Order No:
Serial No:

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

ECHAHID HAMMA LAKHDAR UNIVERSITY OF EL-OUED
FACULTY OF EXACT SCIENCES
COMPUTER SCIENCE DEPARTMENT

End of study thesis

ACADEMIC MASTER

Field: Mathematics and Computer Science
Sector: Computer Science
Speciality: Distributed Systems and Artificial Intelligence (SDIA)

Theme

Analysis of online crime in social networks

Submitted by:
- ✓ HAMMOUDA saadia
- ✓ SAHRAOUI latifa

Supported on:
30/09/2020

Board of examiners:

✓ M. LEJDEL Brahim	Supervisor	Univ. of El Oued
✓ M. BEGGAS mounir	President	Univ. of El Oued
✓ M. BELILA khaoula	Examiner	Univ. of El Oued

Academic year: 2019/2020

ACKNOWLEDGMENT

First of all, we praise Almighty **Allah** without whom we would have not accomplish this humble work and who guided us through the path of science.

On this occasion, we owe thanks, gratitude and appreciation to our supervisor **LEJDEL brahim** who did not spare us his advice and guidance which was the best help to us since the beginning of the research by raising challenge to accomplish the work in its set time despite the difficult circumstances.

Moreover, we would like to express our thanks and appreciation to the honorablemembers of the jury for devoting their precious time and great efforts to evaluate our work, which without whom, it would have not to be refined as it is not preserved fromimperfections.

We extend our deep gratitude to each of **NAOUI Med Anouar**, **NAGOUDI El Moataz Billah** and **LACHRAF Raki** for their assistance and support next to all the teachers in the Department of Information Technology.

Finally, we owe thanks to our colleague and sister **KEBACHE Afaf** for the help in translation.

ABSTRACT

Crime analysis has become a critical area of study to help law enforcement (such as the police) protect civilians. Due to the rapid increase in population, crime rates have increased dramatically. For this, proper analysis in real time is considered a very important option. Indeed, Text mining is an effective tool that can help solve this problem and categorize efficiently crimes. The proposed system aims to detect crimes in the Facebook comments of Algerians to identify which are the dangerous cities and which are not dangerous cities in Algeria. In this dissertation, classification techniques were used to detect crimes and identify their natures in relation to the different classification algorithms. In the experimentation part, we evaluated different algorithms, such as SVM, DT, NB, and RF, in terms of precision of detection in the crime. In addition, different feature extraction techniques were evaluated, including Has Positives Word (HPW), Has Negative Word (HNW), Positives Word Count (PWC), Negative Word Count (NWC) and Comment Length (CL). Finally, we conclude that result obtained by the RF classifier is the best obtained result compared to other classifiers with an accuracy of 96.1%.

Key words:

Crimes, Machine learning, Classification, Features extraction techniques, Algerian comments, Facebook.

RÉSUMÉ

L'analyse de la criminalité est devenue un domaine d'étude critique pour aider les forces de l'ordre (comme les polices) à protéger les personnes civiles. En raison de l'augmentation rapide de la population, les taux de criminalité ont considérablement augmenté. Pour cela, l'analyse appropriée en temps réel est considérée comme une option très importante. En effet, L'exploration de texte est un outil efficace qui peut aider à résoudre ce problème et classer les crimes de manière efficace. Le système proposé vise à détecter les crimes dans les commentaires Facebook des Algériens pour identifier quelles sont les villes dangereuses et les villes non dangereuses en Algérie. Dans ce mémoire, des techniques de classification ont été utilisées pour détecter les crimes et identifier leurs natures par rapport aux différents algorithmes de classification. Dans la partie expérimentation, nous avons évalué différents algorithmes, tels que SVM, DT, NB et RF, en matière de précision de détection dans le domaine de la criminalité. De plus, différentes techniques d'extraction de caractéristiques ont été évaluées, y compris HPW, HNW, PWC, NWC et CL. Enfin, nous concluons que le résultat obtenu par le classificateur RF est le meilleur résultat obtenus par rapport aux autres classificateurs avec une précision de 96.1%.

Les mots clés:

Criminalité, Apprentissage automatique, Classification, Techniques d'extraction des fonctionnalités, Commentaires algériens, Facebook.

ملخص

أصبح تحليلالجريمة مجال دراسة بالغ الأهمية لمساعدة فرق تطبيق القانون (مثل الشرطة) على حماية المدنيين. بسبب الزيادة السريعة في عدد السكان، زادت معدلات الجريمة بشكل كبير. لهذا، يعتبر التحليل المناسب في الوقت الفعلي خياراً مهماًا للغاية. يعد تحليل النصوص أداة فعالة يمكن أن تساعد في حل هذه المشكلة وتصنيف الجرائم بكفاءة. يهدف النظام المقترح في هذه المذكرة إلى الكشف عن الجرائم الواردة في تعليقات الجزائريين على فيسبوك للتعرف على المدن الخطرة والمدن غير الخطرة في الجزائر. في هذه المذكرة تم استخدام تقنيات التصنيف لكشف الجرائم وتحديد طبيعتها من خلال استخدام خوارزميات التصنيف المختلفة. في الجزءالتجريبي، قمنا بتقييم خوارزميات مختلفة مثل SVM, DT, NB و RF من حيث دقة الكشف في مجال الجريمة. بالإضافة إلى ذلك، تم تقييم تقنيات مختلفة لاستخراج الميزات، بما في ذلك وجود الكلمات الموجبة في التعليق HPW، وجود الكلمات السالبة في التعليق HNW، عدد الكلمات الموجبة PWC، عدد الكلمات السالبة NWC وطول التعليق CL. في الأخير، نستنتج أن النتيجة التي حصل عليها مصنف RF هي أفضل نتيجة تم الحصول عليها مقارنة بالمصنفات الأخرى بدقة 96.1 %.

الكلمات المفتاحية:

الجرائم، التعلم الالي، تصنيف، ميزات تقنيات الاستخراج، تعليقات جزائرية، الفيسبوك.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS & CONCEPTS

SVM Support Vector Machine

DT Decision Tree

KNN K-Nearest Neighbors: is a non-parametric method used for classification and regression.

NB The Naïve Bayes

CNB Complement Naive Bayes

TN True Negative

TP True Positive

FP False Positive

FN False Negative

MSA Modern Standard Arabic: is the standardized and literary variety of Arabic used in writing and in most formal speech.

RF Random Forest

HPW Has Positives Word

PWC Positives Word Count

NWC Negative Word Count

HNW Has Negative Word

CL Comment Length

UNICEF the United Nations Children's Fund

USA United States of America

API Application Programming Interface: is a computing interface which defines interactions between multiple software intermediaries.

HTML Hyper Text Markup Language

MLP Multi Layer Perceptron: is a class of feedforward artificial neural network (ANN).

CNN Convolutional Neural Network: is a class of deep neural networks, most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.

TFIDF Term Frequency Inverse Document Frequency

SMOTE Synthetic Minority Oversampling Technique

DEC Different Error Cost

GATE General Architecture for Text Engineering: is an infrastructure for developing and deploying software components that process human language.

JAPE Java Annotation Patterms Engine

GUI Graphical User Interface

ID3 Iterative Dichotomiser 3

UTF-8 Unicode Transformation Format-8bit

JSON Java Script Object Notation

TPR True Positive Rate

TNR True Negative Rate

PPV Positive Predictive Value

NPV Negative Predictive Value

LDA Latent Dirichlet Allocation

ANN Artificial Neural Networks: is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.

LSTM Long Short-Term Memory

BOT Bag-Of-Tokens: is then much more a "bag-of-words". Words (ie Tokens) is the atomic unit of text comparison. If we want to compare two documents, we count how many tokens they share in common.

CM Confusion Matrice: also known as an error matrix,is a specific table layout that allows visualization of the performance of an algorithm,Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).

XML eXtensible Markup Language

XHTML eXtensible Hyper Text Markup Language

AI Artificial Intelligence

RFC Random Forest Classifier

URL Uniform Resource Locator

PC Personal Computer

NLTK Natural Language Tool kit

ANLP Applied Natural Language Processing

POS Part Of Speech

ANNIE A Nearly-New Information Extraction

RBF Radial Basis Function

IR Information Retrieval

DA Dialectal Arabic

CA Classical Arabic

NN Neural Networks

ME Maximum Entropy

ML Machine Learning

TC Text Classification

TM Text Mining

NLP Natural Language Processing: is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language.

BoW Bag-Of-Words: is a simplifying representation used in natural language processing and information retrieval.

P percentage

LingPipe is tool kit for processing text using computational linguistics, it architecture is designed to be efficient, scalable, reusable, and robus.

OpenNLP The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.

CSV is a plain text file that contains a list of data. These files are often used for exchanging data between different applications.

Crawler4J is an open source web crawler which is used widely for simple applications and is easily modifiable to fit into the application. It supports multithreaded and resumable crawling. It is implemented in java so that it can be easily integrated into java projects.

Z crime is depicting the "Zero Crime" in the society.

GENERAL INTRODUCTION

Security is the most important element in our life, in terms of priority. Our most important needs cannot be met unless we are secure. Therefore, security is necessary in human life, which allows us to achieve our goals either collectively or individually.Nowadays, with the 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 with 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.

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 Snap chat. Facebook is one of the most common social networking sites for casual chats, sharing photos and ideas, and transferring information and news through text that is limited to 140 characters which is called "Comment".

The monthly number of active Facebook users worldwide in the first quarter of 2020 is 2.6 billion users. Facebook is considered as one of the biggest social media network worldwide. In the third quarter of 2012, the number of active Facebook users surpassed one billion, making it the first social network to do so ever [19].Many people in Algeria are using Facebook regularly (There were $23,670,000$ Facebook users in April 2020 [20] which makes it suitable for this study). Due in part to its small, readable comments, Facebook has become an important way

for communicating news of criminal activity.

According to their terms of reference, statistics provided by the National Security Representative and the National Gendarmerie representative revealed that Algeria had recorded a total of 207, 000 crimes of various kinds over the past nine months [2019/2020]; approximately 700 per day [21]. The scale of this activity places extreme burdens on our country to adequately analyze crime data .This requires a careful study and analysis of crime, its escalation, and its geographic spread, in order to objectively develop strategies to slow the crime rate.

Rising crime rates along with the spread of these news through social networking sites was a major motivation for the current study. The main objective of this research, therefore, is to extract usable and credible information in order to identify the nature of crimes and to assist law enforcement with future crime prevention, thus contributing to ensure the security of humanity.

Text classification imposes 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. Particularly, this research involves four classification algorithms, including SVM, DT, RF, and NB, in terms of accuracy, speed of training, and execution time. Also, different features extraction techniques are evaluated, including HPW, HNW, PWC, NWC and CL.

The dissertation is divided into four chapters. In the first chapter, we focus on the state of the art of online crime analysis, particularly related works. The second chapter deals with the particularities of the Arabic language. In Chapter three, we study machine learning and we talked about the evaluation system. Chapter four presents the implementation of the model and the experience, in addition to a discussion of the results obtained. Finally, we add a conclusion and some perspectives.

CHAPTER 1

STATE OF THE ART

1.1 Introduction

Social media allows disseminating, distributing, and communicating information, in addition to sharing contents or expressions. As any means of communication the content may either encourage crime or not.

Detecting crime and related comments is crucial for protecting the community. The search for crime is demanding since it is impossible to cover all the words, especially those which are written in local dialects or belong to certain cultures or even writing new words like "ندیروا کاس" (taking a cup) which means (let's have a drink) or "ضرك نقتلك / تو نقتلك" (I will kill you) which means only beating.

In this section, we will discuss the crime theme, how it is detected online? and what it is related to sentiment analysis ,and we will offer some works related to our work.

1.2 Crime

In ordinary language, a crime is an illegal act which shall be punished by the state or another competent authority. In modern criminal law; the term crime does not have any simple or universally accredited definition, one proposed definition is that a crime or offence (or criminal offence) is a harmful act not only to some individual but also to a community, society, or the state ("a public wrong"). Such acts are forbidden and shall be punished by law.[22]

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.

1.3 Types of crimes:

Although, there are many different kinds of crimes, criminal acts can generally be divided into five primary categories: personal crimes, property crimes, inchoate crimes, statutory crimes, and financial crimes.[23]

1.3.1 Personal Crimes:

Personal crimes are those that result in physical or mental harm to another person[23]. This category of crime includes:

- Assault and battery
- Arson
- Child abuse
- Domestic abuse
- Kidnapping
- Rape and statutory rape
- Murder
- Robbery
- Terrorism...[24]

1.3.2 Property of Crimes:

Property of crimes typically involve interference with the property of another. Although they may involve physical or mental harm to others, they primarily result in the deprivation of the use or enjoyment of property. Crimes against property include:

- Burglary
- Larceny-theft
- Robbery
- Motor vehicle theft
- Shoplifting...[23]

1.3.3 Inchoate Crimes:

Inchoate crimes refer to those crimes that were initiated but not completed, and acts that assist in the commission of another crime. Inchoate crimes require more than a person simply intending or hoping to commit a crime. Rather, the individual must take a "substantial step" towards the completion of the crime in order to be found guilty. Inchoate crimes include aid and provocation, attempt, and conspiracy. In some cases, inchoate crimes can be punished to the same degree that the underlying crime would be punished, while in other cases, the punishment might be less severe.[23]

1.3.4 Statutory Crimes:

Statutory crimes include those crimes, in addition to the crimes discussed above, which are proscribed by statute. Four significant types of statutory crimes are alcohol related crimes, drug crimes, traffic offenses, and financial/white collar crimes...[23]

1.3.5 Financial and Other Crimes:

Finally, financial crimes often involve deception or fraud for financial gain. Although white-collar crimes derive their name from the corporate officers who historically perpetrated them, anyone in any industry can commit a white-collar crime. These crimes include many types of fraud and blackmail, embezzlement and money laundering, tax evasion, and cybercrime... [23]

1.4 Discussions:

In order to discover the situation of crime in Algeria, we have studied the development of the crime rate in Algeria. This rate shows the widespreading of crimes in society which is in increase.

These are some statistics:

- According to the report of UNICEF, there were $1,100$ kidnappings of children in Algeria from 2001 to January 2016. [35]
- Algeria says thwarts $40,000$ illegal immigrants per year from reaching Europe. [25]
- The annual report, prepared in cooperation with the Regional Security Office at the USA Embassy in Algeria said, "in 2018, authorities arrested early $45,000$ people, including foreign nationals, for drug-related offenses, a major jump from 2017". [36]
- The judicial police chief announced that there are 5183 persons arrested under judicial orders last year (2019), including 580 persons researched in theft cases, 641 is researched in the cases of forming a association of 20 persons who were researched in murder cases and 73 attempted murder cases, 277 cases of beatings and arson wounds. 253 fraud and fraud, 89 smuggling cases..... [26]

Based on the statistics, it became clear that the most prevalent crimes, especially in the major Algerian cities, are: murder, theft, kidnapping children, trade and smuggling of money and drugs, illegal immigration, etc., and they are constantly increasing.

In addition, the first motive that encouraged us to conduct this study, along with technological development, is the fact that these crimes are expressed by the perpetrator on social media before they are committed. some take vow to kill (himself or others), others to take revenge for their sons or daughters, some make commercial deals for drugs and legally prohibited actions, and some ask how to get out of the country, regardless the way (illegal immigration) ... and so on.

The challenge of this study is to collect the crucial Algerian terms that are considered as secret words by which the perpetrator expresses his emotional state on social media before committing the crime.

1.5 Use of Social Media for Prediction:

The proliferation of social media services has prompted a surge of interest in using the associated data for various predictive purposes.

As shown, many researchers have attempted to use social media to predict or detect disease outbreaks[1], election results [2], macroeconomic processes (including crime) [3], box office performance of movies [4], natural phenomena such as earthquakes [5], product sales [6], and financial markets [7].

