classifiers, including SVM, DT, KNN, and CNB for crimes classification in level three.

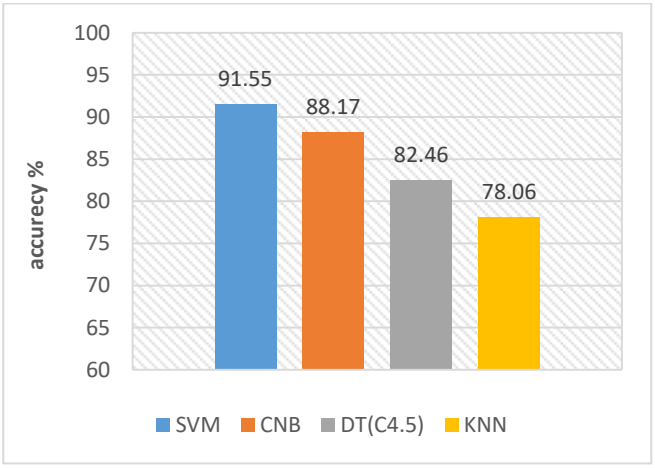

Fig. 2. Classification Accuracy for Classifiers in Classifying Crimes According to Type.

The Naive Bayes usually has a low performance rate with text because of its high dimensionality. Furthermore, the Naive Bayes is usually affected by an unbalanced dataset since it gives a high weight to a class that has more examples. In addition, it assumes that there is no relationship between features. However, we find good results, with 93.29%, 90.88%, and 88.17% in the first, second, and third levels, respectively. These good results are due to our use of a special type of Naive Bayes, which is CNB. CNB works with texts perfectly because it overcomes the weaknesses of Naive Bayes.

KNN got the worst results among the classifiers, with 78.06% in classifying according types. The performance of KNN was affected directly by the feature-extraction techniques because it measures the distance between words. Arabic stemmers have poor performance and return many unrelated words, leading to the poor performance of KNN. Meanwhile, DT results were somewhat low due to the impact of the large number of features in the dataset, which made it more difficult to build the appropriate tree. Fig. 5 illustrates impact of the number of classes on accuracy. The most affected was KNN, which had 89.44% in detection crimes and decreased to 78.06% in classifying crimes, while SVM was the least affected, reflecting its strength and durability. Fig 3 shows the impact of number of classes on classifier accuracy.

Fig. 3. The Impact of Number of Classes on Accuracy.

Accuracy is not enough to ensure the validity of model. Thus, recall and precision are measured for evaluation. Fig. 4 represents the recall and precision of SVM with tri-gram for crime classification.

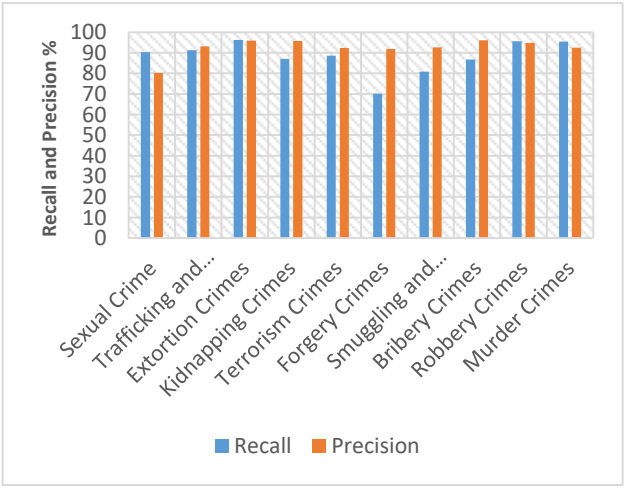

Fig. 4. Recall and Precision for SVM Classifier in Classifying Crimes According to Type.

B. Feature Extraction Reduction

The number of features is inversely correlated with classification accuracy. Because the crime dataset contains 187,765 words, we seek to reduce the number of features. Fig .5 reflects the magnitude of reduction according to each of these methods, as well original dataset after tokenizing and filtering.

