

إقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Investigating Crimes Using

Text Mining and Network Analysis

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه
حيثما ورد، وإن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل لنيل درجة أو لقب علمي أو
بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

DECLARATION

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification

Student's name: *Nael T. Elyezzy*

Signature: *[Handwritten Signature]*

Date: *31/8/2015*

اسم الطالب/ة: *نائل تبسير اليازجي*

التوقيع: *[Handwritten Signature]*

التاريخ: *31 أغسطس 2015*

Islamic University – Gaza
Deanery of Post Graduate Studies
Faculty of Information Technology



**Investigating Crimes Using
Text Mining and Network Analysis**

Submitted by:
Nael T. Elyezjy

Supervisor
Dr. Alaa El-Halees

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master in Information Technology

Sahiwal 1436H - July 2015



نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ نائل تيسير نمر اليازجي لنيل درجة الماجستير في كلية تكنولوجيا المعلومات برنامج تكنولوجيا المعلومات وموضوعها:

التحقيق في الجرائم من خلال تحليل النصوص وتحليل الشبكات

Investigating Crimes Using Text Mining and Network Analysis

وبعد المناقشة التي تمت اليوم الأربعاء 19 شعبان 1436هـ، الموافق 2015/08/26م الساعة الواحدة ظهراً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

.....	مشرفاً و رئيساً	أ.د. علاء مصطفى الهاليس
.....	مناقشاً داخلياً	د. إياد محمد الأغا
.....	مناقشاً خارجياً	د. يوسف نبيل أبو شعبان

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في كلية تكنولوجيا المعلومات / برنامج تكنولوجيا المعلومات.

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله و لزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.



نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤف علي المناعمة

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

قَالَ اللَّهُ تَتَلَوْنَهَا

قَالَ اللَّهُ تَتَلَوْنَهَا

Abstract

In these days, security of citizens is considered one of the major concerns of any government in the world. In every country, there is a huge amount of unstructured texts coming from investigating offenders in police departments. As a result, the importance of crimes analysis is growing day after day. Criminology is one of the hot areas that focuses on the scientific study of crimes and aims to identify crime characteristics and both criminal behaviors and networks. This field of study is one of the most intelligent research areas where text mining is used to process unstructured texts and extract meaningful information which is hidden in the unstructured texts. The knowledge extracted from text mining is very useful to police departments where solving crimes is a very complex task that requires human effort and experience.

There is a little research in methods and techniques that extract criminal networks from unstructured investigations texts especially in Arabic language. Accordingly, the current research proposes a system to identify networks of criminals, and extract useful information relevant to crimes such as offender's connection networks and discover a new hidden relationship between offenders by linking investigation documents with each other. After that, the results of the research are visualized direct and indirect relation between offenders to help policemen find pieces of evidences related to certain crimes and accordingly apply the law.

In our proposed system, we climb three main distinct contributions to discover forensics using investigation documents. The first by extracting offender names from unstructured text. Secondly, by constructing a crime network from real Arabic investigation documents. Finally, we provide analysis of the interaction between offenders in different documents that directly and indirectly related used to discover a new clue used to solve the crime puzzle. To evaluate the performance and effectiveness of the proposed system, real unstructured documents about investigations are obtained from police departments in the Gaza Strip. The experimental results show that the proposed system is effective in identifying proper offender person's name from real Arabic Documents. The average results for our

system using the F-measure is 89% also the average of F-measure in a proposed algorithm for discovery hidden relationship arrive to 92%. In addition, we found that our approach achieves best F-measure results in most cases.

Keywords: Criminology, Text Mining, Crime investigation, Criminal Networks, law enforcement.

التحقيق في الجرائم من خلال تحليل النصوص والشبكات الاجتماعية

في هذه الأيام، يعتبر أمن المواطن من المهام الرئيسية لأي حكومة في العالم، حيث أنه هناك كم هائل من نصوص التحقيقات مع المجرمين. وعلية، فإن أهمية تحليل الجرائم تزداد يوماً بعد يوم. علم الجريمة هو أحد النقاط الساخنة في البحوث العلمية التي يهدف من خلالها التعرف على خصائص الجرائم والسلوكيات الإجرامية وشبكة العلاقات بين المجرمين. تعتبر هذه الدراسة واحدة من البحوث الأكثر ذكاءً حيث تقوم بتحليل النصوص الغير منتظمة ويتم من خلالها استكشاف البيانات من نصوص التحقيقات. المعرفة المستخرجة من نصوص التحقيقات مفيدة جداً لرجال الشرطة في حل جرائم معقدة تتطلب خبرة وكثيراً من الجهد والوقت البشري. على حد علم الباحث، هناك القليل من التقنيات والطرق المستخدمة في استكشاف الشبكات الإجرامية من خلال نصوص التحقيقات وخاصةً في اللغة العربية. وبناءً على ذلك، يقترح البحث الحالي نموذجاً جديداً لاكتشاف شبكات المجرمين من خلال ربط نصوص التحقيق مع بعضها البعض. بعد ذلك يتم عرض النتائج بطريقة سلسلة لمساعدة رجال الشرطة للعثور على بعض الأدلة المتعلقة بجريمة معينة مما يؤدي الى تطبيق القانون.

في نظامنا المقترح، نسعى لتحقيق ثلاث أهداف رئيسية في اكتشاف الأدلة الجنائية باستخدام نصوص التحقيقات، أولهما استخراج أسماء الجناة من نصوص التحقيق. ثانياً، من خلال بناء شبكات المجرمين باستخدام نصوص تحقيق عربية حقيقية. وأخيراً، نقدم تحليل تفاعل الجناة مع بعضهم باستخدام وثائق التحقيق المختلفة لاكتشاف خيوط جديدة تساعد على حل لغز الجريمة. لتقييم أداء وفعالية النموذج المقترح، تم الحصول على نصوص تحقيقات واقعية من إدارات الشرطة في قطاع غزة. النتائج التجريبية تبين أن المقترح المقدم فعال في تمييز أسماء الأشخاص المتهمين من نصوص التحقيقات العربية الحقيقية حيث أن متوسط المعدل وصل الى 89% أيضاً معدل الخوارزمية المقدمة لاكتشاف العلاقات الغير مباشرة وصل إلى 92%. ووجدنا نهجنا حقق أفضل نتائج ودقة أكثر في معظم الحالات.

الكلمات المفتاحية: الكشف عن الجرائم، تعدين النصوص، تحقيق الجرائم، الشبكات الإجرامية وتطبيق القانون.

Dedication

To my beloved mother ...

To my beloved father ...

To my wife and children ...

To my sisters and brothers...

To my best friends ...

To Palestine ...

Acknowledgements

First and foremost, thanks to Allah for giving me the power and help to accomplish this research. Without the grace of Allah, I was not able to accomplish this work.

*Many thanks and sincere gratefulness goes to my supervisor **Dr. Alaa El-halees**, without his help, guidance, and continuous follow-up; this research would never have been.*

Special thanks also to my parents, my wife, my brothers and sisters for their endless support. Without them, I would never have been able to achieve my goals.

Table of contents

ABSTRACT	II
المخلص.....	IV
DEDICATION	V
ACKNOWLEDGEMENTS.....	VI
TABLE OF CONTENTS	VII
LIST OF FIGURES.....	X
LIST OF TABLE	XI
LIST OF ABBREVIATIONS	XII
1. CHAPTER 1 INTRODUCTION	2
1.1 PROBLEM STATEMENT	5
1.2 OBJECTIVES	5
1.2.1 <i>Main Objective</i>	5
1.2.2 <i>Specific Objectives</i>	5
1.3 SIGNIFICANCE OF THE THESIS	6
1.4 CHALLENGES OF THIS THESIS	7
1.5 SCOPE AND LIMITATIONS	7
1.6 RESEARCH METHODOLOGY	8
1.7 THESIS FORMAT	11
2. CHAPTER 2 RELATED WORKS	13
2.1 CRIME DETECTION	13
2.2 CRIMINAL SOCIAL NETWORK ANALYSIS	15
2.3 SUMMARY	17
3. CHAPTER 3 THEORETICAL FOUNDATION	19
3.1 INFORMATION EXTRACTION	19
3.2 NAME ENTITY RECOGNITION	19
3.2.1 <i>Learning method of NER</i>	20
3.3 INTEGRATED DEVELOPMENT ENVIRONMENTS	21
3.4 HYPOTHESIS GENERATION	22
3.4.1 <i>Community discovery problem</i>	23
3.4.2 <i>Discovery Algorithms</i>	23

3.4.3	<i>The Problem of Indirect Relationship Discovery</i>	23
3.5	CRIMINAL NETWORK ANALYSIS AND VISUALIZATION	24
3.5.1	<i>Criminal Network Analysis</i>	25
3.6	DATA VISUALIZATION	25
3.7	PERFORMANCE METRICS	26
3.7.1	<i>Confusion Matrix</i>	26
3.7.2	<i>Accuracy</i>	27
3.7.3	<i>Precision</i>	27
3.7.4	<i>Recall</i>	27
3.7.5	<i>F-measure</i>	27
3.8	SUMMARY	28
4.	CHAPTER 4: THE PROPOSED CRIME DETECTION APPROACH	30
4.1	INTRODUCTION	30
4.2	THE OVERALL CRIME DETECTION SYSTEM (CDS) ARCHITECTURE	30
4.3	INITIAL PREPARATION STAGE ARCHITECTURE	31
4.3.1	<i>Data Gathering</i>	31
4.3.2	<i>Data preprocessing</i>	32
4.3.3	<i>Tokenization</i>	32
4.3.4	<i>Normalization</i>	32
4.4	EXTRACT OFFENDER NAMES STAGE	33
4.4.1	<i>Gazetteers</i>	33
4.4.2	<i>Rule-based approach</i>	34
4.4.3	<i>Criminal Communities Discovery</i>	36
4.5	INDIRECT RELATIONSHIP EXTRACTION	37
4.6	DATA VISUALIZATION	40
4.7	SUMMARY	40
5.	CHAPTER 5 EXPERIMENTS AND RESULTS	42
5.1	EXPERIMENTS SETUP	42
5.1.1	<i>Experimental Environment and Tools</i>	42
5.2	ARABIC INVESTIGATION DOCUMENTS CORPUS	43
5.3	DATA PREPROCESSING STAGE	44
5.4	NAME ENTITY RECOGNITION	44
5.4.1	<i>A Nearly-New Information Extraction system (ANNIE)</i>	45
5.5	INDIRECT RELATIONSHIP DISCOVERY ALGORITHM	48
5.6	DATA VISUALIZER	48

5.7	SYSTEM EFFICIENCY EVALUATION	50
5.7.1	<i>Name Entity Recognition and Human Evaluation</i>	<i>50</i>
5.7.2	<i>Indirect Relationship Discovery Algorithm</i>	<i>54</i>
5.7.3	<i>System Scalability Evaluation</i>	<i>61</i>
5.7.4	<i>Discussion</i>	<i>62</i>
5.8	SUMMARY	63
6.	CHAPTER 6 CONCLUSION AND FUTURE WORK	65
6.1	SUMMARY	65
6.2	CONTRIBUTION	65
6.3	RECOMMENDATIONS.....	66
6.4	FUTURE WORK	66
	REFERENCES.....	67
	APPENDIX A.....	71
	A.1 OFFENDER NAME EXTRACTOR RULES	71
	A.2 NAME ENTITY RECOGNITION EVALUATION	77

List of Figures

Figure 1.1: relationship between investigations documents.....	4
Figure 1.2: The Research Methodology	11
Figure 4.1: System Architecture	31
Figure 5.1: Crime Detection corpus	44
Figure 5.2: Preprocessing techniques.....	44
Figure 5.3: Name Entity Extraction	46
Figure 5.4: Sample OrthoMatcher Results in Co-reference Editor	46
Figure 5.5: Sample OrthoMatcher Results in Annotation List.....	47
Figure 5.6: Sample of offender name extraction.....	47
Figure 5.7: Sample network visualization.....	49
Figure 5.8: Sample Data table presentation	49
Figure 5.9: Key person of Community	49
Figure 5.10: Preprocessing Documents before evaluate by Human expert	50
Figure 5.11: Measure (R, P, F) using Annotation Diff Tool in GATE tool Case 1	52
Figure 5.12: Measure (R, P, F) using Annotation Diff Tool in GATE tool Case 2	53
Figure 5.13: No. of documents vs. execution runtime	62

List of Table

Table 3.1: Simple Confusion Matrix.....	26
Table 4.1: The results of modifying Gazetteer lists	34
Table 5.1: Machine environment properties	42
Table 5.2: Summary of evaluation system for extract offender names from Arabic investigation documents case 1	51
Table 5.3: Summary of evaluation system for extract offender names from Arabic investigation documents case 2	52
Table 5.4: Conclusion results of F- measure calculation	53
Table 5.5: Case (1) for discovery indirect relationship using human expert	55
Table 5.6: Results of compute R, P, F for offender name extraction case 1	56
Table 5.7: Evaluation of hidden relationship discovery algorithm case 1	56
Table 5.8: Average calculation for (R, P, F) in case 1	57
Table 5.9: Case (2) for discovery indirect relationship using human expert	57
Table 5.10 : Results of compute R, P, F for offender name extraction case 2	58
Table 5.11: Evaluation of hidden relationship discovery algorithm case 2	59
Table 5.12: Average calculation for (R, P, F) in case 2	60
Table 5.13: Summarize the results of calculation (R, P, and F) for discovery algorithm	60
Table 5.14: No. of documents vs. execution runtime.....	61
Table A.1: evaluation results of name entity recognition	77

List of Abbreviations

NLP	Natural Language Processing
IE	Information Extraction
IR	Information Retrieval
CDS	Crime Detection System
NER	Named Entity Recognition
NE	Named Entity
CO	co-reference resolution
GATE	General Architecture for Text Engineering
ANNIE	A Nearly-New Information Extraction system
JAPE	Java Annotation Patterns Engine
LHS	Left hand side
RHS	Right hand side
SNA	Social Network Analysis
DM	Data Mining
MUC-6	The 6th Message Understanding Conference
QA	Question Answering
POS	Part-of-Speech
SL	Supervised Learning
SSL	Semi-supervised learning
UL	Unsupervised learning
HMM	Hidden Markov Systems
SVM	Support Vector Machines
ME	Maximum Entropy Systems
CRF	Conditional Random Fields
HG	Hypothesis Generation

Chapter 1

Introduction

Objectives

- Provide introduction to our thesis.
 - Provide the motivation for undertaking this research.
 - Explain the research methodology.
 - Provide scope and limitations for the thesis.
 - Provide the thesis structure.
-

Chapter 1 Introduction

Over the quite a lot of years, especially with growing in terms of populations, it is impossible to find a crime-free society in the world. This makes security of people become one of the most crucial responsibilities of governments all over the world. Thus, the main goal, here, is to reduce crime incidences[1]. The Gaza strip is one of the areas in the world with a society that is growing in intensity and complexity. This leads to an increase of crime incidences such as burglaries, thefts, robberies, vehicle crimes, murders, armed trafficking, sexual crimes, and international crimes, etc.[2-4].

Crime is a deviation in some people's behaviors, from normal habits, that leads people to many harms at the level of spirit, personal properties, environment, etc. [5]. Police officers play a major role in civil administration; they are responsible for preventing and predicting crimes, and enforcing the law as well. Preventing crimes are very important to make people live safer and more stable.

Criminology is the process of identifying the characteristics of crimes [1]. Usually, a crime is not random[6]. Police officers prepare reports manually and in unstructured form. Analysis of criminal networks manually from this unstructured form is time-consuming and investigators can take months to solve a crime puzzle. This is because of the huge amount of data scattered across multiple structured and unstructured documents about current and old investigations. Nevertheless, doing this task manually may consume more time and resources; and it usually depends on the experience of policemen.

The increase of crimes complexity motivates and leads governments and police departments to use technology to reduce the effort and to speed up the process of analyzing and linking information to discover criminals.

The abundance of investigation reports has increased the amount of data available at police departments. These data are usually kept for archiving purposes. But, if this content is mined for specific information, this may lead to other resources that can assist the investigation process and analyze this retrieved information and may help catch offenders in a faster way.

Unstructured text is very common. According to Gupta et al. over 80% of information is stored as texts [7]. Many police departments have a huge amount of

investigations as unstructured text that can be useful to detect and prevent a new crime accident by identifying crime patterns. Text mining techniques played a vital role in the last few years in knowledge extraction from unstructured documents, especially in crime detection and prevention. This is done through the analysis of large amounts of crime-related information to find those who are responsible for each crime [8].

The number of publications and research projects in data mining in law enforcement area is slowly increasing [9]. Most of the tools and software used by police departments utilizes structured databases which are easy for investigators to compute some statistics about the crime, or search for particular information about crime. But unstructured documents of previous investigations are usually saved in an archive. On the other hand, data can also be stored in unstructured texts, but these data will be complex to manipulate. Therefore, the study of criminal networks is increasing day after day to detect suspicious networks and apply the law. Most police departments nowadays have come to realize the detailed knowledge of invisible links and more details about offenders who are involved in crimes [10, 11].

Arabic language is one of the most widely spoken languages in the world. As far as is known, a little research focus on crime domain in Arabic language. The first goal of the current study is to develop a system for extracting useful information in Arabic crime domain from unstructured investigation data in order to mine it [12, 13].

The main purpose of this thesis is to study various approaches to create a new system for crime detection using unstructured text mining techniques. It focuses on solving problems of discovering social relations between offenders from unstructured text investigations and find out useful information by applying text-mining methods. This system will be conducted in an investigation to help police officers to efficiently detect hidden relations between criminals in a large volume of investigation documents.

