AUTHORSHIP ATTRIBUTION OF ARABIC TEXTS

BY

SADAM HUSSEIN MOHAMMED AL-AZANI

A Thesis Presented to the

DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

APRIL, 2014

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Sadam Hussein Mohammed Al-Azani** under the direction his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE.**

mahmoud

Prof. Sabri A. Mahmoud
(Advisor)

Dr. Wasfi Al-Khatib
(Member)

Dr. Lahouari Ghouti
(Member)

عادل

Dr. Adel Ahmed
Department Chairman

Prof. Salam A. Zummo
Dean of Graduate Studies

12/6/14

Date

عمادة الدراسات العليا
DEANSHIP OF GRADUATE STUDIES

DEDICATED TO

I dedicate this dissertation with all of my love to my father, my mother, my wife and my children.

ACKNOWLEDGMENTS

In the name of Allah, Most Kind, Most Merciful. All Praise and thanks be to Allah alone, and His peace and blessings be upon His messenger and bondman our Prophet Muhammad, his family, his companions and his followers.

I would like to thank my **parents** whose prayers are always a great source of strength for me and my **wife** and **children: Anwaar, Al-Hussein and Abraar** who took all the pain to give me time to finish this work.

I wish to express my appreciation to **Prof. Sabri A. Mahmoud**, who served as my major advisor, for his guidance and patience through the thesis, his continuous support and encouragement can never be forgotten. Thanks are due to my thesis committee members **Dr. Wasfi Al-Khatib** and **Dr. Lahouari Ghouti** for their cooperation, comments and support.

I would like to thank **King Fahd University of Petroleum & Minerals (KFUPM)** for supporting this research and providing the computing facilities. I also would like to thank **Thamar University**, which gave me the opportunity for completing my MSC. degree in KFUPM.

I owe thanks to my brother **Khalid** for his help. For everyone who helped me: جزاكم الله خيرًا.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA	:	**Authorship Attribution**
BF	:	**Best First**
CBSE	:	**Consistency-based Subset evaluation**
CFSS	:	**Correlation-based Feature Subset Selection**
ChiSAE	:	**Chi-Square Attribute evaluation**
DF	:	**Document Frequency**
DIA	:	**Darmstadt Indexing Approach**
ED	:	**Euclidean Distance**
GA	:	**Genetic Algorithm**
GS	:	**Genetic Search**
IG	:	**Information Gain**
K-NN	:	**K-nearest Neighbours**
LS-SVM	:	**Least Squares Support Vector Machines**
MLP	:	**Multi-layer Perceptron**
NLP	:	**Natural Language Processing**
OCR	:	**Optical Character Recognition**
OR	:	**Odd Ratio**
PCA	:	**Principal Components-based Subset evaluation**
POS	:	**Part-Of-Speech**
RS	:	**Rank Search**

SMO	:	**Sequential Minimum Optimization based on Support Vector Machines**
SVM	:	**Support Vector Machines**
VFI	:	**Voting Feature Intervals**

ABSTRACT

Full Name : SADAM HUSSEIN MOHAMMED AL-AZANI

Thesis Title : AUTHORSHIP ATTRIBUTION OF ARABIC TEXTS

Major Field : COMPUTER SCIENCE

Date of Degree : APRIL, 2014

Authorship attribution (AA) of Arabic text is addressed by utilizing the state of the art identification techniques, stylometric features, feature selection techniques and classifiers. This is in addition to designing novel stylometric features and techniques in this thesis.

An authorship attribution prototype for Arabic text is designed and developed. As there is no benchmarking corpus for Arabic AA, we first constructed an Arabic corpus of 20 well-known authors for authorship attribution. We investigated several stylometric features including lexical, character and syntactic features. We proposed a set of 309 Arabic function words and new lexical features (viz. word n-grams richness and specific words per author). In addition, we constructed novel stylometric features (viz. Arabic semantic features) and evaluated them on AA.

We tested several feature selection techniques and then applied them in order to optimize the extracted features and to study their effect on Arabic AA. The full and the selected feature vectors are evaluated using several classification methods (viz. Euclidean Distance (ED), K-nearest Neighbours (K-NN), Delta rule, Least Squares Support Vector

Machines (LS-SVM), Multi-layer Perceptron (MLP) and Sequential Minimum Optimization based on Support Vector Machines (SMO)).

The experimental results show that our system can identify the author of Arabic texts successfully such that it achieves best accuracy rate of 99.67%. Our system also compares favorably with the literature.

ملخص الرسالة

الاسم الكامل: صدام حسين محمد العزاني

عنوان الرسالة: تحديد كاتب النصوص العربية

التخصص: علوم الحاسب الآلي

تاريخ الدرجة العلمية: ابريل 2014م

تتناول هذه الرسالة الأساليب المختلفة لمعرفة كاتب النصوص العربية من خلال الاستفادة من التقنيات والسمات والمصنفات الحديثة. كما تساهم هذه الرسالة بإضافة وتصميم سمات وتقنيات جديدة.

تم في هذا البحث بناء وتطوير نظام فعال لمعرفة كاتب النصوص العربية وتمييز خصائص وأسلوب الكتابة لدى الكاتب. ونظرا لعدم وجود قاعدة بيانات لهذا الغرض، فقد شمل هذا العمل بناء قاعدة بيانات لتحديد كاتب النصوص العربية حيث تم اختيار20 من كتاب الأعمدة المشهورين في الصحف العربية. تم استخراج العديد من السمات اللغوية وهي: المفردات المستخدمة، واحتمالات سلسلة المحارف المتتالية (Character n-grams). وقد تم استخدام نوع جديد من السمات وهو ثراء الكلمات المتتالية (Word n-grams richness) والكلمات الخاصة بكل كاتب. وكذلك اقتراح مجموعة تحوي 309 من الكلمات الوظيفية في اللغة العربية. كما تم إنشاء سمات جديدة عالية المستوى وهي السمات ذات الدلالة المعنوية في اللغة العربية وتطبيقها لأول مرة لمعرفة كاتب النصوص العربية.

كما تم اختبار وتقييم مجموعة من تقنيات اكتشاف واستخلاص السمات الأكثر كفاءة من بين السمات المقترحة ومن ثم تطبيق التقنيات الأكثر كفاءة. وقد تم إجراء مجموعة من التجارب بتطبيق السمات المستخلصة والسمات المختصرة على قاعدة البيانات المنشأة باستخدام مجموعة من المصنفات: مصنف المسافة الاقليدية (Euclidian Distance) ومصنف الجيران الأقرب (K-NN)، وقاعدة الدلتا (Delta Rule)، ومصنف الشبكات العصبية (MLP)، ومصنفات دعم الاتجاهات (SMO and LS-SVM).

أجريت العديد من التجارب المختلفة في هذا النظام لمقارنة السمات المختلفة المستخدمة وأشارت النتائج إلى كفاءة النظام في معرفة كاتب النصوص العربية. حقق النظام دقة بلغت نسبتها 99.67 % وأشارت النتائج إلى كفاءة السمات والتقنيات المستخدمة مقارنة مع الأنظمة الأخرى.

CHAPTER 1

INTRODUCTION

Authorship attribution (AA) is the task of deciding the author of a disputed document. It can be seen as a typical classification task. In other words, a set of attributed documents (i.e. documents with known authorship) are used for training; then the problem is to identify the author of unattributed documents. The advent of non-traditional authorship attribution techniques goes back to the 19th century, when Mendenhall (1887) first created the idea of counting features like the length of word on the plays of Shakespeare. This work was followed in the 20th century by the works of (Yule, 1939, 1944) with the use of sentence lengths and vocabulary richness. It is agreed that the work of Mosteller & Wallace (1964) to solve the issues of the disputed Federalist papers is the seminal study on AA. That work is based on function words and Bayesian method. The Federalist Papers are composed of 85 political articles published in 1788 attributed to three authors namely Alexander Hamilton, James Madison, and John Jay. Twelve articles of them are anonymous and it is claimed that they were written by Alexander Hamilton or James Madison. In other words, the Federalist papers are composed of 12 disputed articles and 73 attributed articles.

Authorship attribution can be applied in a wide range of applications, for example to analyze anonymous or disputed documents/books, such as the plays of Shakespeare or Federalist papers (Ebrahimpour et al., 2013; Gill & Swartz, 2011; Jockers & Witten, 2010). It can also be used in plagiarism detection where it can be used to determine whether the claimed authorship is valid. Authorship attribution can also be applied in Forensic investigations to verify the authorship of e-mails and newsgroup messages, or to identify the source of a piece of intelligence. Authorship attribution is also applied in criminal investigation (Bosch & Smith, 1998).

Authorship attribution problems can be divided into three categories: one-class, binary class and multi-class classification (Zhao & Zobel, 2005). In one-class classification, some of the documents are by a particular author while the authorship of the other documents is unspecified and the task is to determine whether the given documents are by the single known author (Zhao & Zobel, 2005). In binary-class classification, the documents written by two authors are provided to identify who of them is the most likely author of unattributed text (Kaster et al., 2005; Seroussi et al., 2011; Shaker & Corne, 2010; Zhao & Zobel, 2005). In multi-class classification, documents by more than two authors are provided (Zhao & Zobel, 2005).

Stylometric features can be classified as lexical, character, application-specific, syntactic and semantic features (Efstathios Stamatatos, 2009). The most commonly used analytical techniques for authorship attribution are statistical and machine learning approaches.