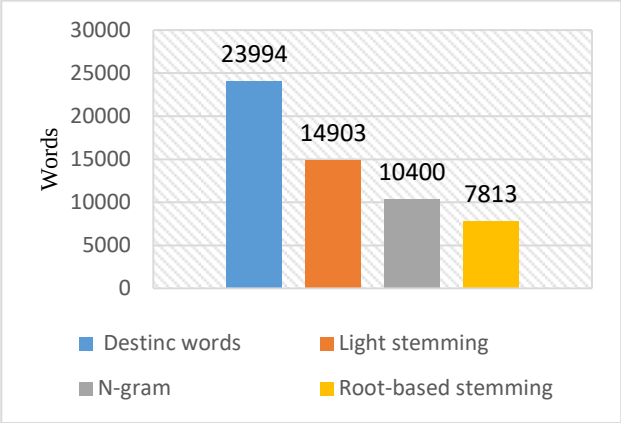

Fig. 5. Impact of Feature-Extraction Techniques on Crime Dataset.

The figure indicates that root-based stemming provides the highest reduction in features with 67.4% due to the fact that a large number of words share the same root. Light stemming results in the smallest reduction, as it removes only prefixes and suffixes. The application of the tri-gram reduction resulted in approximately 56.7% fewer words. The large reduction occurred for two reasons: The nature of the dataset had many similar words; because of that, we specialized in a specific domain, which is the crime domain, as well as took a small n value (three) for the gram that led to generate many similar triple words, which were deleted by RapidMiner.

C. Feature-Extraction Technique's Impact on Accurecy

Three methods have been used to extract features, including the light stem, root-based stem N-gram, as well as original dataset. The N-gram has been tested for both character-based and word-based on different values of N to get the best value so as to ensure the highest accuracy. Table 2 represents the accuracy of bilateral, triple, quadruple, and quintet grams with SVM and CNB.

TABLE II. ACCURACY OF THE BILATERAL, TRIPLE, QUADRUPLE, AND QUINTET GRAMS

classifiers		Accuracy	
		Character-based	Word-based
SVM	2-gram	87.82	78.28
	3-gram	91.55	73.74
	4-gram	89.57	70.10
	5-gram	87.15	67.63
CNB	2-gram	79.41	81.14
	3-gram	88.17	80.12
	4-gram	87.08	78.95

The results demonstrate low performance for words-based n-gram in contrast to characters-based n-gram. It takes only 78.28 and 81.84 when N=2 for SVM and CNB, respectively. Also, the table shows the supremacy of the tri-gram with character-based over other values because 85% of the words in Arabic have triple roots. Despite the fact that the tri-gram generates many incomprehensible words, it often leads to the appearance of the root directly for different words that have the same root. For example, "شرعيه" means "legitimacy" and becomes "شرع, رعي, عيه," and "يشرعن" means "legitimizes" and becomes "يشر, شرع, رعن." The root produced for both words directly is "شرع." Also, "القبض," meaning "the arrest," becomes "الق, لقب, قبض." The root produced directly is "قبض." This fact improves classification accuracy. Fig. 6 represents the impact of different feature-extraction on accuracy.

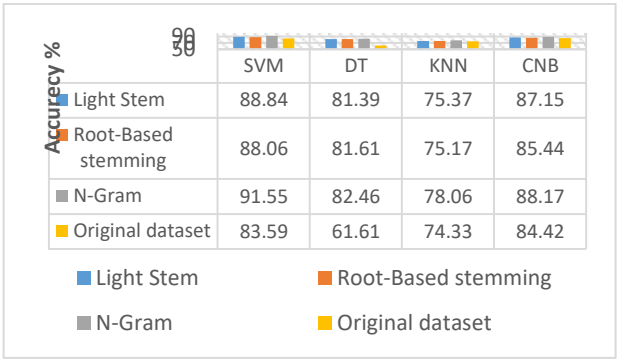

Fig. 6. Impact of Feature-Extraction Techniques in Classifying Crimes According to Type.