The use of social media for illegal activities and measures to control it is also the subject of the study of several researchers. There are researchers who argue that Internet provides new opportunities for all types of criminal activities including organized crime. Major criminal activities on social networks include: [8]

- Terrorism and extremism,
- Unrest and violence,
- Sex and drug trafficking, and
- Cyber-bullying and victimization.

Criminals use social networks for recruiting, organizing, and carrying out illegal activities. In the past, law enforcement agencies have relied upon Twitter and other social networks for intelligence gathering. The other authors emphasize the need for research in modeling to detect criminal activities on social networks. [8]

For that, we will seek to use Facebook comments to predict crime in major cities of Algeria. To the best of our knowledge, this study have been conducted in Saudi Arabia before, however, it can be considered as new interest in Algeria since there are only few previous works on this subject.

1.6 Sentiment analysis:

The crime's nature reflects the criminal's emotional state. Nowadays, the criminals have become technologically sophisticated and often expressing their emotions on the web. The World Wide Web's phenomenal growth has resulted in the increase of the users who tend to express their opinions online. For this reason, we have highlighted this element:

Sentiment analysis was used to determine a writer's/speaker's attitude toward a topic or an overall contextual polarity of a text. Researchers use this analysis to measure emotions in online texts. The rise of social media fueled interest in using sentiment analysis to identify public opinions and interests. Several open source software tools utilize machine learning, statistics, and natural language processing techniques to automate sentiment analysis on a large collection of texts that have been gathered from various sources. [9]

Sentiment analysis is a two-step process that includes both subjectivity classification and sentiment classification. The term subjectivity classification , is defined as distinguishing factual sentences from those used to present opinions, before analyzing sentiments. Paragraphs that present facts are typically removed so the researcher can focus on those paragraphs in which the author expresses opinions. Both NB classification and Cut Based classification are used for subjectivity classification.[9]

The term sentiment classification , is defined as detecting sentiment polarity of the subjective sentences. This sentiment classification is also divided into two categories: binary sentiment classification and multi-class sentiment classification. Binary sentiment classification involves classifying sentiments either positive or negative. Multi-class sentiment classification involves classifying sentiments into one of five categories: strong positive, positive, neutral, negative and strong negative.[9]

The most common machine learning techniques used for sentiment classification include NB, ME, and SVM. Most sentiment analysis algorithms use simple terms to express sentiment. However, the cultural factors, linguistic nuances, and differing contexts prevent researchers from drawing the sentiment accurately.[9]

1.7 Related Works:

Sentiment Analysis and its applications have spread throughout many languages and domains. Regarding to Arabic and its dialects, we see an increasing interest simultaneously with increase of Arabic texts volume in social media. However, there is little research in the Arabic language on less concerning specific crimes.

This section will address two fields: Sentiment analysis and Crime analysis in social media or web sites in general.

1.7.1 Sentiment analysis:

Assia Soumeur and al have developed an automatic analyzer of the sentiments of Algerian Facebook users.

They have written some python codes to interact with the Facebook graph API. they have selected more than 20 brand pages of companies in Algeria that have the greatest numbers of subscribers. For each page, they downloaded an average of 250 publications and for each of these, they downloaded the text, the image (if used), and a maximum of 350 comments.[10]

At the end of this phase, they managed to collect more than 100,000 comments. After collecting their raw corpus of Facebook users' comments, and for best quality, they manually annotated them as positive, neutral or negative. To facilitate the task, they created a small application to help the annotators. Each comment is annotated by two native, university-educated annotators.[10]

If they give the same polarity, then it is taken and no validation is necessary; otherwise, a third annotator is needed and the vote of the majority is taken. As a result, they obtained a sentiment-annotated corpus of comments by Algerian Facebook users as described in Table1.1

Class	**Number of Comments**
Positive	19313
Negative	3808
Neutral	2352
All	25475

Table 1.1: Statistics on the sentiment-annotated corpus.

The comments are directly collected from the Facebook (API) which contains HTML addresses, hashtags (#), stop words, etc. A preprocessing step to clean up the texts was carried out as shown in Figure 1.1 [10]

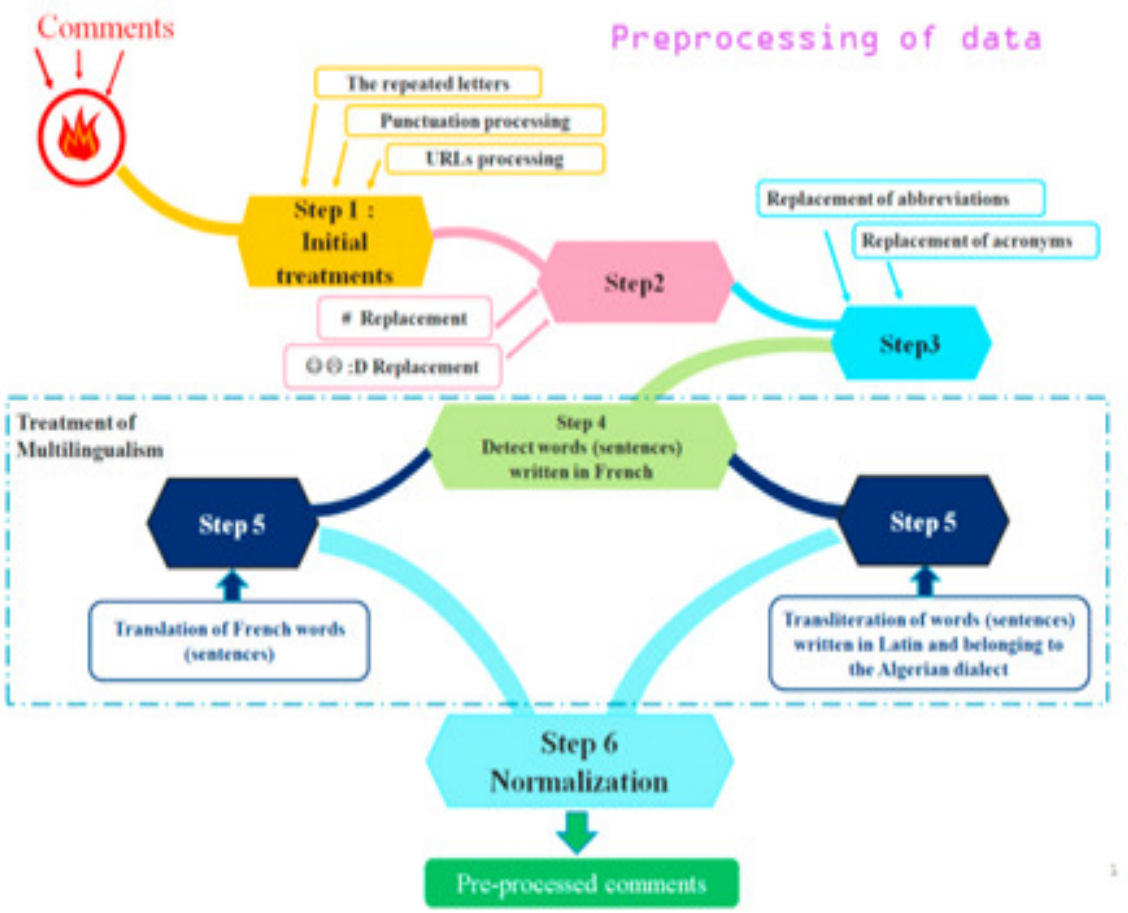

Figure 1.1: Corpus preprocessing steps.

The researchers execute the following process:

- They decided to remove any letter that is repeated more than twice in sequence.
- The punctuation was removed and they only kept the question mark ("?") and the exclamation mark ("!"), since they can affect the sentiment orientation of the comment.
- They have also handled numbers and tags: each number has been replaced by the word “meaning number“, and each tag has been replaced by the word "tag".
- They have decided to keep emoticons in the comments and they have chosen to organise them in 6 different categories (happy, sad, angry, disgusted, frightened and surprised). As such, each emoticon/emoji was replaced by the name of the category to which it belongs.
- A list of 112 abbreviations and their equivalents as illustrated was created in Table 1.2

- In order to handle Multilingualism, they have applied the following steps:
  - Identification of the French Language Words.
  - Translation
  - Transliteration
- They applied a normalization .[10]

Abbreviation	**Word**	**Language**	**Meaning**
Pk	Pourquoi	French	why
dac	d'accord	French	ok/agreed
bcp	beaucoup	French	a lot
jms	jamais	French	never
tjr	toujours	French	always
dsl	désolé	French	sorry
bzfl	bzf	Algerian dialect	a lot/ many

Table 1.2: Excerpt of the list of abbreviations.

In this work, they have applied two ML methods: a (MLP) NN and a (Deep) CNN. The best (MLP) and (CNN) architectures accuracies are given in Table 1.3 [10]

Architecture	**N°of hidden layers**	**N°of neurons per layer**	**Accuracy**
MLP	3	50/50/50	81.6
CNN	3	50/50/50	89.5

Table 1.3: Best accuracy results for the MLP and CNN models.

The accuracy results, showing clearly that the (CNN) model outperforms the (MLP) one with 89.5 accuracy against 81.6 accuracy for the latter. [10]

1.7.2 Crime analysis:

Several similar work has been carried out in English language in the area of online crime detection, however, there is a little research in Arabic for specific crimes.

During this work, we have had the opportunity to address several projects related to our research work. In this section, we will present various works which related into our work. These works are divided into two categories, which use foreign languages and other use Arabic language.

1.7.2.1 Foreign Languages:

Isuru Jayaweera and al have proposed an online crime analysis and detection system in Sri Lanka. The proposed system performs crime analysis operations such as hotspot detection, crime comparison and crime pattern visualization.[11]

The system collects articles from various newspapers using the Crawler4J.

The proposed system consists of seven major components. They are crawler, classifier, entity extractor, duplicate detector, data base handler, analyzer and graphical user interface. Highlevel architecture of the system including the above components is given in Figure 1.2[11].

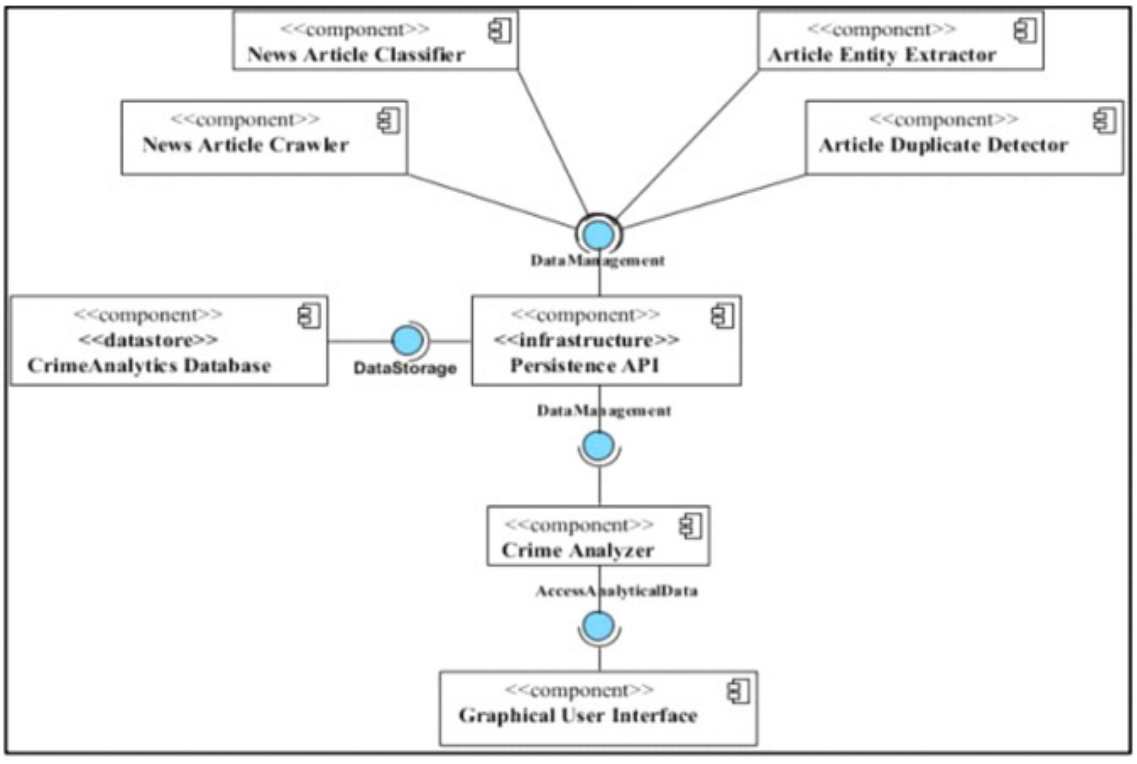

Figure 1.2: High level architecture of the proposed system.

The content required for crawled articles is stored in the database for further processing. A center crawler was implemented by expanding a public crawler known as Crawler4J. It has been used to extract the required data (page content, etc.) from the HTML crawled pages.

The Weka Library was used to process documents before they were used in the classification process. Documents have been transformed to feature vectors while removing stop words from them using TFIDF transformation.[11]

Stemming/ lemmatization has been performed in order to reduce words into their base forms. LibSVM library has been used to implement the classifier. CSVC SVM type and RBF kernel has been used for classification. Grid search has been performed in order to find appropriate values for cost and gamma values. Training article collection was highly unbalanced because there are large numbers of non-crime articles compare to crime articles when we consider a particular sample from an article population.[11]

Initially hybrid sampling technique was used in order to handle data unbalanced problem. Minority class has been up sampled using SMOTE and majority class has been down sampled using support vectors. However, it was not very effective so that different error cost DEC approach has been used. In that approach a higher error cost has been assigned to the minority class in order to reduce the effect of the majority class. [11]

Entity Extractor is used to extract important entities from the classified newspaper articles, such as crime date, location, police, court, victim count...etc. Several ancillary processes have been carried out to do the required preprocessing on the document corpus to prepare documents for named entity extraction.[11]

General Architecture for Text Engineering: is an infrastructure for developing and deploying software components that process human language. (GATE) has been used for text processing as well as for entity extraction. It consists of several ready-made applications such as ANNIE, LingPipe and plugins such as date normalizer, OpenNLP and others.[11]

Text content of each article has been tokenized using ANNIE English tokenizer and each single word has been considered as a separate token. Sentences have been segmented using ANNIE sentence splitter. The weaknesses that it possessed have been minimized using custom JAPE rules. ANNIE POS tagger has been used to perform POS tagging and identify phrases, referential proper nouns etc. ANNIE POS tagger has been combined with Stanford POS tagger in order to enhance the POS tagging process. [11]

Dates given in article have been normalized with respect to the published date using GATE date normalizer plugin. After identifying the location of a crime, the corresponding district has been identified using Google Maps (API). Extracted addresses are sent to Map (API) to obtain the details of the district. Proposed hybrid approach is compared with conventional ML and context based named entity extraction approaches. [11]

These measurements were taken based on the crime location entity extraction. The results clearly display that the proposed hybrid approach is better than the two conventional methods.

The main purpose of Duplicate Detector is to identify exact/near duplicates of newspaper articles and remove them from the database.

All database transactions are handled using Database Handler. This has been implemented using Hibernate framework .[11]

Analyzer module will perform crime analysis operations on processed crime articles. It will perform following crime analysis steps :

- Hot spot detection.
- Crime comparison.
- Crime pattern visualization.

In this work , they used Web based GUI to visualize crime statistical details of the previous years. Then law enforcement officers and other interested users will be able to use this system effectively and efficiently for crime analysis processes.[11]

The article classifier is evaluated to measure its classification accuracy. Different classification accuracy measurements such as accuracy, sensitivity, specificity and Fscore are used. Results are obtained using cross validation method. They are given in Table 1.4[11].

Accuracy Measurement	**Value**%
Accuracy	95.7163
Sensitivity	96.22
Specificity	95.21
G-mean	95.71
ROC	95

Table 1.4: Evaluation Results of Different Classification Accuracy Measurements.

