

# Investigating Crimes using Text Mining and Network Analysis

Nael T. Elyezjy

The Islamic University of Gaza  
Department of Postgraduate Studies  
Faculty of Information Technology  
Gaza, Palestine

Alaa M. Elhalees

The Islamic University of Gaza  
Department of Postgraduate Studies  
Faculty of Information Technology  
Gaza, Palestine

## 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 crime analysis is growing day after day.

There is a little research; in methods and techniques that extract criminal networks from unstructured investigations texts especially in Arabic language. 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.

## 1. INTRODUCTION

The security of people one of the most crucial responsibilities of governments all over the world. Thus, the main goal, here, is to reduce crime incidences [1]. There are many crime incidences types 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.

Usually, police officers prepare reports about crime 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.

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. [8].

The number of publications and research projects in data mining in the 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. [10, 11].

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

The main purpose of this paper focus on solving the problem of how to discover social relations between criminals from unstructured text investigations and extract useful information using textual mining methods. This system will be used in the last stage of an investigation to help police officers to efficiently detect hidden relations between criminals and others from a large volume of investigation documents.

## 2. RELATED WORK

Many researchers 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 section, 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.

### 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. Also they 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.

The same authors in [2] 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.

Alkaff, et al in [3] 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. But our system concern in building crime social network and discover the relation between offenders.

### Criminal social network analysis

Chen, et al in [4] developed a system called COPLINK. The system allowed different police departments to exchange data in an easy way. 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 which built the network with known information. However, the system used structured database.

Baumgartner, et al; in [5] 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.

Al-Zaidy, et al; in [6] 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 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.

D. Prakash, et al; in [7] 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.

## 3. PROBLEM DESCRIPTION

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 problem of criminal network analysis can be divided into three problems. The first one is to extract Arabic offender names from unstructured texts. The second one is to discover prominent communities in Arabic document set and extract direct link between offenders in the same community. The third is to generate hypothesis of indirect relationships between offender names and communities. These three problems are formally defined as follows:

### The problem of extract offender names

The problem of identifying the proper offender names from Arabic investigation documents is particularly difficult since they do not start with capital letters so we cannot mark them in the text by just looking at the first letter of the word. Hence, we adopt rule based approach and modifying Gazetteer.

### 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  $U = \{p_1, \dots, p_n\}$  denoted all offender names in  $D$ . each  $d \in D$  is represented as set of offender names such that  $d \in D$ . 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 4.1 (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.

### The problem of indirect relationship hypothesis generation

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  $R(\cdot) \subset D$  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 4.2 (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  $[t1, \dots, tn]$  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.

**4. THE PROPOSED CRIME DETECTION APPROACH**

In this section, we present the proposed crime detection system (CDS) framework. An overview of the framework’s stages, as depicted in Figure 1. It is divided into four types of activities.

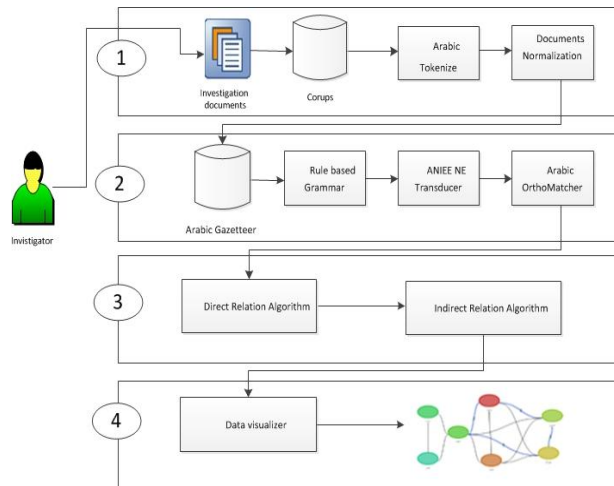


Figure 1: System Architecture

**Initial Preparation Stage Architecture**

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

**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 documented 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 get our corpus from police departments in the Gaza strip.

**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.

**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.

**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 "أحمد" Alef without Hamza above or "احمد" Alef with Hamza above. Therefore, to make the data more consistent, this process is applied.

**Sentence Splitter**

It segments the input text into several sentences; boundaries of sentence can be recognized by full stops, punctuation, end of the line, etc.

**Extract offender names Stage**

The first step in our paper 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 [8] but most of them used to handle English language. However, we have implemented 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.

**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 modifying Arabic Gate Gazetteer by adding a new distinct names using high secondary school results in Palestine from period between 2012 and 2014. After that modify GATE Gazetteer lists as shown in **Error! Reference source not found.**

Table 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

**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 [9]. In our paper, we focus on only extract offender person names.

### 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. ، م. 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 عيد" "العزیز" which means "Ahmed Abdel Aziz St."

### 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. 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. However, we use JAPE GATE tool to extract community.

### Indirect Relationship Extraction

We propose a new algorithm to discover the evidential trails between community identified in previously and other offenders who are not in the community. The trails extracted as chain of intermediate names that link a 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 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 1, shows the full steps of the indirect relationship discovery.