As can be seen in Fig. 6 the tri-gram achieved the greatest accuracy in classifying crimes according types, with 91.55%, 82.46%, 78.06%, and 88.17% for SVM, DT, KNN, and CNB, respectively. The results of root-based and light stemming are approximately similar for DT, SVM and KNN, while light stemming is better for CNB. The lowest result was achieved by DT in classifying crimes for original dataset, at 61.61%.

The overall results reflect the conspicuous superiority of the N-gram to other methods in all classifiers. This is due to reducing the number of features to more than half and also because of our use of the triple-gram. The results of light stemming and root-based stemming are somewhat convergent, although the results of light stemming are better because, as we mentioned earlier, it keeps the word meaning, while root-based stemming produces the same root for many words that have different meanings. The worst results are for raw text due to the large number of variations in words that reflected negatively on the performance of the classifier.

The study [44] and [45] involved completing an experiment to compare light stemming and root-based stemming. The results confirm that light stemming achieves a better level of accuracy than does root-based stemming. Also, the study [7] supports the idea that, despite the fact that convergent results of two types were received. The author reported that the root-based stemming achieves a high level of accuracy because it works perfectly with the triple root and most of Arabic words have triple root, whereas light stemming is better from a linguistic and semantic viewpoint.

D. Classifiers' Training Time

Training time is an important factor for building classifiers, especially for the high dimensionality of a text dataset. Fig.7 shows the training time of different classifiers with different feature-extraction techniques, for crime dataset.

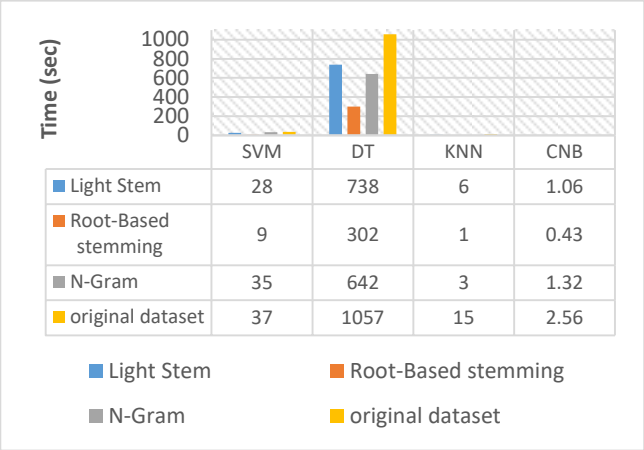

Fig. 7. Training Time for Classifiers in Classifying Crimes According to Type.

CNB was the fastest among the algorithms, taking just a few seconds of training time. The high speed of CNB is due to the simplicity of the account conditional probability, which contains just addition and division operations. KNN has few amount of time for training due to its nature, as there is no need to build a model; simply just store features in the memory for later use in classifying. The impact of feature-extraction techniques on DT, CNB, and SVM was similar, where root-based stemming took less time because it reduced the number of distinct words to 67.4% of total words, original set consumed more time to build a model for a large number of distinct words.

E. *Classifiers' Execution Time*

Executing time is another factor used to evaluate the classifiers. Executing time includes the total time for preprocessing, training time, and applying the model time to evaluate the validity of the model. Fig. 8 shows the execution times of different classifiers with different feature-extraction techniques, as well as those for raw text.

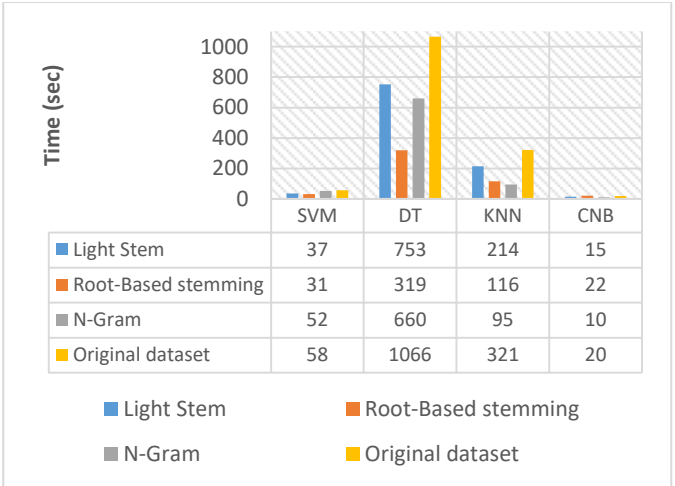

Fig. 8. Execution Time for Classifiers in Classifying Crimes According to Type.