Article entity extractor is evaluated for measuring its accuracy and comprehensiveness of identifying entities. Entities such as crime location, crime date and crime type are used to obtain values for accuracy measurements. Results of the evaluation are given in Table 1.5 [11].

Entity Category	**Accuracy of Extractions%**
Crime location	82
Crime Type	82
Crime Date	74

Table 1.5: Evaluation Of Crime Type and Crime Date Entites.

The study covered the crimes occurred between 2012 and 2014. By using "SVM", the articles were classified into two categories: crime and non-crime, using SVM. The results were extremely positive, with an accuracy of 95.71 %.[11]

In addition to this study,we find :

Raja Ashok Bolla study, who 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.[9]

M. Sharma and al have used classification techniques to construct new software called "Z crime" based on the DT ID3 algorithm. The software was used to detect any terrorist attacks that might occur via email analysis. [12]

S. Sathyadevan and al 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 NB classifier and the results were extremely positive, with an accuracy of 90%. [13]

1.7.2.2 Langue arabe classique:

Unlike the previous works, which detects crime in different languages such as English and Indian, the previous work is very limited when it comes to Arabic language. Because of its complexity and lexical changes, the Arabic language creates interesting challenges. There is only one work that tackles the subject of crime and its detection which is:

Hissah AL-Saif and Hmood Al-Dossari used various ML algorithms, to detect crime-related tweets in the Saudi Arabian region. They investigated them in different features in the form of light stemming and stemming, together with N-grams.[14]

In the data collection stage, Twitter crime tweets were collected using various tools, such as Twitter (API) and Topsy, and then extracted to an Excel file using UTF-8 Unicode to support Arabic text. Each tweet is marked with its corresponding category based on the contents used to create the training group. The data set consists of more than $8,000$ Arabic tweets from 2013 to 2016. The search targeted specialized news accounts on Twitter. The data set includes nearly $4,000$ different crime news items and $4,000$ news items on new political, health, and technology issues. The data set contains approximately $187,748$ words, with $23,994$ distinct words.[14]

In Their system, preprocessing contains four steps:

- Normalization.
- Tokenization.
- Filtering.
- Stemming.

There were three methods that 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 dataset.Also, they have adopted TFIDF in order to create vectors.

The different text classification algorithms that are used in this study are NB, DT(C4.5), SVM, and KNN. [14]

In this work, they have measured 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 this experiment, they have selected K as 10, and then the original data set was divided into 10 folds, each one containing the same amount of data.[14]

The performance of the model was evaluated by different measures, including accuracy, recall, and precision. Accuracy is the most important measure for evaluating the model. Figure 1.3 represents the accuracy of different classifiers, including SVM, DT, (KNN), and CNB for crimes classification in level three.[14]

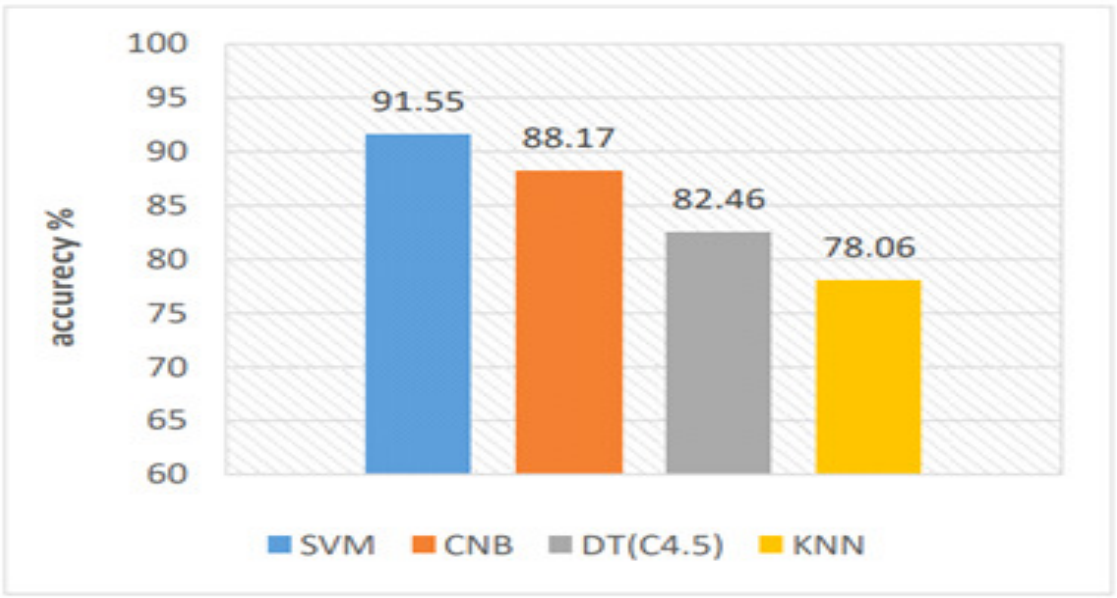

Figure 1.3: Classification Accuracy for Classifiers in Classifying Crimes According to Type.

The results, revealed that 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). However, the most affected of the class numbers on speed was DT.

Specifically, the results indicate the superiority of SVM with trigram over other classifiers, with a 91.55% accuracy. [14]

In the other detailed work, Mohammed and al presented a system that acquires, preprocesses, trains, classifies, and tests Arabic tweets to detect suspicious messages.Figure 1.4 shows the architecture diagram of the system. [8]

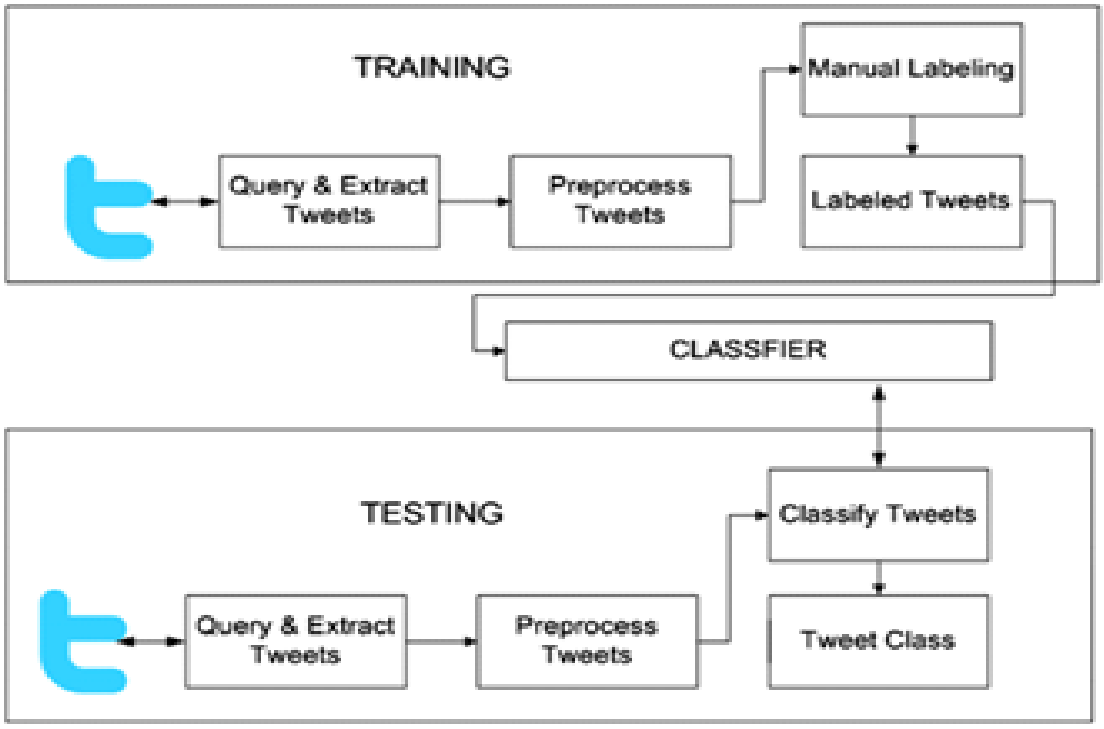

Figure 1.4: System Architecture.

They deliberately acquired a part of dataset that composed of tweets that come under the category of suspicious. tweet dataset consists of total 1555 tweets, out of which 826 tweets are suspicious, and 729 are not-suspicious. The acquired data are in JSON format. they saved the tweet's data in Microsoft Excel using UTF-8 that supports the Arabic character set.[8]

In Their system, preprocessing contains four steps:

- Data Filtration,
- Data Tokenization,
- Stemming, and
- Lemmatization.

During the training stage, they performed manual labeling of tweets into two classes, suspicious (label 1) and not-suspicious (label 0). A tweet is labeled as suspicious if it needs further analysis by a security/intelligence analyst.That latter analysis includes looking further deep into the tweet and related data such as information about the person who tweeted, context, and time line. However,the study is limited only to label a tweet suspicious or not-suspicious. Fleiss's Kappa measure has been applied in that work with three experts' annotators[17].

During the classification stage, an unlabeled tweet is identified as suspicious or not-suspicious using the pre-trained classifier.

A total of 72% of the tweets are categorized by the three experts with the same class, while two experts agree to classify 24% of the tweets. The remaining (4%) of the tweets are not agreed upon by the three annotators and are excluded from the data. A few samples of Arabic tweets with their English translations and labels are shown in Table 1.6[17].

Arabic Tweet	**English Translation**	**Label**
جرمة يجب ان يعاقب علها فاعلها	Acrime's perpetrator must be punished	1(suspicious)
عائلة في مدريد تتعرض للسطو في منزلهم من قبل مجرمين	A family in Madrid is being robbed in their home by criminals	1(suspicious)
الموعد اللي ماتت احلامنا فه	The data when our dreams died	0(not-suspicious)
احانا اللطف مع الناس جرمة ضد النفس	Sometimes kindress with people is a crime against the self	0(not-suspicious)

Table 1.6: Examples of Arabic tweets with translations and labels. Suspicious (label 1) and not-suspicious (label 0).

The following terminologies were used to measure the accuracy of the system:[17]

- True Positive (TP) are those tweets that correctly classified as suspicious.
- True Negative (TN) are those tweets that correctly classified as not-suspicious.
- False Positive (FP) are those tweets that incorrectly classified as suspicious.
- False Negative (FN) are those tweets that incorrectly classified as not-suspicious.
- Accuracy is the ratio of tweets that are correctly classified
- The total number of tweets classified, mathematically,

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

.

- True Positive Rate (TPR) or sensitivity is the probability of correctly detecting a tweet as suspicious, mathematically,

$$TPR = \frac{TP}{(TP + FN)}$$

.

- True Negative Rate (TNR) or specificity is the probability of correctly detecting a tweet as not-suspicious, mathematically,

$$TNR = \frac{TN}{(TN + FP)}$$

.

- Positive Predictive Value (PPV) is the probability that a tweet label "suspicious" is actually suspicious, mathematically,

$$PPV = \frac{TP}{(TP + FP)}$$

.

- Negative Predictive Value (NPV) is the probability that a tweet label "not-suspicious" is actually not-suspicious, mathematically,

$$NPV = \frac{TN}{(TN + FN)}$$

.

They evaluated the system accuracy using DT, KNN, LDA, SVM, ANN, and LSTM classifiers, and they used the BoW model (also known as a term frequency counter) to represent tweets. Table 1.7 shows classification accuracy at various values of a holdout.[17]

Holdout%	**DT**	**KNN**	**LDA**	**SVM**	**ANN**	**LSTM**
20	85.53	73.92	77.23	86.72	80.96	74.34
30	85.36	75.38	76.99	86.09	80.64	72.66
40	84.59	73.44	76.65	85.61	81.54	7453
50	84.07	72.92	76.40	85.52	80.87	72.74
60	84.75	71.51	75.61	85.32	79.74	72.95
70	84.68	70.98	76.39	85.21	78.46	69.04
80	84.43	69.31	77.62	85.39	75.76	71.95

Table 1.7: Mean accuracy of classification at various values of a holdout.

The P% holdout means that the original tweet dataset is randomly partitioned into two sets containing P% of the tweets as the testing set and remaining (1−P)% of tweets as the training set.[17]

For each classifier, they performed the simulations 100 times for each holdout value and then took the mean accuracy value. Among the 100 iterations, they computed the CM of an arbitrary iteration for each classifier. Different classifiers may achieve maximum accuracy in a different iteration.[17]

Figure 1.5 shows accuracy and execution time of classifiers as line graphs at various values of holdout. Execution time includes the sum of partitioning, training, testing, and accuracy evaluation times in seconds.[17]

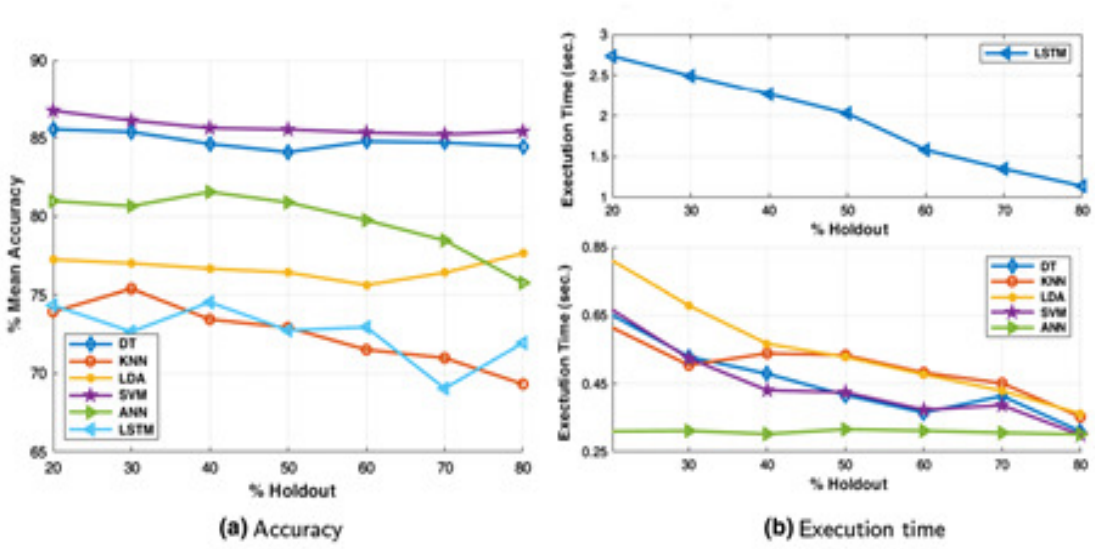

Figure 1.5: Mean accuracy and mean execution time of classifiers as line graphs at various values of the holdout.

Since the execution time of LSTM is comparatively high compared to other classifiers, they have to show it in a separate subplot.[17]

Figure 1.6 shows CM of six classifiers at 40% holdout value (i.e., 40% test data). Each CM consists of $3 * 3$ cells, there is the distribution of information in the cells of a CM:

- The top left four cells $(2 * 2)$ of a CM show number of tweets and the percentage of the total number of tweets classified. In this $2 * 2$ array of cells, rows correspond to the predicted class (determined by the classifier) and the columns correspond to the true class (as given in the known dataset of tweets). The diagonal cells correspond to tweets that are correctly classified (i.e., TP + TN). The off-diagonal cells correspond to incorrectly classified tweets (i.e., FP+FN).