### Algorithm 1: Indirect relation discovery algorithm

**Input:**

- *List of person names in d where  $d \in D$*   
e.g. : array([10]=>offender name<sub>1</sub>, offender name<sub>2</sub>, .. [15]=>offender name<sub>2</sub>,offender name<sub>3</sub>,.. )

*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. }
    
```

### Data Visualization

In this paper 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 communities of offenders are mapped to node and the relationships are visualized as edges, each edge contains offender names in it. In this phase the end user is provided with network graph and data table as detailed view figure.

### 5. EXPERIMENTS AND RESULTS

In this section, we present and analyze the experimental results to provide evidence that our approach can identify offenders' names from Arabic investigation documents. Also, 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.

#### Arabic Investigation documents Corpus

We used real investigation documents about theft crime as a source of the corpus, where we get an investigations text about theft from the police department in the Gaza strip.

The dataset is divided into two sets, the first dataset used as training phase in order to build rule-based approach and modify the Gazetteer lists. The second dataset which used by our system as testing to extract offender names from text and creating communities.

We perform all text preprocessing techniques on the corpus, including tokenizing string to word and normalizing process to initialize the text. However, to implement this phase, we used GATE tool developer.

#### Name Entity Recognition

Most researchers in NLP use GATE tool 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.

#### 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 [10].

**Gazetteer** Is a list build Name Entity Recognition (NER) describe in previous section are add as ANNIE Gazetteer. It used to identify proper name within documents.

**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 extraction.

**Figure 2** shows name entity extraction components using GATE tool.

	GazetterTraining	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 2: Name Entity Extraction**

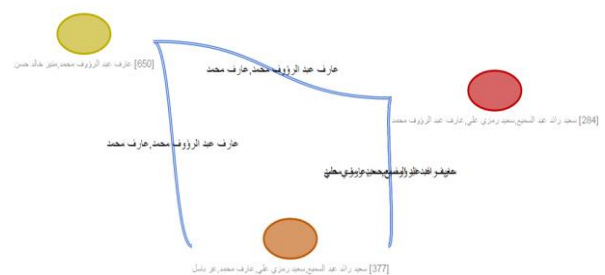
After apply the ANNIE process in crime corpus, we are building a JAPE rule to extract offender person name out of GATE tool. This rule response on building community for each Arabic investigation document in corpus and determine the count of appearing of offender person name in text to determine the 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.
- We are ignored all communities have only one offender name on it

#### 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 1 can identified unlimited levels of relationship. **Figure 3** shows an sample for using an indirect relationship algorithm. However, if no link found in one or more community represents as a direct relationship between offenders in the same community. The algorithm was implemented using PHP scripting language.



**Figure 3: Sample network visualization using indirect relation discovery algorithm**

#### Data visualizer

In this phase, we utilized to visualize technique to assist in the crime data analysis and to be better understood. We build our visualizer using Dracula Graph Library [53]. Where each

community represents as a node in crime network and link between two communities using offender names as intermediate between them. The graph was build using JavaScript with PHP.

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.

### System Efficiency Evaluation

To ensure that the system works well, 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, recall, and F-measure.

### Name Entity Recognition and Human Evaluation

To evaluate our name extraction methodology in our system, we used human references to extract offender names from 100 Arabic investigation documents was selected randomly. For each document we compute the three measurements precision (P), recall (R), and F-measure. **Table 2** show the conclusion of results for all documents has been computed. The average of F-measure for all the chosen cases is considered the system's performance in ability to extract a proper names.

**Table 2: 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.

### Indirect Relationship Discovery Algorithm

In order to evaluate the effectiveness of indirect relationship discovery algorithm used in our system, we used human references to extract hidden relation between individual and communities. However, we choose 45 Arabic investigation documents from our crime corpus, and divide the documents that chosen to 15 groups, then we are given the groups to human expert to extract offender name from text and create community for each document, then discover the relationship between communities and individual and draw crime social network. After that, we apply our methodology to the same groups and compare the results for each group with human results to compute the three measurement of precision (P), recall (R), and F-measure two times. Firstly, for offender name extraction as make in the previous section. Secondly, for all crime, social network draw from human and that draw using system. Finally, we compute the average results of (R, P, and F) for each case and compute the average for all cases.

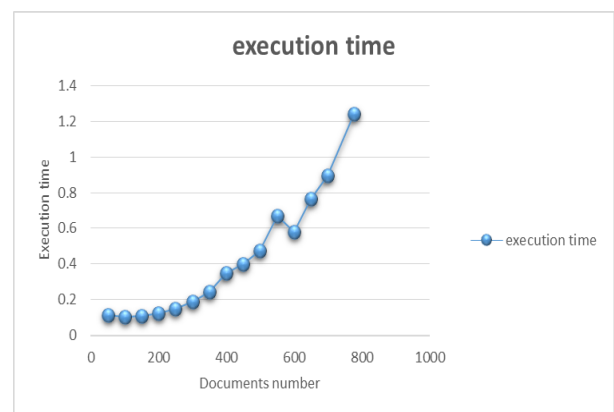
**Table 3** shows the results for all cases has been computed. The average of F-measure for all the chosen cases is considered the system's performance in ability to discover hidden relationships between communities and individual.

**Table 3: Average results of calculation (R, P, and F) for discovery algorithm**

Recall	Precision	F-measure
0.94	0.90	0.92

### System Scalability Evaluation

We evaluate the scalability of our proposed methods by measuring the runtime required for an indirect relationship discovery algorithm on datasets of various sizes. **Figure 4** 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 . In general, the total runtime increase as number of documents increase as shown in Figure 4.



**Figure 4: No. of documents vs. execution runtime 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.
- Our system 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 4.

## 6. CONCLUSION AND FUTURE WORK Summary

This paper has presented to Crime Detection System, 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.

## **Contribution**

Developing crime detection system for Arabic language within the crime domain has been the main aim of this paper. The main contribution of this paper 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 can use to help police officers to discover a new relationship and enforcing law.

## **Future Work**

In this paper, 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 Co-reference.

## **7. ACKNOWLEDGMENTS**

Our thanks to the experts who have contributed towards evaluated our system efficiency, also to the leader of police officer in Gaza strip where help us to use the real investigation documents in our experiment.

## **8. 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] 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.
- [3] A. ALKAFF and M. MOHD, "EXTRACTION OF NATIONALITY FROM CRIME NEWS," *Journal of Theoretical & Applied Information Technology*, vol. 53, 2013.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.