Understating relationships between crime investigations can help investigators to detect hidden information in order to identify trends and patterns, and even predict new actors of the crime. For example, suppose in Figure 1.1 that an investigator interrogates an offender named (Ali) about an accident such as a theft, and the

investigator has an archive of unstructured investigation documents $d = (d_1, d_2, \dots, d_n)$ where n is the number of documents. Where, someone need to find the relation between (Ali) and other offenders using archive documents. First, all names of persons will be extracted from each document. After that, stemming for those names will follow criteria that unify the results of finding names of offenders. Accordingly, direct and indirect relations between offenders will be extracted. In the end, results will be visualized as shown in Figure 1.1.

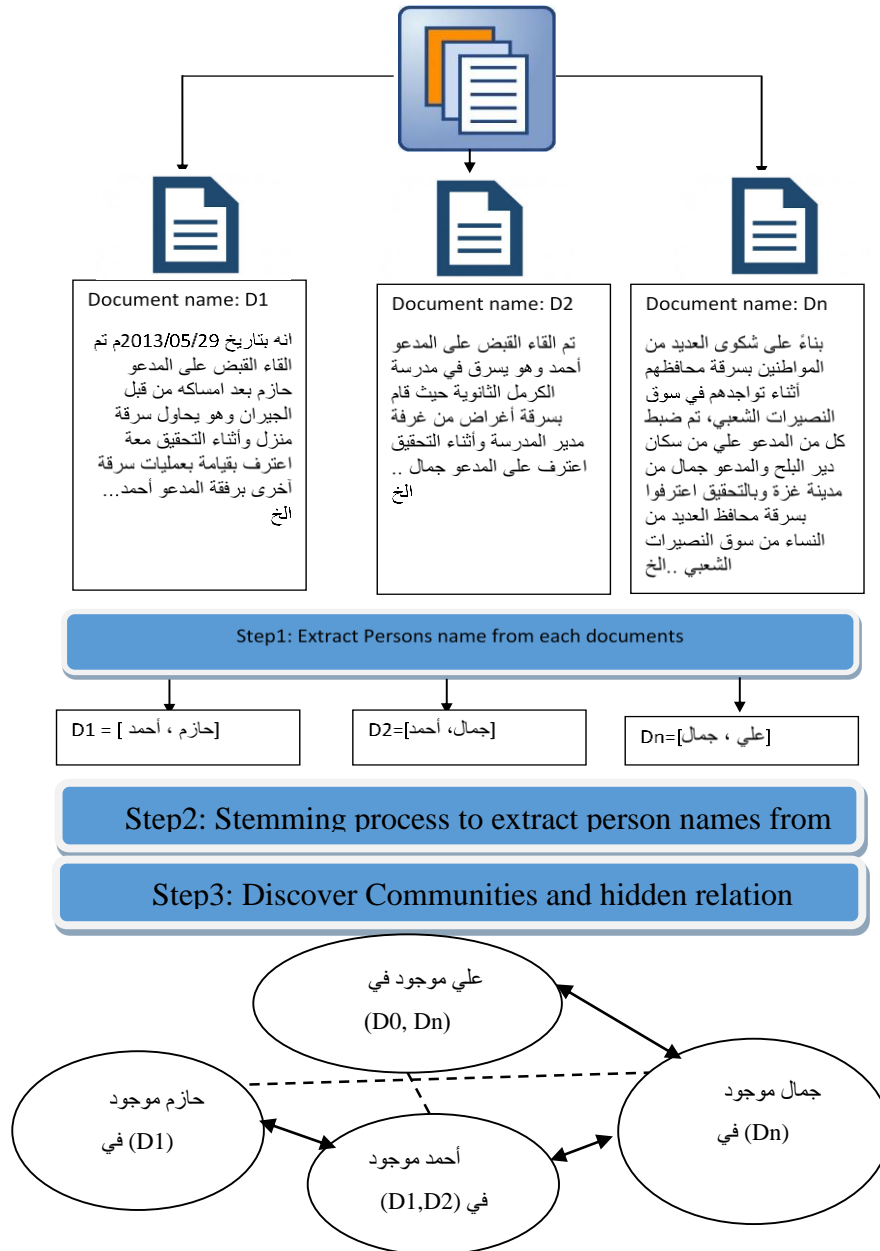


Figure 1.1: relationship between investigations documents

As shown in Figure 1.1,

- 1- Ali has direct relation to Jamal
- 2- Jamal has direct relation with Ahmed, so Ahmed may know Ali.
- 3- Ahmed has direct relation with Hazem, so Hazem may know Jamal and Ali.
- 4- And so on.

In our system, a three-stage approach to crime detection is suggested. In the first stage, named entities in the target texts are automatically recognized. In the second stage, social relationships between criminals from multiple investigations documents are identified. In the third stage, hidden links between criminals are extracted by analyzing the criminal social network. The goal, here, is to discover more evidences about the crime.

1.1 Problem Statement

Police departments need a system to handle a huge number of investigation documents to reduce effort and time in order to find hidden relationships between the actors. These relations are very important to detect dangerous links between networks and extract useful information from investigation documents that can be used as an evidence.

Therefore, the current research attempts to answer the following question.

How can we investigate a relationship between offender actors of different crimes using text mining based on investigation documents written in Arabic?

1.2 Objectives

1.2.1 Main Objective

The main objective of this thesis is to develop an effective system using text-mining techniques to analyze Arabic-based crime investigation documents. This help identify direct and indirect criminal relationships between actors in order to assist investigators to detect hidden communities and identify other participants.

1.2.2 Specific Objectives

The specific objectives of the project are to:

- Collect real data from police departments.
- Perform preprocessing phases and classification efficiently
- Find a suitable algorithm to extract Arabic named entities from unstructured text.
- Propose an approach to build a criminal network.
- Analyze crime networks and trace patterns of behaviors.
- Identify the key person of offenders in the same investigation document.
- Visualize the resulting network to the user of the proposed system.
- Evaluate the performance of the proposed approach with real investigation documents using recall and precision.

1.3 Significance of the Thesis

The importance of this study springs from the fact that crime ratio and complexity of criminal accidents rapidly increase. And usually police officers, when follow a crime, depend on the expertise of other police officers. It is also important because safety of people is considered a major objective and concern in any country in the world. Therefore, police needs rapid and effective techniques to predict crime patterns and find a criminal network between offenders to improve strategic decision-making. Moreover, text-mining methods have become a key feature for homeland-security technologies [14]. To sum up, the current study is important because of the following.

- Helps police officers predict crimes and extract relations between actors from unstructured investigations texts; and this may lead to new clues and criminals tracking.
- To the best of the researcher's knowledge, the current system will be one of the least attempt in this field using Arabic language in crime detection to find invisible links between actors.
- There is no comparable study on real crime investigations documents.

- Mines hidden relationships on criminal social networks by discovering ambiguous relations between offenders.

1.4 Challenges of this thesis

There are several challenges in this area:

- Data set is written in Arabic language without grammar rules and unstructured Arabic language.
- Existing techniques are developed for English language and do not always work for other languages such as Arabic.
- There are many basic assumptions about capitalization and tokenization that would not work for other languages especially Arabic.
- Most of the algorithms that extract names from unstructured text depend on English language and not support Arabic Language.
- The name of the same person may be referred to more than one time.
- Collecting sensitive data about investigations from police departments in Gaza strip.
- Due to the sensitive nature of real crime datasets, they are not easily available for academic research because they involve problems and difficulties[15].

1.5 Scope and Limitations

There are some limitations, which should be considered during the phases of this research. They are as follows:

- The research is only concerned about defining criminal networks between offenders.
- The data set does not contain misleading or noise value.
- The dataset collected from police departments is only about theft incidents in Gaza strip and limited to the period between the beginning of 2008 and the end of 2013.

- The research only considers documents in Arabic language.
- We assume there is a relationship between any two offenders which might be found in the same investigation document.

1.6 Research Methodology

The research methodology, as shown in Figure 1.2 employed in this thesis is described and summarized in the following points:

Background

To start this thesis, a great amount of relevant literature review is studied. This leads the researcher to establish a road map for collecting appropriate materials needed to cover this stage. Several sources are also used such as IEEE Explore, Google search engine, ACM, etc.

Data Collection

The proposed work is carried out using a real dataset related to crime incidents. This dataset is collected from police departments in the Gaza strip. In addition, the target investigation documents about theft incidents in Gaza strip are limited to the period between the beginning of 2008 and the end of 2013.

Document Pre-processing

This phase includes preparing the target data to convert them to some standard format suitable for text mining. The preparation of such data consists of several steps: tokenization, sentence splitting, foreign word/symbol removal, stop word removal, and normalization [16, 17].

Extract person names using Arabic NER

Named Entity Recognition is one of the main natural language processing (NLP) tasks [18]. First, the question that should be asked is: what kind of system or tools can best extract Arabic names from text?

In this thesis, attempts are made to answer this question. Each person name is a candidate to be a node in the crime network.

Co-Reference

Co-Reference play a vital role for several Natural Language Processing (NLP) applications such as Text Summarization, Machine Translation or Information Extraction (IE) [19]. It is used when two or more names refer to the same person. For example, (Ahmed) and (he said); the proper noun (Ahmed) and the pronoun (he) refer to the same person, namely Ahmed. In this step, a standardization of names is generated.

Normalization

The goal of the normalization process is to eliminate duplicate names that refer to the same person. This reduces redundancy in the identified communities.

Extract Criminal Communities

A community consists of nodes and edges between these nodes [20]. The main purpose of this phase is to extract communities from set of investigations documents by extracting person's names. This leads to build a strong linkage between criminals. It is assumed that a relationship exists between two offenders if they have participated in the same crime. For that, FP-growth algorithm[21] may be used as it is an efficient algorithm to find frequent pattern in a set of documents.

Extract Key person of Community

At the same time when a community is extracted from investigations documents, if there is more than one person in documents, the frequency of a person's name is counted. The person with the highest frequency is labeled as the key person of that community of criminals.

Detect Invisible Relations that Link Communities

Most contributions in developing social networks have exclusively focused on direct links between actors [22]. However, it is an important step to discover the unexpected actors in a crime. This can be by achieved finding a friend of a friend relationship between person's names in different communities. In fact, actors may belong to different communities. Therefore, in this phase, an efficient algorithm is

needed to detect the overlapping communities. In this phase, a system is developed to discover hidden relations between actors using a new discovery algorithm.

Visualize Criminal Networks

To evaluate the accessibility of the current system, visualization will be used as one of the technical possibilities after unstructured information has been transformed into a structured representation (i.e., the crime network). The visualization is used to simplify the process of results reading and drawing conclusions. While the criminal names are mapped to nodes, the relationships are mapped to edges between nodes; and we will distinct between direct and indirect edges.

Evaluation

System evaluation is a difficult task because there is no ideal crime network for a given documents or set of documents and this type of evaluation depends on human experts. However, in the current approach, a real data set is used to evaluate the performance. In order to evaluate the effectiveness of approach, we measure the precision of both direct and indirect relation between offender's discovery modules using some of actual real-life data. By manually expecting the result where obtained from experiment data set; where expert human can verify the results and evaluate the accuracy of the system. Also the scalability of the proposed system will be evaluated by measuring runtime execution per data size.

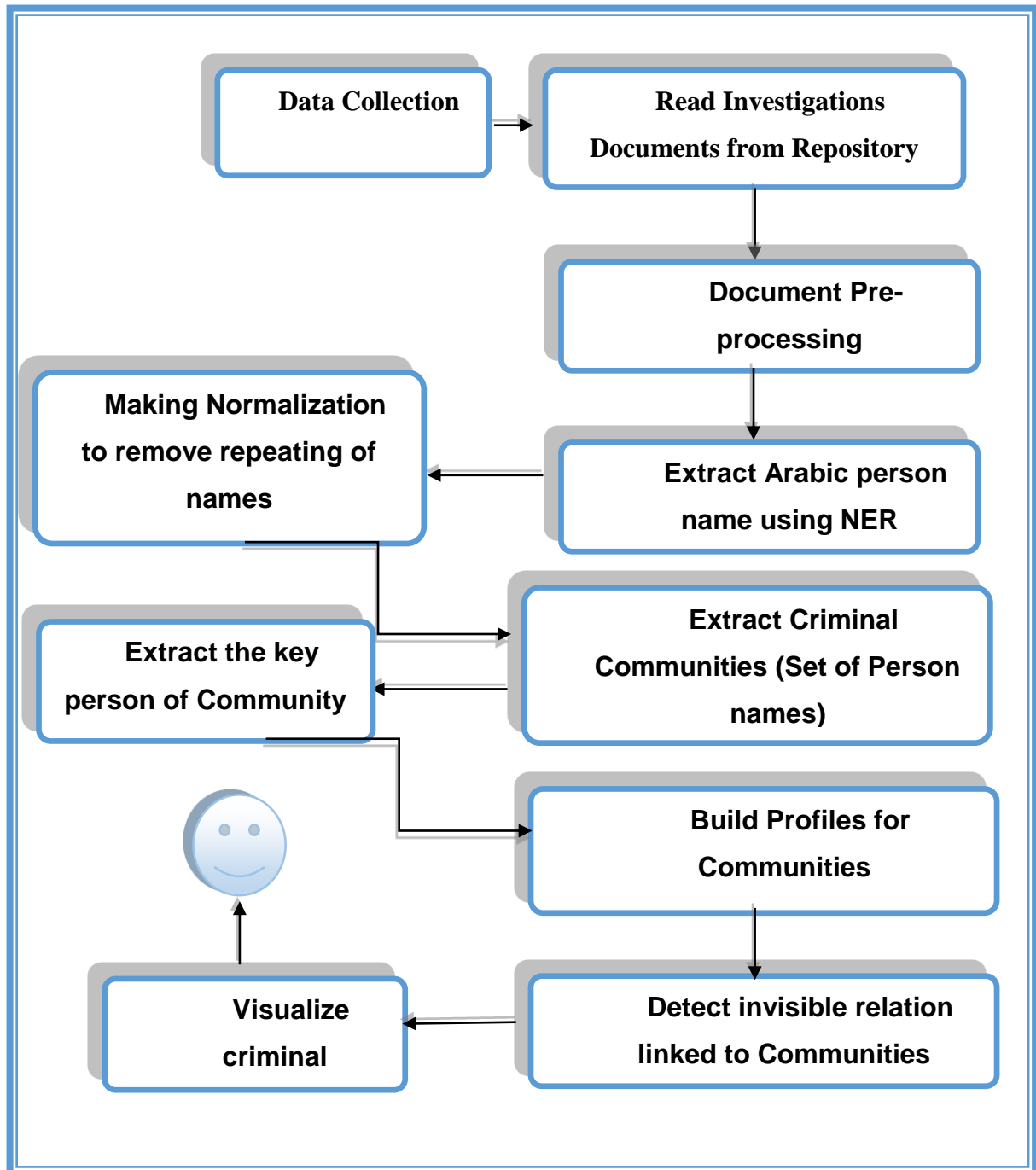


Figure 1.2: The Research Methodology

1.7 Thesis Format

The rest of the thesis is organized into 6 chapters, as follows: Chapter 2 discusses the state of the art and literature survey. Chapter 3 includes the theoretical foundation of the thesis. Chapter 4 presents the proposed crime detection approach. Chapter 5 presents the experimental results and evaluation. Finally, Chapter 6 presents the conclusions, recommendation and future work.

Chapter 2

Related Works

Objectives

- Discuss related research and current approaches.
 - Provide Arabic crime detection literature review.
 - Provide criminal social network analysis literature review.
 - Discuss drawback and advantage of literature review.
-

Chapter 2 Related Works

Many researches gave a great attention to criminal network analysis, but a few of them proposed systems for a criminal network analysis to handle Arabic language. In this chapter, a number of research works that focused on criminal network analysis and crime detection is reviewed. This literature review is divided into two sections: literature on crime detection, and a criminal network analysis.

2.1 Crime detection

There are some researches in Arabic crime detection such as: Alruily, et al in [1] built software that is used to determine types of crime from free text. The main approach of their paper is basically based on building a predefined dictionary that contains some important keywords that can be used to classify the crime domain. The researchers used two techniques in their paper. First; they used direct recognition using gazetteers to recognize patterns of crime types. The second technique used set of rules based on a predefined set of keywords for performing recognition function. They also demonstrated an initial prototype for determining crime types from crime news; but as it is noted, this prototype depended on a predefining dictionary that was manually built and this may have caused a lack of system where it only looked up on keywords that were previously defined in a dictionary list to classify a crime type.

The same authors in [13] extracted types of crime documents in a crime domain using a rule- based approach, and a cluster of Arabic crime documents based on crime types. The system had an ability to extract keywords based on syntactic standard. However, the main drawback of their paper was that it did not extract networks.

Sharef, et al in [17], tried to solve the problem of how to recognize an Arabic name and use it in a crime domain. The system had three phases. First, the linguistic preprocessing split the text in this system after the process of tokenization; then both splitting and tokenization used Part of speech tagging which was used in [23]. The second phase used in their approach was called Name Entity Identification and Classification. It used pattern rules, a set of predefined gazetteer which included a list of (people's names, organization and location names), and set of

grammatical rules. The final phase of their approach included extraction and classification of Name Entities. A corpus was also used to evaluate Sharif's approach aggregate from newspapers.

Alkaff, et al in [24] was built systematically collect information about the nationality of criminal from crime news. Additional references are used to identify the nationalities of suspects, victims, and witnesses. They evaluate the direct and indirect extraction of nationality from crime news. Their model is based on gazetteers and rule-based extraction, as well as a co-reference resolution to link the references. However, these types of systems play an important role in gathering crime information nowadays. However, our system concern in building crime social network and discover the relation between offenders.