- The rightmost column $(1 * 3)$ of a CM shows the percentages of all the tweets "predicted" by the classifier.
- The bottom row $(3 * 1)$ of a CM shows the percentages of all the tweets that "actually" belong to each class.
- The bottom right cell $(1 * 1)$ of a CM shows the overall accuracy of the classifier.[17]

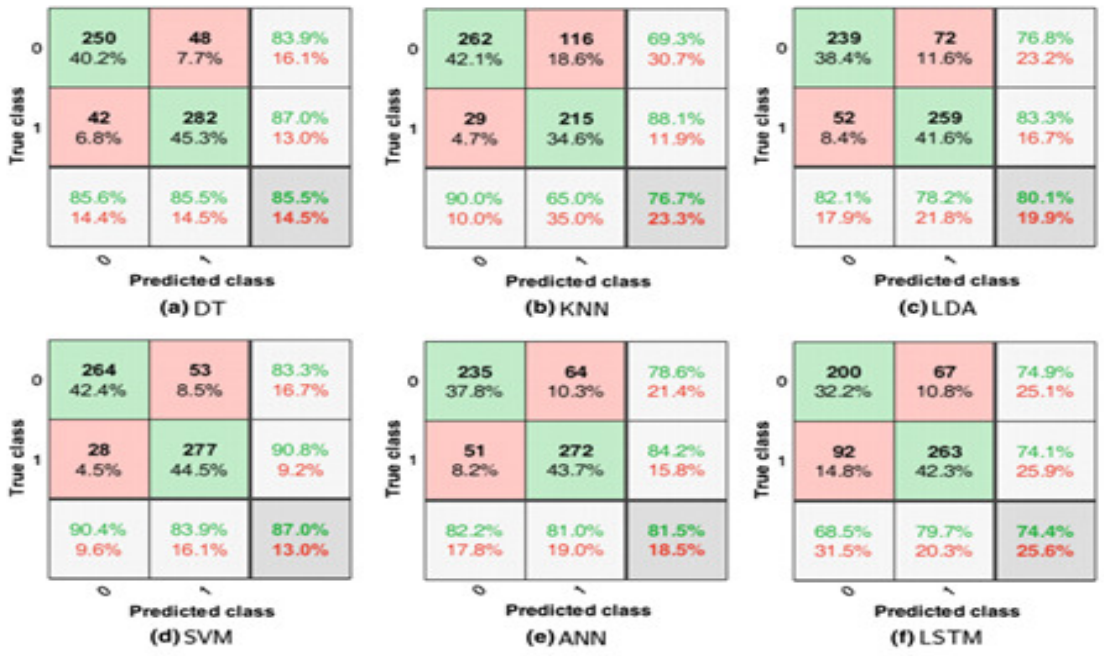

Figure 1.6: Confusion matrices of test dataset at 40% holdout for any iteration (1-100) where the accuracy is maximum. Labels: $1 \rightarrow suspicious$,$0 \rightarrow not - suspicious$.

From the experiments, it is evident that the accuracy of the SVM classifier is best, followed by DT, while the accuracy of LSTM and KNN is poor.[17]

1.8 Criticism of Related Works:

- In the study **Assia Soumeur and al**[10], the results were weak in the range of 80% despite the use of a large number of comments (25475). Only two algorithms were used for ML (MLP and CNN), and they summarized all emotions into six categories.
- The studies of **Isuru Jayaweera and al**[11], **M. Sharma and al**[12], and **S. Sathyadevan and al**[13] only used one algorithm, using SVM, DT and NB respectively.
- In the study of **Isuru Jayaweera and al**[11], the accuracy of the study of the place, time and type of crime was not encouraging, as it was between 70 and 80 percent.
- In the study of **Hissah AL-Saif and Hmood Al-Dossari**[14], despite the use of four ML algorithms, the results were weak, not exceeding 90 percent.
- In the study of **Mohammed and al**[8], they used a small dataset (1, 555 tweets) although six ML algorithms were applied that gave poor results.
- The accuracy of all Related work was poor due to the lack of algorithms used or the limited data set collected.

1.9 Conclusion:

In this chapter we discussed the crime subject, specifically in Algeria, the possibility of using social media to detect it and its relation with the analysis of feelings. In addition, we presented some works that are related to our study. The next chapter tackles the distinctiveness of the Arabic language and the Algerian one.

CHAPTER 2

STANDARD ARABIC LANGUAGE AND DIALECTS:

2.1 Introduction

Arabic language belongs to the family of Semitic languages, precisely, in the southern section. It represents a collection of languages spoken since the earliest times in the Middle East and Near East as well as in North Africa, and it is one of the branches of the Afro-Asian language family, where the origin and geographical expansion of these languages is not clear the origin and geographical expansion of these languages, either from Asia to Africa or the opposite.

The fact that Arabic is the language of the Qur'an extended beyond the Persian Arabian Gulf, reaching North Africa and Asia Minor. In addition, the regional expansion of the Islamic Empire has made Arabic as language of culture and science. Moreover, the diversity of the Arab population and their cultures has led to the emergence of various forms of Arabic, from the classical Arabic used in the Qur'an to the MSA.

2.2 Text Mining (TM):

Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform the data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. It can be applied on a variety of data types. Data types such as structured data (relational), multimedia data, free text, and hypertext as shown in Figure 2.1. We can strip hypertext from XML/XHTML tags to get free text.[15]

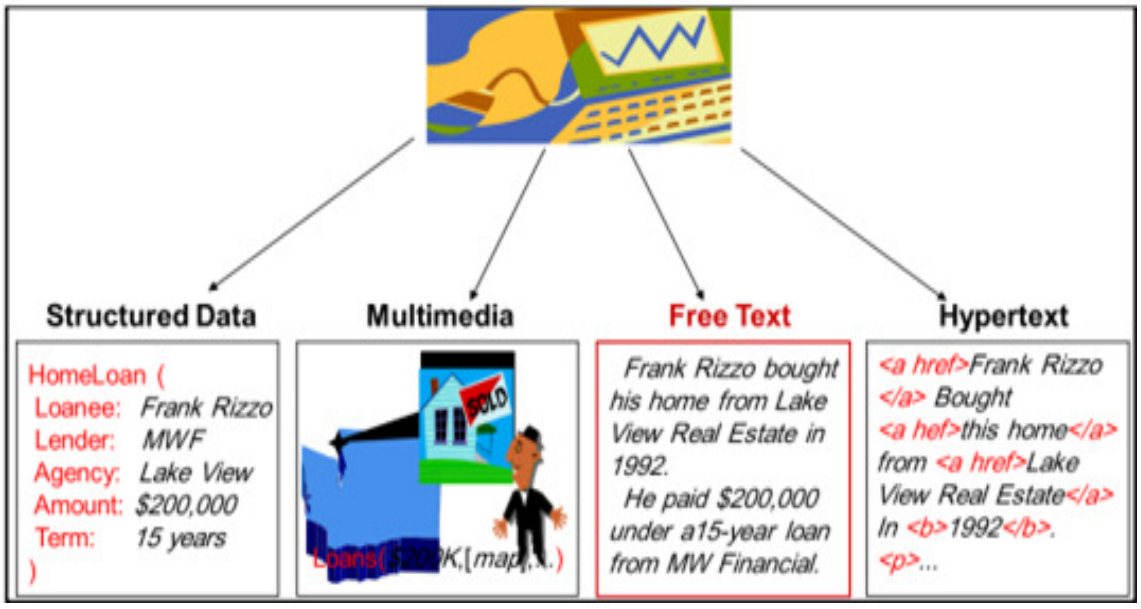

Figure 2.1: Data mining over verity of data.

TM, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. TM usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output as shown in Figure 2.2. 'High quality' in TM usually refers to some combination of relevance, novelty, and interestingness.[15]

Typical TM tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).[15]

The purpose of TM is to process unstructured (textual) information, extract meaningful numeric indices from the text, and make the information contained in the text accessible to the various data mining algorithms. The information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, we can analyze words, clusters of words used in documents, etc., or we could analyze documents and determine similarities between them or how they are related to other variables of interest in data mining. In the most general terms, TM will "turn text into numbers" (meaningful indices), which can then be incorporated in other analyses such as predictive/descriptive data mining. TM is well motivated, due to the fact that much of the world's data can be found in text form (newspaper articles, emails, literature, web pages, etc.).[15]

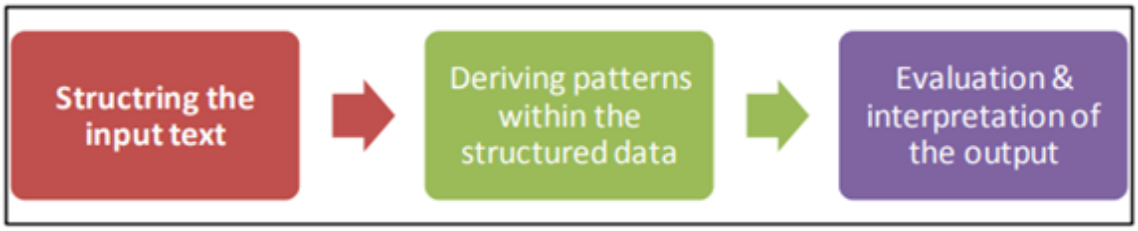

Figure 2.2: Text Mining Process.

2.3 Text Classification (TC):

Text Classification (TC)(also known as text categorization, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set. This task, which falls at the crossroads of Information Retrieval (IR) and ML, has witnessed a booming interest in the last ten years from researchers and developers alike.

TC can provide conceptual views of document collections and has important applications in the real world. Automatic text categorization can significantly reduce the cost of manual categorization.[15]

Making the text at human level understanding to machines is not trivial task. The process includes deriving linguistic features from text to be at human like interpretation to be mined.[15]

TM must overcome a major difficulty that there is no explicit structure. Machines can reason relational data well since schemas are explicitly available. However, text encodes all semantic information within natural language. TM algorithms, then, must make some sense out of this natural language representation. Humans are great at doing this, but this has proved to be a problem for machines.[15]

The TC problem is composed of several sub problems, which have been studied intensively in the literature such as the document indexing, the weighting assignment, document clustering, dimensionality reduction, threshold determination and the type of classifiers. Several methods have been used for TC such as: SVM, KNN, NN, NB, DT, ME, N-Grams and Association Rules.[15]

The main consecutive phases of building a TC system involve compiling and labeling text documents in corpus, selecting a set of features to represent text documents in a defined set classes or categories (structuring text data), and finally choosing a suitable classifier to be trained and tested using the compiled corpus, see Figure 2.3. The constructed classifier system then can be used to classify new (unlabeled) text documents as shown in Figure 2.4.[15]

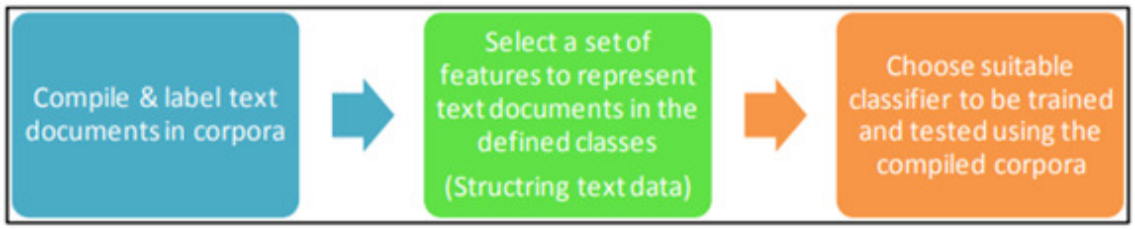

Figure 2.3: Building Text Classification System Process.

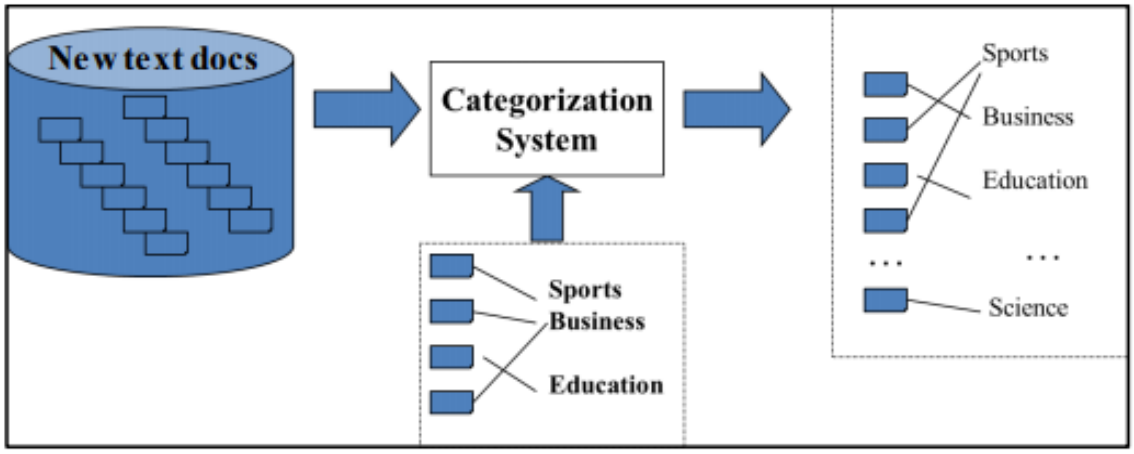

Figure 2.4: Classifying new text documents using text classification system.

Structuring text data is a process to view text as a BOT(words). This is the same approach as IR. Under that model, we can already summarize, classify, cluster, and compute co-occurrence statistics over free text. These are quite useful for mining and managing large volumes of free text. However, the (BOT) approach loses a lot of information contained in text, such as word order, sentence structure, and context; these are precisely the features that humans use to interpret text. (NLP) attempts to understand document completely (at the level of a human reader). General (NLP) has proven to be too difficult because of the high ambiguity of the text. Natural Language is meant for human consumption and often contains ambiguities under the assumption that humans will be able to develop context and interpret the intended meaning. Figure 2.5 shows the process of structuring text data as SVM[15]

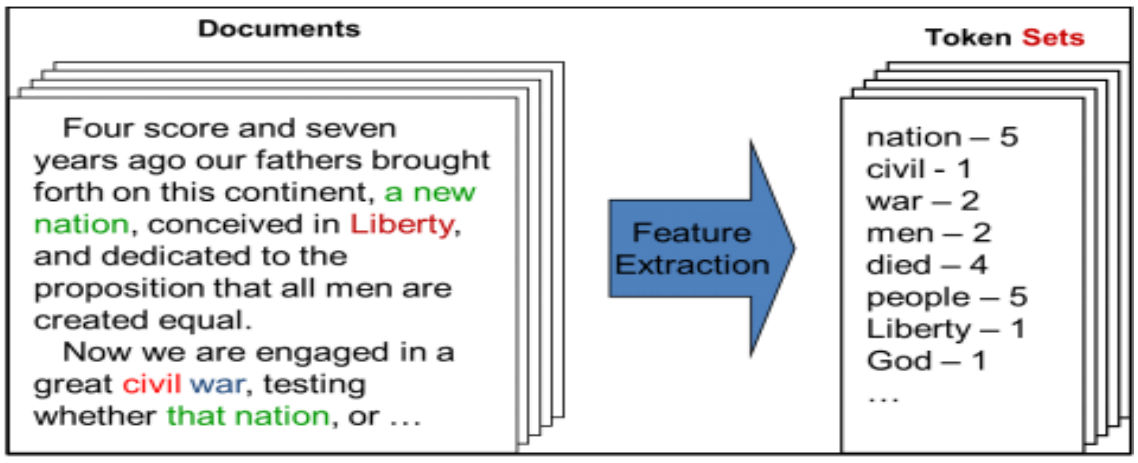

Figure 2.5: Structuring text data as SVM.

2.4 Arabic Language:

Arabic Language is the 5^{th} widely used languages in the world. It is spoken by more than 422 million people as a first language and by 250 million people as a second language. Arabic Language belongs to the Semitic language family. Semitic languages are commonly written without the vowel marks, which would indicate the short vowels. Semitic languages can get away with this because they all have a predictable root pattern system. Arabic alphabet consists of the following 28 letters (أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي) in addition to the Hamza (ء) . There is no upper or lower case for Arabic letters like English letters. The letters (أ و ي) are vowels, the rest are constants. Unlike Latin-based alphabets, the orientation of writing in Arabic is from right to left.[15]

The Arabic script has numerous diacritics, including ijam (إعجام) , consonant pointing, and tashkīl ،(تشكيل) supplementary diacritics. The latter include the áarakāt (حركات) singular áaraka (حركة) ,vowel marks. The literal meaning of taškīl is "forming". As the normal Arabic text does not provide enough information about the correct pronunciation, the main purpose of tashkīl (and áarakāt) is to provide a phonetic guide or a phonetic aid; i.e. show the correct pronunciation (double the word in pronunciation or to act as short vowels). The áarakāt, which literally means "motions", are the short vowel marks. There is some ambiguity as to which tashkīl are also áarakāt; the tanwīn, for example, are markers for both vowels and consonants.