1.1 Problem Statement

Authorship Attribution has a broad range of applications. Authorship attribution technology for English has advanced a lot over the past few decades. Unfortunately, there has been a lack of effort in the field of Arabic authorship attribution. The aim of this thesis is to fill this gap and to conduct advanced research in the field of Arabic authorship attribution. Therefore, in this thesis we conducted research in automated authorship attribution of Arabic texts. To evaluate the performance of the techniques developed in this thesis, we implemented a prototype system of Arabic AA.

Most of the previous works built or collected their own corpora for authorship attribution and a few of them are based on some benchmarking datasets that contain different writing styles from many authors. Unfortunately, to our knowledge, there is no such corpus for Arabic authorship attribution texts which can be used as a benchmarking dataset. Therefore, it is a necessity to build such a corpus for Arabic authorship attribution. For this reason, we built our corpus of Arabic authorship attribution texts. It will be made available to the researchers of authorship attribution.

1.2 The Contributions of the Thesis

The main contributions of this thesis can be listed as follows.

1. We conducted research in the area of automated Arabic authorship attribution. This results in developing the theory of Arabic authorship attribution as well as producing software tools/modules.

2. A literature survey of AA is conducted with exploring all Arabic AA researches that we are aware of and those which address non-Arabic AA researches with more focus on those done during the period of 2010 until now. To our knowledge, there are no surveys of AA research since 2010 have been published.

3. A corpus of Arabic AA texts is built. The corpus includes 1000 documents that cover several topics written by 20 authors. To our knowledge, this is the first benchmarking corpus for Arabic AA. We aim to make the corpus freely available to the research community. This is expected to provide a platform for researchers to compare their results with other researchers.

4. Several types of stylometric features (viz. lexical, character and syntactic features) are extracted for Arabic AA using our feature extractor. Additionally, we applied new lexical features such as specific words per authors and word n-grams richness features. We constructed novel stylometric features (viz. Arabic semantic lexicon) and evaluated it on AA.

5. We proposed a collection of Arabic function words that we used in AA.

6. Several feature selection techniques are applied to Arabic AA.

7. We developed a prototype system for automated authorship attribution of Arabic texts.

1.3 The Research Methodology

Authorship attribution of Arabic texts, in this thesis, can be broadly divided into a number of phases (viz. building the corpus, pre-processing, feature extraction, feature selection, training and classification). Figure 1-1 illustrates the process of our authorship attribution to Arabic text. The following methodology is followed in the course of the thesis in order to achieve our objectives.

- ***Phase 1: Literature review***

  In this phase, a literature survey of AA is conducted with exploring all Arabic AA researches.

- ***Phase 2: Corpus building***

  A corpus of Arabic AA texts is built. The corpus includes 1000 documents that cover several topics written by 20 authors.

- ***Phase 3: Features Extraction***

  We built our feature extractor to extract several types of stylometric features. Lexical, character, syntactic and semantic features are used for Arabic AA. In this phase the model of the full features are generated.

- ***Phase 4: Features Selection***

  In this phase, several feature selection techniques are applied to Arabic AA and the optimized features are selected.

- ***Phase 5: Author Attribution***

  In this step we are able to use the developed prototype of the previous steps to identify the author of an unknown Arabic text.

- ***Phase 6: Experimental results and Comparisons***

The experimental results of Arabic AA have been addressed. To show the effectiveness of our work, we compared our work with the most related works.

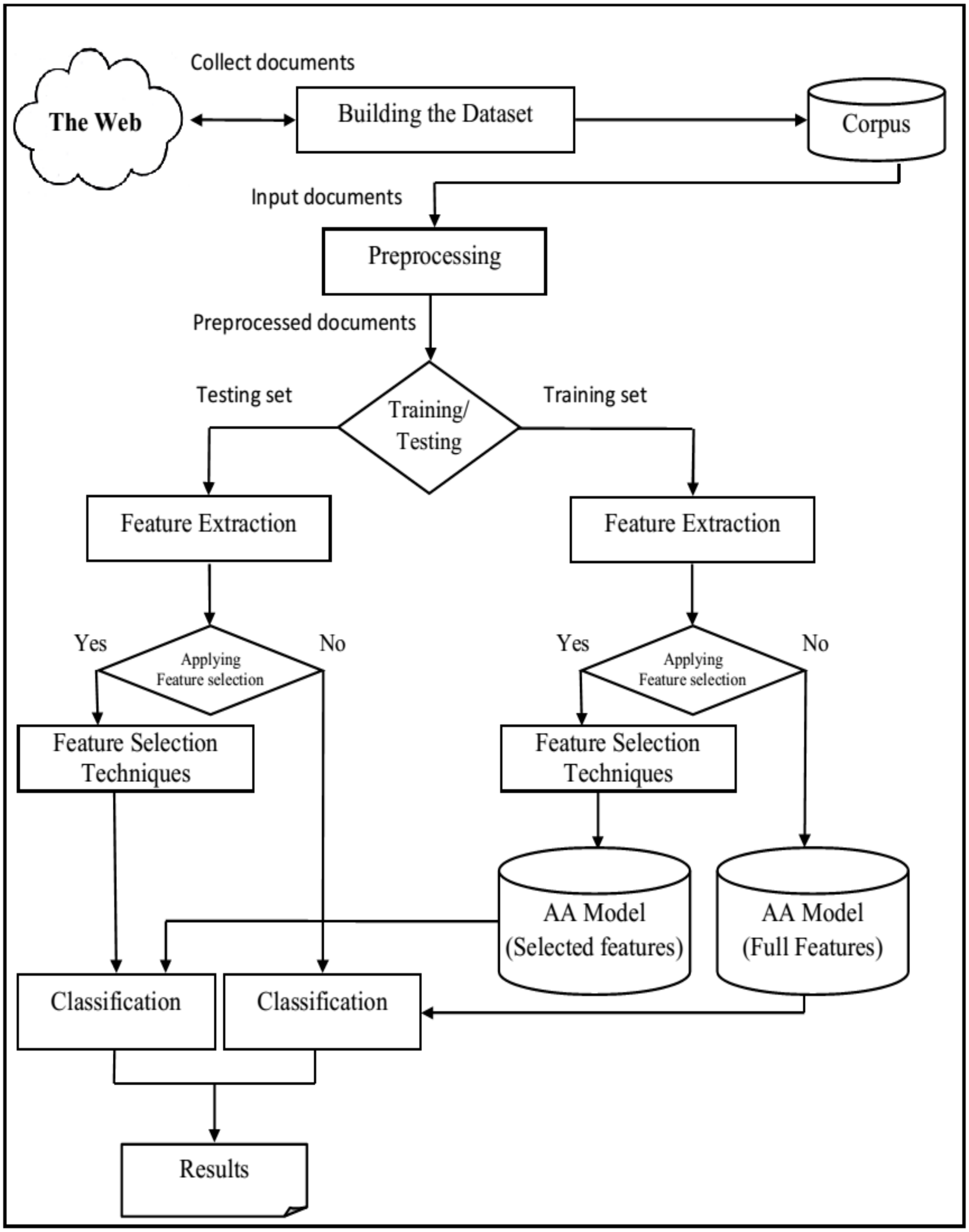


Figure 1-1 The process of authorship attribution of the thesis

1.4 Thesis outline

The thesis is organized as follows: Chapter 2 provides a survey of Authorship attribution researches. It surveys the contributions, strengths and drawbacks of the related works of non-Arabic AA researches done since 2010 and all Arabic AA researches that we are aware of. These works are classified based on the types of stylometric features, AA classification methods and techniques, selection feature techniques and corpora used by researchers. We also described the characteristics of Arabic and its challenges from the point of view AA.

Chapter 3 explores the corpora that are used in most related and recent works of non-Arabic AA and all corpora used by Arabic researches. In addition, we presented the design of our corpus. Chapter 4 discusses the extracted features and the feature selection techniques. We conducted a case study to select the most efficient feature selection techniques. Our Arabic semantic lexicon construction and extraction is detailed in Chapter 5. We discussed the results of our experiments in chapter 6; finally, the conclusions and feature works are presented in chapter 7.

CHAPTER 2

LITERATURE SURVEY AND THEORY

In this Chapter, we present the literature review of authorship attribution. The considerations here are for the most recent works that we are aware of in the field of authorship attribution especially those published since 2010 as the earlier works are surveyed by (Efstathios Stamatatos, 2009). In addition, we surveyed all works of Arabic AA texts.

Those previous works that we reported are classified based on the extracted features, the used AA methods and classifiers, the feature selection techniques and the used corpora including comments on contributions, strengths and limitations. Moreover, we explored the characteristics of Arabic and the challenges of authorship attribution research for Arabic.

The remainder of this chapter is organized as follows. Section 2.1 presents the used stylometric features of the surveyed works while section 2.2 addresses the used AA methods and classification techniques; in Section 2.3 we present the features selection techniques whereas in Section 2.4 we describe the used corpora in AA researches; Section 2.5 discusses some characteristics of Arabic and its challenges. We surveyed Arabic AA in section 2.6 and finally the summary of the chapter is presented in section 2.7.

2.1 Stylometric Features

Stylometry can be defined as the statistical analysis of literary style of authors based on the characteristics of expression in their writings. Therefore, attempting to capture the creative, unconscious elements of language is an important matter to discriminate authors and reflect or characterize the authors’ styles.

Authorship attribution features, or stylometric features, are classified into lexical, character, application-specific, syntactic, and semantic features (Efstathios Stamatatos, 2009), as shown in Figure 2-1. In this section we survey the related works based on these types of features.