There are lots of work that has been done for named entity recognition such as, Aboaga, et al in [16] used some other techniques to extract an Arabic name recognition. Where the focus was on extracting a person name in Arabic texts using four techniques where, they used predefined keywords for persons names found in different domains such as politics, sports and economy. Their approach consisted of three phases. First, the pre-processing which was a preparation of the dataset using a sentence splitter, tokenization and then normalization of the different forms of an Arabic letter. The second phase was the annotations in which the authors used a dictionary of person's names to identify a person name in a text. The third phase was the application of rules which was used to meet the named entities that were not found in the used dictionary. However, they used four rules to identify a person name. First, they used Introductory Words Person List (IWPL) which uses a name as a defined word. In relation to that, Buckwalter Arabic Morphological Analyzer (BAMA) was used to get the Part of Speech (POS) tagging for the words that appeared after the IPWL. The third rule recognized person's names that occurred before the Introductory Verbs Person List (IVPL). The forth rule was offered here to recognize the persons names that appeared before the introductory person verb list (trigger key-words).

2.2 Criminal social network analysis

Chen, et al in [25] developed a system called COPLINK. The system allowed different police departments to exchange data in an easy way. The system had the ability to make data migration and access to different database types through one user interface. The goal of their system was to develop knowledge management systems and methodology for accessing, analyzing, visualizing, and sharing law enforcement-related information in social and organization contexts. The drawback of researchers' system was that it was concerned about visualization and exploration that built the network with known information. However, the system used structured database.

Yang and Ng in [26] present a method to retrieve the criminal networks from a website that provide a mechanism for blogging service using a specific topic exploration. In addition, they are utilizing the web crawlers to identify the actors in the network where participated in a discussion related to some criminal topics and classify the performers in the network. After the network is constructed, they utilized some text classification techniques to analyze the content of the documents. Finally, they visualize the results to social network view or concept network view. However, our proposed work is different from these works in three aspects. First, our study focuses on unstructured Arabic text investigation documents obtained from police departments. Second, most the work in this paper focus on identifying direct relationships, but our work provides an algorithm to identify indirect relationships between offenders in different communities.

Baumgartner, et al; in [27] employed Bayesian Network modeling utilizing the fact that most offenders had previous criminal accidents . However, their system aided in the suspect prioritization process with positive results. Nevertheless, their approach was still limited because the research used a small sample for individual crime network predictions.

Hosseinkhani et al in [10] provided a framework to analyze the web browsing history sequences and links in scientific methodology to arrive the suspects, thereby used in the investigation process by combining data mining technique with web log analysis to obtain relevant information to help discovery of cybercrime. However,

our work is different from this work where focuses on extracting information for investigation from text files.

Al-Zaidy, et al; in [28] focused on the identification of invisible social group or individuals from textual files using social mining methods. Hence, the main contribution of their approach can be summarized in two points. First, it discovered and identified the eminent communities in a set of documents and extracted useful knowledge from it. Second, the researchers generated hypothesis of indirect relationships between main offenders and other people names in the set of documents. The process started by fetching documents from a suspect's machine and then extracting a profile of data from available data, such as tagging names of persons; then it implemented normalization of names. After that, it extracted prominent communities to demonstrate contact information, city names, and perform text summarization on the extracted text. Later, the process built community profiles used to detect names that had hidden relations in the communities and visualize criminal networks. The drawback of their paper was that it used the Stanford NER which was trained to deal with English newswires and handle only with English documents. In addition, unstructured textual data were obtained from offenders hard drives while the data of the current study were obtained from real investigations documents. Also, their method did not analyze the interaction between indirect documents and criminals. Moreover, the authors focused on extracting profile data from available criminal networks.

D. Prakash, et al; in [15] the contribution of their paper was to try to find hidden relationships in criminal social networks, and discover invisible relations between actors and match nodes that were related to others. The authors of this paper used mining algorithms such as Min-cut and Regression-Based for community mining where it was able to detect an acceptable number of invisible societies. The main drawback of their paper was that it only utilized a real social network dataset to extract hidden relationships and analyze networks. However, the research under study aims to build a criminal network and discover hidden relationships between offenders and between communities.

Iqbal , et al; in [29] in their paper, the authors utilized chat log to build a framework to analyze messages to detect crimes. The framework was able to extract

criminal networks from chat log, extract topic of chat without a prior knowledge, and identify the information about crimes. After that, they made visualization of the knowledge for investigation.

We can conclude that studying the papers it was clear they did not tackle the relationship in extracting criminal names network in Arabic language. In addition, the current thesis focuses on unstructured Arabic language texts and obtained real data sets from police officers. Therefore, the new approach will extract social networks using Arabic texts. In addition, features from Semantic networks will be used to provide more features about crime networks that can help investigators to (1) discover community, (2) determine indirect connections between offenders in different documents and other criminals, (3) provide the capability to discover the key person of a community, finally, (4) measure the importance of nodes by measuring incoming and outgoing links to the node to determine the important person in the network.

2.3 Summary

In this chapter, we presented a review of existing works closely related to our research and identified the advantage and drawbacks of existing approaches; we classified the previous works into two categories: The first category includes approaches used in Arabic crime detection. The second category includes approaches used in criminal social network analysis.

In the next chapter, we present the theoretical foundation underlying our research.

Chapter 3

Theoretical Foundation

Objectives

- Present theoretical information used in this thesis.
 - Provide tools used to extract person names from text.
 - Provide methodology used to create discovery algorithm using hypothesis generation.
 - Provide how data visualized.
 - Provide methods for performance metrics.
-

Chapter 3 Theoretical Foundation

In this chapter, the fundamental concepts which represent the basis for understanding our research are presented. First, Named Entity Recognition is introduced, followed by providing an overview to Name Entity Recognition (NER) and development environment used in this thesis. Then provide an introduction about hypothesis generation used in creating our discovery algorithm. After that, we introduce criminal network analysis and data visualization. Finally, we present an overview of used performance metrics and classification measures.

3.1 Information Extraction

Information Extraction (IE) is used to automatically extract structured information from unstructured or semi-structured readable documents. Information Extraction is a subfield of Natural language processing (NLP) [30].

Natural language processing is the field of Artificial Intelligence (AI) that concerned with interactions between computers and natural human languages. In the past few decades, many of the application were developed to handle by the NLP field such as Information Retrieval (IR), Machine Translations, Text Mining (TM), Question Answering (QA). The main focus of NLP or IE in general is Name Entity Recognition (NER) [31, 32].

3.2 Name Entity Recognition

NER is also known as entity extraction or entity identification. The concept of NER was born in MUC (Message Understanding Conference) in 1990s. It is a very important subtask of information extraction that aims to find and classify the name in unstructured text, the main task of NER was broken down into three subtask included:

- **Name entities (NE) - ENAMEX tag** to identify proper names including locations (cities, countries, rivers, etc.), persons, and organizations (company, government, committees, etc.).
- **Temporal Expression - TIMEX tag** to identify dates and times.

- **Number Expression - NUMEX tag** to identify number and percentages and money in documents [33].

In our work we only need Name Entities.

3.2.1 Learning method of NER

The essential part of NER system is to identify previously unknown persons. This ability depends on whether the detection and classification rules triggered by features with positive and negative examples assigned. While early studies were mostly on craft rules that use the most recent monitoring Machine learning as a way to induce automatic systems or rule-based sequence Labeling algorithms based on a collection of examples of training [34].

There are three primary methods of learning NE: Supervised Learning (SL), Semi-supervised learning (SSL) and unsupervised learning (UL). The main imperfection of SL is the requirement of a large annotated corpus. The lack of such resources and the high cost of creating them lead to two other alternative learning methods [35]. In our work we will concentrate on SL because of the availability of resources.

3.2.1.1 Supervised Learning

The idea of supervised learning is to learn automatically from large collection of annotated documents and then supervised by human [36]. Examples of systems that are based on this approach of SL techniques include Decision Trees [37], Hidden Markov Models (HMM) [38], Support Vector Machines (SVM) [39], Maximum Entropy Systems (ME) [40], and Conditional Random Fields (CRF) [41] [42]. The main SL method, which is often suggested, consists of tagging words of a test corpus, if they are marked as entities in the training data. The efficiency of the system relies on the baseline to be passed into the vocabulary, with the percentage of words that appear without repetition, both in training and test corpus.

3.2.1.2 Semi-supervised Learning

As the name implies, enclose a small degree of supervision for the learning process. This type of learning is still comparatively young and it has still improved and tested with NER tasks. The famous techniques used for this approach is called bootstrapping, that only requires minimal supervision, namely, a set of seeds in order to initiate the learning process [43].

3.2.1.3 Unsupervised Learning

The last dominant technique, for NER learning methods, depends on clustering approach. Basically, the techniques based on existing semantic lexical resources such as WordNet, on lexical patterns and on statistics computed on a large unannotated corpus [44].

3.3 Integrated Development Environments

To generate NE from text, a tool can be used this tool called Integrated Development Environments, common Environments are:

1. **GATE** The General Architecture for Text Engineering

This is one of the most popular tools used to dealing with NLP. It is free and open source tools developed at the University of Sheffield in 1996. GATE is built based on JAVA used by the researcher as infrastructure for developing and deploying software components that process human language such as NER projects, coreference resolution, and others [45]. GATE handles with Multilanguage such as Arabic, English, Chinese, and Hindi, etc. GATE supports many text file formats such as XML, HTML, PDF, RTF, email, and plan text [46]. GATE provides many essential tools such as gazetteers, chunkers, Pos taggers, tokenizers and parser. Also, GATE has features to build rule-based NER system which help the researcher and development to build their own grammatical rules as a finite state transducer using JAPE (a Java Annotation Patterns Engine). Also, GATE has many build in plugins used for specific language such as Arabic plug-in that contains many components such as gazetteers, toknizers,

orthoMatcher and other. Therefore, many researcher used GATE tool such as Elsebai and et al in [47], Shoaib in [48], and other.

2. **LingPipe** is a Java based natural language processing tool kit founded by Alias-I in 2006, it is a free version with limited production for text processing using computational linguistics where need to upgrade to obtain full production. It supports different natural language processing such as POS tagging. NE recognition, spelling correction. NER in LingPipe components based on hidden Markov system interface and the learned system can be evaluated using k-fold cross validation over annotated data set. LingPipe is multi-lingual such as Arabic, English, Chinese [45, 49].

Many researchers used LingPipe tool such as S Abdel Rahman and et al in [50] and others.

3. **NooJ** is a free tool for linguistic development multi-lingual environment. NooJ enables the developer to build, test, and maintain large coverage lexical resources, as well as applied morpho-syntactic tools for Arabic processing. However, there are many researchers used NooJ tools such as W Brini and et al in [51], Mesfa in [52] and others.

However, in this research we used GATE tool to extract named entity from an Arabic unstructured text because it have many facility such as:

1. Easy to use.
2. Have most tools such as gazetteer, tokenizer, etc.
3. Easy to build JAPE rules.
4. Have many plugins used for Arabic language.

3.4 Hypothesis Generation

In this thesis, we have used hypothesis generation term to discover a new relationship between offenders in different communities. A new hypothesis generation required prior knowledge, experience and intuition [53]. Many researchers used data mining techniques or other metric analyses to start generate hypothesis such as Swanson and et al in [54] where he used text mining techniques to propose several hypothesis generation. However, in our thesis, we used hypothesis

generation to discover community from investigation documents and also we used it in building our discovery hidden relationship algorithm as follows.

3.4.1 Community discovery problem

The problem of defining community is to identify groups of offenders from investigation documents that's obtain from police departments. Let D set of investigation text documents. Let $U=\{p_1, .. , p_n\}$ o denoted all offender names in D . each $d \in D$ is represented as set of offender names such that $d \in U$. Each document in D has community $C \subset d$ where C is grouping of person names founded in d if and only if number of persons in $C > 1$

Definition 3.1 (Community discovery)

Let D be a set of Arabic investigation documents where each document $d \in D$ represent a set of person name k where $k \subseteq U$. Community is a set of person names k in document d if $k > 1$ person else community for document will ignore.

3.4.2 Discovery Algorithms

In normal state when the user query about topic A should return terms B that related to it. If the user is interested to study relation between two topics such as A and C ; he should query to finding one or more collection of terms B that intermediate between A and C .

3.4.3 The Problem of Indirect Relationship Discovery

Let D be a set of Arabic investigation documents and let C and E be a prominent community in D . Let U be the set of distinct offender person names in D . The problem of indirect relationships discovery between two communities C and E where p sets of the intermediate chain of individual $p \in U$ called T that identify the relationship between C and E . Intuitively, we need to determine a set of terms that connect two community with others. Using the concept of hypothesis generation, we can present the problem of extracting indirect relationships as follow:

Consider a criminal prominent community C, E and an individual p in D . Let $R(.) \subset D$ indicate the set of documents containing the enclosed argument where the enclosed argument is a community. The problem of detecting hypothetical, conceptual linkages between communities C and E uses intermediate individuals p in

U is creating the tuple (C; E) from D. This tuple is generated for each community in D by identifying connecting terms t that conceptually link C and E. Where terms t represent intermediate individual between C and E and occur at least once in both two communities.

Definition 3.2 (Indirect Relationship)

Let D be a set of documents. Let U be a set of unique names in D. Let C and E be a prominent community and p and k be an individual where $C \subseteq U$, $E \subseteq U$ and $p, k \in U$. Let R(C) and R(E) be two community extracted from two documents in D. An indirect relationship, between C and E is defined by the tuple (C, E) generating by identifying terms $[t_1, \dots, t_n]$ such as:

- 1- $R(C) \cap R(E) = p$
- 2- $(p \in R(C)) \wedge (p \in R(E))$
- 3- $(R(C) \cap R(E) = p) \wedge ((R(E) \cap R(M) = k) \therefore (R(C) \cap R(M)) = p \text{ and } k$

According to this definition, an indirect relationship between two communities C and E using intermediate of individual p if the following conditions apply:

- 1- Community C and E have indirect relation if intersect in intermediate individual p between them and this called first level of indirect relation.
- 2- If community C has relation with community E using individual p, and E community has relation with community M using individual k. The results community C will have hidden relation with M using individual p and k as follows:

$C \rightarrow E$ using individual p and $E \rightarrow M$ using individual K the results will be: $C \rightarrow M$ using individual p and k and this called second level of indirect relation and so on.

3.5 Criminal Network Analysis and Visualization

In this days criminal network analysis has long used in intelligent law enforcement as a task that is related to organized crime. The analysis is performed by collecting data from various incidents and sources connected to the case under investigation. When analyzing such crimes investigators not only explore the

characteristics and behavior of individual offenders but also pay much attention to the organization, structure, and operation of groups and the overall network. [53, 55]

A good to begin the analysis process is to map the criminal activities to a visualized graph that displays the associations between the criminals. In this section, we present the patterns in criminal network analysis methods and tools. We also discuss visualizing criminal networks processing to help police solve crime.

3.5.1 Criminal Network Analysis

Recently, the research on social network has received increasing attention. The main feature of criminal network analysis is to discover the relationship between entities and discover a new relationship. Graph theory looks at object as nodes and the relationships as edges between nodes. These objects may be people, buildings, companies, etc. while the corresponding relationships could be family ties, roads, or competitive relation. However, Criminal network analysis (CNA) needs three points of data- actors A, actors B, and the link or tie between them where actors called node and can be people, organizations, buildings, computers, etc. for the purposes of this thesis, each node in network refer to one community. Where community is a group of individuals name as defined in Chapter 1. The linkage between nodes represented as a set of persons names. Therefore, the main objective of graph theory attempt to understand these networks where used to help detect and describe changes in criminal organizations.

Criminal network analysis, therefore requires the ability to integrate information from multiple crime incidents or even multiple sources and discover regular patterns about the structure, organization, operation, and information flow in criminal networks [56].

3.6 Data Visualization

Data visualization is the process of visually depicting data. The Visualization method usually used to enhance the analysis of data and help in discovering a valuable information. To visualize a social network, an appropriate layout algorithm must be chosen to assign locations to nodes. Traditionally, Bellman–Ford algorithm [57] was used for computes shortest paths from a single source node to all other

nodes in graph, can run on graphs with negative edge weights as long as they do not have any negative weight cycles. Dijkstra algorithm [58] better than Bellman-Ford for sparse graphs, but cannot handle negative edge weights. However, in this thesis we used Dracula Graph Library [59] where was built using JavaScript. JavaScript is light and can be easily customized and integrated in a web page. We used Dracula Graph Library for drawing a graph, for the web interface we used HTML through PHP.

3.7 Performance Metrics

The performance metrics are a measure of system performance. There are several performance metrics such as: efficiency and scalability, and many classification measures like: accuracy, precision, recall, and F-measure. The performance metrics are a measure of system performance. They will be used in later to evaluate the effectiveness of our proposed approach.

3.7.1 Confusion Matrix

The confusion matrix [60] is one of popular tools to evaluate the performance of a system in tasks of classification or prediction. The confusion matrix is represented by a matrix with each row representing the instances in a predicted class, while each column representing in an actual class as shown in Table 3.1

Table 3.1: Simple Confusion Matrix

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- **True Positive (TP):** refers to the number of positive instances that are correctly labeled by the classifier.
- **True Negative (TN):** refers to number of negative instances that are correctly labeled by the classifier.

- **False Positive (FP):** refers to the number of positive instances that are incorrectly labeled by the classifier.
- **False Negative (FN):** refers to number of negative instances that are incorrectly labeled by the classifier.