Arabic diacritics include: Fatha, Kasra, Damma, Sukūn, Shadda, and Tanwin. The pronunciations of aforementioned diacritics for the Arabic letter (ب) are presented in Table 2.1 . Arabic words may also have Tatweel as shown in figure 2.6.[15]

Double Constant	**No Vowel**	**Nunation**			**Vowel**		
بّ	بْ	بٍ	بٌ	بً	بِ	بُ	بَ
/bb/	/b/	/bin/	/bun/	/ban/	/bi/	/bu/	/ba/

Table 2.1: Diacritics.

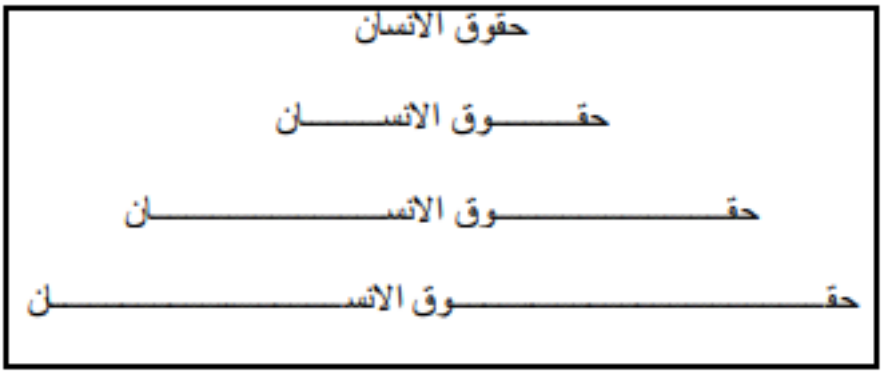

Figure 2.6: Tatweel.

Arabic words have two genders, masculine (مذكر) and feminine (مؤنث)); three numbers, singular (مفرد) ,dual (مثنى) , and plural(جمع) ; and three grammatical cases, nominative (الرفع),accusative (النصب), and genitive (الجر) . A noun has the nominative case when it is subject (فاعل) ,accusative when it is the object of a verb (مفعول) ,and the genitive when it is the object of a preposition (مجرور بحرف جر) .Words are classified into three main parts of speech, nouns (اسماء) (including adjectives (صفات) and adverbs (ظروف)),verbs(افعال), and particles (ادوات).

Arabic has three forms; Classical Arabic (CA), MSA, and Dialectal Arabic (DA). CA includes classical historical liturgical text, MSA includes news media and formal speech, and DA includes predominantly spoken vernaculars and has no written standards.[15]

2.4.1 Complexity of Arabic Language:

Arabic is a challenging language for a number of reasons:

1. Orthographic(الاملاء) with diacritics is less ambiguous and more phonetic in Arabic, certain combinations of characters can be written in different ways.

2. Arabic language has short vowels that give different pronunciation. Grammatically they are required but omitted in written Arabic texts.

3. Arabic has a very complex morphology compared to English language.

4. Synonyms are widespread. Arabic is a highly inflectional and derivational language.

5. Automatic TC depends on the contents of documents, a huge number of features or keywords can be found in Arabic text such as morphemes that may generated from one root, which may lead to a poor performance in terms of both accuracy and time.

6. Lack of publically freely accessible Arabic Corpora.[15]

2.4.2 Examples from Arabic to show the complex nature of Arabic Language:

2.4.2.1 Word meanings :

It is possible to identify the different meanings associated with a word, due to one word; you may have more than one meaning in different contexts.In addition, by using corpus, this kind of ambiguity can be authentically detected. Table 2.2 shows the Arabic word (قلب) which has three meaning as a noun.[15]

Word meaning	**Sentence**
Core	في قلب الاحداث
Heart	اجرى عملية قلب مفتوح
Center, middle	الكرة في قلب الملعب

Table 2.2: Meaning of word (قلب) as a noun.

2.4.2.2 Variations in lexical category:

One word may have more than lexical category (noun, verb, adjective, etc.) in different contexts as shown in Table 2.3. The Morphological analysis of a given corpus includes investigating word frequency of a word as a lexical category.[15]

Word meaning	**Word Category**	**Sentence**
Ain	Proper-Noun	عين جالوت
Wellspring	Noun	عين الماء
Eye	Noun	عين الانسان
Delimitate/be delimitate	Verb/passive Verb	عين وزيرا للخارجية

Table 2.3: Lexical Category of word (عين).

2.4.2.3 Synonyms:

Languages have many words that are considered as synonymous. Through a given corpus, the researchers can use morphological analysis tools to know: the synonyms of a word, the frequency of each word of those synonyms, and which one of them is more common. Examples of synonyms in Arabic are (بذل ، منح ، اعطى ، وهب) which means (give), (اسرة ، عائلة) which means (family), and (صف ، فصل) which means (classroom).[15]

2.4.2.4 Word form according to its case:

The form of some Arabic words may change according to their case modes (nominative, accusative or genitive). For instance the plural of the word (مسافر) which means (traveler) may become the form (مسافرون) in the case of nominative (مرفوعة) and the form (نُسافر) in the case of accusative/genitive (مجرورة منصوبة). Arabic light stemming can handle these cases.[15]

2.4.2.5 Morphological characteristics:

An Arabic word may be composed of a stem plus affixes and clitics. The stem consists of a consonantal root (جذر صحيح) and a pattern morpheme (اصغر كلمة ذات معنى) .The affixes include inflectional markers (علامات او حركات اعرابية) for tense, gender, and/or numbers. The clitics include some prepositions (حروف الجر) conjunctions(حروف العطف) , determiners(محددات) ,possessive pronouns (ضمائر الملكية), and pronouns (ضمائر) The clitics attached to the beginning of a stem are called proclitic and the ones attached to the end of it are called enclitics. Most Arabic morphemes are defined by three consonants, to which various affixes can be attached to create a word. For example, from the tri-consonant "ktb" (كتب) , we can inflect (يصرف) several different words concerning the idea of writing as (wrote كتب) ,(book كتاب) ,(the book الكتاب) ,(books كتب) ,(he writes يكتب) ,(author كاتب) ,(library مكتبة) . Moreover an Arabic word may correspond to several English words. Because of the variability of prefixes and suffixes, the morphological analysis is an important step in Arabic text processing. For example, the Arabic word (وبنفوذها) and its equivalence in English "and with her influences" . This makes segmentation of Arabic textual data different and more difficult than Latin languages.[15]

Affixes of Arabic	**Examples**
Prefixes of length three	ولل، وال، وكال، وبال
length two Prefixes	ال، لل
length one Prefixes	ل، ب، ف، س، و، ي، ت، ن، ا
length three suffixes	تمل، همل، تان، تين، كمل
length two suffixes	ون، ات، ان، ين، تن، كم ، هن، نا، يا، ها، تم ، كن، ني، وا، ما، هم
length one suffixes	ة، ه، ي، ك، ت، ا، ن

Table 2.4: Affix set in Arabic Language.

Affixes set in Arabic are shown in Table 2.4, and Arabic patterns (الاوزان) and roots are shown in Table 2.5. The word (علم) may give various meanings by adding different affixes (prefixes, infixes, or suffixes) as shown in Table 2.6. [15]

Arabic Pattern and roots(الاوزان)	**Example**
Length four Pattern	فاعل، فاعول، فعلة، فعال، مفعل
Length five Pattern and Length three roots	فعالة، فعلان، تفاعل، افتعل، افعال، فعولة، تفعلة، تفعيل، مفعلة، مفعول، فاعول، فواعل، مفاعل، مفعيل، افعلة، فعائل، منفعل، مفتعل، فاعلة، مفاعل، مفلاع، يفتعل، تفتعل، فعلالي، انفعال
Length five Pattern and Length four roots	تفعلل، افعلل، مفعلل، فعللة، فعلان، فعالل
Length six Pattern and Length three roots	استفعل، مفاعلة، افتعال، افعوعل، انفعل، مستفعل
Length six Pattern and Length four roots	انفعلل، افعلان، متفعلل

Table 2.5: Arabic Patterns and Roots.

Meaning	**Suffix**	**Infix**	**Prefix**	**Word**
Scientific	ية	***	***	علمية
Learned us	تنا	***	***	علمتنا
his science	ه	***	***	علمه
Scientists	اء	***	***	علماء
Teaching	***	ي	ت	تعليم
Sciences	***	و	***	علوم
Informative	يه	ا	است	استعالمية

Table 2.6: Versions of the word (علم) and its meaning when adding affixes.

2.4.3 Richness of Arabic Language:

The Arabic Language is rich in conveying meanings of a phrase, and one word is described in many styles of articulation, meaning, and definition. For instance, for the word' love' (حب) ,there are at least 11 words and hundreds for 'camel' (جمل).[27]

Besides, each of the Arabic words for love expresses a different point in the process of falling in love. The word 'Hawa' (هوى),(attraction), for example, illustrates the first attraction towards another. The term comes from the core word 'h-w-a' (هواء)a transitory breeze that can increase and plummet.[27]

Also, 'Alaaqa' (علاقة) ,(which comes from the main word ('a-l-q) (علق) ,(meaning 'to grip on to' which portrays the stage when someone's heart begins to attach itself to the other, before progressing into a sightless desire 'Ishq' (عشق) ,(desire) and all devouring love 'Shaghaf'(شغف) (affliction).[27]

The last stage of falling in love, 'Huyum' (هيام) (insanity), tells the entire mind loss of explanation. Remarkably, the most popular word for love in Arabic, 'Hub' (حب) (love) comes from the same root as the word 'seed' which has the potential to develop into something charming.[27]

Moreover, Arabic is said to have many words for 'camel'(جمل) .For instance, 'Al-Jafool' (الجفول) means a camel that is terrified by anything; 'Al-Harib' (الهريب) is a female camel that moves ahead of others by a significant distance so that it seems to be escaping. [27]

2.4.4 Algerian Arabic:

The Algerian Arabic or Algerian dialect, which is a group of North African Arabic, is dialects mixed with different languages spoken in Algeria. The friction of several languages throughout the history of the region has produced a complex and rich language including words, expressions and linguistic structures, such as Berber, French, Italian, Spanish and Turkish as well as other Mediterranean Romance languages. The Algerian dialect is strongly influenced especially by French where we can find the switching of codes and borrowing.[16]

The lyrics of this dialect can be categorized as follows:

- MSA encoded in Arabic letters, such as:« عيدكم مبارك كل عام وانتم بخير »
- MSA encoded in Romanized letters, such as:« ana said hada elyaoum »
- Dialect encoded in Arabic letters, such as:« كيراك شريكي »
- Dialect encoded in Romanized letters, such as: « maysoumouch bla shour »
- Foreign language encoded in Romanized letters, such as: « je vais à l'univ »
- Foreign language encoded in Arabic letters, such as: « برافوووو »

- For that, every text belongs to one of these categories or a mixture of them. Moreover, we may find another form of writing where a number replaces a letter, such as: « b1- بيا -bien », «3adi- عادي -normal », « 5amsa-خمسة -cinq », « 6ariq- طريق -route », « sa7b- صاحب -ami », « 9raya- قراية -étude »...

2.4.5 Algerian Dialect on Facebook:

On the social media, Algerian users, hereafter Facebook users (since we only consider texts on Facebook), use Algerian Dialect as well as French and English. This poses a major problem known as "Code Switching", that is the use of several languages and dialects within the same document, even sentence.[10]

We can thus find sentences like: "Bonjour, كيف حالك اليوم Brother?" (Good morning, how are you today brother?) where French, Arabic, and English are used consecutively in the same sentence.

In addition, each region in Algeria has its own dialect variations, or even accent, which can affect written words.For example, the sentence قال لي (" He told me") may be found written in different ways depending on the Algerian region: "قالي" (qAlly), "آلي" (Ally), "ڤالي" (gAlly), etc.[10]

On the other hand, Algerian Facebook users may use Latin letters to write Arabic words and phrases. This is known as Arabizi. For example, the sentence "كيف هي حالة الطقس" (how is the weather?) could be found as (kayf hiya halat eltaks?) in Latin letters.One can also find Arabic letters used to write French words or phrases. For example, "Bonjour!" (Good morning) may be found as "بونجور".[10]

The absence of spelling rules has rendered the textual content of social networks characterized by high orthographic heterogeneity: a word can be written in so many ways, based on the users' preferences or even habits. To illustrate this phenomenon, Table2.7 shows different ways of writing "إن شاء الله" (God Willing) which was found, whether written with Arabic or Latin letters.[10]

With Latin letters	With Arabic letters
In shaa llah	إن شاء الله
In chaa Allah	ان شاء الله
In chaa ellah	انشاالله
Inchaallah	نشالله

Table 2.7: Different spellings of "إن شاء الله" in Algerian Dialect using Arabic or Latin letters.

Another common phenomenon on social networks is the repetition of letters in a given word. For example, the word "3laaaaaaaach" is equivalent to "3lach" ("why"). [10]

The fact that Algerian Dialect has its origins from formal Arabic with added vocabulary from Berber, Turkish, French, and Spanish what makes it difficult enough to process. If, especially on social networks, we also consider the problem of code switching, with various possible spelling variations, writing errors and newly coined words, then it is out of question that Algerian Dialect is difficult to understand for the non-native user and complex to be automatically processed.[10]

2.4.6 Difficulties of Arabic language and dialect:

We may cite the following difficulties:

- One dialect, such as the Algerian Dialect, may contain several sub-dialects.
- The great distance between MSA and some dialects.
- A root may take many forms depending on the context.
- Repeating a letter several times to emphasize the meaning or feeling (bzeeeef).
- Arabic has various diacritic signs, the presence or absence of such signs, can completely change the meaning of words.
- The negation words used to negate verbs in past or present, which change the meaning to the opposite [18].

All the pre-mentioned difficulties highlight challenge of processing Algerian Dialect especially the use of the non-Arabic origin words.

2.5 Conclusion:

In this chapter, we investigated the history of our language, its vocabulary richness, its complexity and its features in terms of construction and composition in order to highlight its complex and difficult automatic processing, to end up with the Algerian dialect and its origin.

In the next chapter, we will see some theoretical studies on machine learning, its algorithms and lexical diving.

CHAPTER 3

MODELLING AND IMPLEMENTATION:

3.1 Introduction

Crime detection is a difficult task, but it is an important task for security agencies using on line social media. Our work is based on the use of the main automated learning methods, including classifying Facebook comments, to reveal the nature of the crimes that may occur and the its perpetrators. For this reason, we will give definition ML and its algorithms, we give some measurements that will be needed for evaluation. We will implement ML algorithms and we will use the comments generated from Facebook that have been unified to be ready to use. In addition, we will discuss the results of analysis and classification of different comments. In general, our system consists of four stages, as presented in Figure3.1.