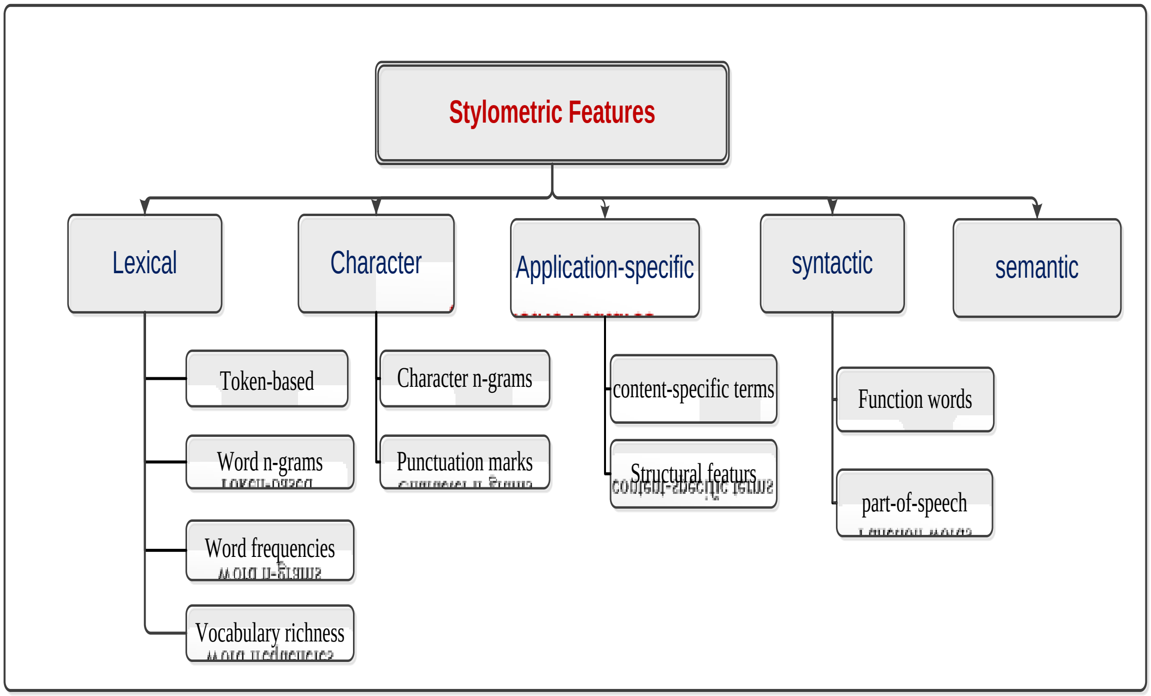


Figure 2-1 **Stylometric Features based on** (Efstathios Stamatatos, 2009)

2.1.1 Lexical features

The **lexical features** include token-based features (like word length, sentences length,…), vocabulary richness, word frequencies, word n-grams and spelling errors (Efstathios Stamatatos, 2009). The advantages of lexical features are that they are independent language features. Vocabulary richness features include three types (viz. type-to-token ratio, Hapax legomena and Hapax dislegomena)(Abbasi & Chen, 2005; Türkoğlu, Diri, & Amasyalı, 2007; Zheng, Li, Chen, & Huang, 2005). Type-to-token ratio is presented as V/N such that V is the size of the unique tokens (vocabulary) of the text, and N is the total number of tokens of the text. Hapax legomena refers to words that occur once in a given body of text while Hapax dislegomena refers to words that occur twice in a given body of text. Jockers and Witten (2010) used all common words and word bigrams of all authors to discriminate the most probable author of the disputed articles of Federalist papers.

In more recent studies, word frequencies are the main applied stylometric feature (Arun, R., Saradha, R., Suresh, V., Murty, M., & Madhavan, 2009; Ebrahimpour et al., 2013; Savoy, 2012a, 2012b, 2013a, 2013b; Seroussi et al., 2011).

Savoy (2012b) analyzed AA accuracy rates that are obtained from word types and lemmas as features. Lemma can be defined as the base form of the verb, to recognize the difference between word type and lemma. In the case of word type, each type for a word is considered as different feature. For example ‘go’, ‘goes’, ‘went’ and ‘gone’, which are forms of the verb ‘go’, are considered as four different features whereas with lemmas all of these forms are considered as one feature. Savoy (2012b) reported that the performances of both word types and lemmas seem to be similar within Delta and Z-score-based approaches; lemmas are slightly better than word types. However, lemmas require an advanced NLP tools to detect common homographic forms (Lemmatizer); such tool is available just for some natural languages. For French and German, the part-of-speech tagger (POS-tagger) is able to derive the lemmas automatically while for English they need to do some preprocessing operation for the POS-taggers such as change all plural nouns to singular nouns (e.g. Authors/NNS➔ author/NN). Therefore, applying such features is still rare. Savoy (2013a) chose word frequencies to evaluate and to compare the use of Latent Dirichlet allocation as an approach to authorship attribution with other statistical and machine learning methods.

Lexical features necessitates tools like tokenizer, sentence splitter, stemmer, spell checkers for extracting token-based features, vocabulary richness, word frequencies, word n-grams and spelling errors (Efstathios Stamatatos, 2009).

Researches differ in the size of lexical features. For example, De Vel et al. (2001) used 170 lexical features while Zheng et al. (2005) considered 87 lexical features to identify the author of English online messages and 16 lexical features for Chinese messages. The 87-lexical features are also used by Abbasi & Chen (2005) to identify the author of English online messages while they considered 79 lexical features for Arabic messages.

2.1.2 Application-specific features

Structural features and content-specific terms (keywords) are considered as application-specific features. Having a greeting acknowledgment, using a farewell acknowledgment, containing signature text and number of attachments are examples of the structural features (Abbasi & Chen, 2005; Zheng et al., 2005). Structural features are useful in the case of authorship attribution of on-line messages because some of those features are related to such type of documents (Abbasi & Chen, 2005; De Vel et al., 2001; Zheng et al., 2005). The limitation of such features is that they depend on the application and the genres of data. Content-specific features are important words (keywords) within a specific topic domain. They are important discriminative features to represent specific application domains since special words for a specific topic might be helpful to differentiate authors. Intuitively, this kind of features has a positive effect in the case of in-domain corpora (i.e. both training and testing documents belong to the same topic). Zheng et al. (2005) identified 11 English keywords and 10 Chinese keywords. Additionally, other features were used for documents in HTML format such as measures related to tag of HTML distribution (De Vel et al., 2001), counts of font size and count of font color (Abbasi & Chen, 2005). Identifying the author of a computer source code is another type of application specific features which uses features related to the source code such as code metrics (Bandara & Wijayarathna, 2013; S. Burrows, 2012). For example, Bandara and Wijayarathna (2013) used the number of characters in one source code line, the number of words in one source code line, the relative frequency of access levels (public, protected, private), the whitespaces that occurs on the interior areas of non-whitespace lines, the length of each identifier, etc. as the measures to identify the most likely author of a given source code.

2.1.3 Character features

Character features are considered as the easiest type of stylometric features to be extracted because this family of features requires a computationally simplistic approach without the need of any complicated Natural Language Processing (NLP) tools. This kind of features includes the character n-grams, the number of alphabetic characters, the number of digit characters, the number of uppercase and lowercase characters, the number of punctuation marks, etc. (Abbasi & Chen, 2005; De Vel et al., 2001; Zheng et al., 2005). Character level n-grams have been applied in previous researches and achieved good accuracy rates in authorship attribution (Eder, 2010, 2013; Escalante, Nicol, Garza, & Montes-y-g, 2011; Jamak, Savatić, & Can, 2012; Ouamour & Sayoud, 2012; Türkoğlu et al., 2007). Escalante et al. (2011) used local histograms of character n-grams. They reported that local histograms of character n-grams are more discriminating measures than the usual global histograms of words or character n-grams. They compared their work with the work of (Plakias & Stamatatos, 2008) using the same corpus of 10 authors with 100 English articles per author and SVM. They reported that they obtained higher accuracy rates. Although using a high n-grams values would better capture lexical and contextual information (Stamatatos, 2009), it results in increasing the dimensionality of the representation substantially.

Türkoğlu et al. (2007) considered character bi- and tri-gram features of Turkish datasets and showed that tri-grams have better performance than bi-grams while combining them together give better performance. Liu et al. (2013) used character n-grams of variable-length (n=1-5). Eder used character tri and quad-grams (Eder, 2010, 2013).

The success of such features is due to the ability of the character n-grams to capture nuances in lexical, syntactical, and structural level as well as their ability to handle limited data (Eder, 2010; Ouamour & Sayoud, 2012). Additionally, such features may do the same role of more complex features. For example, extracting roots of words and lemmas in Arabic require advanced NLP tools. In Arabic most of the roots of words have the length of three to four characters so when extracting character n-grams and taking the most n tri-grams frequencies are equivalent to extracting roots of words. For example, the root of “يكتبون /Yaktoobona/ (They are writing)” is “كتب /Kattaba/ (he wrote)” and it has many different lemmas such as “يكتب /Yakktobo/ (he wrote)”, “اكتب /Okktobb/ (Write!)”, etc. Another positive characteristic of such features is that the character features are not sensitive to misspelling or noise corpora (Eder, 2013; Efstathios Stamatatos, 2009) especially when using large value for n (i.e. tri-grams ore more). Character level n-grams are powerful features. However, their high dimensionality is the drawback of this type of features.