3.7.2 Accuracy

Refer the percentage of test set instances that are correctly classified by the classifier.

$$\text{Overall Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3.1)$$

3.7.3 Precision

Refer to the number of correctly predicted items as a percentage of the number of items identified for a given topic. The higher the precision, the better the system is at ensuring that what is identified is correct.

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (3.2)$$

3.7.4 Recall

Refer to the number of correctly predicted items as a percentage of the total number of correct items for a given topic.

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (3.3)$$

3.7.5 F-measure

It is a standard statistical measure that is used to measure the performance of system. The F-measure is a conjunction parameter based on precision and recall.

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.4)$$

3.8 Summary

In this chapter, we presented an overview of the basic theoretical foundation related to our research. We present Name Entity Recognition (NER), integrated development environment, Hypothesis Generation (HG), criminal network analysis and visualization. Finally, we described performance metrics and classification measures that are used to evaluate the effectiveness of a crime detection approach.

In the next chapter, we provide a detailed description of the proposed crime detection approach.

Chapter 4

The Proposed Crime Detection Approach

Objectives

- Present detailed information about the proposed architecture phases.
 - Describe the components of each phase.
 - Show how the components of the system's architecture interact.
-

Chapter 4: The Proposed Crime Detection Approach

4.1 Introduction

In this chapter, we present the proposed crime detection system (CDS) framework. Section 4.2 provides an overview of the framework's architecture. Section 4.3 describes the corpus collection and preprocessing stage, which is comprised of three components: data gathering, tokenization and normalization. Section 4.4 presents the steps of extracting offender names from investigation documents, which is dedicated to building crime communities. Section 4.5 provides a new algorithm used to discover hidden relationships between communities. Finally, Section 4.6 presents the visualization stage.

4.2 The Overall Crime Detection System (CDS) Architecture

An overview of the framework's stages, as depicted in

Figure 4.1. It is divided into four types of activities:

1. Initial Preprocessing Stage

This stage contains four components: data gathering, data preprocessing, tokenization, normalization.

2. Extract offender names Stage

Extraction of offender names plays a vital role in building our system, at this stage we create many rules based grammar using GATE tool and made updates to gazetteer lists by adding a new person names in it to extract person names from documents by utilizing the proposed corpus that were presented in the first activity. After that, we build communities of crimes.

3. Extract hidden relationships

In this phase, we build an algorithm to discover hidden relations between offenders in different communities.

4. Visualize the results

In order to assist in analyzing the discover relation, a visualization process is employed.

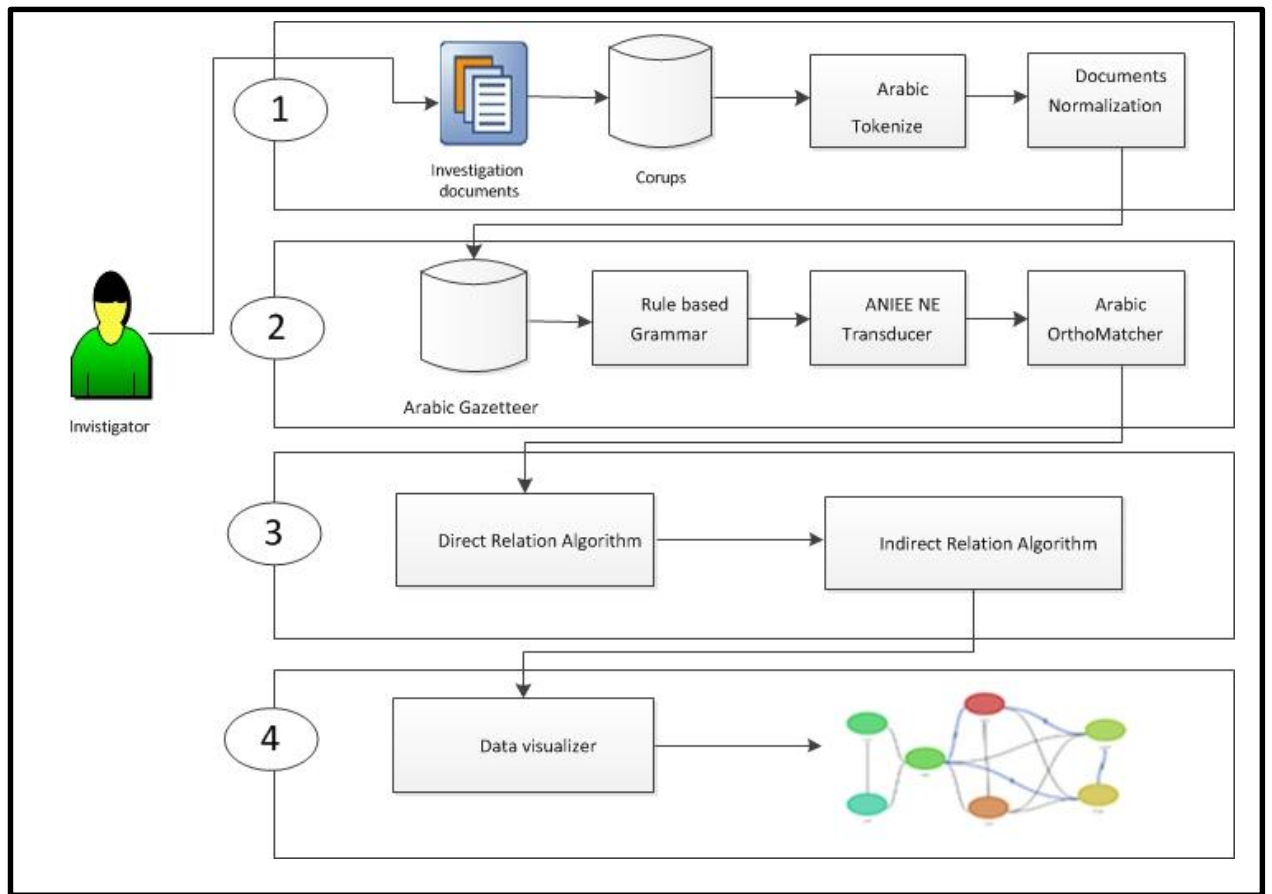


Figure 4.1: System Architecture

Detailed explanations of each stage are provided in the following sections.

4.3 Initial Preparation Stage Architecture

This stage contains four main components: data gathering, data preprocessing, tokenization, normalization. Each component described as follows:

4.3.1 Data Gathering

One of the difficulties that encountered this work in the field of getting real Arabic investigating documents where it is unavailable to public. The first step is documents gathering, conducted in order to build a corpus. A corpus used to collect documents in one place and allow run analysis in all documents at the same time. However, we got our corpus from police departments in the Gaza strip about theft incidents, where the investigation documents limited to the period between the beginning of 2008 and the end of 2013 and the total number of documents were 777

cases. It is well known that text mining research relies heavily on the availability of a suitable corpus.

4.3.2 Data preprocessing

The corpus is collected from real investigation documents used by text mining techniques for performing various tasks, such as text preprocessing are applied to remove non-Arabic text.

4.3.3 Tokenization

An important step in the processing of textual documents, which takes place before an information extraction is tokenization. Here the words in the documents are separated out into individual words that are identified by the blank spaces or special character between them. This step performed using GATE tool.

4.3.4 Normalization

Usually this process used before Applying the text mining techniques in order to avoid or reduce data scatters in the data being processed. In Arabic, it is possible to write Ahmed in two different ways "احمد" *Alef* without *Hamza* above or "أحمد" *Alef* with *Hamza* above. Therefore, to make the data more consistent, this process is applied. Normalization process includes the following steps:

- Replace some characters:
 - As vowels like *Alef* with *Hamza* (أ), (إ) with *Hamza* below or *madda* (آ) becomes simply *Alef* (ا).
 - The second character is "ة" which may write at the end of words as "ة" or "ة"; this will be normalized to "ة"
 - The third character is "ى" which may write at the end of words as "ي" or "ى"; this will be normalized to "ى"
- Delete special characters: as "(" and ")"

However, this process makes the corpus more consistent.

4.4 Extract offender names Stage

It plays a main role in our research, where one of our main goal is to discover hidden relationships between offenders in different documents. However, the first step in our research is to identify offender names from unstructured crime investigation documents. There are many tools and methods in market to extract named entity recognition from text such as Stanford NER [61] but most of them used to handle English language. However, we have applied this stage using GATE tool to identify proper names we used two methods: firstly, used predefined list or Gazetteers and add a new names to it. Secondly, we adopt many rule based approach to develop our system.

4.4.1 Gazetteers

The gazetteer lists are plain text files with one entry per line, each list represents a set of names such as names of cities, organizations, locations etc. This type of gazetteer is built manually. Therefore, for extracting proper name we implement the following preprocessing steps to increase the quality of the result:

- Modify Gazetteer list

We collect the results of High secondary school (Tawjehi) results in Palestine from period between 2012 and 2014, where the researcher separate the person names to three categories:

- Male names
- Female names
- Surname or family names

- Gazetteer Normalization

Before modifying Gazetteer lists in GATE by adding a new person names, we apply normalization steps explained in previous subsection 4.3.4.

After that, we remove the duplicates of names in each categories and modify GATE Gazetteer lists as shown in Table 4.1.

Table 4.1: The results of modifying Gazetteer lists

#	List	# of records before	# of records after	New records
1	male_names.lst	2,780	4,431	1,651
2	female_names.lst	708	2,220	1,512
3	surnames.lst	198	8,239	8,041

4.4.2 Rule-based approach

The rule-based approach applies a set of grammar rules are implemented as regular expressions to relies on linguistic knowledge in order to extract pattern base for location name, person name, organization, etc. these rules mostly depend on large lists of lookup gazetteers [62]. In our research, we focus on only extracting offender person names. Implementing rule-base algorithms in GATE Developers requires expert knowledge in Java Annotation Patterns Engine (JAPE).

4.4.2.1 Java Annotation Patterns Engine (JAPE)

JAPE grammar provides pattern matching in GATE, where each JAPE rules consists of a set of phases, each of which consists of a set of pattern/action rules. However, each JAPE rule consists of the left hand side (LHS) contain pattern to match and right hand side (RHS) which contains annotation or features to be created [63].

4.4.2.2 Rules for Offender Person Names Extractor

We implement many JAPE rule-based algorithm using GATE tool to improve nominating the correct names from unstructured text. So we divided the rule base objective into two sections. Firstly, we built many rule-based algorithms to choose the proper offender name from Arabic investigation documents such as follows:

- We built a rule used to annotate each offender name in investigation documents which is preceded by the Arabic word "المدعو" which means "The named".
- Another rule used to annotate that each offender name is preceded by the word "المتهم" which means the "Accused".

- Another rule used to annotate each name by the position of the person in Arabic such as: "م. د. ، أ. ، الخ" which is : “Eng., Dr, Mr., etc.”
- Another rule used to choose strange of offender name where assume that each name in investigation document preceded by nickname such as "ابو" which means “The father of” or "ام" which means “The mother of” will be considered as a name.

Secondly, we built many rules based to remove all annotations about non-offender names such as follows:

- Rule based that used to remove annotation about all personal names which preceded by the word "المواطن" which refer to the citizen name who provided the complaint and not the offender name.
- Another rule built to remove annotation about names of facilities or buildings such as a name of a mosque which refers to a real person name such as "مسجد علي بن أبي طالب" which means “The mosque of Ali bin Abi Talib”, where Ali is a name of a person that refers to the name of the mosque.
- Another rule that used to remove annotation about name of a location or a residential neighborhoods that carry names of persons such as: "شارع احمد عبد العزيز" which means “Ahmed Abdel Aziz St.”

An example explaining applying the JAPE rule: First we create in List 4.1, which aims at removing the annotation of the person who provided the complaint such as: "تقدم المواطن/ خليل حسين", Khalil Hussein is the complainant (non-offender name).

Rule: nonOffenderNameRule

Priority:10

```
(
  ( {Token.string =~ "مواطن"} ({Token.string == "/"}) )
  ({Lookup.majorType == person})+ :nonOffender
):ignore
-->
{
  GATE.AnnotationSet lookup = (GATE.AnnotationSet) bindings.get("nonOffender");
  GATE.Annotation ann = (GATE.Annotation) lookup.iterator().next();
  Long startOffset = ann.getStartNode().getOffset();
  Long endOffset = ann.getEndNode().getOffset();
  GATE.AnnotationSet toGo =
  GATE.AnnotationSet)inputAS.get("Lookup",startOffset,endOffset);
  inputAS.removeAll(toGo);
}
```

List 4.1: Rule 1 remove annotation from non-offender person name

Another rule used to remove annotation from name of residential neighborhoods such as "منطقة الشيخ رضوان" which means "Elshekh redwan area", as described in List 4.2.

```
Rule: nonPersonRule2
Priority:20
(
//سكان الشيخ رضوان
(
{Lookup.majorType != person}
{Token.string =~ "شيخ"}
{Token.string =~ "رضوان"}
)
/
(
{Lookup.majorType == Per_desc}
{Lookup.minorType == surname }
)
/
(
{Lookup.majorType != person}
{Token.string == "بو"}
{Token.string =~ "سكندر"}
)
):ignore
-->
{
GATE.AnnotationSet lookup = (GATE.AnnotationSet) bindings.get("ignore");
GATE.Annotation ann = (GATE.Annotation) lookup.iterator().next();
Long startOffset = ann.getStartNode().getOffset();
Long endOffset = ann.getEndNode().getOffset();
GATE.AnnotationSet toGo =
GATE.AnnotationSet)inputAS.get("Lookup",startOffset,endOffset);

inputAS.removeAll(toGo);
}
```

List 4.2: Rule 2 remove annotation from non-offender person name

The complete implementation of Person Names Extractor is listed in Appendix A.

4.4.3 Criminal Communities Discovery

After identify offender names, the next step is to identify all prominent criminal communities. Criminal communities' discovery is a major component in the system. We assume that each offender name in the same investigation document will be in the same community as defined in Section 3.4.1. Furthermore, variants of the same offender name in the same community are represented as one name. For instance,

"محمد خالد حسين", "محمد حسين", "محمد" are transform to the common form: "محمد خالد حسين".

The community contain a group of offenders who interact frequently with each other in the same investigation text document. Therefore, each individual in the same community have a strong linkage and direct relation with others. Moreover, it generates hypothesis for potential indirect relationships between individuals across the data set.

In some real life criminal cases, only one offender founded in investigation text document. Therefore, to get the best results, we ignore all communities have only one individual name. However, we use JAPE GATE tool to extract community as show in Appendix A, List 8.

4.4.3.1 Extract Key person of Community

Key person is the individual with highest repeated his name in the same community of criminals. This may leads to new clues for further investigation. In Appendix A, List 8 used to extract key person of community by counting the number of repeating names in the same community using name matching techniques.

4.5 Indirect Relationship Extraction

Our goal in the previous section is to identify all prominent criminal communities in each Arabic investigation documents for evidence extraction. In this section, we examine the communities' contents to further analyze the offenders' criminal social networks. Hidden and indirect relationships discovery play a vital role of analyzing criminal communities relationships.

In this section, we propose a new algorithm to discover the hidden relationship between community identified in pervious section and other offenders who are not in the community. The linkage extracted as chain of intermediate names that link a community as define in section 3.4. Thus, let C be a set of prominent communities. Let $U = \{p_1, \dots, p_n\}$ where U denote all distinct offender names in C . Each community $c \in C$ is represent a set of offender names such that $c \subseteq U$, the indirect relationship discovery algorithm identifies unlimited of intermediate offender names between two communities. For instance, $R(c_i) \cap R(c_j) = p$ indicates community c_i has indirect

relationship with c_j community through the individual p , which p is one or more intermediate names between communities. The indicate relation may be considered valuable from an investigator's perspective since it indicates the relationship between community and offenders in these community.

The proposed algorithm is to find unlimited of intermediate terms between two communities using recursive function. The algorithm applied for each community c_i that previously defined as inputs, where each community c_i has an Id of a community and a list of offender names. In addition, the algorithm needs a list of all distinctive offender names as a U to be considered as an input. The following explains how the algorithm works:

- The first step is to find all matching names found in U list. The algorithm for this step uses OrthoMatcher plugin in Gate tool.
- The next step is to find all repeated offender names using the results from the first step. After that, the repeated names list will be used to find the intermediate of offender names between communities and this will be the first level of indirect relationship.
- Consequently, the algorithm applies the recursive function to find all intermediate offender names between communities. Where each new discovered indirect relationship between communities increases the level variable plus one. Where the level refers to the depth of the relationship between communities.

The Algorithm 4.1 shows the full steps of the indirect relationship discovery.

Algorithm 4.1: Indirect relation discovery algorithm**Input:**

- *List of person names in d where $d \in D$*
e.g. : array([10]=>offender name₁, offender name₂, offender name₃
[15]=>offender name₂,offender name₅
... Etc.)

Where array index represent community id and array values represent person name for each community.