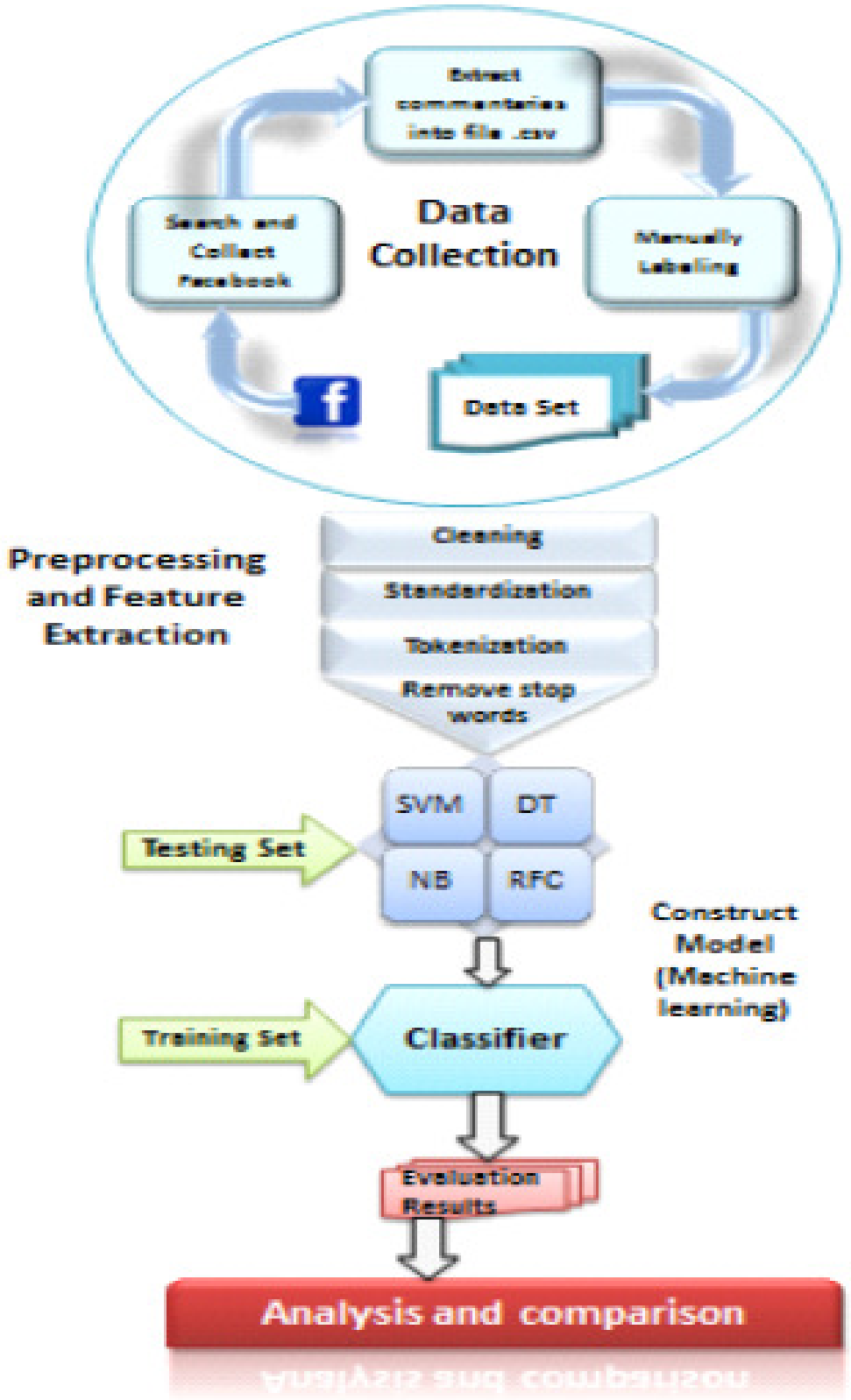

Figure 3.1: Four Stages of The Construct Model.

3.2 Natural Language Processing (NLP):

It is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.[40]

3.3 Machine learning:

Algorithms are a sequence of instructions used to solve a problem. Algorithms, developed by programmers to instruct computers in new tasks, are the building blocks of the advanced digital world we see today. Computer algorithms organize enormous amounts of data into information and services, based on certain instructions and rules. It's an important concept to understand, because in ML, learning algorithms – not computer programmers – create the rules.[38]

Instead of programming the computer every step of the way, this approach gives the computer instructions that allow it to learn from data without new step-by-step instructions by the programmer. This means computers can be used for new, complicated tasks that could not be manually programmed. Things like photo recognition applications for the visually impaired, or translating pictures into speech.[38]

The basic process of ML is to give training data to a learning algorithm. The learning algorithm then generates a new set of rules, based on inferences from the data. This is in essence generating a new algorithm, formally referred to as the ML model. By using different training data, the same learning algorithm could be used to generate different models. For example, the same type of learning algorithm could be used to teach the computer how to translate languages or predict the stock market.[38]

Inferring new instructions from data is the core strength of ML. It also highlights the critical role of data: the more data available to train the algorithm, the more it learns. In fact, many recent advances in AI have not been due to radical innovations in learning algorithms, but rather by the enormous amount of data enabled by the Internet.[38]

Although a ML model may apply a mix of different techniques, the methods for learning can typically be categorized as three general types:

- Supervised learning: The learning algorithm is given labeled data and the desired output. For example, pictures of dogs labeled "dog" will help the algorithm identify the rules to classify pictures of dogs.
- Unsupervised learning: The data given to the learning algorithm is unlabeled, and the algorithm is asked to identify patterns in the input data. For example, the recommendation system of an e-commerce website where the learning algorithm discovers similar items often bought together.
- Reinforcement learning: The algorithm interacts with a dynamic environment that provides feedback in terms of rewards and punishments like the self-driving cars that are being rewarded to stay on the road. [38]

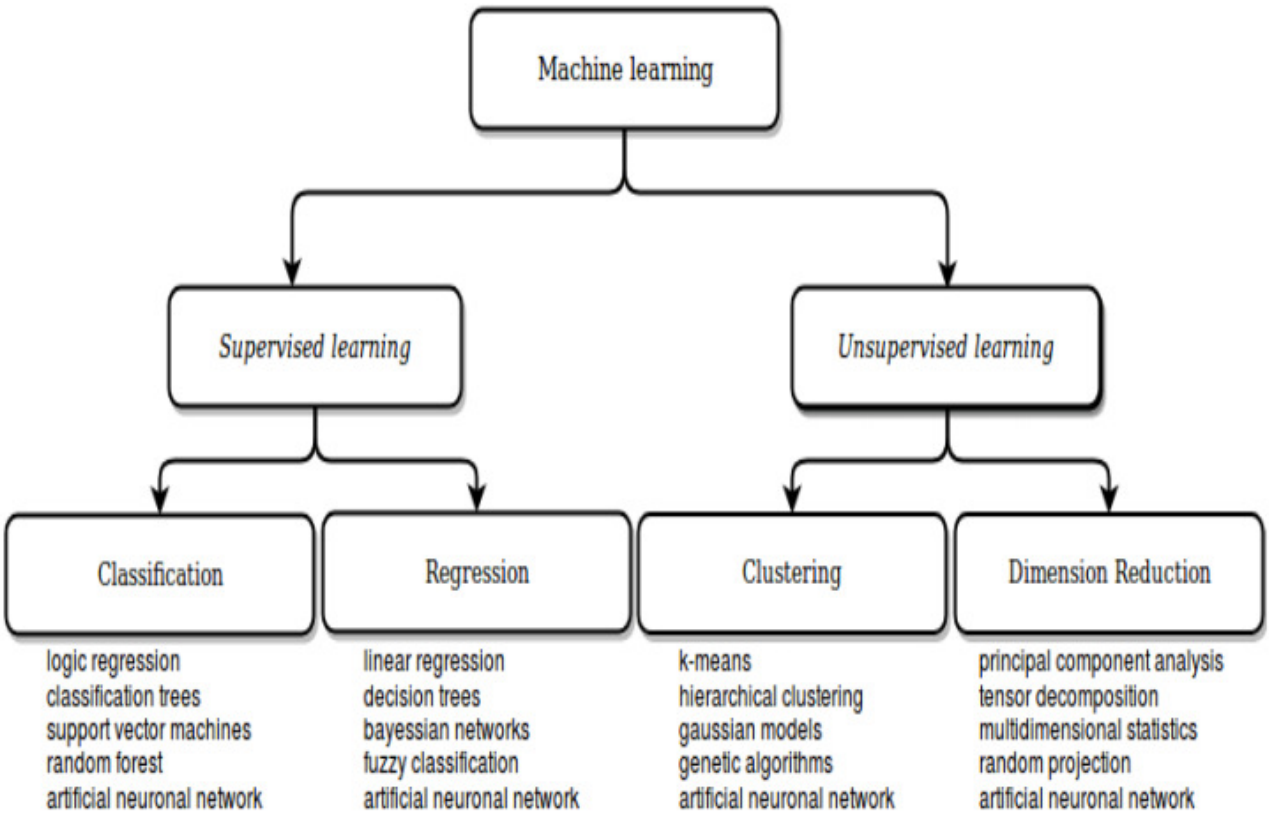

Figure 3.2: Machine learning Algorithms.

3.4 Algorithms of classification:

3.4.1 Support Vector Machine:

Support Vector Machine (SVM)) is a supervised ML algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems.

In the SVM algorithm, we plot each data item as a point in n-dimensional space (where "n" is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line). [39]

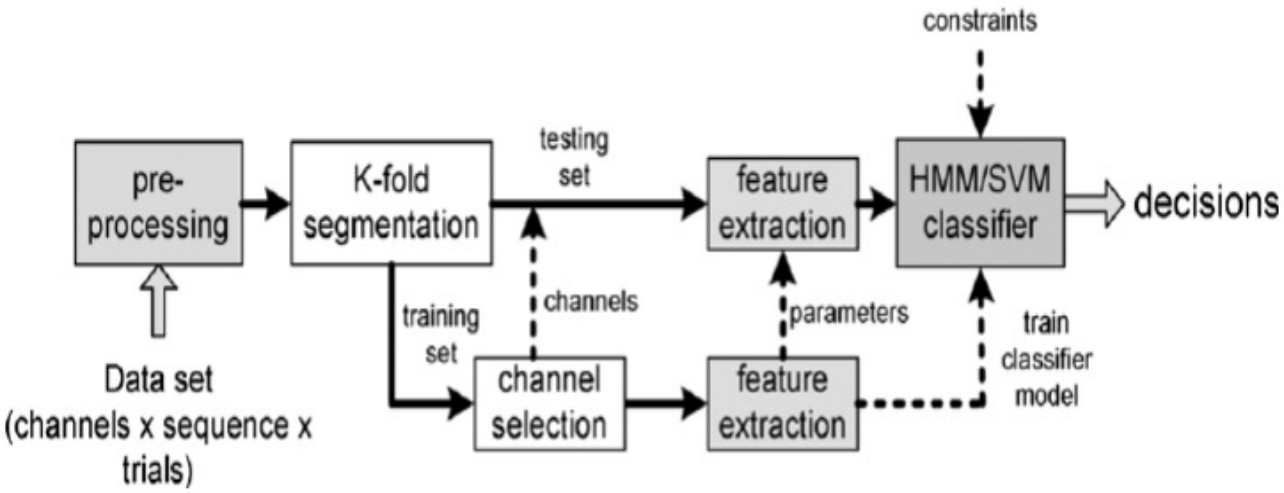

Figure 3.3: Support Vector Machine.

3.4.2 Decision Tree:

A Decision Tree (DT) is a graph that uses a branching method to illustrate every possible outcome of a decision.

DT can be drawn by hand or created with a graphics program or specialized software. Informally, DT are useful for focusing discussion when a group must make a decision. Programmatically, they can be used to assign monetary/time or other values to possible outcomes so that decisions can be automated. DT software is used in data mining to simplify complex strategic challenges and evaluate the cost-effectiveness of research and business decisions. Variables in a decision tree are usually represented by circles.[28]

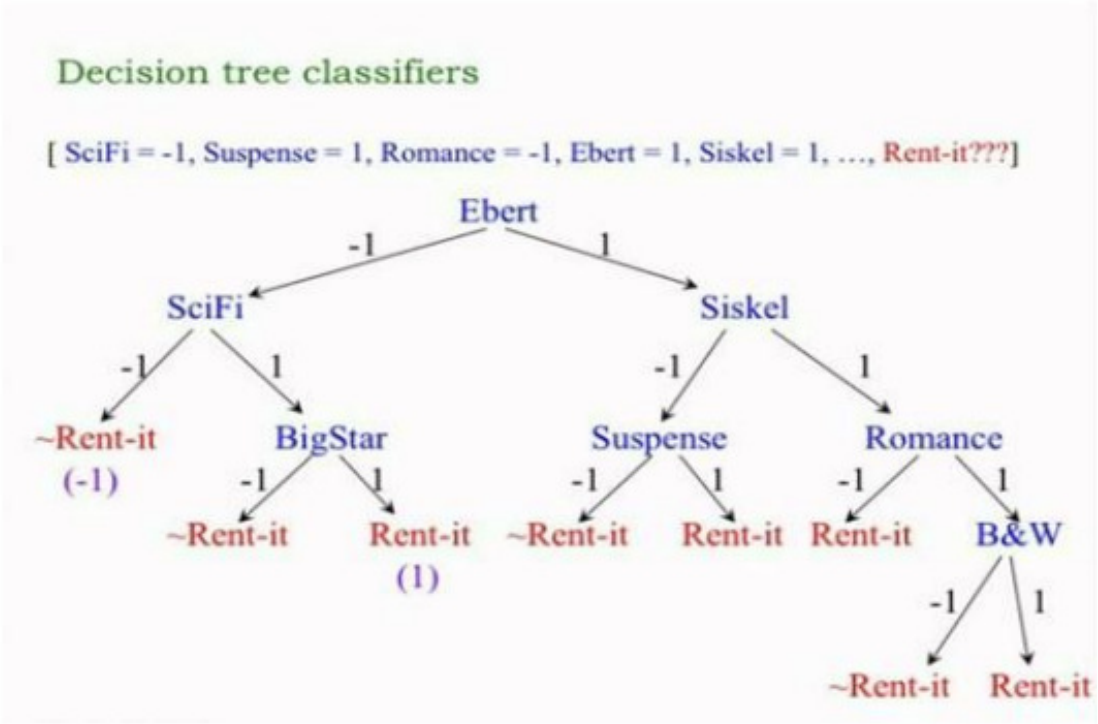

Figure 3.4: Decision Tree.

3.4.3 Random Forest Classifier:

The Random Forest (RF) is a supervised learning algorithm that randomly creates and merges multiple DT into one "forest". The goal is not to rely on a single learning model, but rather a collection of decision models to improve accuracy. The primary difference between this approach and the standard DT algorithms is that the root nodes feature splitting nodes are generated randomly. [29]

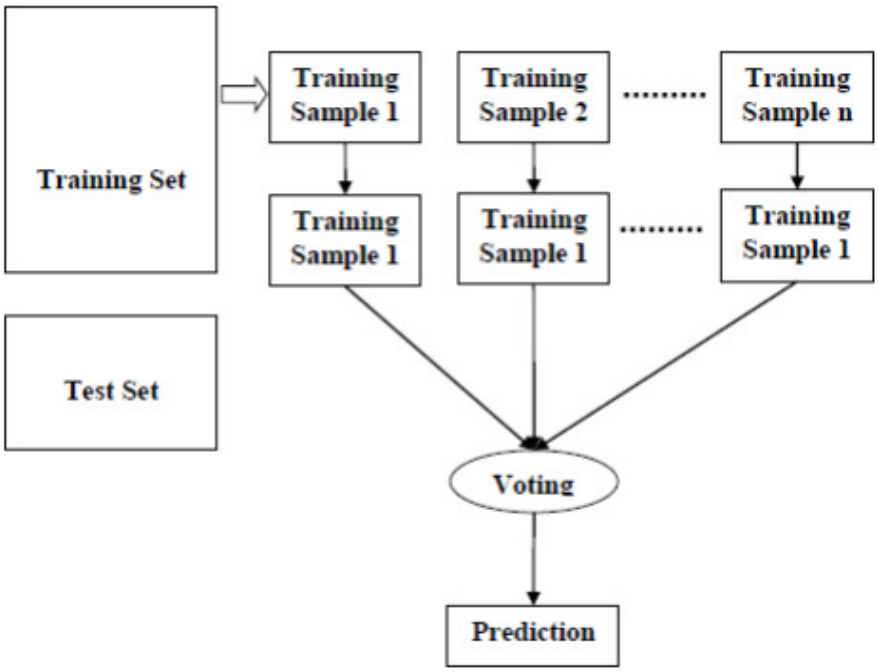

Figure 3.5: Random Forest Classifier.