2.1.4 Syntactic features

Syntactic features are suggested as more reliable authorial fingerprint than lexical features as they reflect or represent the characteristic and styles of authors. Syntactic information is more powerful as they are not under the conscious control of the writer. This type of features includes function words and part-of-speech tags. However, syntactic features are language dependent regarding feature extraction tools.

Function words, also called “content-free” features, are terms that do not contain information about the documents content. They serve to express the grammatical relationships with other words within a sentence. Prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles are examples of function words. The advantage of using function word features lies in that they are topic-independent (i.e. they are writing style markers). It should be mentioned that, for nearly all natural languages, the researchers have not come up with standard lists of function words yet. Kaster et al. (2005) considered any word in texts other than nouns, verbs, and adjectives as function words, while other researchers suggest several variety of lists. Zheng et al. (2005) used a set of 150 English function words and a set of 69 Chinese function words while Zhao and Zobel (2007) used a list of 363 English function words. In addition, Abbasi and Chen (2005) used the same set of English function words used in Zheng et al. (2005) as well as they used a set of 200 Arabic function words. Argamon et al. (2007) used a set of 627 English function words.

Türkoğlu et al. (Türkoğlu et al., 2007) formed a list of 620 Turkish function words. The function word list created by Türkoğlu et al. (Türkoğlu et al., 2007) is corpus dependent, though. In other words, they considered the function words that only appear in their datasets. Pavelec et al. (2008) used conjunctions and adverbs (94 adverbs and 77 conjunctions) of the Portuguese language. These sets of conjunctions and adverbs are developed through adding a set of the most commonly used Portuguese verbs and a set of pronouns by (Varela, Justino, & Oliveira, 2010) to improve the accuracy rates by 4% which are also used in the work of (Varela, Justino, & Oliveira, 2011). Schaalje et al. (2013) used a set of 70 function words of the Federalist paper corpus. Arun et al. (2009) used Latent Dirichlet allocation approach on content words, stopwords and hybrid of content words and stopwords over a dataset of English novels. Surprisingly, their experiments demonstrated that the best performance is obtained for 25 topics with stopwords alone. Chandrasekaran and Manimannan (2013) used 24 function words and 18 morphological variables to identify the possible author from three contemporary Tamil scholars.

Part of speech tags are another type of syntactic features which are robust and accurate. The first use of part of speech tags in authorship attribution is attributed to Baayen et al. (1996). Such type of information was also used to identify the authorship (Eder, 2010, 2013; Kaster et al., 2005; Shaker & Corne, 2010; Solorio & Al., 2011; Efstathios Stamatatos, Fakotakis, & Kokkinakis, 2000; Zhao & Zobel, 2007). To the best of our knowledge, part of speech tagging has not been applied to Arabic authorship attribution.

2.1.5 Semantic features

Another type of stylometric features is **semantic features**. We are not aware of any attempt to apply such advanced features to AA since 2010. In general the researches of applying semantic features are rare and are surveyed in (Efstathios Stamatatos, 2009). These studies obtained poor results. An advanced study that apply semantic feature is done in (Argamon et al., 2007). Argamon et al. (2007) defined the semantic of the set of lexical features (words and phrases) according to the theory of Systemic Functional Grammar (SFG) to develop new stylistic features. They used CONJUNCTION, MODALITY, and COMMENT. The CONJUNCTION system network in SFG refers to words and phrases (like ‘and’, ‘but’, ‘on the other hand’) that combine clauses. Elaboration, Extension and Enhancement are the top-level options of CONJUNCTION scheme that are considered. The MODALITY system network might refer to modal verb (such as ‘can’, ‘might’, ’may’), use of projective clauses (like ‘It seems that…’, ’ I think that…’) or an adverbial adjuncts(such as, ‘probably’, ’preferably’). This system enables authors to express the likelihood, typicality or necessity of the events in the text. The top-level options that are considered are Type, Value, Orientation and Manifestation. In addition, the COMMENT system network refers to the status of a message with respect to textual and interactive context in a discourse. They considered eight options of COMMENT system (viz. Admissive, Assertive, Desiderative, Evaluative, Predictive, Presumptive, Tentative, and Validative). Their experiments demonstrate that best accuracy of less than 90% was obtained from the combined set of 627 English function words and these semantic features.

2.2 Authorship Attribution Methods and Techniques

There are two types of discriminative methods in modern authorship attribution research, the statistical approach and the computational or text categorization approach. The statistical approach involves statistical analysis and comparison of texts while the computational or text categorization approach involves machine learning for classification.

Support Vector Machines (SVMs) seem to be the most accurate classifier for AA studies and were used in many previous works (Arun, R., Saradha, R., Suresh, V., Murty, M., & Madhavan, 2009; Ebrahimpour et al., 2013; Eder, 2010, 2013; Escalante et al., 2011; Jockers & Witten, 2010; Z. Liu et al., 2013; Ouamour & Sayoud, 2012, 2013; Pavelec et al., 2008; Seroussi et al., 2011; Varela et al., 2010, 2011). SVMs were also applied in other earlier studies (Diederich et al., 2003; Kaster et al., 2005; Koppel et al., 2006;

Stamatatos, 2008).This is followed by multi-layer perceptron (MLP) which also achieves high accuracy rates (Chandrasekaran & Manimannan, 2013; Jamak et al., 2012; Ouamour & Sayoud, 2013; Türkoğlu et al., 2007). Other classification methods are applied to AA such as decision tress (Abbasi & Chen, 2005; Pillay & Solorio, 2010; Türkoğlu et al., 2007; Zhao & Zobel, 2005; Zheng et al., 2005), Naïve Bayes (Pillay & Solorio, 2010; Savoy, 2012a, 2013a; Schaalje et al., 2013) and Bayesian Networks (Pillay & Solorio, 2010; Türkoğlu et al., 2007; Zhao & Zobel, 2005). Chandrasekaran and Manimannan (2013) used the Generalized Regression Neural Network.

Common statistical classifiers were, also, applied on authorship attribution such as discriminate analysis (Baayen et al., 2002; Chaski, 2005; Shaker & Corne, 2010; Stamatatos et al., 2000) and principal component analysis (Jamak et al., 2012). Delta method (J. Burrows, 2002) is designed specifically for authorship attribution that used the differences of normalized term frequencies, where such frequencies were normalized using Z-score. Delta rule achieves high accuracy rates and it is used in many researches (Eder & Rybicki, 2013; Eder, 2010, 2013; Savoy, 2012a, 2012b, 2013a). k-nearest neighbour (K-NN) is used in (Eder, 2010; Jockers & Witten, 2010; Savoy, 2012b). Other distance based classification methods are used such as Mahalanobis distance (Ebrahimpour et al., 2013), Manhattan distance, Cosine distance and Stamatatos distance (Ouamour & Sayoud, 2013), Chi-square (χ^2) and Kullback–Leibler divergence (KLD) (Savoy, 2012a, 2013a, 2013b).

Zhao and Zobel (2005) used naïve Bayesian, Bayesian networks, nearest-neighbour, and decision trees. Their experiments demonstrated that Bayesian networks are the most effective while decision trees are particularly poor. Zheng et al. (2005) employed decision trees, back-propagation neural networks and support vector machines (SVMs) to identify the authors of English and Chinese messages. SVMs outperform the remaining classifiers. However, neural networks also have significantly better performance compared to decision trees. This finding is confirmed by the study of (Abbasi & Chen, 2005) where they reported that SVMs achieved higher accuracy than those obtained by decision trees for both Arabic and English on-line message datasets. This is also true regarding the study of (Türkoğlu et al., 2007) where they applied naïve Bayes, SVMs, random forest, MLP, and K-NN classifiers. Their experiments showed that the most successful classifiers are MLP and SVM.

Bandara and Wijayarathna (2013) applied an unsupervised feature learning technique, called, sparse auto-encoder for source-code author identification on five datasets. To identify whether a given source belongs to a particular author, the learnt features are used as inputs for the logistic regression. They used nine code metrics to generate a feature vector of 642 features as a set of token frequencies.

Savoy (2012a) introduced a technique for computing a standardized Z-score that is able to define the specific vocabulary found in a text compared to that of an entire corpus. He also used the Delta rule method, the chi-square distance, KLD scheme and the naive Bayes approach. He reported that his suggested classification scheme tends to perform better than other classification methods. This finding is confirmed by other work when Savoy (2012b) used principal component analysis (PCA) with K-NN, the Delta approach and the authorship attribution method based on specific vocabulary in (Savoy, 2012a).

Based on three English, French and German corpora and using word types and lemmas as features, he reported that the suggested classification method performs better than the PCA method, and slightly better than the Delta method. However, these approaches (Savoy, 2012a, 2012b) were not compared with higher classification methods such as SVM.

Arun et al. (2009) was the first to apply a generative probabilistic topic model (called Latent Dirichlet Allocation approach) of Blei et al. (2003) for authorship attribution. This approach was followed by the works of (Savoy, 2013a; Seroussi et al., 2011). Savoy (2013a) employed Latent Dirichlet allocation approach to determine the possible author of a disputed text using English and Italian corpuses. In addition, he compared his scheme with the Delta method, χ^2 approach, and KLD. The experiments demonstrated that Latent Dirichlet allocation based authorship attribution method outperforms the Delta method and the χ^2measure for the English corpus. For the Italian corpus, the Latent Dirichlet allocation scheme performs better than the χ^2 metric but at a lower performance level than the Delta method. KLD scheme performs significantly better than Latent Dirichlet allocation-based authorship attribution method. However, this approach is not compared with other Latent Dirichlet allocation-based approaches. When evaluating their scheme with an authorship attribution method based on Naive Bayes, they reported that Naive Bayes performs better in most cases. However, Latent Dirichlet allocation approach has some limitations in authorship attribution as it requires cross-domain and large size corpora (Savoy, 2013a; Seroussi et al., 2011). This is in addition to poor results obtained when applied to a large number of authors (Seroussi et al., 2011).