Output: indirect relation between communities and persons

```

1. Find all name matching in List of person names.
2. Duplicate_Name[] = all repeated person in Step 1 where count > 1
3. Foreach(Duplicate_Name as name ) loop
4.     Check if name exists in List of person
5.     If True then
6.         tmp [] = doc_key
7.     End if
8. If count(tmp) > 1 then
9.     Result[name] = tmp
10. End if
11. Clear tmp list
12. End loop
13. If count(Result) > 0 then
14.     namesList = all names in Results
15.     Hidden = call hiddenRelation(namesList,Result,2)
16. End if
17. Function hiddenRelation(namesList,Result,level){
18.     Duplicate = get all duplicate name in namesList
19.     If duplicate not have value then
20.         If level > 2 then
21.             Return result
22.         Else
23.             Return null
24.         End if
25.     Else
26.         Loop foreach item in duplicate
27.             Loop foreach list in result
28.                 If item exist in list then
29.                     Indx += key of list
30.                     newList = Remove item from list
31.                     Tmp[] = newList
32.                 End if
33.             End loop
34.             If count(tmp) > 1 then
35.                 Tmp = unique tmp
36.                 Result[indx] = tmp
37.             End if
38.         End loop
39.         namesList[] = all item in result
40.         Return hiddenRelation(namesList,Result,level+1)
41.     End if
42. }
```

4.6 Data Visualization

In this thesis we used Dracula Graph Library [53] where was building using JavaScript. We used Dracula Graph Library for drawing a graph, for the web interface, we used HTML through PHP. In this graph, each community of offenders map to node and the relationships present as edges, each edge contains offender names in it. In this phase, the end user provide with network graph and data table as detailed view.

In a data table detailed view, users can view criminal communities of offenders and show the relationship between two communities and the level of the relationship. The investigator can do a quick search of all fields in the data table and specify offender name or community to view the potential hypothetical links.

4.7 Summary

In this chapter, we presented the proposed crime detection approach. We give an overview about the system and then present the stages of the system beginning from initial preparation stage architecture, then we presented the state of extract offender names from Arabic unstructured text using GATE tool, after that we take about an indirect relationship algorithm. Finally, we presented data visualization library and tools that used to represent the results to the end user.

In the next chapter, we present and discuss the experiments carried out to realize and evaluate the proposed approach.

Chapter 5

Experiments and Results

Objectives

- Explore Crime Detection System (CDS).
 - Extract offender names from unstructured text.
 - Discover hidden relationship between communities and individuals.
 - Visualize the results.
-

Chapter 5 Experiments and Results

In this chapter, we present and analyze the experimental results to provide evidence that our approach can identify offenders' names from Arabic investigation documents. In addition, it has the capability of community identification. In addition, we evaluate the performance of the proposed algorithm in discovering hidden relation between communities and individual. Finally, we visualize the results in a graphical representation to provide views of final user to show the results of proposed methods.

5.1 Experiments Setup

In this section, a description about the experimental environment, tools used in experiments, measures of performance evaluation of named entity recognition (NER) to extract offender names and the indirect relationship discovery algorithm has been provided.

5.1.1 Experimental Environment and Tools

We applied experiments on a machine with properties that is as shown in Table 5.1

Table 5.1: Machine environment properties

System Model	Dell
Processor	2.30 GHz Intel Core i5-2410M
Memory Modules	4 GB
Hard disk (HD)	500 GB
Operating System	Windows 7

To carry out our proposed system that presented in Chapter 4 (including the experimentation), special tools and programs was used:

1. **GATE tool:** used to manipulate natural language processing techniques in our approach, and conduct experiments practical and extracting the offender names from Arabic unstructured investigation text.
2. **PHP and JavaScript:** used to apply indirect relationship discovery and visualize the results.

5.2 Arabic Investigation documents Corpus

We used 777 real investigation documents about theft crimes from the police department in the Gaza strip for the period between 2008 and 2013 as a source of the corpus. The average size of a document by words is about 300 words/document.

The dataset is divided into two sets: The first dataset contains 50 investigation documents used as a training phase in order to build rule-based approaches and modifies the Gazetteer lists.

The second dataset contains 727 investigation documents which used by our system to test extracting offender names from a text and creating communities. In addition to discover the hidden relationships.

We perform all text-processing techniques on the corpus; including tokenizing string into words and normalizing process to initialize the text as in Section 5.3 which provides more details about data preprocessing stage. However, to implement this phase, we use GATE tool developer.

We create three corpora in GATE as shown in Figure 5.1, where the first crime corpus used to upload the Arabic investigation documents which used to extract offender names and build communities for each document. The second corpus which named Real Results used by human experts to identify offender names manually from documents selected randomly from the first corpus, consequently these results will be used to calculate Precision, Recall and F-measure as described in Section 3.7 via using Annotation diff tool. Finally, the third corpus is used in the hidden relationship discovery algorithm to identify matching names between communities as described in Section 4.5.

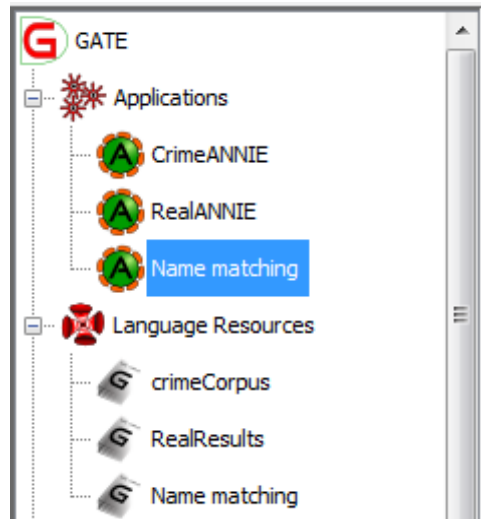


Figure 5.1: Crime Detection corpus

5.3 Data Preprocessing Stage

GATE Developer tools has collection of operation that are suitable for text mining. In this phase, crime corpus that is identified in pervious section 5.2 are prepared to make them a standardized format for the text mining process. There are many of preprocessing techniques such as cleaning, document normalization, tokenization and others. For more details about in Figure 5.2 show preprocessing methods used in our system using GATE tool.

Selected Processing resources		
!	Name	Type
	Document Reset PR_00012	Document Reset PR
	Arabic Crime Doc normalizer	Document normalizer
	Arabic Crime Doc Tokeniser	Arabic Tokeniser
	ANNIE Sentence Splitter	ANNIE Sentence Splitter

Figure 5.2: Preprocessing techniques

The document reset resource enables the document to be reset to its original state, by removing all the annotation sets and their contents, apart from the one containing the document format analysis.

5.4 Name Entity Recognition

Most research in NLP use GATE to create their own programs and pipelines. GATE comes with pre-load plugins handle many fields and Multilanguage. In this

phase we use an ANNIE application (A Nearly-New Information Extraction system) to tag previous crime corpus with names entities.

5.4.1 A Nearly-New Information Extraction system (ANNIE)

ANNIE is a ready-made information extraction system for English by default, is provided as part of GATE tool. Application ANNIE is made up a chain of Processing Resources. However, ANNIE consists many component used finite state techniques to implement various tasks from tokenization to semantic tagging or verb phrase chunking [64].

In this research we create a new ANNIE to handle our Arabic crime corpus describe in Section 5.2. ANNIE components from a pipeline as shown in Figure 5.3 as part of ANNIE used to extract named entity recognition. In addition there are several processing resources such as Gazetteer are not part of ANNIE itself but it come with GATE installation [47].

Gazetteer Is a list build Name Entity Recognition (NER) describe in Section 4.4 are add as ANNIE Gazetteer. It used to identify proper name within documents. Also,

Arabic Main Grammar Used Java Annotation Patterns Engine (JAPE) to implement regular expression base on rules, we identify many JAPE rules to satisfy high accuracy in offender names extract for more details about research JAPE rule list describe in Section 4.4. The complete implementation of Offender Names Extractor is listed founded in Appendix A.1






	GazetteerTraning	Arabic Inferred Gazetteer
	Arabic Gazetteer	Arabic Gazetteer
	Crime Arabic Main Grammar	Arabic Main Grammar
	ANNIE NE Transducer	ANNIE NE Transducer
	Arabic OrthoMatcher	Arabic OrthoMatcher

Figure 5.3: Name Entity Extraction

ANNIE NE transducer

Gazetteer used to find terms that suggest entities, but usually entity is ambiguous e.g. “May 2015” vs “May I can help you?”. Therefore, NE transducer used to classify the terms that the lookup is identify, these classify can help disambiguate. The NE transducer use one or more grammars written in the JAPE language [65].

Arabic OrthoMatcher

OrthoMatcher module used to solve the problem of coreference and name matching in text. The orthoMatcher module detects orthographic coreference between named entities in the text e.g. "خالد حسين" and "خالد" usually refer to the same person name in the same. OrthoMatcher can used to improve the name classification process by classifying unknown proper name [66]. In this thesis, we used Arabic OrthoMatcher to identify a key of person in investigation text by extracting the large number of name matching. Figure 5.4 and Figure 5.5 show the sample results of using Arabic OrthoMatcher in GATE tool.

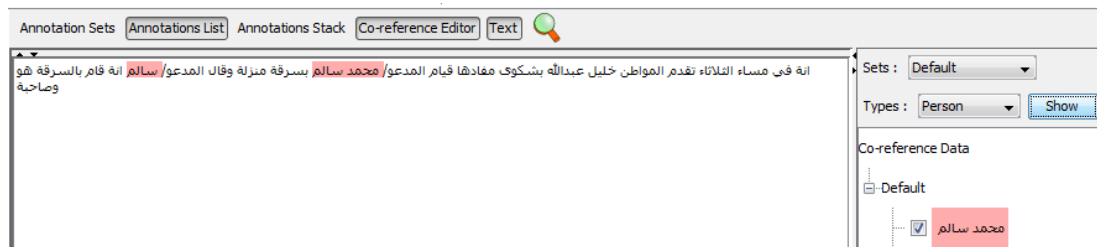


Figure 5.4: Sample OrthoMatcher Results in Co-reference Editor

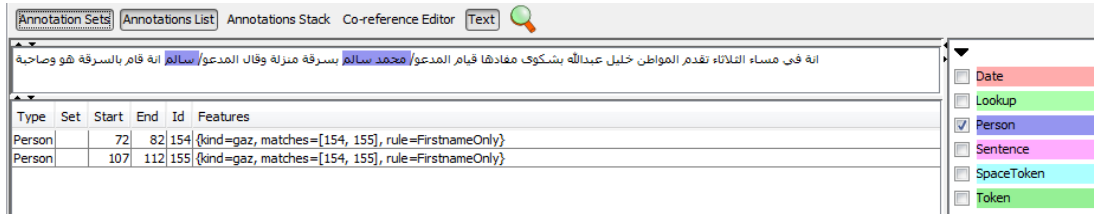


Figure 5.5: Sample OrthoMatcher Results in Annotation List

After selecting PR, for pipeline, the application running and the result display as annotations.

Annotations

One of the main features in GATE is to represent information about the text, and allowing user to obtain various information about the texts being processed.

However, different processing module such as tokeniser and NE transducer running over text, represent as show in Figure 5.6 using annotations features.

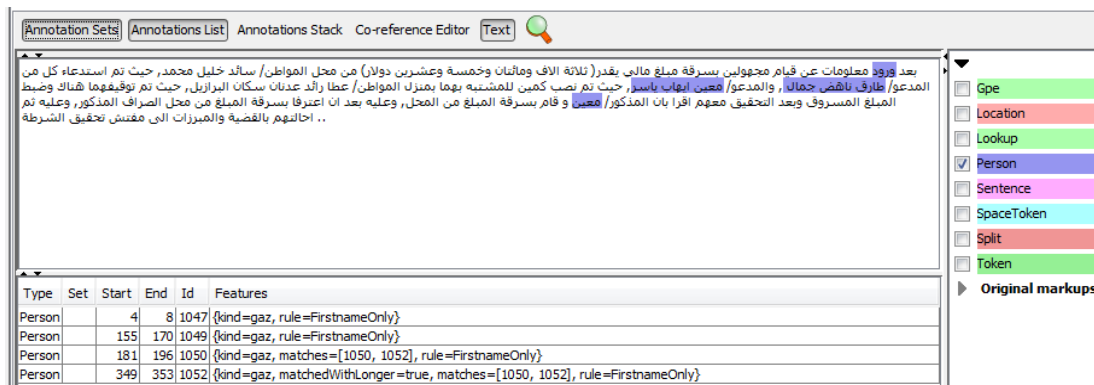


Figure 5.6: Sample of offender name extraction

After apply the ANNIE process in crime corpus, we are building a JAPE rule to extract offender person name out of GATE tool. This rule response on building community for each Arabic investigation document in corpus and determine the count of appearing of offender person name in text to determine the key of person.

In this JAPE rule we put some constrains in exporting names out of GATE tool to satisfy our goal in discovery hidden relationship and to get the best knowledge as follow:

- We are ignored all names in document that contains only one name such as "ورود" as in Figure 5.6.

- We are ignored all communities have only one offender name on it

5.5 Indirect relationship discovery algorithm

The next step in this experiment is to identify indirect relationships between different communities and discover hidden relation between individual. The algorithm 4.1 can identified unlimited levels of relationship. However, if no link found in one or more community represents as a direct relationship between offenders in the same community. Shows an example for using an indirect relationship algorithm. The algorithm was implemented using PHP scripting language. The algorithm consists of two phases. In the first phase, we discover all intermediate offender name between all communities by finding name matching between different communities. In the second phase, we used a recursive function to find all possible relationships between different communities.

5.6 Data visualizer

In this phase, we utilized to visualize technique to assist in the crime data analysis and to be understood better. In this stage, we present the results to graph, visualization is considered useful for allowing the results of the relationship between communities be more readable. This framework is designed to allow the end users to visualize summary results for all relationships between communities and individual. We build our visualizer using Dracula Graph Library [53], described in section 3.6. Where each community represents as a node in crime network and link between two communities using offender names as intermediate between them as shown in Figure 5.7, this graph built using JavaScript with PHP. The user can redesign the different views in crime visualization by selecting the node in the graph and move it to make the visualization more readable.



Figure 5.7: Sample network visualization

Another way to represent the results we preview the results in data table this facility allow the end used to search in crime results and concerns on a specific offender to discover more knowledge about his hidden relationship with others. Figure 5.8 depicts the sample of using data table. In addition, we present the key for each community in the social crime network as shown in Figure 5.9 .

details of relation

Show entries :Search

level	intermediate	Second document offender names	first document offender names	Relation Type	docId#2	docId#1
1	محسن جهاد منصور	يوسف صلاح تيسير	سليم سلمان محمود	غير مباشرة	702	554

Showing 1 to 1 of 1 entries Next Previous

Figure 5.8: Sample Data table presentation

المشتبه الأكثر ظهوراً في التحقيق

Key person of Community

Show entries :Search

#repeted	offender key	document offender names	#docId
2	عارف عبد الرؤوف محمد	عارف عبد الرؤوف محمد سعيد رمزي طي سعيد راند عبد المنيع	284
2	مصطفى محمد مندوح	هادي محمد حسين جمال عبدالله مصطفى محمد مندوح سائد انور	532
1	لا أحد	سليم سلمان محمود محسن جهاد منصور	554

Showing 1 to 3 of 3 entries Next Previous

Figure 5.9: Key person of Community

5.7 System Efficiency Evaluation

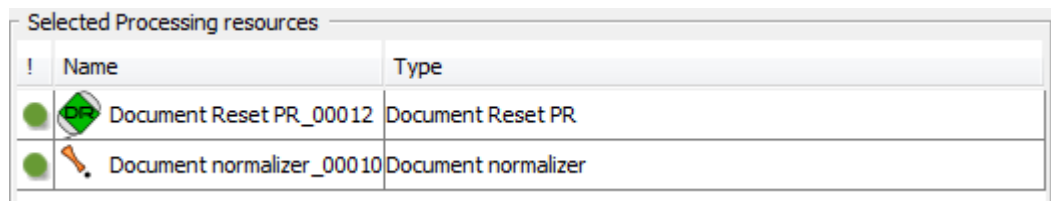
System evaluation is a hard task as mentioned in section 1.6 because it is sometimes difficult to find names for a given document or set of documents and usually depends on human experts.

To ensure that the system works well with tested Arabic investigation documents, we used human expert as reference to extract offender names from text and build crime social networks. After that, we get the results and calculate precision (Eq. 3.2), recall (Eq. 3.3), and F-measure (Eq. 3.4).

5.7.1 Name Entity Recognition and Human Evaluation

To evaluate our name extraction methodology in Crime Detection System (CDS) we used human references to extract offender names from 100 Arabic investigation documents which were selected randomly from the main dataset that contained 727 investigation documents, and uploaded them in the Real Results corpus as described in Section 5.2. For each document we computed the three measurements precision (P), recall (R), and F-measure: Table 5.2 and Table 5.3 show an example of this evaluation in each:

- We replaced the real names in the documents to virtual names to keep privacy of offenders and personal information.
- We applied initial processing stage described in Section 4.3 on the Real Results corpus to normalize the text documents before human expert evaluation as shown in Figure 5.10.
- The extracted names from the 100 documents have been experimented by the system and human experts by computing the P, R and F-measure.
- We used the Annotation Diff tool that found in GATE tool to calculate precision (P), recall (R), and F-measure as shown in Figure 5.11







Selected Processing resources		
!	Name	Type
	 Document Reset PR_00012	Document Reset PR
	 Document normalizer_00010	Document normalizer

Figure 5.10: Preprocessing Documents before evaluate by Human expert

Table 5.2 shows an evaluation of Arabic investigation document by comparing with human results and compute R, P, F-measure manually.