Compared with the training time, we found that the results are somewhat convergent for DT, SVM, and CNB because the application of the model for classifying consumed only a few seconds or a few fractions of a second in some cases. In addition, the preprocessing time consumed only a few seconds. On the contrary, the KNN results differ entirely because, as mentioned earlier, it does not build a model, and the training time is close to zero. This demonstrates the need to spend a lot of time measuring the similarity to the K-nearest neighbors' instances.

CNB was the fastest among the algorithms in all levels even with the execution time, while the slowest was DT. The results of the effect of extracting features of the execution time are dramatically different from the impact on the training time. The impact of feature-extraction techniques on DT, KNN, and SVM depends on the number of words where root-based stemming takes less time and the raw text consumes more time. Meanwhile, CNB is different from the rest due to its high speed of building the model and testing, which take fewer than three seconds. The speed of CNB is linked directly to the speed of the extraction-feature techniques. CNB with root-based stemming was the slowest because root-based stemming usually consumes time to remove the affix and then extract the root, while raw text is the second because a large number of words need to be tokenized.

The study [44] makes a comparison between light stemming and root-based stemming in term of execution time with KNN. The results show that light stemming takes more time, which reinforces the validity of our findings.

Note that the speed of the execution model is affected by three factors: the algorithm used; total number of classes; and the types of feature-extraction techniques. Fig. 9 shows effect of number of classes on classifiers' time. DT was the classifier most affected by the number of classes: 258 seconds in detecting crimes to 660 seconds in classifying crimes according to types. A tree with a great depth had to be built to cover the 11 classes. SVM and CNB rose slightly, while for KNN, there was a disparity between the raising and going down, which mean KNN not affected by the number of classes.

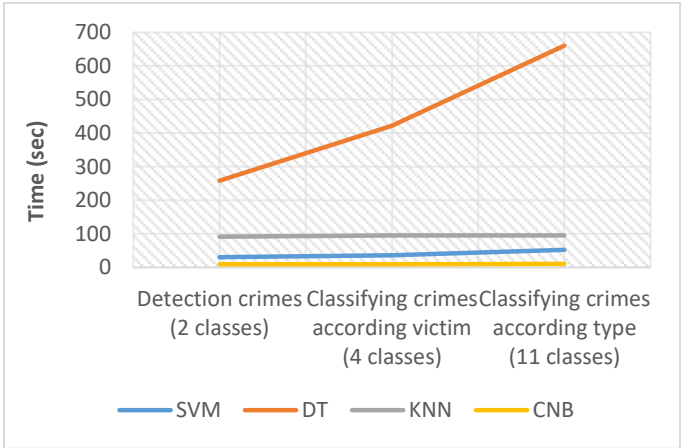

Fig. 9. The Effect of Number of Classes on Classifiers' Execution Time.

IX. Conclusion

In this paper, we put in place, utilizing various machine learning algorithms, a system that is capable of detecting crime-related tweets. Then, we conducted an investigation of different features in the form of light stemming and stemming, together with N-grams. The results indicate that root-based stemming yielded the best results in terms of feature reduction, while the character-based N-gram obtained the best level of accuracy. Based on the findings, we recommend tri-gram to use for the Arabic language, particularly in classification in specific domain due to the nature of the Arabic language, which often relies on the triple roots.

SVM had the best accuracy among the classifiers while the worst accuracy was achieved by KNN. In terms of speed, CNB was the fastest among the classifiers in both training time and execution time, while DT was the slowest, especially in classifying the crimes due to the large number of classes. The most affected of the class numbers in terms of accuracy was KNN. While the most affected of the class numbers on speed was DT.

In future work, we plan to evaluate other machine learning algorithms, such as neural networks, association rules, and others. Also, we intend to expand our analysis to include spatial and temporal analysis to find out when and where crime has spread in the past, and when and where it is most likely to spread in the future.