Ebrahimpour et al. (2013) compared the accuracy rates of Mahalanobis distance (using Multiple Discriminate Analysis (MDA) as feature selection technique) to SVM classifier. The reported accuracy of both methods is in excess of 90% for a corpus of English short stories written by 7 authors. However, the comparison is unfair as the most discriminative features are used in the first method (MDA with Mahalanobis distance) while the entire features are used for the second classifier (SVM).

Jockers and Witten (2010) compared the accuracy of five classification methods, namely Delta, K-NN, SVM, nearest shrunken centroids (NSC), and regularized discriminant analysis (RDA) based on the Federalist Papers. They reported that the 12 disputed articles are written by Madison. This finding is confirmed by the work of (Schaalje et al., 2013). Schaalje et al. (2013) introduced a specific Bayesian AA model based on the beta-binomial distribution with an explicit inverse relationship between extra-binomial variation and text size on the Federalist papers. They used regularized multinomial logistic regression (RMLR), SVM, neural nets (NN), and (NSC). Jockers and Witten (2010) made their suggestions based on NSC classifier with all common wards and word bigrams while Schaalje et al. (2013) based on the specific Bayesian method with 70 function words. These findings of assigning the 12 distributed papers to Madison is agreed by other earlier studies (Holmes & Forsyth, 1995; Mosteller & Wallace, 1964).

Gill and Swartz (2011) introduced an approach to AA based on a Bayesian Dirichlet process mixture model using multinomial word frequency data. The word frequency data is the stylistic features to identify the most probable author (‘word print’). Bayesian

Dirichlet process mixture model is based on model-based clustering of the vectors of the probability values of the multinomial distribution.

Compression model is another method for AA (Khmelev & Teahan, 2003; Kukushkina, Polikarpov, & Khmelev, 2001; Marton, Wu, & Hellerstein, 2005; Oliveira, Justino, & Oliveira, 2013). File compressors take a file and try to transform it into the shortest possible file. Such method is a good way to represent an approximation of the file. Therefore, Compression algorithms are exploited to define measures of similarity or dissimilarity between pairs of sequences of characters such as Normalized Compression Distance (NCD) (M. Li, Chen, Li, Ma, & Vitányi, 2004), Conditional Complexity of Compression (CCC) (Benedetto, Caglioti, & Loreto, 2002; Malyutov, 2005). Oliveira et al. (2013) compared Lempel-Ziv type (GZip), block sorting type (BZip) and statistical type (PPM) compressors along with two different similarity measures (viz. Normalized NCD and CCC that are based on compression). Moreover, they used both instance- based and profile-based attribution methods. They reported that the best performance rate of 99% is obtained using the first corpus of 20 authors and 30 articles per author and a combination of PPM and NCD. The performances obtained from the second corpus of 100 authors and 30 articles per author are below 77%. They compared their work with the other authorship attribution studies (AA based on feature extraction and classification methods), that used function words-based AA (Pavelec et al., 2008; Varela et al., 2011). Their obtained accuracy rates outperform the accuracy rates of (Pavelec et al., 2008; Varela et al., 2011). In general, however, we observed that nearly all performance rates obtained from the studies based on function words using any classification method don’t exceed 90%. So, it is preferable to compare their work with higher discriminate features (character level n-grams) as compression algorithms define measures of similarity based on characters.

2.3 Feature Selection

Intuitively, applying stylometric features such lexical and character features generate high dimensionality (large number of features). The large number of features slow down the process while they, perhaps, give similar results as obtained with much smaller feature subset. Sometimes, these reduced features give better accuracy than the original ones. Some extracted features might not be necessarily all relevant for the inductive learning which leads to reduced quality of the induced model. The process of reducing the extracted large number of features by selecting the most effective ones is called *feature selection*. The idea behind feature selection is selecting the most discriminating features. The easiest method to select features in authorship attribution is document frequency (DF) selection function. With DF, the n most frequent terms is taking into account (Eder & Rybicki, 2013; Eder, 2010, 2013; Jockers & Witten, 2010; Ouamour & Sayoud, 2012, 2013; Savoy, 2013a, 2013b).

There are several feature selection techniques that are used to reduce the number of stylometric features such as Odd ratio (OR) (Savoy, 2013a, 2013b), Principle component analysis (PCA) (Jamak et al., 2012; Z. Liu et al., 2013; Schaalje et al., 2013), Information Gain (IG), Chi-square, Darmstadt Indexing Approach (DIA), pointwise mutual information (Savoy, 2013b), Genetic Algorithms (GA) (Varela et al., 2011), and Correlation-based Feature Subset Selection (CFSS) (Türkoğlu et al., 2007). In a study conducted by (Forsyth & Holmes, 1996), they found that selecting features of character n-grams is more distinctive than by frequency. Jockers and Witten (2010) reduced the original feature set to include only words meeting a minimum relative frequency.

Liet al. (2006) optimized a set of 270 features to a subset of 134 using genetic algorithms. The optimized subset achieved somewhat higher accuracy than the original set. Varela et al. (2011) employed the genetic algorithm to optimized a set of 408 Portuguese function words to 58 function words. As they reported, the accuracy rate is improved from 58% with full features to 74% using the selected features. Savoy (2013a) compared the OR and the DF selection functions using Naive Bayes such that the achieved accuracy rates with DF selection function are higher than those obtained with OR. This finding is confirmed by the study of (Savoy, 2013b) where he reported that the highest accuracy rate is achieved by the DF strategy. Savoy (2013b) compared six feature selection functions (viz. IG, pointwise mutual information, OR, Chi-square, DIA, and the DF) using KLD on Italian newspapers written by four authors. Savoy (2012b) used DF to reduce the space feature vectors. Türkoğlu et al. (Türkoğlu et al., 2007) utilized CFSS that is implemented on WEKA (Witten & Frank, 2005) with many stylometric features extracted from Turkish corpora. Liu et al. (2013) used a semi-random subspace (Semi-RS)[1] method to overcome the high redundancy of the feature set and non-robustness to identify the authorship. They compressed the original feature space using PCA, and divided the selected subspace (PCA subspace) into several individual-author feature set (IAFSs) by computing the divergence between different authors' training sets. Then, they constructed a set of base classifiers (BCs) on different feature subsets which are

[1] Semi-RS is the random sampling in individual-author feature set (IAFS) partitioned from the whole author-group feature set (AGFS) instead of random global sampling in random subspace method (RSM) is performed (Z. Liu et al., 2013).

randomly sampled from each IAFS. For the final decision they combined all BCs using a combination rule.

2.4 Authorship Attribution Corpora

Authorship attribution is applied on most natural languages and for several genres. Many publications addressed English (Argamon et al., 2007; Ebrahimpour et al., 2013; Eder & Rybicki, 2013; Eder, 2010, 2013; Jockers & Witten, 2010; Z. Liu et al., 2013; Savoy, 2013a, 2012a, 2012b; Seroussi et al., 2011; E Stamatatos, 2008), Italian (Eder & Rybicki, 2013; Eder, 2010, 2013; Savoy, 2012a, 2013a, 2013b), German is considered by (Eder & Rybicki, 2013; Eder, 2010, 2013; Savoy, 2012b), Greek (Ebrahimpour et al., 2013; Eder, 2010, 2013). Other languages are considered such as French (Eder & Rybicki, 2013; Savoy, 2012b), Bosnian (Jamak et al., 2012), Polish (Eder & Rybicki, 2013; Eder, 2010, 2013), Persian (Mehri, Darooneh, & Shariati, 2012) and Arabic (Ouamour & Sayoud, 2012, 2013; Shaker & Corne, 2010).

Some studies have considered one language while others have taken into account more than one language. For example, (Z. Liu et al., 2013) conducted AA on English newspapers. Jamak et al (2012) applied AA methods on Bosnia novels. Mehri et al. (Mehri et al., 2012) applied AA applied the complex network approach for AA of books. Ouamour and Sayoud (2012) and Shaker and Corne (2010) considered Arabic books. Savoy (2013b) applied AA feature selection techniques on Italian newspapers. Another works analyzed several different genres such as (Ebrahimpour et al., 2013) (English short stories and Federalist papers) and (Seroussi et al., 2011) (English judgments, movie reviews and blog bots).

Abbasi and Chen (2005) applied AA methods to Arabic and English Web forum messages associated with known extremist groups. The accuacy rates of AA obtained from English outperform these obtained from Arabic. Stamatatos (2008) selected newswire stories in English and newspaper reportage in Arabic to conduct experiments on several different multiclass imbalanced cases. An explicit reason for the better performances obtained within Arabic may be due to the larger average size of Arabic corpora. Savoy (2012a, 2013) compared the quality of the different authorship attribution methods on English and Italian newspapers. The First object of the study of (Savoy, 2012b) is to evaluate AA methods based on English, French, and German novels. The number of authors and the number of works per author are variant for all languages. The best performances are obtained with English which is followed by French.

Eder and Rybicki (2013) introduced a method to choose and verify the appropriate training samples for AA. To choose the training and testing samples randomly, they used a bootstrap-like approach with 500 iterations. The corpus is presumed to be very sensitive to the permutations of the training samples if the density function shows widespread results. Five corpora in English, French, German, Italian, and Polish are selected to test this methodology using Delta method. They reported that, the English corpus is insensitive to permutations while other corpora are sensitive to permutations. such study cannot be generalized as they just considered one genre (novels).