Table 5.2: Summary of evaluation system for extract offender names from Arabic investigation documents case 1

<p>تم انجاز كتاب البحث والتحري الوارد من مفتش تحقيق الشرطة بخصوص البحث عن ملابس شكوى <u>الموطن سعيد عادل حسين</u> 28 عام سكان الشيخ رضوان والتي مفادها بانه بتاريخ 2013/8/9م قيام مجهول بالسطو على منزله وسرقة عصافير وتم سرقتهم من داخل بلكونة المنزل في الطابق الاول والمسروقات عبارة عن 16 عصفور من نوع كنانير بالوان مختلفة حيث بعد البحث تم الاشتباه بالمدعو/محمود علي محمد سعيد 24 عام سكان الشيخ رضوان بالقرب من البركة وباستدعائه والتحقيق معه اعترف بالواقعه وافاد بانه دخل منزل المشتكي عن طريق البلكونة وقام بسرقة 8 ثمانية عصافير وعند هروبه من منزل المشتكي فقد منهم ثلاثة عصافير وبقي خمسة فقام ببيعهم للمدعو/اشرف رياض حسن فخري 22 عام سكان الشيخ رضوان بالقرب من البركة بمبلغ 200 شيكل وباستدعاء الاخير والتحقيق معه افادته افاد بشراء 5 خمسة عصافير من المدعو/سعيد بمبلغ 200ش وهو يعلم انهم مسروقات وايضاً يعلم بتفاصيل حادثة السرقة وكما تم ضبط 4 اربعة عصافير والعصفور الخامس قد مات وعليه تم احالة القضية بالكامل طرف مفتش تحقيق الشرطة</p>		
System Results		Human Results
1. repeated name : 1 name	سعيد عادل حسين	1. محمود علي محمد سعيد
2. repeated name : 2 name	محمد علي محمد سعيد	2. اشرف رياض حسن فخري
3. repeated name : 1 name	اشرف رياض حسن فخري	
Recall (R)	Precision (P)	F-measure (F)
$2/2 = 1$	$3/4 = 0.75$	$[(1+1)*(2/2) *(3/4)]/[(2/2)+(3/4)] = 0.86$

Figure 5.11 shows the measures of (R, P, F) by using the Annotation diff tool founded in GATE tool for case 1 shown in Table 5.2.

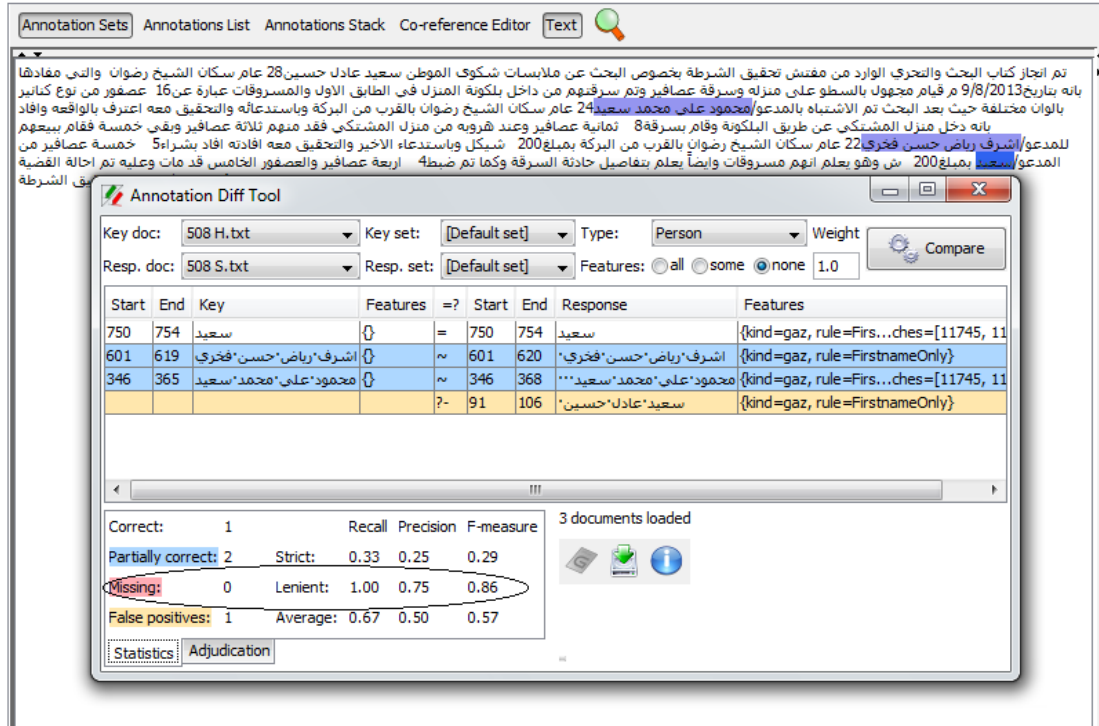


Figure 5.11: Measure (R, P, F) using Annotation Diff Tool in GATE tool Case 1

Table 5.3 shows another sample of evaluation for an Arabic investigation document by comparing with human results and compute R, P, F-measure manually.

Table 5.3: Summary of evaluation system for extract offender names from Arabic investigation documents case 2

<p>3. بناءً علي شكوي المواطن/ سائد عودة جميل التي مفادها السطو علي المخازن الخاصة به وسرقة 20 كيلو بهارات و300 كيلو فلفل اسود، وبعد البحث والتحري تم استدعاء المشتبه بهم: المدعو/ جابر جواد ، والمدعو/ يحيي غازي عبدالله ، والمدعو/ مصطفى محمد بسام، والمدعو/ ابراهيم سليمان سليم، والمدعو/ هاني ابوخليل وبعد التحقيق الشفوي معهم تم اقر كل من: المدعو/ يحيي عبدالله، والمدعو/ مصطفى بسام بالواقعة المنسوبة اليها، وعليه تم احالتهما الي مفتش التحقيق لاستكمال الاجراءات القانونية بحقهما، وتم تسليم المدعو/ هاني الي مباحث البلاد (القرارة) لوجود قضية سرقة عليه.</p>	
System Results	Human Results
<ul style="list-style-type: none"> repeated name : 2 name : مصطفى محمد بسام repeated name : 2 name : هاني ابوخليل repeated name : 1 name : فلفل اسود repeated name : 1 name : جابر جواد 	<ul style="list-style-type: none"> جابر جواد يحيي غازي عبدالله مصطفى محمد بسام ابراهيم سليمان سليم هاني ابوخليل

<ul style="list-style-type: none"> repeated name : 2 name يحيى : غازي عبدالله repeated name : 1 name ابراهيم : سليمان سليم 		
Recall (R)	Precision (P)	F-measure (F)
8/8=1	8/9 = 0.89	$[(1+1) * (8/8) * (8/9)] / [(8/8) + (8/9)] = 0.94$

Figure 5.12 shows the measures of (R, P, F) by using the Annotation diff tool founded in GATE tool for case 2 shown in Table 5.3.

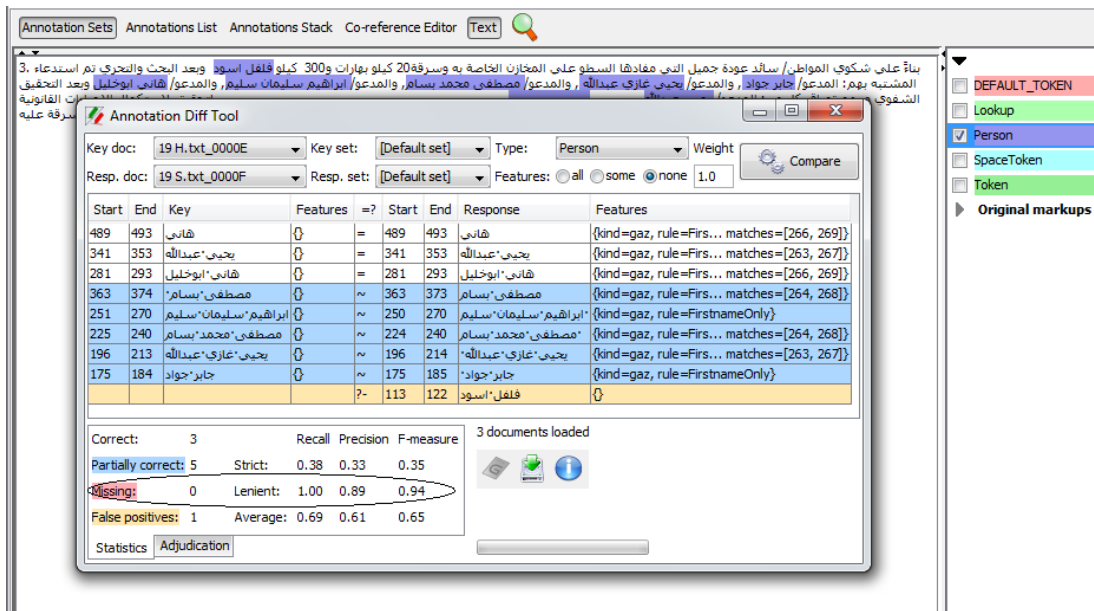


Figure 5.12: Measure (R, P, F) using Annotation Diff Tool in GATE tool Case 2

Table 5.4 show the conclusion of results for all documents has been computed. The complete results show in Appendix A. The average of F-measure for all the chosen cases is considered the system's performance in ability to extract a proper names.

Table 5.4: Conclusion results of F- measure calculation

Recall (R)	Precision (P)	F-measure (F)
0.97	0.84	0.89

The results show that, the average F-measure 89%, we note in the general F-measure of most cases is better than precision where precision refers to the number of correctly predicted items as a percentage of the number of items identified for a given topic. For instance, recall result is 97%, while the F-measure is 89% where recall refer to the number of correctly predicted items as a percentage of the total number of correct items for a given topic. Because, usually the system have the ability to extract correct person name from documents because most of offenders names in Gazetteer and strange name that not founded on Gazetteer is annotated using Rule-based technique that review in section 4.4

5.7.2 Indirect Relationship Discovery Algorithm

In order to evaluate the effectiveness of our indirect relationship discovery algorithm used in Crime Detection System (CDS) we used human experts to extract hidden relations between offenders and communities. We chose 45 Arabic investigation documents from our crime corpus defined in Section 5.2, and divided the documents that chosen to 15 groups where each group has an indirect relation as system-resulted, then we submitted the groups to the human expert to extract offender names from the text and to create a community for each document, then to discover the relationship between communities and offenders and to draw a crime social network. After that, we applied our methodology to the same groups and compared the results for each group with human results to compute the three measurements of precision (P), recall (R), and F-measure for offender names extraction as done in the previous section and for all crime social network draw from the expert and from the system. Finally, we compute the average results of (R, P, and F) for each case and compute the average for all cases. Table 5.5 and Table 5.9 show examples of this evaluation in each one:

- We replaced the real names in the documents to virtual names to keep privacy of offenders and personal information.
- We applied initial processing stage described in Section 4.3 on the Real Results corpus to normalize the text documents before human expert evaluation as shown in Figure 5.10.

- We computed the P, R and F-measure for each group of documents that solved by the system and the expert.
- We used the Annotation Diff tool that found in GATE tool to calculate precision (P), recall (R), and F-measure for each name extracted as shown in Figure 5.11, and calculate P, R, F-measure manually for each crime social network.

Table 5.5 show the case 1 of an evaluation of Arabic investigation document by comparing with human results and compute R, P, F-measure manually to discover hidden relationships between communities and individual.

Table 5.5: Case (1) for discovery indirect relationship using human expert

DocId	Investigation text
554	تم الرد على تكليف النيابة العامة الوارد في محضر الاستدلال 2013/06/17 والخاص بشكوى المواطن/ شادي رأفت جابر سكان دير البلح، حيث تقدم بشكوى مفادها قيام مجهول بسرقة دراجته الهوائية من داخل منزله، وبالبحث والتحري تم ضبط الدراجة مع المدعو/ سليم سلمان محمود 22 عام سكان دير البلح، حيث تم ضبط الدراجة معه، وتم احضاره والدراجة طرفنا وبالتحقيق معه أفاد أنه اشترى الدراجة من المدعو/ محسن جهاد منصور 20 عام سكان دير البلح، وباحضار الاخير والتحقيق معه أفاد أنه اشترى الدراجة من شخص لا يعرفه، ومن طرفنا تم إحالة المذكورين والمضبوطات لمفتش تحقيق الشرطة لاتخاذ الاجراءات القانونية.
702	تم الرد على التكليف الوارد لنا من مفتش تحقيق شرطة ديرالبلح والخاص بشكوى المواطن/ زياد خالد عبدالمجيد سكان ديرالبلح الحكر حول قيام مجهولين بالدخول لمنزله وسرقة دراجة هوائية وبعد البحث والتحري من طرفنا تم ضبط الدراجة بحوزة كل من المدعو/ محسن جهاد منصور 20 عام سكان ديرالبلح البركة والمدعو/ يوسف صلاح تيسير 19 عام سكان ديرالبلح السلام، حيث تم احضارهم وبالتحقيق الشفوي معهما أقروا بقيامهم بالدخول لمنزل المشتكي وسرقة الدراجة الهوائية، وعليه تم من طرفنا احالتهم والمضبوطات لمفتش تحقيق شرطة ديرالبلح لإستكمال باقي الإجراءات القانونية بحقهما.

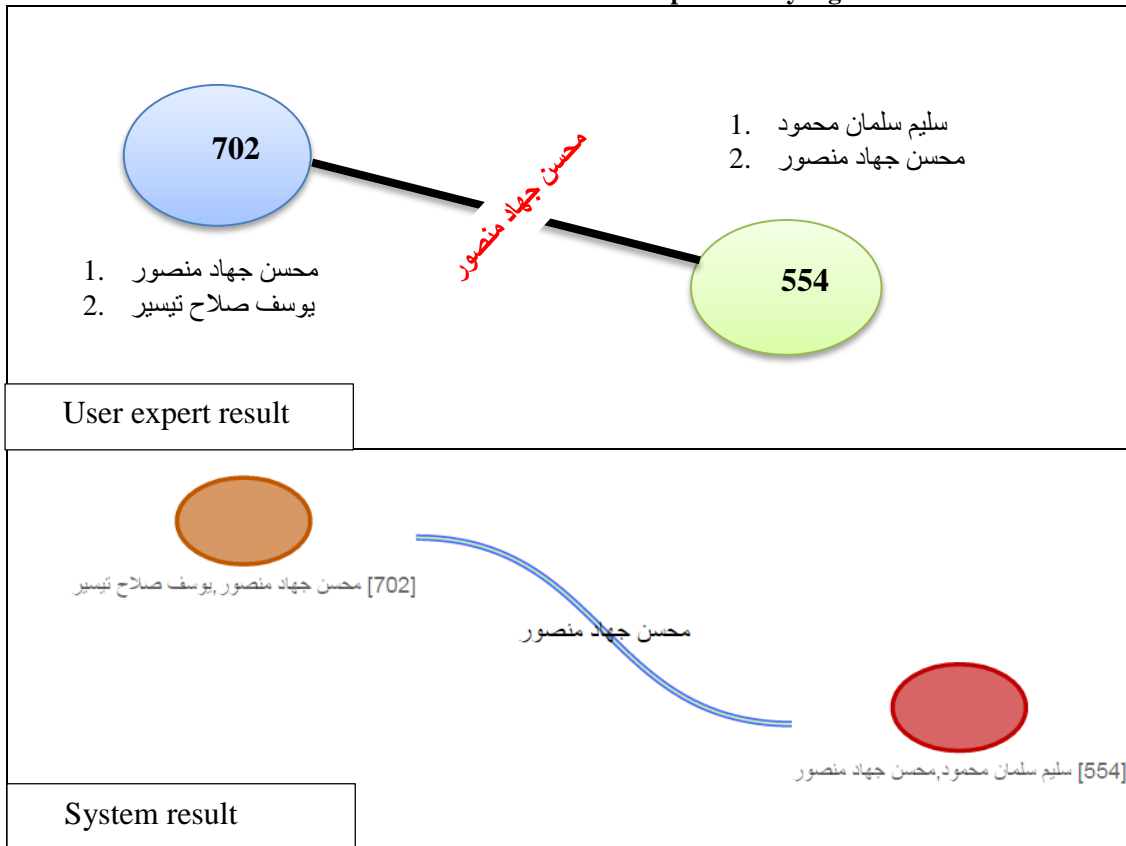
Firstly, human was determined the offender names for each document in the group that show in Table 5.5, then we apply our approach using the system for the same group, in final we compute R, P, F-measure as shown in Table 5.6

Table 5.6: Results of compute R, P, F for offender name extraction case 1

docId	System Results	Human Results
554	1. repeated name : 1 name: سليم سلمان محمود 2. repeated name : 1 name: محسن جهاد منصور	1. سليم سلمان محمود 2. محسن جهاد منصور
702	1. repeated name : 1 name: محسن جهاد منصور 2. repeated name : 1 name: يوسف صلاح تيسير	1. محسن جهاد منصور 2. يوسف صلاح تيسير
Recall (R)		F-measure (F)
4/4		$[(1+1)*(4/4)*(4/4)]/[(4/4)+(4/4)] = 1$
Precision (P)		
4/4		

The second step, human expert should extract discovery relation between documents in the same group, after that we apply our discovery algorithm for extract-hidden relationship with the same group of documents. Finally, we compute we compute R, P, and F-measure as show in Table 5.7.

Table 5.7: Evaluation of hidden relationship discovery algorithm case 1



details of relation						
Show	10	entries	<input type="text"/> :Search			
level	intermediate	Second document offender names	first document offender names	Relation Type	docId#2	docId#1
1	محسن جهاد منصور	يوسف صلاح تيسير	سليم سلمان محمود	غير مباشرة	702	554
Showing 1 to 1 of 1 entries				Next	1	Previous
Recall (R)	Precision (P)	F-measure (F)				
2/2	2/2	$[(1+1)*(2/2)]/[(2/2)+(2/2)] = 1$				

Finally, we calculate the average of Recall (R), Precision (P), F-measure between two results as shown in Table 5.8.