References

[1] A. Shamany, “Interior: 270 crimes every day. Committed by unemployed and juveniles,” 2015.

[2] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining,” LDV Forum - Gld. J. Comput. Linguist. Lang. Technol., vol. 20, pp. 19–62, 2005.

[3] V. Gupta and G. S. Lehal, “A survey of text mining techniques and applications,” J. Emerg. Technol. Web Intell., vol. 1, no. 1, pp. 60–76, 2009.

[4] A. Farghaly and K. Shaalan, “Arabic natural language processing : Challenges and solutions,” ACM Trans. Asian Lang. Inf. Process., vol. 8, no. 4, pp. 1–22, 2009.

[5] H. K. Al Ameed, S. O. Al Ketbi, a. a. Al-Kaabi, K. Al Shebli, N. Al Shamsi, N. H. Al Nuaimi, and S. S. Al Muhairi, “Arabic light stemmer: A new enhanced approach,” Second Int. Conf. Innov. Inf. Technol., pp. 1–9, 2005.

[6] F. A. Allah, S. Boulaknadel, a. El Qadi, and D. Aboutajdine, “Arabic Information Retrieval System Based on Noun Phrases,” 2006 2nd Int. Conf. Inf. Commun. Technol., vol. 1, no. April, pp. 720–725, 2006.

[7] M. Saad, “The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification,” Comput. Eng., no. August, p. 172, 2010.

[8] K. S. and G. R, “Stemming Arabic text,” Lancaster, 1999.

[9] L. Larkey, L. Ballesteros, and M. Connell, “Light stemming for Arabic information retrieval,” Arab. Comput. Morphol., pp. 221–243, 2007.

[10] N. P. Katariya and M. S. Chaudhari, “Text Preprocessing for Text Mining Using Side Information,” vol. 3, pp. 3–7, 2015.

[11] L. Khreisat, “Arabic text classification using N-gram frequency statistics a comparative study,” Conf. Data Mining| DMIN’06, pp. 78–82, 2006.

[12] S. H. Mustafa and Q. A. Al-Radaideh, “Using N-grams for Arabic text searching,” J. Am. Soc. Inf. Sci. Technol., vol. 55, no. 11, pp. 1002–1007, 2004.

[13] B. Baharudin, L. H. Lee, and K. Khan, “A Review of Machine Learning Algorithms for Text-Documents Classification,” J. Adv. Inf. Technol., vol. 1, no. 1, pp. 4–20, 2010.

[14] “Languages of the World (18th ed.).,” 2015. [Online]. Available: http://www.ethnologue.com. [Accessed: 16-Sep-2016].

[15] Z. M, “The crown bride of the jewels dictionary.” 1965.

[16] A. Shanqity, “The treatment of Holy Qur’an to crime.” Madinah.

[17] Ministry of the Interior in Saudi, “Statistical Yearbook of crimes,” Riyadh, 2014.

[18] F. B. of I. (FBI), “The Uniform Crime Reporting (UCR),” 2014.

[19] A. M. A. Meshel, “Chi square feature extraction based SVMs Arabic language text categorization system,” J. Comput. Sci., vol. 3, no. 6, pp. 430–435, 2007.

[20] H. Noaman and S. Elmougy, “Naive Bayes Classifier based Arabic document categorization,” 2010 7th Int. Conf. Informatics Syst., pp. 1–5, 2010.

[21] F. Harrag and E. A.- Qawasmah, “Improving Arabic Text Categorization using Neural Network with SVD,” J. Digit. Inf. Manag., vol. 8, no. 4, pp. 233–239, 2010.

[22] S. Al-Harbi, a Almuhareb, and a Al-Thubaity, “Automatic Arabic text classification,” 9es Journées Int. Anal. Stat. des Données Textuelles, pp. 77–84, 2008.

[23] M. Faqeeh, N. Abdulla, M. Al-Ayyoub, Y. Jararweh, and M. Quwaider, “Cross-lingual short-text document classification for facebook comments,” Proc. - 2014 Int. Conf. Futur. Internet Things Cloud, FiCloud 2014, pp. 573–578, 2014.