Eder (2013) verified the impact of unwanted noise that were carried out on English, German, Polish, Ancient Greek, and Latin prose texts corpora. He run a procedure to damage these texts by selecting some characters randomly. These selected characters are then replaced by different characters determined randomly to generate noise texts similar to those texts generated using machine-readable text (mistakes generated when applying optical character recognition (OCR) techniques). The aim of this study is to attempt to test the impact of optical character recognition (OCR) on the attribution accuracy. In general, they found there is significant drop in the attribution accuracy. He reported that there is an impressive robustness of character-based markers. These findings may be inferred intuitively as damaging the corpus impacts the styles of authors. Besides, in the case of character level n-grams (they used n=3, 4) the character n-grams are not affected because of their huge size and the damages do not appear on the same n-grams many times. Moreover, the accuracy of quad-grams is more than tri-grams that means the probability to damage the similar sequence of tri-grams is more than quad-grams. So this is why they did not use uni-grams and bi-grams. To simulate the impact of OCR on authorship accuracy, in our opinion, it is preferable to damage the texts through using OCR confusion characters. In other words, changing the original characters with misrecognized ones gives more realistic analysis impact of OCR errors on attribution accuracy instead of making the damage in texts randomly.

Eder (2010) discussed the problem of finding minimum size of text samples for authorship attribution that would support sufficient information for author’s styles. They reported that the minimum sample length differed from 2,500 words in the case of Latin prose to 5,000 or more words in the case of English, German, Polish, and Hungarian novels. Such outcomes can be inferred intuitively.

Chandrasekaran and Manimannan (2013) took up the works of three contemporary Tamil scholars , who contributed their articles (32 articles) by attributing their names. Later, these authors wrote other unattributed articles (23 articles) on the same theme. By applying the General Regression Neural Network as classifier (using 14 function words and 18 morphological variables) they reported all the 23 unattributed radicals are assigned to the claimed author. However, the number of authors is a limitation of such study. Authorship attribution methods are also applied to non-natural languages (i.e. programming languages) to identify the possible author of the source codes in (Bandara & Wijayarathna, 2013; S. Burrows, 2012)(S. Burrows, 2012)(S. Burrows, 2012)(S. Burrows, 2012)(S. Burrows, 2012)(S. Burrows, 2012).

2.5 Characteristics and Challenges of Arabic

Arabic is a Semitic language belonging to the Arabic Afro-Asian group that poses structural and stylistic unique challenges. Arabic alphabet is the second most widely used alphabet after Latin. It is used for writing other languages such as Urdu, Persian, etc.

Arabic is written in cursive style from right to left, so the shape of a letter vary based on its position in a word (i.e. beginning, middle, end or isolated). Arabic letters have one case so no capitalization in Arabic. Arabic Alphabet consists of 28 basic letter and 8 secondary letters, 10 digits, 11 punctuation marks, eight diacritical marks, and special symbols (viz. Kashida and Maddah) as shown in Table 2-1.

Table 2-1 Arabic Alphabet, digits and especial symbols

Type	Symbols
Basic Alphabets	أ، ب ، ت ، ث ، ج ، ح ، خ ، د ، ذ ، ر ، ز ، س ، ش ، ص ، ض ، ط، ظ، ع ، غ ، ف ، ق ، ك ، ل ، م ، ن ، هـ ، و، ي
Secondary Alphabets	ا ، آ ، إ ، ئ ، ؤ ، ء ، ى ، ة
Digits	9 ،8 ،7 ،6 ،5 ،4 ،3 ،2 ،1 ،0
Punctuation	،؛؟!.:-", ([{
Diacritical marks	ـَ ،ـِ، ـُ ، ـْ، ـّ ، ـً ، ـٌّ، ـٍ
Kashida	ـ
Maddah	~
Affixes letters (حروف الزيادة)	ء، ا، ت، س، ل، م، ن، هـ، و، ي

The meaning of a word may change with different diacritic and could be ambiguous without diacritics. For example, “كتب” may be: “كَتَبَ (he wrote)”, “كُتِبَ (it is written)”, “كَتَّبَ (he teaches him the writing)”, “كُتُبْ (books)”, etc. as shown in Figure 2-2. Arabic native readers are used to deducing the meaning from the context.

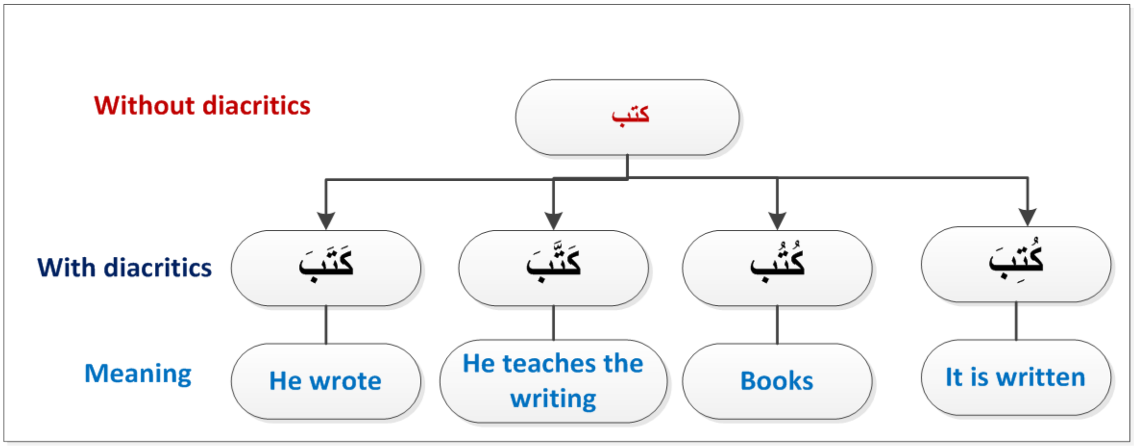


Figure 2-2 Example of Diacritics

Another challenge of Arabic is inflection; inflection is the derivation of several forms from the base form of a word. Each derived form has its own pattern which refers to the main meaning of the root and may differ with other forms in the function. For example the root كتب (he wrote) can generate many forms such as “كاتب (writer)”, “يكتب (he write)”, “كتاب (book)”, and “مكتب (office)”, etc. Each of these forms has a pattern such that the pattern “فاعل” for “كاتب” which has the meaning of the doer of the verb action and the pattern “مفعل” for the form “مكتب” which has the meaning of location (or the place noun). These forms have different functions that refer to the base form; see Figure 2-3. These functions do not change when the base form changes.

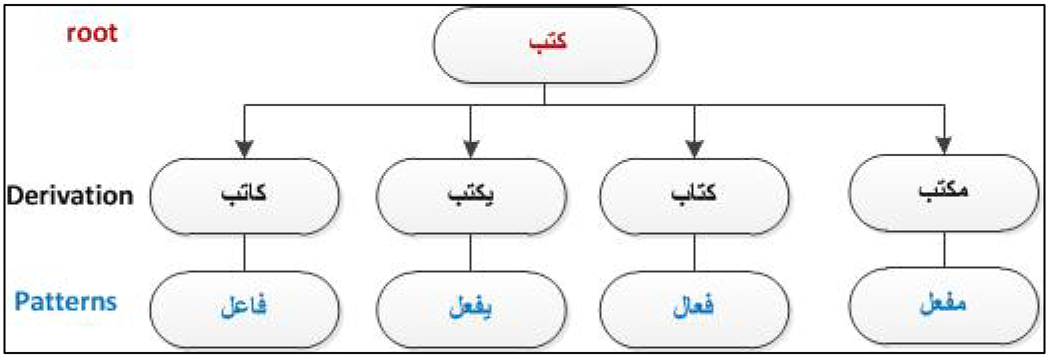


Figure 2-3 Examples of inflection in Arabic

Arabic has ten affixational letters (حروف الزيادة) which are the letters of the clause “سألتمونيها”. These letters may be used as suffixes and prefixes to any token word (verb, noun, particle, adverb, etc.) such as “سنذهب (we will go)”. In addition, the character ك as the addressee pronominal clitic (ضمير المخاطب المتصل) can be added to verbs like “علمتك (I taught you)” or to prepositions and nouns such as “منك (from you)” and “أسلوبك (your style)”. Other characters are prefixed to a word such as “ف” and “ك”.

While the previous challenges are general for Arabic natural language processing the following characteristics might affect Arabic authorship attribution. Length of words is a lexical feature and such feature may discriminate the probable author as authors differ in the use of words. Most Arabic words, however, have short length which may not have impressive effects to capture authors’ styles because the range of such length distribution is small (Abbasi & Chen, 2005). Abbasi and Chen (2005) also consider stretching out (or elongating) Arabic words as an AA issue which is called elongation. Elongation is done using Kashida symbol for purely stylistic reasons for example “كتـــــــــب (wrote)”. It can be stated that such characteristic may discriminate authors if the authorship is typed the by author himself/ herself like messages. In other genre of data like articles, elongation does not reflect the author style since they are normally typed by another person.