3.4.4 Naïve Bayes:

The The Naïve Bayes (NB) classifier is one of the simplest classification methods. It is part of the probabilistic family based on Bayes' Theorem. The NB 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 TC based on the NB 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. [14]

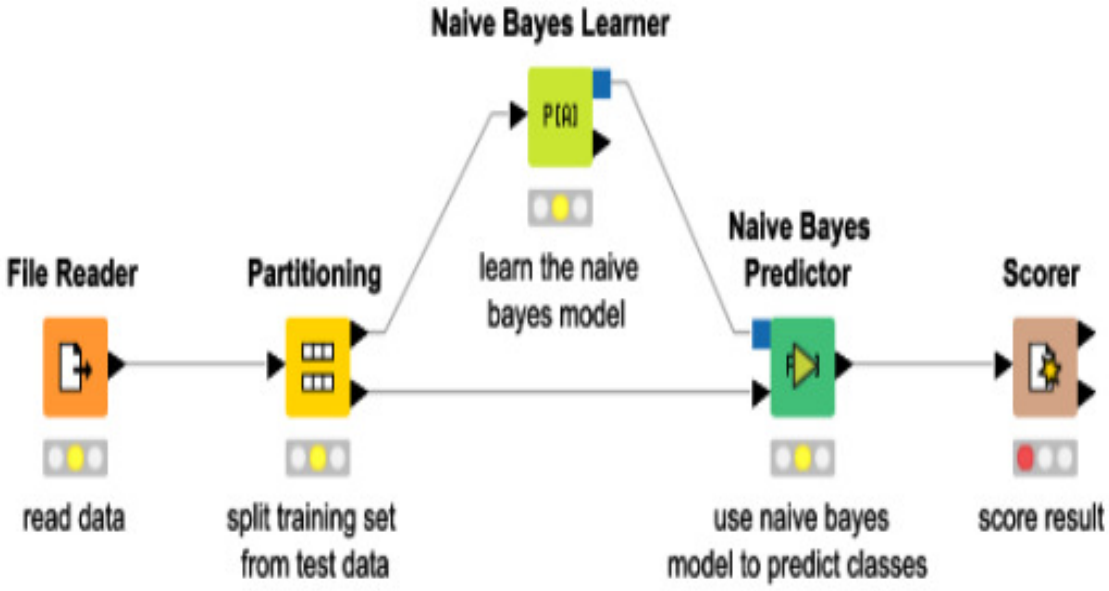

Figure 3.6: Naïve Bayes.

3.5 System Evaluation:

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:

$$\frac{(TP + TN)}{N}$$

- Recall represents the number of items correctly classified as positive divided by the total number of positive. It is calculated using the following equation:

$$\frac{TP}{(TP + FN)}$$

- 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:

$$\frac{TP}{(TP + FP)}$$

  - Where TP represent all tuples that were correctly predicted as crimes.
  - FN represent all tuples that were incorrectly predicted as non-crimes.
  - FP represent all tuples that were incorrectly predicted as crimes.
  - TN represent all tuples that were correctly predicted as non-crimes. [14]

3.6 Data Collection:

The crime is divided into individual and collective crimes, committed by the perpetrator against himself or against others. Many Algerian pages that talk about all kinds of crime have been collected from Facebook.

We have created three Python code in our system. The first can extract all posts from a predefined page where each topic has its own URL, so that we create the secondly Python code that can extract all comments of the specified post through its URL. To simplify the extraction process, we created a third Python code which can extract all the comments from a specific page. Finally, we have a large Dataset which contains various topics.

More than 550 posts have been collected which includes different topics such as illegal immigration, theft, assassination, abduction, etc.

In addition, more than 500 comments have been collected which covers different topics. Due to the limited time and the many types of crime, we have focused only on the topic of illegal immigration since it is the most common, recently, in the major Algerian cities recently. At this stage, we collect more than 2000 comments.

3.7 Manually Labeling:

3.7.1 Annotation:

Annotation of opinions is a complex and difficult operation that is proceed manually by the experts. This operation can take great time because each expert must treat all comments one by one, and it needs, sometimes great discussion between experts to reach the final decision.

Using two experts, we take from three to four weeks to annotate 2127 comments. We classified this comments into three categories, suspicious, neutral and not-suspicious polarities which corresponding respectively to three values $-1, 0$ and 1. Table 3.1 shows the different categories of our dataset according to the three polarity values.

In Table 3.7, we annotate some examples according to their polarities.

Polarity	**Not-suspicious**	**Suspicious**	**Neutral**	**Total**
Number of comments	8	185	1934	2127

Table 3.1: Number of comments by polarity.

In Table 3.7 we have described some examples with their polarities.

Comments	**Polarity**
انا ضد بوطي اصلا	Not-suspicious
نبيعو ونشري هاربيل و نخدم على روحي	Not-suspicious
شوفناا خلناا نحرقوو ول اعطناا نمروك	Suspicious
خصني حرقة تاع ٣ سوايع راني في اسبانيا الي عنده خيط في وهران يوصلني بيه	Suspicious
راني نحضر رسالة دكتورا عنوانها الهجرة غير الشرعية	Neutral
ليس شرط أن تدخن المخدرات لكي تكون أسطورة	Neutral

Table 3.2: Example of annotation some comments.

3.7.2 Creation of the dictionary :

This is a set of words that a classifier can use to evaluate the polarity of the text. In literature, we did not find a specific dictionary for the words of the Algerian dialect regarding the crime. Due to the sensitivity and seriousness of the subject, it was difficult to collect terminology, which may have several meanings other than crime.

For that, we have had to create our dictionary. It took us several weeks of work to collect 350 suspicious words from websites. We also asked for the assistance of several people from different regions of Algeria.

Figure 3.7 shows a few words from the our dictionary:

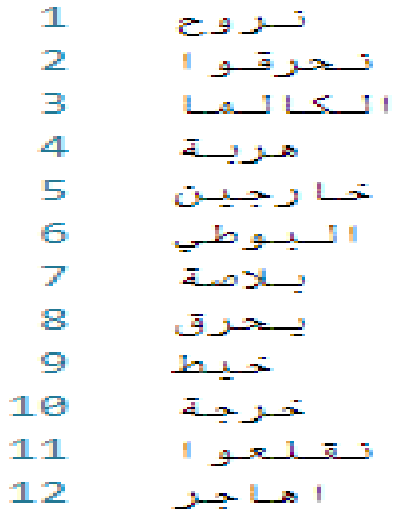

Figure 3.7: Few words from the initial list.

3.8 Pre-processing:

The objective of this step is the extract atomic element that can be word or sequences of words. These elements are used in the step of the classification. Thus, for extract these elements, we must pass by three main streps which are clean, standardization and tokenization. Standardization of the text is the deletion of signs, symbols, repeated letters, empty words or words that do not provide any information on the subject being studied. The tokenization is the final step in which the comments are divided into lexical units (tokens). Finally, we remove the stop marks. Table 3.8 presents an example of pre-treatment of one comment:

Task	**Result**
Initial text	كاش جديد يا خاوتي (˜ .˜) (˜ .˜) حبيت نخرج من هاااد لبلاااد
Cleaning	كاش جديد يا خاوتي حبيت نخرج من هاااد البلاااد
Standardization	كاش جديد يا خاوتي حبيت نخرج من هاد البلاد
Tokenization	’كاش ’,’ جديد ’,’ يا ’,’خاوتي ’,’حبيت ’,’ نخرج ’,’ من ’,’ هاد ’,’ البلاد ’
Remove stop words	’كاش ’,’ جديد ’ , ’خاوتي ’,’حبيت ’,’ نخرج ’,’ البلاد ’

Table 3.3: Example of preprocessing a comment.

3.9 Extraction of features:

Feature extraction is a process which used to reduce the dimensionality of Dataset by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. We use ML algorithms for this process.

Feature extraction is the name for methods that can select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. In our work, we use the features described in table 3.9.

Features	**Abbreviation**	**Meaning**
Has Positve Word	HPW	0 ou 1
Has Negative Word	HNW	0 ou 1
Positive Word Count	PWC	>=0
Negative Word Count	NWC	>=0
Comment Length	CL	>0 (Digital)

Table 3.4: Different categories of Features used in our work.

3.10 Construct model and Analysis:

ML algorithms use to process Training and Test. Training process used to construct models that can classify new comments, but the testing process is used to evaluate the constructed model. In this study, we use principally four algorithms, including SVM, DT, NB and RF. The classification process is influenced by many characteristics of our dataset in the training process, which are the diversity, the best selection of features, the type of classifier, and targeted language.

3.11 Experiments and results:

In this work, we use the supervised learning method and the lexicon-based approach, we have divided our corpus into two parts, 80% for training and 20% for testing. We have made several tests, the Accuracy results are presented in table 3.10:

Test	**Feature/Classifier**	**SVM**	**DT**	**RF**	**NB**
1	All features	**95.6** %	**95.8** %	**96.1** %	7.4 %
2	HPW, HNW, PWC, NWC	95.3 %	95.3 %	95.3 %	7.6 %
3	PWC, NWC	95.3 %	95.3 %	95.3 %	7.6 %
4	PWC, NWC, CL	94.1 %	95.8 %	96.1 %	7.4 %
5	HPW, HNW	95.3 %	95.3 %	95.3 %	7.6 %
6	HPW, HNW, CL	93.1 %	95.8 %	95.8 %	7.4 %
7	NWC,CL	94.6 %	95.8 %	96.1 %	**94.8** %

Table 3.5: Classification results.

3.12 Discussions:

- The best result in all tests is **96.1** %, which is carried out by RF when we use of all Features described in Table3.9. It is the same for SVM and DT, their best results were with the first test (**95.6** % and **95.8** % respectively). In the test 7, we remark the best result is carried out by NB which is **94.8** %. In the other tests, we observe that the same result is carried out using SVM, DT and RF.

- According to tests (2), (3) and (5) we noticed that the two couples of features (PWC, NWC) and (HPW, HNW) has almost the same value of influence, that is logical because the number of mixed comments (which contain at the same time positive words and negative words) is usually small.

- According to tests (3) and (4), we found that when we add the feature (CL) to features (PWC, NWC), the results of classifiers will be improved.

- According to tests (5) and (6), we found that when we add the feature (CL) to features (HPW, HNW) the results of classifiers has been increased except in the case of SVM and NB which results has been decreased.

To test the model RF (which gives the best result in all tests), we have put the test data into our model and compare our result with the predicted. Figure 3.8 represents the comparison of results in CM.

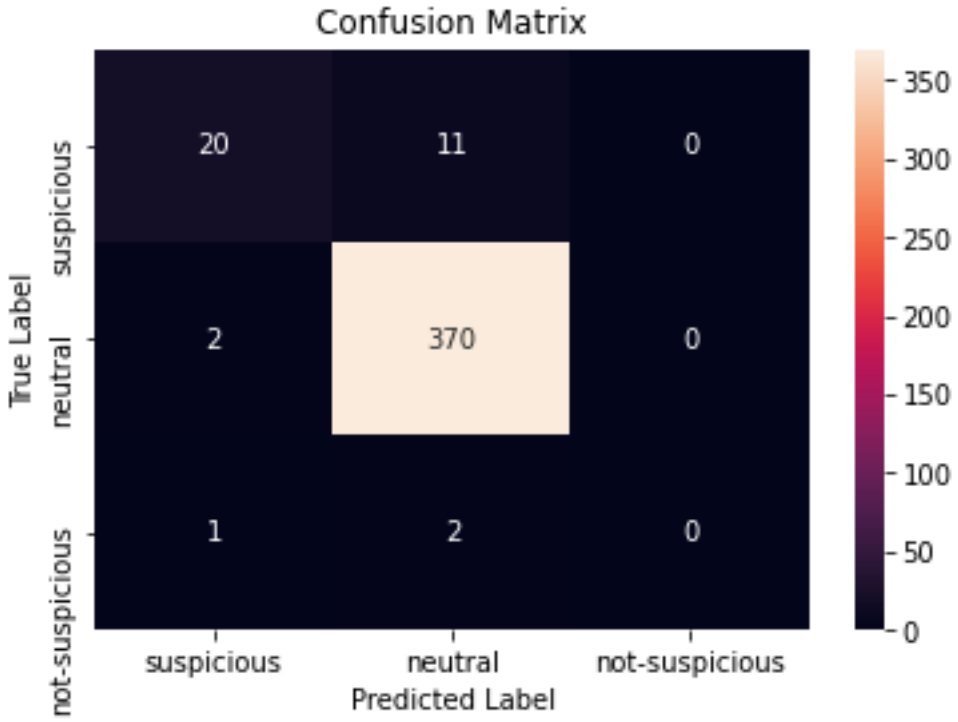

Figure 3.8: CM of the RF testing result.

- We can see that we classify correctly 20 suspicious comments (out of the total of 31 comments used in the test) and missed 11 suspicious comments that was wrongly predicted as “neutral ” comments.

- We classify correctly 370 neutral comments (out of the total of 372 comments used in the test) and missed 2 neutral comments that was wrongly predicted as “suspicious” comments.

- But, we couldn’t classify correctly not-suspicious comments, we missed 3 not-suspicious comments that was wrongly predicted as 1 “suspicious” comments , and 2 neutral comments.

Losing 11 suspicious comments by classifying them as neutral or not-suspicious comments or the opposite is something that cannot be neglected. Table 3.11 represents the classification errors of system.

Example	**Comments**	**Our annotation**	**System result**
1	خوكم زوالي من مسيلة كاش خرجا راني معكم.	−1	0
2	نحرقو البحر راه شباب.	−1	0
3	ليس شرط أن تدخن المخدرات لكي تكون أسطورة.	0	−1
4	انا ضد الحرقة اصلا.	1	−1

Table 3.6: Examples of classification errors of system.

- In example (1), the error is among the unexpected errors for the system because of the spelling mistakes in addition to the writing style for facebook users (this was tackled in chapter 2). For that, we find that the system did not classify the word "خرجا" as negative word since it is not written in its normal form "خرجة ". As a result, it is hard for the model to analyze it (same error in example 2 concerning the word "نحرقو" which is expected to be written "نحرقوا") .

- In example (3), the system have classified the comment as suspicious one since it recognized the word "المخدرات" as negative word. However, we have classified the comment as neutral comment since it doesn't contains any criminal content. This shows that the existence of negative word in a text is not necessary means that the comment is suspicious.

- In example (4), the same error have occurred. The system have classified the comment as suspicious one since it recognized the word "الحرقة" , but it did not take into consideration the previous word "ضد" which indicates that the comment is not-suspicious. For that, one of the solutions is to use the bigrame so that the word "ضد الحرقة" is considered as a positive one.

for improving the result of our model, we could training it with sufficient data.

3.13 Required tools:

In our experiment, we used two PCs, the first was from the HP Pavilion brand, with multi-core I5 processors, 2.20 GHZ frequency clocks and 4 GB RAM. And the second was from the Dell brand, with multi-core I5 processors, 1.80 GHZ frequency clocks and 4 GB RAM.

For programming the application, we used the following environments and libraries:

3.13.1 Python:

Python is an interpreted, high-level, and general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.[30]

Figure 3.9: Python.

3.13.2 Scrapy:

Scrapy is a fast high-level web crawling and web scraping framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing. [31]

Figure 3.10: Scrapy.

3.13.3 The Facebook Crawler:

The Facebook Crawler crawls the HTML of a website that was shared on Facebook via copying and pasting the link or by a Facebook social plugin on the website. The crawler gathers, caches, and displays information about the website such as its title, description, and thumbnail image.[40]

3.13.4 Natural Language Tool Kit:

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.[32]

Figure 3.11: Natural Language Tool Kit.