[24] A. H. Wahbeh and M. Al-Kabi, “Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text,” Abhath Al-yarmouk “Basic Sci. Eng.,” vol. Vol. 21, no. 1, pp. 15– 28, 2012.

[25] I. Hmeidi, M. Al-Ayyoub, N. a. Abdulla, a. a. Almodawar, R. Abooraig, and N. a. Mahyoub, “Automatic Arabic text categorization: A comprehensive comparative study,” J. Inf. Sci., vol. 41, no. 1, pp. 114–124, 2014.

[26] H. Froud, A. Lachkar, and S. Ouatik, “A comparative study of root-based and stem-based approaches for measuring the similarity between arabic words for arabic text mining applications,” Adv. Comput. An Int. J., vol. 3, no. 6, pp. 55–67, 2012.

[27] N. Khasawneh, R. Duwairi, and M. N. Al-Refa, “Feature ReductionTechniques for Arabic Text Categorization,” J. Am. Soc. Inf. Sci. Technol., vol. 14, no. 4, pp. 90–103, 2009.

[28] A. Ayedh, G. TAN, K. Alwesabi, and H. Rajeh, “The Effect of Preprocessing on Arabic Document Categorization,” Algorithms, vol. 9, no. 2, p. 27, 2016.

[29] R. Duwairi and M. El-Orfali, “A study of the effects of preprocessing strategies on sentiment analysis for Arabic text.,” J. Inf. Sci., vol. 40, no. 4, pp. 501–513, 2014.

[30] B. Brahimi, M. Touahria, and A. Tari, “Data and Text Mining Techniques for Classifying Arabic Tweet Polarity 1,” vol. 14, no. 1, pp. 15–25, 2016.

[31] S. Mine, “Crime pattern detection using online social media,” 2014.

[32] I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, and A. Wijayasiri, “Crime analytics: Analysis of crimes through newspaper articles,” 2015 Moratuwa Eng. Res. Conf., no. April, pp. 277–282, 2015.

[33] M. Sharma, “Z - CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree,” 2014 Int. Conf. Data Min. Intell. Comput. ICDMIC 2014, 2014.

[34] S. Sathyadevan, M. S. Devan, and S. Surya Gangadharan, “Crime analysis and prediction using data mining,” no. August 2014, pp. 406–412, 2014.

[35] F. Sebastiani, “Machine learning in automated text categorization,” ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, 2002.

[36] V. Korde and C. N. Mahender, “Text Classification and Classifiers: A Survey,” Int. J. Artif. Intell. Appl., vol. 3, no. 2, pp. 85–99, 2012.

[37] G. Sanguinetti, “Text Classification using Naive Bayes,” no. February, 2012.

[38] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification,” AAAI/ICML-98 Work. Learn. Text Categ., pp. 41–48, 1998.

[39] and K. D. T. Rennie J, Shih L, Teevan J, “The Poor Assumptions of Naive Bayes Classifiers,” no. In Proceedings of the Twentieth International Conference on Machine Learning (ICML)., 2003.

[40] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Data MiningConcepts Tech., pp. 3–26, 2000.

[41] I. Dilrukshi and K. De Zoysa, “Twitter news classification: Theoretical and practical comparison of SVM against Naive Bayes algorithms,” 2013 Int. Conf. Adv. ICT Emerg. Reg., no. December, pp. 278–278, 2013.

[42] A.Sheshasaayee and G. Thailambal," Comparison of Classification Algorithms in Text Mining", International Journal of Pure and Applied Mathematics,pp.425-433,2017.

[43] C. Aggarwal and C. Zhai, “A Survey of Text Classification Algorithms,” Min. Text Data, pp. 163–222, 2012.

[44] M. Al-refai, “Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization Rehab Duwairil Department of Computer,” no. September 2007, pp. 446–450, 2008.

[45] D. A. Said, N. M. Wanas, N. M. Darwish, and N. H. Hegazy, “A Study of Text Preprocessing Tools for Arabic Text Categorization,” Proc. Second Int. Conf. Arab. Lang. Resour. Tools, pp. 330–336, 2009.