2.6 Arabic Authorship Attribution

Authorship attribution is applied to few Arabic texts (Ouamour & Sayoud, 2012, 2013; Shaker & Corne, 2010; E Stamatatos, 2008). Some limitations may be attributed for these works. For example they use corpora with 10 or less authors (Ouamour & Sayoud, 2012, 2013; Shaker & Corne, 2010; E Stamatatos, 2008). Also, the type of training and testing documents is either extracted from the same source (i.e. the same book for the author) (Ouamour & Sayoud, 2012, 2013) or is manipulated in a method that affects the style of the authors (E Stamatatos, 2008). AA is also applied to Arabic messages in (Abbasi & Chen, 2005). The types of stylometric features that are applied for Arabic AA are lexical features (Ouamour & Sayoud, 2012, 2013), character features (Ouamour & Sayoud, 2012; E Stamatatos, 2008) and syntactic features (Abbasi & Chen, 2005; Shaker & Corne, 2010). Abbasi and Chen also used word root features. Abbasi and Chen reported that best accuracy rate of 85.43% is obtained using SVM (Abbasi & Chen, 2005).

Stamatatos presented methods to deal with imbalanced multi-class textual datasets in order to produce a fairer classification model by segmenting the training texts into text samples according to the size of the class (E Stamatatos, 2008). Toward this end, he segmented the training set by producing many short text samples for the minority classes and less with longer samples for the majority classes. He conducted several experiments on authorship attribution based on newswire stories in English and newspaper reportage in Arabic. He used the most frequent character tri-grams as features and applied support vector machines. We believe that this method is not suitable for author's identification as performing balanced training dataset using such methods affects the effectiveness of measures that reflect the personality of the author. It may be useful for other classification problems but not for authorship attribution. In the best cases an accuracy rate of 93.6% is reported. Shaker and Corne proposed a set of Arabic function words using a dataset of 14 novels by six authors (2010). They used a set of 104 function words based on creating a collection of common prepositions and conjunctions. They utilized a hybrid evolutionary search and linear-discriminate analysis to show the performance. Highest accuracy of 93.82% is reported for identifying two authors. However, increasing the number of authors affects the recognition rates and the performance of the features. In general, the performances of works conducted on corpora less than ten authors may be questionable. Ouamour and Sayoud (2012) applied authorship attribution technique to Arabic texts written by ten ancient Arabic travelers. They used both character and word n-gram features as input of a Sequential Minimal Optimization based Support Vector Machine (SMO). They reported that character n-grams outperformed the word n-grams. In their later work (Ouamour & Sayoud, 2013) they applied less accurate classification methods using the same word based features, dataset and methodology. They obtained poor results in most cases and in best cases an accuracy rate of 80% is reported. We emphasize that, taking the training and testing set from the same source make such results less reliable.

It can be noticed that there is a lack of work on Arabic authorship attribution and there is a need to fill this gap and to conduct advanced research in the field. To fill this gap, we started with building a benchmarking corpus as presented in Chapter 3. We used several stylometric features including lexical, character and syntactic features. We investigated new lexical features such as word n-gram richness and specific words per author. Word n-gram richness features are extending vocabulary richness where we consider variant n-grams (n=1-4) as defined in Chapter 4. To overcome the large dimensionality size of feature vectors, we applied advanced feature selection techniques. The techniques that we used are described in Chapter 4 and they have not been applied and evaluated on AA. Generally there is no agreement to define a set of function words for AA. We proposed a set of Arabic function words for AA. In general, rare researches focus on advanced features such as semantic features. We are not aware of any attempt of applying Arabic semantic features. We created novel Arabic semantic features and defined our algorithm to extract them as described in Chapter 5.

2.7 Conclusions

We surveyed related works and classified them based on the used stylometric features, the classification methods, the feature selection techniques and the corpora. Authorship attribution features, are classified into lexical, character, application-specific, syntactic, and semantic features. Lexical and syntactic features tend to be the most applied features while character features might be the most discriminate features. Both statistical and machine learning classification methods are applied to identify the possible author of disputed or unattributed texts. Support vector machines, multi-layer perceptron, Bayesian Networks, Naïve Bayes are examples of machine learning classification methods whereas K-nearest neighbours, Delta Chi-square, Kullback–Leibler divergence and other distances-based classifiers are examples of statistical classifiers. Delta rule is a special classification method for AA which achieves good accuracy rates in some researches while SVM still outperforms other classifiers. The high dimensionality is a concern of stylometric features that slows down the process since some extracted features might not be necessarily all relevant. This results in reduced quality of the induced model. The process of reducing the extracted large number of features by selecting the most effective features is called feature selection. There are several techniques that are applied for stylometric features reduction such as χ^2, PCA, IG, DF, GA, CFSS, and DIA. This is in addition to considering the most frequent features based on selected thresholds. DF, χ^2 and IG tend to achieve suitable performances. Authorship attribution is applied on several natural languages such as English, Italian, German, Greek, Arabic, Bosnian, Polish, and Persian. Most of the work addressed English and then Italian whereas Arabic AA is limited. The number of benchmarking AA datasets is limited and some available corpora are composed of a small size set of authors and texts per author such as Federalist papers. Authorship attribution is also applied on non-natural language processing to identify the authors of source codes. There are different sources to collect corpora such as newspapers, books and novels. There are several different genres such as messages, articles, novels, prose and epic poems. The texts may either cover one subject which is called in-domain topic or several topics which is called cross-domain topics.

We presented some challenges and characteristics with respect to authorship attribution of Arabic including inflection, diacritics, word length, and elongation.

Table 2-2 The summary of related works classified based on defined criterion

Approach	Features					Classifiers									Other	Feature selection Technique	Corpus
						ML					Sta.						
	Lexical	Character	Appl. Spe	Syntactic	Semantic	SVM	MLP	DT	NB	BN	K-NN	Delta	X2	KLD			
(Savoy, 2013a)	√	√							√			√	√	√		DF,OR	English, Italian newspapers
(Savoy, 2013b)	√	√												√		DF,OR,IG,Chi,..	Italian newspapers
(Z. Liu et al., 2013)		√				√										PCA	English newswire stories
(Ebrahimpour et al., 2013)	√					√									√	MDA	English short stories + Federalist papers, Greek
(Eder & Rybicki, 2013)	√											√				DF	English, French ,German, Italian, Polish (novels)
(Eder, 2013)	√	√		√		√						√				DF	English, German, Polish, Latin, Greek
(Eder, 2010)	√	√		√		√					√	√				DF	English, German, Polish, Hungarian, Latin, Greek (novels, prose, epic poems)
(Chandrasekaran & Manimannan, 2013)				√			√									-	Tamil magazine articles
(Oliveira et al., 2013)																-	English newspaper articles
Schaalje et al. (2013)				√					√							PCA	Federalist papers
(Ouamour & Sayoud, 2013)	√					√	√								√	DF	Arabic Books
(Savoy, 2012a)	√	√							√			√	√	√	√	-	English and Italian newspapers
(Savoy, 2012b)	√			√							√	√			√	DF, PCA	English, French and Germany novels

Approach	Features					Classifiers									Other	Feature selection Technique	Corpus
						ML					Sta.						
	Lexical	Character	Appl. Spe	Syntactic	Semantic	SVM	MLP	DT	NB	BN	K-NN	Delta	X2	KLD			
(Jamak et al., 2012)	√	√					√									PCA	Bosnia novels
(Ouamour & Sayoud, 2012)	√	√				√										DF	Arabic Books
(Seroussi et al., 2011)	√					√									√	-	English (judgments.+ Movie review+ blog posts)
(Gill & Swartz, 2011)	√														√	-	Federalist papers
(Varela et al., 2011)				√		√										Genetic Algorithm	Portuguese newspapers
(Escalante et al., 2011)		√				√										-	English newswire stories
Varela et al. (2010)				√		√										-	Portuguese newspapers
(Shaker & Corne, 2010)				√											√		Arabic novels
(Jockers & Witten, 2010)	√					√					√	√			√	DF	FP
Arun et al. (2009)	√			√		√									√	-	English novels + Federalist papers
(E Stamatatos, 2008)		√				√										-	English (stories)+Arabic(reportage)
(Pavelec et al., 2008)				√		√										-	Portuguese newspapers
(Argamon et al., 2007)				√	√	√										-	English novels
(Türkoğlu et al. 2007)	√	√		√		√	√	√		√	√					CFSS	Turkish newspaper
(Zhao & Zobel, 2005)				√				√	√	√	√					-	English novels
(Kaster et al. 2005)	√			√		√										-	English books
(Zheng et al., 2005)	√		√	√		√	√	√								-	English +Chinese messages
(Abbasi & Chen, 2005)	√		√	√		√		√								-	English +Arabic messages

CHAPTER 3

Arabic Authorship Attribution Corpus

In this chapter, we present our work on authorship attribution of Arabic texts corpus. We also survey the corpora used for most recent AA of non-Arabic studies. In addition, we survey all reported AA researches of Arabic that we are aware of. A total of 20 regular and well-known authors are selected to build our Arabic authorship attribution corpus. In order to capture features that characterize or reflect the style of authors, sufficient works per author are collected from popular Arabic newspapers. These articles cover different topics that were published during the period from 2011 to 2013.

3.1 Introduction

The number of benchmarking AA datasets is very limited. Hence comparing reported performances is problematic unless the same data is used. In addition, some corpora are limited in the number of authors and the number of articles per author such as the Federalist papers (Ebrahimpour et al., 2013; Gill & Swartz, 2011; Jockers & Witten, 2010).

Since there is no corpus for authorship attribution of Arabic text research, this thesis aims at building such corpus. The corpus is used in our research of automated authorship attribution of Arabic text and will be made publicly available as a benchmarking dataset to other researchers. This is expected to aid in the research of authorship attribution of

Arabic text as researchers will be relieved from the task of generating their own Arabic authorship attribution corpus for their research. This will, also, enable researchers to compare the results of their techniques with published work on Arabic authorship attribution using this corpus.