Table 5.8: Average calculation for (R, P, F) in case 1

Recall	Precision	F-measure
$(1+1)/2 = 1$	$(1+1)/2 = 1$	$(1+1)/2 = 1$

Another example, as shown Table 5.9 for evaluation of effectiveness of discovery algorithm by comparing human results with system results and compute R, P, F-measure manually.

Table 5.9: Case (2) for discovery indirect relationship using human expert

DocId	Investigation text
284	تم انجاز كتاب البحث والتحري الوارد الينا من مفتش تحقيق الشرطة بخصوص البحث عن ملايسات شكوى المواطن/ سعدي هاني عصام 31 عام سكان منطقة الكرامة والتي مفادها قيام مجهولين بسرقة قالونات سولار من شاحنته التي يضعها امام منزله بشارع خميس بمنطقة الكرامة وتقدر قيمة المسروقات بمبلغ وقدره 1750 شيكل حيث تم البحث وتبين بقيام الاخوة في قسم مباحث الشجاعية بضبط شبكة لصوص وهم كل من المدعو/ عارف عبد الرؤوف محمد 24 عام ويحمل هوية رقم 123456789 والمدعو/ سعيد رمزي علي 23 عام ويحمل هوية رقم 987654321 والمدعو/ سعيد راند عبد السميع 30 عام يحمل هوية رقم 123654789 وجميعهم سكان الشاطئ حيث تم التوجه لقسم مباحث الشجاعية وبتدوين افادة المذكورين اعترف المدعو/ عارف محمد انه قام بسرقة جالونات السولار من شاحنة المشتكي هو المتهمين المذكورين اعلاه وعليه تم احالة ملف القضية طرف مفتش تحقيق الشرطة.
377	تم ورود معلومات من احد مصادرنا مفادها الاشتباه بالمدعو عارف محمد 27 عام سكان الشاطئ قرب نادي الصداقة بتنفيذ عدة سرقات في منطقة الاختصاص وعليه تم البحث والتحقيق مع المذكور فاعترف بعدة سرقات نفذها المذكور حيث أفاد أن قام بسرقة مولد كهربائي أصفر اللون سعة 7ك من محلات خليل للسجاد

	بشارع الوحدة ومحفظة جوال من محل قرب العباس وأقفاص دجاج عدد 6 بعضها من مؤسسة جميل وبعضها قرب دوار دراييه وجالونات سولار عدد 9 من سيارات لشخص يدعى عز باسل سكان التوام وخضروات من سوق الشيخ رضوان وقد أفاد أنه نفذ السرقات بمشاركة شقيقه جهاد والمدعو سعيد رمزي علي 23 عام سكان السودانية والمدعو سعيد راند عبد السميع 32 عام سكان الشاطئ واعترف المذكورون على الواقعة وعليه تم احالة القضية الى مفتش التحقيق.
650	تم إنجاز كتاب مفتش تحقيق شرطة مركز الزمال بخصوص البحث عن ملابس شكوى المواطن/ جمال يوسف هاني حول قيام مجهولين بسرقة بطارية السيارة واسطوانة الغاز من داخل السيارة الخاصة به اثناء توقفها أمام جمعية نادر في منطقة النصر ، حيث بعد البحث والتحري حول الواقعة تم الاشتباه بالمدعو/ عارف عبد الرؤوف محمد 28 عام سكان النصر وبإحضاره وبالتحقيق معه اعترف بما نسب اليه وذلك بالاشتراك مع المدعو/ منير خالد حسن 25 عام سكان الشجاعية وبإحضار الاخير وبالتحقيق معه اعترف بذلك وتم احضار المضبوطات واحالة القضية الى مفتش تحقيق الشرطة لاستكمال الاجراءات القانونية .

Firstly, human was determined the offender names for each document in the group that show in Table 5.9, then we apply our approach using the system for the same group, in final we compute R, P, F-measure as shown in Table 5.10.

Table 5.10 : Results of compute R, P, F for offender name extraction case 2

docId	System Results	Human Results
284	1. repeated name : 2 name: عارف عبد الرؤوف محمد 2. repeated name : 1 name: سعيد رمزي علي 3. repeated name : 1 name: سعيد راند عبد السميع	1. عارف عبد الرؤوف محمد 2. سعيد رمزي علي 3. سعيد راند عبد السميع
377	1. repeated name : 1 name: سعيد راند عبد السميع 2. repeated name : 1 name: عارف محمد 3. repeated name : 1 name: عز باسل 4. repeated name : 1 name: سعيد رمزي علي	1. عارف محمد 2. جهاد 3. سعيد رمزي علي 4. سعيد راند عبد السميع
650	1. repeated name : 1 name: عارف عبد الرؤوف محمد 2. repeated name : 1 name: منير خالد حسن	1. عارف عبد الرؤوف محمد 2. منير خالد حسن
	Recall (R)	Precision (P)
	8/9=0.89	8/9=0.89
	F-measure (F)	
	[(1+1)*(8/9) *(8/9)]/[(8/9)+(8/9)] = 0.89	

The second step, human expert should extract discovery relation between documents in the same group, after that we apply our discovery algorithm for extract hidden relationship with the same group of documents. Finally, we compute we compute R, P, and F-measure as shown in Table 5.11.

details of relation						
Show	10	entries	<input type="text"/> :Search			
level	intermediate	Second document offender names	first document offender names	Relation Type	docId#2	docId#1
1	عارف عبد الرؤوف محمد, عارف محمد	سعيد رائد عبد السميع سعيد رمزي علي عز باسل	سعيد رائد عبد السميع سعيد رمزي علي	غير مباشرة	377	284
1	عارف عبد الرؤوف محمد, عارف محمد	منير خالد حسن	سعيد رائد عبد السميع سعيد رمزي علي	غير مباشرة	650	284
1	سعيد رائد عبد السميع, سعيد رمزي علي	عارف محمد عز باسل	عارف عبد الرؤوف محمد	غير مباشرة	377	284
1	عارف عبد الرؤوف محمد, عارف محمد	منير خالد حسن	سعيد رائد عبد السميع سعيد رمزي علي عز باسل	غير مباشرة	650	377

Showing 1 to 4 of 4 entries

Next 1 Previous

Recall (R)	Precision (P)	F-measure (F)
3/3=1	3/3=1	$[(1+1)*(3/3)]/[(3/3)+(3/3)] = 1$

Finally, we calculate the average of Recall (R), Precision (P), and F-measure between two results as show in Table 5.12 .

Table 5.12: Average calculation for (R, P, F) in case 2

Recall	Precision	F-measure
$(1+1)/2 = 1$	$(1+1)/2 = 1$	$(1+1)/2 = 1$

Table 5.13 shows the results for all cases has been computed. The average of F-measure for all the chosen cases is considered the system's performance in ability to discover hidden relationships between communities and individual.

Table 5.13: Summarize the results of calculation (R, P, and F) for discovery algorithm

Case NO.	Recall	Precision	F-measure
1	1	1	1.00
2	0.92	1	0.96
3	1	1	1.00
4	0.84	0.84	0.84
5	1	1	1.00
6	0.93	0.9	0.91
7	0.88	0.76	0.82
8	1	1	1.00

9	1	0.94	0.97
10	1	0.79	0.88
11	0.95	1	0.97
12	1	0.86	0.92
13	0.87	0.81	0.84
14	0.83	0.78	0.80
15	0.9	0.83	0.86
Average	0.94	0.90	0.92

5.7.3 System Scalability Evaluation

We evaluate the scalability of our methods by measuring the runtime required for an indirect relationship discovery algorithm on datasets of various sizes. Table 5.13 Shows the runtime of our proposed algorithm with respect to count of documents from 50 documents to 777 documents, adding 50 documents for each runtime, the time spend excludes the reading documents from hard disk and visualize the results also the first step in an algorithm for determining the name matching between different communities . Table 5.14 shows the runtime takes per document number process. In general, the total runtime increase as number of documents increase as shown in Figure 5.13.

Table 5.14: No. of documents vs. execution runtime

#	no. of documents	execution time(s)
1	50	0.1128
2	100	0.1051
3	150	0.1101
4	200	0.1237
5	250	0.1462
6	300	0.1902
7	350	0.2418
8	400	0.3476
9	450	0.3981
10	500	0.4757
11	550	0.671
12	600	0.581
13	650	0.7641
14	700	0.8948
15	777	1.2431

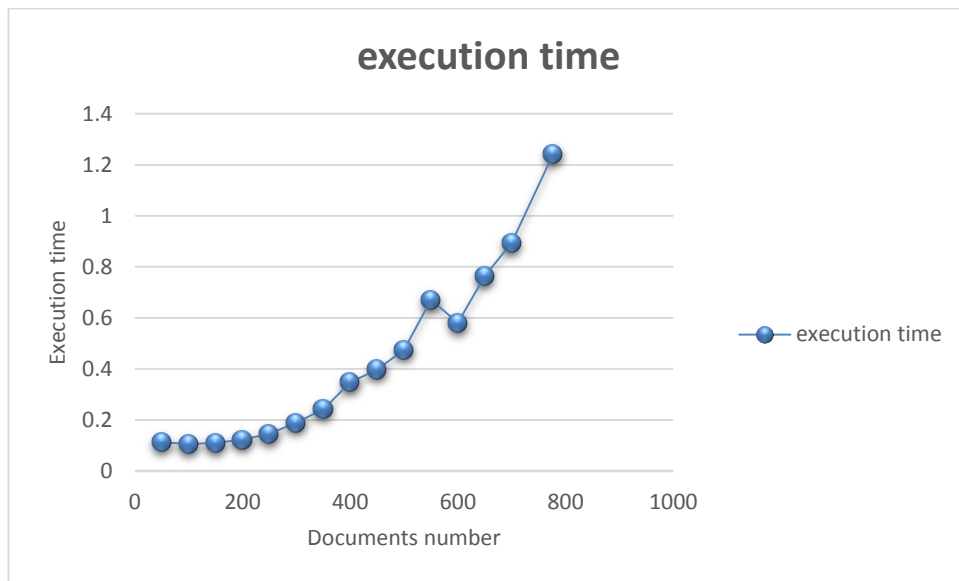


Figure 5.13: No. of documents vs. execution runtime

5.7.4 Discussion

From the previous experiments and comparisons, we can find that:

- Name Entity Recognition in our system is specialized to fetch offender name in Arabic crime investigation documents, so we ignored all person names come after "المواطن" etc. more details about rule based founds in Section 4.4 and Section A.1.
- Our system (CDS) is very similar to human results as it achieves similarity of 89% F-measure for offender name extraction and 92% for discovery hidden relation algorithm.
- Indirect relationship has the ability to find unlimited relationships between communities and individual.
- Execution time of discovery hidden algorithm increase with document number increase as shown in Figure 5.13.
- Crime detection system (CDS) has some drawbacks as:
 - A. There is no special method for the system to do name matching specially when using with an indirect relationship discovery algorithm.

- B. Name entity recognition in the system depends on predefine Gazetteer and special rule based used to extract offender names from the Arabic text document, so need to add another machine learning method to increase the ability of the system to extract proper offender name from text.

5.8 Summary

This chapter presented and analyzed the experimental results. It explained the experimental setup where presented the corpus characteristics, and data preprocessing stage, and implementation of the Name Entity Recognition (NER) using GATE tool. In addition, it presented indirect relationship discovery algorithm used to predict hidden relationships between offenders and communities. After that, we presented the data visualization. Finally, we presented the experimental results of crime detection and its performance. The evaluation of the efficiency of the crime detection system during sets of experiments.

Chapter 6

Conclusion and Future Work

Objectives

- Present the summary of the thesis.
 - Provide some recommendations.
 - Provide future work.
-

Chapter 6 Conclusion and Future Work

6.1 Summary

Text mining plays a vital role in information extraction, where used to extract particular information from unstructured text. This information may discover a new knowledge and help in making decisions. The importance in this field has been growing because difficult mining a grate data is stored as free text. The abundance of investigation reports has increased the amount of data available at police departments. There is an urgent need for intelligent tools to deal with such data. There are few of the tools used text mining techniques deal with Arabic language especially in crime detection.

Accordingly, this thesis has presented to Crime Detection System (CDS), which have developed to discover a new relationship between offenders and communities using Arabic investigation document and visualize the results to assist crime data analysis.

As already seen through this thesis, the proposed system has answered the research question, we have shown that the CDS have the ability to chive the following task:

- Extract offender names from Arabic crime documents.
- Create community for each crime documents.
- The system able to discover unlimited hidden relationships between communities and offenders.
- Various visualization methods used to present the relationships, such as crime social network (graphs), and Data table.

6.2 Contribution

Developing crime detection system (CDS) for Arabic language within the crime domain has been the main aim of this thesis. The main contribution of this thesis as follows:

- Automatically extract offender names from real unstructured crime text, while the traditional system used to structured database systems and need to save the identification number (ID) for all offender names Analysis and

discover hidden relationships between offenders usually depend on the crime investigator experts and spend a lot of time and may be difficult to review all investigation documents. For that, the Crime Detection System (CDS) can use to help police officers to discover a new relationship and enforcing law.

6.3 Recommendations

Based on previous studying of the thesis the real investigation documents provided from Gaza police department reviewing by the researcher during the field study. However, there are some of the recommendations can be formulated to adopt the goal for this thesis, as the following:

- The police departments, especially in the Gaza strip should computerize the full investigation documents to extract more knowledge leads to help in applying law enforcement.
- The top management in police department in the Gaza strip should support IT field and developing systems used text mining process and artificial intelligence to help the policeman to find a new mesh about crime.

6.4 Future Work

In this thesis, we apply many ideas as presented and this lead to extend our work. The following is a summary of the future work:

- Using more methods in machine learning to extract a proper offender name in order to enhance the accuracy of the system.
- Studying other types leads to discover hidden relation using another identification such as street, mobile number and other types of data that may be useful to the investigator to lead to new clues and criminal tracking.
- Modify or create name matching methods to be more efficient in determining a proper coreference.
- Create a profile for each offender from pervious investigation documents and indicate the crime actor's properties using historical investigation.

References

- [1] M. Alruily, A. Ayesh, and H. Zedan, "Crime type document classification from arabic corpus," in *Developments in eSystems Engineering (DESE), 2009 Second International Conference on*, 2009, pp. 153-159.
- [2] V. Grover, R. Adderley, and M. Bramer, "Review of current crime prediction techniques," in *Applications and Innovations in Intelligent Systems XIV*, ed: Springer, 2007, pp. 233-237.
- [3] A. Malathi, "An Enhanced Algorithm to Predict a Future Crime using Data Mining," *International Journal of Computer Applications*, vol. 21, 2011.
- [4] P. Thongtae and S. Srisuk, "An analysis of data mining applications in crime domain," in *Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on*, 2008, pp. 122-126.
- [5] E. Polat, "Spatio-temporal crime prediction model based on analysis of crime clusters," MIDDLE EAST TECHNICAL UNIVERSITY, 2007.
- [6] J. Song, V. Spicer, P. Brantingham, and R. Frank, "Crime Ridges: Exploring the Relationship between Crime Attractors and Offender Movement," in *Intelligence and Security Informatics Conference (EISIC), 2013 European*, 2013, pp. 75-82.
- [7] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of Emerging Technologies in Web Intelligence*, vol. 1, pp. 60-76, 2009.
- [8] K. Bogahawatte and S. Adikari, "Intelligent criminal identification system," in *Computer Science & Education (ICCSE), 2013 8th International Conference on*, 2013, pp. 633-638.
- [9] J. S. De Bruin, T. K. Cocx, W. A. Kusters, J. F. Laros, and J. N. Kok, "Data mining approaches to criminal career analysis," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*, 2006, pp. 171-177.
- [10] X. Tang and C. C. Yang, "Terrorist and criminal social network data sharing and integration," in *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*, 2009, pp. 230-230.
- [11] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 93-106.
- [12] M. Alruily, "Using Text Mining to Identify Crime Patterns from Arabic Crime News Report Corpus," 2012.
- [13] M. Alruily, A. Ayesh, and A. Al-Marghilani, "Using Self Organizing Map to Cluster Arabic Crime Documents," in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, 2010, pp. 357-363.
- [14] S. Decherchi, P. Gastaldo, J. Redi, and R. Zunino, "A text clustering framework for information retrieval," *Journal of information Assurance and Security*, vol. 4, pp. 174-182, 2009.
- [15] D. Prakash and S. Surendran, "Detection and Analysis of Hidden Activities in Social Networks," *International Journal of Computer Applications*, vol. 77, pp. 34-38, 2013.