3.13.5 Scikit-learn:

Scikit-learn is a Python module integrating classical machine learning algorithms in the tightly-knit world of scientific Python packages (numpy, scipy, matplotlib).

It aims to provide simple and efficient solutions to learning problems that are accessible to everybody and reusable in various contexts: machine-learning as a versatile tool for science and engineering.[33] (See [34] for complete documentation).

Figure 3.12: Scikit-learn.

3.14 Examples of source codes:

In this section, we describe some examples of codes which have been implemented during our work:

Figure 3.13 the extraction code of all the posts.

```
>scrapy crawl fb -a email="EMAILTOLOGIN" -a password="PASSWORDTOLOGIN" -a page="NAMEOFTHEPAGETOCRAWL" -a lang="en" -o DUMPFILE.csv
```

Figure 3.13: Extraction code of all the posts.

Figure 3.14 shows the extraction code of all comments of specific posts from its URL.

```
>scrapy crawl comments -a email="EMAILTOLOGIN" -a password="PASSWORDTOLOGIN" -a post="LINKOFTHEPOSTTOCRAWL" -o DUMPFILE.csv
```

Figure 3.14: Extraction code of all comments.

However, Figure 3.15 shows the extraction of all comments from one page.

```
>pageCommentsCollector("./", "DUMPFILE6.csv", "email", "password")
```

Figure 3.15: Extraction of all comments from one page.

Figure 3.16 presents the code used to read the DataSet.

```
1  # Read Dataset and Dictionary
2  import pandas as pd
3  data = pd.read_csv('data/datasetDico.csv',sep=',')
4  data.head()
5
```

	id	categore	text
0	0	-1	اني نحوس نروح شحال راهي سوما
1	1	-1	شحال راهي بلاصة
2	2	0	اخي من المغرب ولا من الجزائر
3	3	-1	حراقة وهران اسبانيا ياخوي بكم سعر الرحلة من ا...

Figure 3.16: Read Dataset.

Figure 3.17 presents the code used to read the negative words.

```
1  # Read Negative words
2  Negative =pd.read_csv("data/motsNegative.csv", encoding ="utf-8")
3  Negative.columns=['text']
4  Negative.head()
```

	text
0	نحرقوا
1	الكالما
2	هربة
3	خارجين
4	البوطي

Figure 3.17: Read Negative words.

Figure 3.18 presents the code used to read the positive words.

```
# Read Positive words
Positive =pd.read_csv("data/motsPositive.csv", encoding ="utf-8")
Positive.columns=['text']
Positive.head()
```

	text
0	الزواج
1	العمل
2	الحلال
3	تجارة
4	الشرعية

Figure 3.18: Read Positive words.

After the reading of dataset, we have to check the existence of negative and positive words.

Figure 3.19 presents the code used to check the existence of negative words.

```
#Feature Engineering: Feature Creation
#Test the existence of a negative word
ON = []
for t in data['text']:
    if any(W in t.split() for W in Negative.text) :
        ON.append(1)
    else :
        ON.append(0)

data["one_Negative"] = ON
data.head()
```

Figure 3.19: Existence of Negative Words.

Figure 3.20 presents the code used to check the existence of positive words.

```
#test the existence of a positive word
OP = []
for t in data.text:
    if any(W in t.split() for W in Positive.text) :
        OP.append(1)
    else :
        OP.append(0)

data["one_Positive"] = OP
data.head()

```

Figure 3.20: Existence of Positive Words.

Figure 3.21 shows the code used to calculate the length of comment.

```
# Introduce the functionality of languor
word_count = []
for i in data.text:
    word_count.append(len(i.split()))

data["word_count"] = word_count
data.head()

```

Figure 3.21: Introduce the feature "length".

Figure 3.22 presents the code used for dividing the dataset into Training/Test.

```
# Split into train/test
from sklearn.model_selection import train_test_split

X=data[['one_Negative', 'one_Positive' , 'Negative_count', 'Positive_count','word_count' ]]#1
#X=data[['one_Negative', 'one_Positive' , 'Negative_count','Positive_count' ]]#2
#X=data[[ 'Negative_count', 'Positive_count' ]]#3
#X=data[[ 'Negative_count', 'Positive_count','word_count' ]]#4
#X=data[['one_Negative', 'one_Positive' ]]#5
#X=data[['one_Negative', 'one_Positive' , 'word_count' ]]#6
#X=data[['Negative_count' , 'word_count']]#7

y=data['categore']

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2, random_state=5)
```

Figure 3.22: Preparation for Analysis.

Figure 3.23 presents the call of SVM classifier.

```
[19] # Final evaluation of models
     # Call from SVM Classifier
     from sklearn.svm import SVC
     svm = SVC()
     svm.fit(X_train, y_train)
     acc1 = svm.score(X_test, y_test)
     pr1 = svm.predict(X_test)
```

```
from sklearn.metrics import precision_recall_fscore_support as score
from sklearn.metrics import accuracy_score as acs
precision, recall, fscore, train_support = score(y_test, pr1, pos_label='1', average='micro')
print('Precision: {} / Recall: {} / F1-Score: {} / Accuracy: {}'.format(
    round(precision, 3), round(recall, 3), round(fscore,3), round(acs(y_test,pr1), 3)))
```

```
Precision: 0.956 / Recall: 0.956 / F1-Score: 0.956 / Accuracy: 0.956
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/_classification.py:1321: UserWa
  % (pos_label, average), UserWarning)
```

Figure 3.23: Call of SVM Classifier.

Figure 3.24 presents the call of DT classifier.

```
[23]  # Call from DT classifier
      from sklearn.tree import DecisionTreeClassifier
      dt = DecisionTreeClassifier()
      dt.fit(X_train, y_train)
      acc2 = dt.score(X_test, y_test)
      pr2 = dt.predict(X_test)
```

```
from sklearn.metrics import precision_recall_fscore_support as score
from sklearn.metrics import accuracy_score as acs
precision, recall, fscore, train_support = score(y_test, pr2, pos_label='1', average='micro')
print('Precision: {} / Recall: {} / F1-Score: {} / Accuracy: {}'.format(
    round(precision, 3), round(recall, 3), round(fscore,3), round(acs(y_test,pr2), 3)))
```

```
Precision: 0.958 / Recall: 0.958 / F1-Score: 0.958 / Accuracy: 0.958
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/_classification.py:1321: UserW
  % (pos_label, average), UserWarning)
```

Figure 3.24: Call of DT Classifier.

Figure 3.25 presents the call of RF classifier.

```
[30]  # Call from RF classifier.
      from sklearn.ensemble import RandomForestClassifier
      rf = RandomForestClassifier(n_estimators=150, max_depth=None, n_jobs=-1)
      rf.fit(X_train, y_train)
      acc3 =rf.score(X_test, y_test)
      pr3 = rf.predict(X_test)
```

```
from sklearn.metrics import precision_recall_fscore_support as score
from sklearn.metrics import accuracy_score as acs
precision, recall, fscore, train_support = score(y_test, pr3, pos_label='1', average='micro')
print('Precision: {} / Recall: {} / F1-Score: {} / Accuracy: {}'.format(
    round(precision, 3), round(recall, 3), round(fscore,3), round(acs(y_test,pr3), 3)))
```

```
Precision: 0.961 / Recall: 0.961 / F1-Score: 0.961 / Accuracy: 0.961
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/_classification.py:1321: UserW
  % (pos_label, average), UserWarning)
```

Figure 3.25: Call of RF Classifier.

Figure 3.26 presents the call of NB classifier.

```
[33]  # Call from NB classifier
      from sklearn.naive_bayes import GaussianNB
      nb = GaussianNB()
      nb.fit(X_train, y_train)
      acc4 = nb.score(X_test, y_test)
      pr4 = nb.predict(X_test)
```

```
from sklearn.metrics import precision_recall_fscore_support as score
from sklearn.metrics import accuracy_score as acs
precision, recall, fscore, train_support = score(y_test, pr4, pos_label='1', average='micro')
print('Precision: {} / Recall: {} / F1-Score: {} / Accuracy: {}'.format(
    round(precision, 3), round(recall, 3), round(fscore,3), round(acs(y_test,pr4), 3)))
```

```
Precision: 0.074 / Recall: 0.074 / F1-Score: 0.074 / Accuracy: 0.074
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/_classification.py:1321: UserW
  % (pos_label, average), UserWarning)
```

Figure 3.26: Call of NB Classifier.

3.15 Conclusion:

This chapter is devoted to highlighting the importance of ML and its mechanism for classifying texts in order to be able to classify comments of social networks to detect probable crime. In this chapter, we give a definition for ML and its category. Then, we discover the different algorithms that will be used to classify comments, we talked about the evaluation system in which we show the main measurements that will be needed to evaluate the accuracy of the model, we proposed an approach which can detect crimes on Facebook using Algerian Dialect.

We have found that the very best result of F-measurement is **96.1**%, which carried out by theRF model compared with the other ML algorithms as SVM, DT and NB which can reach respectively the values**95.6**%, **95.8**% and **7.4**%. In our experimentation, we use five features which are: HPW, HNW, PWC, NWC, and CL.

GENERAL CONCLUSION:

This dissertation aims at revealing crimes that are characterizing the posts on the social media sites such as Facebook in accordance with three ways; positive post, negative post, and neutral post.

This work aims to implement an application in Python which uses dataset CSV that contains classified texts according to the values $(1, -1,$ and $0)$. In addition, we create a list of words in Algerian dialect which used to classify these texts. Firstly, we defined some of the aspects that are used in this work, as Algerian dialect, crime...etc. Then, we present some previous works and we focused mainly on six relevant works. After that, we studied the linguistic characteristics of the Algerian dialect.

In this work, we have contributed in three mainly point:

- We create an Initial list, which includes 350 words with Algerian dialect. This list contains relevant crime from different regions.
- We have collected Dataset from more than 500 posts on Facebook.
- We use more than 2500 comments on Facebook; 500 for different crimes and 2000 for illegal emigration crimes.
- We implement an application which can detect the crime.

In this field of study, we use four classifiers which are NB, RF, DT, and SVM. In addition, the results that we have obtained were encouraging results where we have a better accuracy with the classifier RF which is 96.1%.

As future work, we will evaluate the other MLs Algorithms such as NN, association rules, and others, for more accurate results. In addition, we intend to expand our analysis to include time and space analysis to know when, where, and how much crimes had spread in the past, and predict where and when it may spread in the future. In addition, we will expend the dataset and dictionary to contain different Algerian local accents and different terminologies in the Algerian dialect that are considered as passwords that criminals use before the crime.

BIBLIOGRAPHY

[1] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, Proceedings of the First Workshop on Social Media Analytics, 2010, New York, USA, pages: 115-122.

[2] F. Franch, Wisdom of the crowds 2 : 2010 UK election prediction with social media, Journal of Information Technology & Politics, 10(1), 2012, pages: 57-71.

[3] X. Wang, D. Brown, and M. Gerber, Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information, Intelligence and Security Informatics, IEEE international Conference on Intelligence and Security informatics, Arlington, USA, pages:36-41.

[4] S. Asur and B. Huberman, Predicting the future with social media, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, Toronto, ON, Canada, pages: 1-8.

[5] P.S. Earle, D.C. Bowden, and M. Guy, Twitter earthquake detection: earthquake monitoring in a social world, Annals of Geophysics, 54(6), 2012, pages: 708-715.

[6] H. Choi and H. Varian, Predicting the present with Google Trends, The Economic Record, Wiley Online Library, 88(1), 2012, pages: 2-9.

[7] J. Bollen, H. Mao, and X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science, Elsevier, 2(1), 2011, pages: 1-8.

[8] Mohammed A. AlGhamdi and Murtaza Ali Khan, Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages, Arabian Journal for Science and Engineering, 45(8), 2020, pages: 6021-6032.

[9] Bolla Raja Ashok, Crime pattern detection using online social media, Missouri Unverdity of Science and Technology, 2014.

[10] Assia S.,Mhani M., Ahmed G., and Amina D., Sentiment Analysis of Users on Social Networks: Overcoming the challenge of the Loose Usages of the Algerian Dialect, Procedia Computer Science, ELsevier, 142, 2018, pages: 26-37.

[11] I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, and A. Wijayasiri, Crime analytics: Analysis of crimes through newspaper articles, Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 2015.

[12] Mugdha Sharma, Z - CRIME, A data mining tool for the detection of suspicious criminal activities based on decision tree, International Conference on Data Mining and Intelligent Computing (ICDMIC), New Delhi, India, 2014.

[13] S. Sathyadevan, M. S. Devan, and S. Surya Gangadharan, Crime analysis and prediction using data mining, First International Conference on Networks Soft Computing (ICNSC2014), Guntur, India, 2014.

[14] Hissah AL-Saif, Hmood Al-Dossari, Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques, International Journal of Advenced Computer Science and Applications, 9(10), 2018, pages: $377 - 385$.

[15] Motaz K. Saad, The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification, Master thesis, the islamic university, 2010.

[16] Wafia Adouane, Simon Dobnik, Identification of Languages in Algerian Arabic Multilingual Documents, Proceedings of the Third Arabic Natural Language Processing Workshop, Valencia, Spain, 2017, pages: 1-8.

[17] Fleiss, J.L, Measuring nominal scale agreement among many raters, Psychological Bulletin, 76(5), 1971, pages: 378-382.

[18] Rehab M. Duwairi, Raed Marji, Narmeen Sha'ban, and Sally Rushaidat, Sentiment Analysis in Arabic Tweets, 5th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2014, pages; 1-6.

[19] www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/, "J. Clement",Accessed 04/2020.

[20] www.napoleoncat.com/stats/facebook-users-in algeria/, Accessed 04/2020.

[21] www.echoroukonline.com, "Nawara Bachoush / Ilham Bothalji", Accessed 04/2020.

[22] www.en.wikipedia.org/wiki/Crime, Accessed 02/2020.

[23] www.justia.com/criminal/offenses, Accessed 02/2020.

[24] www.study.com/academy/lesson/crime-definition-types.html,"Jessica Schubert", Accessed 02/2020.

[25] www.xinhuanet.com/english/2018-11/14/c_129993383.htm, "Mu Xuequan",Accessed 02/2020.

[26] www.echoroukonline.com, "Nawara Bachoush", Accessed 03/2020.

[27] www.arabamerica.com/the-richness-of-the-arabic-language,"Ala-Abed-Rabbo/Arab America Contributing Writer", Accessed 03/2020.

[28] www.whatis.techtarget.com/definition/decision-tree, "Margaret Rouse", Accessed 03/2020.

[29] www.deepai.org/machine-learning-glossary-and-terms/random-forest, Accessed 03/2020.

[30] www.en.wikipedia.org/wiki/Python_(programming_language), "Guido van Rossum", Accessed 04/2020.

[31] www.pypi.org/project/Scrapy/2.0.0/, " Scrapy developers", Accessed 03/2020.

[32] www.nltk.org/, "Sphinx 2.4.4", Accessed 04/2020/

[33] www.kite.com/python/docs/sklearn, Accessed 04/2020.

[34] www.scikit-learn.org, Accessed 04/2020.

[35] www.alaraby.co.uk, "Hossam El Din Fadil", Accessed 05/2020

[36] www.osac.gov/Country/Algeria/Content/Detail/Report/bbd53147-8798-447f-8620-15f4aeb5908b, "the Regional Security Office at the U.S. "Embassy in Algiers, Algeria.", Accessed 05/2020

[37] www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html, Accessed 06/2020 .

[38] www.super-ai-diasceative.net/ai-and-machine-learning-policy-paper, Accessed 05/2020.

[39] www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ , Sunial Ray, Accessed 03/2020.

[40] www.developers.facebook.com/docs/sharing/webmasters/crawler/, Accessed 10/2020