- [16] M. Aboaga and M. J. Ab Aziz, "Arabic person names recognition by using a rule based approach," *Journal of Computer Science*, vol. 9, p. 922, 2013.
- [17] M. Asharef, N. Omar, and M. Albared, "ARABIC NAMED ENTITY RECOGNITION IN CRIME DOCUMENTS," *Journal of Theoretical & Applied Information Technology*, vol. 44, 2012.
- [18] R. 5555 Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach."
- [19] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 2011, pp. 28-34.
- [20] D. Chen, Y. Fu, and M. Shang, "An efficient algorithm for overlapping community detection in complex networks," in *Intelligent Systems, 2009. GCIS'09. WRI Global Congress on*, 2009, pp. 244-247.
- [21] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *ACM SIGMOD Record*, 2000, pp. 1-12.
- [22] A. Hassan, A. Abu-Jbara, and D. Radev, "Extracting signed social networks from text," in *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, 2012, pp. 6-14.
- [23] M. Albared, N. Omar, and M. J. Ab Aziz, "Improving arabic part-of-speech tagging through morphological analysis," in *Intelligent Information and Database Systems*, ed: Springer, 2011, pp. 317-326.
- [24] A. ALKAFF and M. MOHD, "EXTRACTION OF NATIONALITY FROM CRIME NEWS," *Journal of Theoretical & Applied Information Technology*, vol. 53, 2013.
- [25] H. Chen, J. Schroeder, R. V. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, *et al.*, "COPLINK Connect: information and knowledge management for law enforcement," *Decision Support Systems*, vol. 34, pp. 271-285, 2003.
- [26] C. C. Yang and T. D. Ng, "Terrorism and crime related weblog social network: Link, content analysis and information visualization," in *Intelligence and Security Informatics, 2007 IEEE*, 2007, pp. 55-58.
- [27] K. Baumgartner, S. Ferrari, and G. Palermo, "Constructing Bayesian networks for criminal profiling from limited data," *Knowledge-Based Systems*, vol. 21, pp. 563-572, 2008.
- [28] R. Al-Zaidy, B. Fung, A. M. Youssef, and F. Fortin, "Mining criminal networks from unstructured text documents," *Digital Investigation*, vol. 8, pp. 147-160, 2012.
- [29] F. Iqbal, B. Fung, and M. Debbabi, "Mining criminal networks from chat log," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, 2012, pp. 332-337.
- [30] C. C. Aggarwal and C. Zhai, *Mining text data*: Springer Science & Business Media, 2012.
- [31] A. Elsebai, "A rules based system for named entity recognition in modern standard Arabic," University of Salford, 2009.
- [32] P. Sondhi, "A Survey on named Entity Extraction in the Biomedical Domain."
- [33] N. Chinchor and P. Robinson, "MUC-7 named entity task definition," in *Proceedings of the 7th Conference on Message Understanding*, 1997, p. 29.

- [34] U. Singh, V. Goyal, and G. S. Lehal, "Named Entity Recognition System for Urdu," in *COLING*, 2012, pp. 2507-2518.
- [35] S. Auer and J. Lehmann, "Creating knowledge out of interlinked data," *Semantic Web*, vol. 1, pp. 97-104, 2010.
- [36] F. Graliński, K. Jassem, M. Marcińczuk, and P. Wawrzyniak, "Named Entity Recognition in Machine Anonymization," *Recent Advances in Intelligent Information Systems*, pp. 247-260, 2009.
- [37] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81-106, 1986.
- [38] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *ASSP Magazine, IEEE*, vol. 3, pp. 4-16, 1986.
- [39] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, pp. 18-28, 1998.
- [40] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, pp. 39-71, 1996.
- [41] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.
- [42] S. Sekine and E. Ranchhod, *Named entities: recognition, classification and use* vol. 19: John Benjamins Publishing, 2009.
- [43] M. Althobaiti, U. Kruschwitz, and M. Poesio, "A Semi-supervised Learning Approach to Arabic Named Entity Recognition," in *RANLP*, 2013, pp. 32-40.
- [44] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, pp. 3-26, 2007.
- [45] K. Shaalan, "A survey of Arabic named entity recognition and classification," *Computational Linguistics*, vol. 40, pp. 469-510, 2014.
- [46] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, C. Ursu, M. Dimitrov, *et al.*, *Developing Language Processing Components with GATE Version 5:(a User Guide)*: University of Sheffield, 2009.
- [47] A. Elsebai, F. Meziane, and F. Z. Belkredim, "A rule based persons names Arabic extraction system," *Communications of the IBIMA*, vol. 11, pp. 53-59, 2009.
- [48] M. Shoaib, "Using Machine Learning to Improve Rule based Arabic Named Entity Recognition," 2011.
- [49] B. Carpenter, "Character language models for Chinese word segmentation and named entity recognition," in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 2006, pp. 169-172.
- [50] S. AbdelRahman, M. Elarnaoty, M. Magdy, and A. Fahmy, "Integrated machine learning techniques for Arabic named entity recognition," *IJCSI*, vol. 7, pp. 27-36, 2010.
- [51] W. Brini, M. Ellouze, O. Trigui, S. Mesfar, H. Belguith, and P. Rosso, "Factoid and definitional Arabic question answering system," *Post-Proc. NOOJ-2009, Tozeur, Tunisia, June*, pp. 8-10, 2009.
- [52] S. Mesfar, "Named entity recognition for arabic using syntactic grammars," in *Natural Language Processing and Information Systems*, ed: Springer, 2007, pp. 305-316.

- [53] R. Alzaidy, "Criminal Network Mining and Analysis for Forensic Investigations," Concordia University, 2010.
- [54] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspectives in biology and medicine*, vol. 30, pp. 7-18, 1986.
- [55] J. Xu, B. Marshall, S. Kaza, and H. Chen, "Analyzing and visualizing criminal network dynamics: A case study," in *Intelligence and security informatics*, ed: Springer, 2004, pp. 359-377.
- [56] J. Xu and H. Chen, "Criminal network analysis and visualization," *Communications of the ACM*, vol. 48, pp. 100-107, 2005.
- [57] D. P. Bertsekas and R. G. Gallager, "Distributed asynchronous bellman-ford algorithm," *Data networks*, p. 4, 1987.
- [58] S. Skiena, "Dijkstra's Algorithm," *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*, Reading, MA: Addison-Wesley, pp. 225-227, 1990.
- [59] Dracula Graph Library, "Dracula Graph Library " May 2015 2012.
- [60] M. H. Dunham, *Data mining: Introductory and advanced topics*: Pearson Education India, 2006.
- [61] R. Al-Zaidy, B. C. Fung, A. M. Youssef, and F. Fortin, "Mining criminal networks from unstructured text documents," *Digital Investigation*, vol. 8, pp. 147-160, 2012.
- [62] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach," *International Journal of Machine Learning & Computing*, vol. 4, 2014.
- [63] D. Thakker, T. Osman, and P. Lakin, "Gate jape grammar tutorial," *Nottingham Trent University, UK, Phil Lakin, UK, Version*, vol. 1, 2009.
- [64] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, *et al.*, "Developing language processing components with gate version 6 (a user guide)," *University of Sheffield, UK, Web: <http://gate.ac.uk/sale/tao/index.html>*, 2013.
- [65] R. Lacson, N. Sugarbaker, L. M. Prevedello, I. IP, W. Mar, K. P. Andriole, *et al.*, "Retrieval of radiology reports citing critical findings with disease-specific customization," *The open medical informatics journal*, vol. 6, p. 28, 2012.
- [66] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham, "Shallow methods for named entity coreference resolution," in *Chaines de références et résolveurs d'anaphores, workshop TALN*, 2002.

Appendix A

A.1 Offender Name Extractor Rules

This section of appendix contains the lists of recognized offenders names using rule-based approach explained in Section 4.4

```
Rule: nonPersonRule
Priority:60
(
  (
    ({Lookup.majorType == location} | {Lookup.majorType == Location})
    ({Lookup.majorType == person})+ :nonoffender
  )
):ignore
-->
{
  GATE.AnnotationSet lookup = (GATE.AnnotationSet) bindings.get("nonoffender");
  GATE.Annotation ann = (GATE.Annotation) lookup.iterator().next();
  Long startOffset = ann.getStartNode().getOffset();
  Long endOffset = ann.getEndNode().getOffset();
  GATE.AnnotationSet toGo =
(GATE.AnnotationSet)inputAS.get("Lookup",startOffset,endOffset);
inputAS.removeAll(toGo);
}
```

List 0.1: Rule 3 remove annotation from non-offender person name

```
// مسجد علي ابن ابي طالب
Rule: nonPersonRule3
Priority:70
(
  ({Token.string =~ "مسجد"}
  |
  {Token.string =~ "جامع"})
  ({Lookup.majorType == person})+ :nonoffender
):ignore
-->
{
  GATE.AnnotationSet lookup = (GATE.AnnotationSet) bindings.get("nonoffender");
  GATE.Annotation ann = (GATE.Annotation) lookup.iterator().next();
  Long startOffset = ann.getStartNode().getOffset();
  Long endOffset = ann.getEndNode().getOffset();
  GATE.AnnotationSet toGo =
(GATE.AnnotationSet)inputAS.get("Lookup",startOffset,endOffset);
inputAS.removeAll(toGo);
}
```

List 0.2: Rule 4 remove annotation from non-offender person name

```

// check "المدعو"
Rule: PersonRule
Priority:80
(
  (
    (
      {Token.string =~ "مدعو"}
      |
      {Token.string =~ "المتهم"}
    )
    ({Token.string == "/"})?
  )
  //if token match
  ({Token , !Lookup})
  //the next should be person name
  :AnyPerson
):pers
-->
:AnyPerson.Person = {kind = gaz, rule="PersonRule"}

```

List 0.3: Rule 1 determine offender name

```

// المهندس، د. ، م.أ. الخ
Rule: TitlePersonRule
Priority:90
(
  (
    {Lookup.minorType == title}
  ):personTitle

  ({Token , !Lookup})
  //the next should be person name
  :AnyPerson
):pers
-->
:AnyPerson.Person = {kind = gaz, rule="TitlePersonRule"}

```

List 0.4: Rule 2 determine offender name

```

// ابو
Rule: NicknamePersonRule
Priority: 100
(
(
//we use doucment normalization and convert each !, , , ! to /
{Token.string == "ابو"}
|
{Token.string == "م"}
):Nickname
({Lookup.majorType == person})*
{Token , !Lookup}
)
//the next should be person name
:AnyPerson
):pers
-->
//:Nickname.Person = {rule="NicknamePersonRule"},
:pers.Person = {kind = gaz, rule="NicknamePersonRule"}

```

List 0.5: Rule 3 determine offender name

```

Imports: {
import static GATE.Utills.*;
import java.util.Scanner;
}

Phase: CrimePersonName
Input: Person
Options: control = once

Rule: OutputAnnotations
(
{Person}
)
-->
{
Set<Annotation> set = new HashSet<Annotation>();
set.addAll(inputAS.get("Person"));

String results = "";
String res = "";
String currDoc = doc.getName();

List<String> tmpList = new ArrayList<String>();
boolean isfound = false;

//direct relation

```

```

List<String> tmpPersonList = new ArrayList<String>();
String DirectRelation = "";

List<Object[]> rowList = new ArrayList<Object[]>();
String allPers="";
int index = 0 ;

try {
for (Annotation annotation : set) {

String type = annotation.getType();
String per = doc.getContent().getContent(
    annotation.getStartNode().getOffset(),
    annotation.getEndNode().getOffset()).toString();
per = per.replaceAll("\n", "");
per = per.trim();

int cnt = 1 ;
FeatureMap entityFeatures = annotation.getFeatures();
List matches = (List) entityFeatures.get("matches");

if(matches != null){
    if(!(tmpList.containsAll(matches))){
        //add to List
        tmpList.addAll(matches);

        // number of repeated name in documents
        cnt = matches.size();
        Annotation antecedent = null;
        for (Object id : matches) {
            antecedent = inputAS.get((Integer) id);
            String cor_per = doc.getContent().getContent(
                antecedent.getStartNode().getOffset(),

                antecedent.getEndNode().getOffset()).toString();
            if(per.length() <= cor_per.length()){
                per = cor_per;
            }
        }
    }else{
        per = "";
    }
}

// remove spaces from start and end of name
per = per.replaceAll("\s+$", "");

if( per.indexOf(" ") >= 0 ){
    results += "repeated name : "+ cnt+" name: "+ per +'\n';
    res += per+ ",";
}

```

```

}

//-----find direct relation-----
if(!(tmpPersonList.contains(per))){
    if(tmpPersonList.isEmpty() || tmpPersonList == null){
        tmpPersonList.add(per);

        }else{
            for (String str : tmpPersonList) {
                DirectRelation += per+ "->" +str+"\n";
            }
            tmpPersonList.add(per);
        }
    }

//-----end find direct relation-----

}
} catch(InvalidOffsetException ex) {
    throw new GATERuntimeException(ex.getMessage());
}
try {

    BufferedWriter out = new BufferedWriter(new
FileWriter("c:/CrimePersonName/"+currDoc.substring(0,currDoc.lastIndexOf(".")
)+".txt"));
    out.write(results);
    out.close();

//append to text file...
try
{

if(res.length() > 1 ){
    res = res.substring(0,res.lastIndexOf(","));

    //check if doc have more than one person name
    /* check if per have spaces means that it may contain first name and family name*/
    int cnt = res.lastIndexOf( ",");

    if(cnt > 1 ){
        FileWriter fw = new FileWriter("c:/CrimePersonName/test1.txt",true); //the
true will append the new data
        fw.write(res+"\n");//appends the string to the file
        fw.close();

        FileWriter fw1 = new FileWriter("c:/CrimePersonName/test.txt",true); //the
true will append the new data
        fw1.write(currDoc.substring(0,currDoc.lastIndexOf("."))
)+"|"+res/*.substring(0,res.length()-1)*+"/"++"\n");//appends the string to the file
        fw1.close();
    }
}
}

```

```

    }
    res = "";

}
catch(IOException ioe)
{
    System.err.println("IOException: " + ioe.getMessage());
}

    out = new BufferedWriter(new
FileWriter("c:/CrimePersonName/"+currDoc.substring(0,currDoc.lastIndexOf("."))
)+"_relation.txt");
    out.write(DirectRelation);
    out.close();

} catch (IOException e) {
    System.out.println("Could not write in the file:
"+currDoc.substring(0,currDoc.lastIndexOf("."))+".txt");
}
}

```

List 8: JAPE code to extract criminal community and Extract Key person of Community

```

// ابن بنت ابي
Rule: NicknamePersonRule2
Priority:300
(
(
//we use doucment normalization and convert each !,!,! to !
{Token.string == "ابن"} |
{Token.string == "بنت"} |
{Token.string == "ابي"} |
)
({Token , !Lookup})
//the next should be person name
:AnyPerson
):pers
-->
:AnyPerson.Person = {kind = gaz, rule="NicknamePersonRule2"}

```

List 0.7: Rule 4 determine offender name

```

Rule: FirstName
Priority:50
(
({Lookup.minorType == male}|{Lookup.minorType == female}|{Lookup.minorType ==
surname})+
):tag
-->
:tag.Person = {kind = gaz, rule = "FirstnameOnly"}

```

List 0.8: Rule 5 determine offender name

A.2 Name Entity Recognition evaluation

This section of appendix contains the table of name entity recognition evaluation results explained in Section 5.7.

Table 0.1:evaluation results of name entity recognition

#	DocId	Recall (R)	Precision (P)	F-measure (F)
1	9	1	1	1.00
2	10	0.9	0.9	0.90
3	23	1	1	1.00
4	31	0.83	0.71	0.77
5	38	1	1	1.00
6	53	0.92	0.5	0.65
7	68	1	1	1.00
8	75	1	1	1.00
9	89	1	1	1.00
10	93	1	0.75	0.86
11	106	0.83	0.83	0.83
12	110	1	1	1.00
13	117	1	0.8	0.89
14	126	1	1	1.00
15	135	0.88	1	0.94
16	148	1	0.83	0.91
17	160	1	0.67	0.80
18	172	1	1	1.00
19	184	1	0.83	0.91
20	196	1	1	1.00
21	211	1	0.8	0.89
22	224	1	0.75	0.86
23	235	1	1	1.00
24	248	1	1	1.00
25	260	0.87	0.71	0.78

26	272	0.83	0.67	0.74
27	286	1	1	1.00
28	307	0.88	0.43	0.58
29	319	1	1	1.00
30	328	1	1	1.00
31	340	0.9	0.45	0.60
32	352	0.83	0.73	0.78
33	367	1	0.33	0.50
34	397	1	0.67	0.80
35	412	0.88	1	0.94
36	439	1	0.43	0.60
37	454	1	0.67	0.80
38	463	1	0.62	0.77
39	475	1	0.75	0.86
40	481	1	0.83	0.91
41	490	1	0.5	0.67
42	499	1	1	1.00
43	508	1	0.57	0.73
44	520	1	0.75	0.86
45	529	1	0.75	0.86
46	541	1	0.8	0.89
47	548	1	1	1.00
48	552	1	0.71	0.83
49	554	1	1	1.00
50	557	1	1	1.00
51	559	1	1	1.00
52	565	1	1	1.00
53	567	1	0.8	0.89
54	570	1	0.83	0.91
55	574	1	1	1.00
56	581	1	1	1.00

57	582	1	1	1.00
58	584	1	1	1.00
59	603	1	0.88	0.94
60	610	1	1	1.00
61	614	1	0.5	0.67
62	616	1	0.38	0.55
63	633	1	0.75	0.86
64	642	1	0.89	0.94
65	655	1	1	1.00
66	664	1	0.86	0.92
67	676	1	1	1.00
68	694	1	1	1.00
69	709	0.92	1	0.96
70	726	1	1	1.00
71	727	1	1	1.00
72	742	1	1	1.00
73	750	1	0.86	0.92
74	751	1	0.82	0.90
75	757	1	0.67	0.80
76	765	1	1	1.00
77	767	1	0.38	0.55
78	768	1	0.6	0.75
79	769	1	0.5	0.67
80	770	0.85	0.8	0.82
81	771	1	1	1.00
82	772	1	1	1.00
83	773	1	0.75	0.86
84	774	1	1	1.00
85	775	1	0.67	0.80
86	776	1	0.5	0.67
87	777	0.7	0.62	0.66

88	4	1	1	1.00
89	5	1	0.9	0.95
90	6	0.82	0.7	0.76
91	8	1	1	1.00
92	184	0.75	0.75	0.75
93	172	1	1	1.00
94	160	1	0.83	0.91
95	148	1	1	1.00
96	135	0.88	0.83	0.85
97	126	1	1	1.00
98	117	1	1	1.00
99	110	1	1	1.00
100	106	0.83	0.83	0.83
		0.97	0.84	0.89