We present the corpora used in the most recent research of AA for non-Arabic languages and all reported corpora of Arabic AA. The described corpora are either benchmarking datasets that are available to researchers or private datasets that are used in AA.

This chapter is organized as follows; Section 3.2 addresses the corpora used in non-Arabic AA; Arabic corpora used in AA researches are discussed in Section 3.3; our Arabic corpus for AA is presented in Section 3.4; the preprocessing operations are presented in section 3.5, and finally the conclusions are reported in Section 3.6.

3.2 Corpora For AA Of Non-Arabic Languages

In this section, we survey the corpora used in AA reported by researches of non-Arabic languages during the period from 2010 to 2013. For the corpora used before this period, reference may be made to (Koppel, Schler, & Argamon, 2009; Efstathios Stamatatos, 2009). Savoy (2012a, 2013) compared the quality of the different authorship attribution methods of English and Italian languages. He chose 20 authors either as well-known columnists or having published numerous papers and collected their works to yield an English corpus made of 5408 articles. The texts are extracted from the Glasgow Herald (GH) that are published in 1995. The extracted texts cover different subjects such as Business, Sports, Social Politics and Arts & Film headings. With regards to the Italian corpus, he chose 20 authors either as well-known columnists or as authors having published numerous papers and collected their works to yield a set of 4326 articles. The texts are extracted from the ELRA web site and published in *La Stampa* in 1994. The extracted texts cover different subjects such as Business, Sports, Social and Politics.

Savoy (2012b) evaluated AA methods based on three datasets written in English, French, and German languages. For English, he extracted 52 English text excerpts from 16 novels written by nine writers in the 19th century. For French corpus, he selected 44 segments from French novels written by 11 authors published mostly in the 19th century. Regarding German corpus, he extracted 59 German text excerpts from novels written by 15 authors published mainly during the 19th and the early 20th century. All texts were extracted from Gutenberg Project and each is around 10,000 word tokens in length.

The top 50 authors in a large corpus for the English language, Reuter corpus volume 1 (RCV1), are selected by (Houvardas & Stamatatos, 2006; Z. Liu et al., 2013; E. Stamatatos, 2007). For each author, they collected 100 articles (50 for training and 50 for testing) belonging to corporate and industrial topics (CCAT). The texts are from newswire stories and range from 450 to 550 words. Other researchers (Escalante et al., 2011; Plakias & Stamatatos, 2008) selected the top 10 authors from RCV1.

The corpus used by (Ebrahimpour et al., 2013) is obtained from Gutenberg Project archives. The corpus is composed of 168 short stories in English written by seven authors in the same period from late 19th century to early 20th century. They shorten each book to approximately the first 5000 words in order to achieve the texts balance.

Eder & Rybicki (2012) and Eder (2013) evaluated AA methods based on five datasets written in English, French, German, Italian and Polish. They are extracted from novels published during the 19th and/or the 20th century. Each dataset is analyzed three times, as entire dataset, with reduced number of texts per author, and with reduced number of writers. In the case of the entire corpora, the English corpus is composed of 63 texts written by 17 authors, the French corpus contains of 71 texts written by 25 authors, the German corpus is composed of 66 texts written by 21 authors, the Italian corpus consists of 77 texts written by 9 authors and finally the Polish corpus contains of 68 texts written by four authors. The dataset is collected from several sources and some of those texts were prepared for other projects.

The Federalist Papers are used in recent works (Ebrahimpour et al., 2013; Gill & Swartz, 2011; Jockers & Witten, 2010; Schaalje et al., 2013).

Mehri, Darooneh, & Shariati (2012) selected 63 books written by five well-known Persian authors from Ganjoor’s website during various periods to study the time evolution of the Persian language network structure.

Eder (Eder, 2010) collected several datasets for various natural languages, namely, English, Polish, German, Hungarian, Latin and Ancient Greek with different genres including novels, epic poetry and prose. The texts are collected from several sources including Perseus Project, The Latin Library, Bibliotheca Augustana, Project Gutenberg, Literature.org, Ebooks@Adelaide. In addition, they utilized some texts that were prepared for other projects and some used in (Eder & Rybicki, 2013); more details are shown in Table 3-1.

Chandrasekaran and Manimannan (2013) selected the literary works of three contemporary Tamil scholars written in the Pre–Independence period. The first author contributed with 19 articles, the second one wrote seven articles, and the third scholar contributed six articles such that all articles are in the same subject (viz. India's Freedom Movement), which is published in India magazine. The three authors then were requested to write 23 texts on the same topic and the same theme without declaring their names.

Varela et al. (2010) selected 20 authors with 30 articles per author. The articles have the average length of 600 tokens and are of several topics. They are collected from two Brazilian newspapers, namely, Gazeta do Povo and Tribuna do Paraná newspapers. Varela et al. (2011) developed this dataset by increasing the number of authors to 100 authors with the same number of documents per author. These texts are of 10 different topics and they are collected from 15 different Brazilian newspapers. Oliveira et al. (2013) used the two corpora of (Varela et al., 2010, 2011).

3.3 Used Corpora In Arabic Authorship Attribution

In this section we survey the reported corpora in Arabic AA researches. Abbasi and Chen (2005) applied authorship identification techniques to Arabic web forum messages extracted from Yahoo groups. This dataset is composed of 20 authors and 20 messages per author which covered political ideologies and social issues in the Arab world. Ouamour & Sayoud (2012) built Authorship attribution of Ancient Arabic Texts database which is collected from 10 ancient Arabic books written by 10 authors. They extracted three different texts per author two for training and one for testing. The texts are related to travelling which were collected in 2011 from “Alwaraq library”. Shaker & Corne (2010) collected a dataset of 14 novels by six different writers. The books ranged in size from 13,987 words to 37,567 words, with a mean of 23,942 words. For each author either one or two books are used for training and one for testing. Stamatatos (2008) collected a dataset of newspaper reportages in Arabic written by 10 authors downloaded from the website of Al-Hayat. The texts cover several topics. He segmented the corpus by producing many short text samples for the minority classes and less number with longer samples so that the majority classes to be with 50 texts for training and 50 texts for testing per author.

We summarized the most recent corpora used in AA researches of non-Arabic languages and those that are used in Arabic AA in Table 3-1. They are classified according to the type of natural language, the type of attributions (genre), the number of authors, the number of texts per author, the size of corpus, the period of published or written texts, the subject (either one topic or different topics) and the source of the texts of the corpus.

Table 3-1 The most recent corpora used in AA researches

Paper	Language	Type of authorships	# author	#Texts per author	Size of corpus	Period	Subject	Source
(Savoy, 2012a, 2013a)	English	articles	20	variant	5408	1995	different	Glasgow Herald
	Italian	articles	20	variant	4326	1994	different	the ELRA web site
(Houvardas & Stamatatos, 2006; Z. Liu et al., 2013; E. Stamatatos, 2007)	English	Newswire stories	50	100	5000	-	corporate/ industrial	Reuter corpus
Escalante, Nicol, Garza, & Montes-y-g, 2011; Plakias & Stamatatos, 2008			10	100/ and less	1000	-		
(Ebrahimpour et al., 2013)	English	stories	7	-	168	19-20 century	-	Project Gutenberg
(Ebrahimpour et al., 2013; Gill & Swartz, 2011; Jockers & Witten, 2010)	English	Federalist Papers (article)	3	-	85	1788	politic	Federalist Papers
(Savoy, 2012b)	English	novels	9	-	52	19century	-	Project Gutenberg
	French		11	-	44			
	German		15	-	59	19-20 century		
(Eder & Rybicki, 2013; Eder, 2013)	English	novels	17	-	66	19-20 century	-	-
	French		25	-	71			
	German		21	-	66			
	Italian		9	-	77			
	Polish		8	-	68			
Eder (Eder, 2010)	English	(novels/ epic poems)	(17/ 6)	-	(63/ 32)	-	-	Several sources
	German		21	-	66			
	Polish		13	-	69			
	Hungarian	novels	9		64			
	Latin	(Prose/ epic poems)	(20/6)		(94/ 32)			
	Greek	(Prose/ epic poems)	(8/ 8)		(72/ 30)			
(Mehri et al., 2012)	Persian	books	5	-		different	Literature	Ganjoor’s website

Paper	Language	Type of authorships	# author	#Texts per author	Size of corpus	Period	Subject	Source
Chandrasekaran and Manimannan (2013)	Tamil	Articles	3	-	55	Same period	India's Freedom Movement	India magazine
(Oliveira et al., 2013; Varela et al., 2010)	Portuguese	Newspaper articles	100	30	3000	-	different	Brazilian newspapers
(Oliveira et al., 2013; Varela et al., 2010)	Portuguese	Newspaper articles	20	30	600		different	Brazilian newspapers
(Abbasi & Chen, 2005)	Arabic	messages	20	20	400 messages		Politics and socials	Yahoo groups
(Ouamour & Sayoud, 2012)		books	10	3	30 texts		travelling	Alwaraq library
(Shaker & Corne, 2010)		novels	6	-	14 books		-	Arab Writers Union
(E Stamatatos, 2008)		reportages	10	100	1000 texts		several topics	website of Al-Hayat

3.4 Arabic Authorship Attribution Corpus Collection

To our knowledge, there is no benchmarking corpus for Arabic AA, so we decided to build a corpus for Arabic AA. It can be noticed, based on the survey in the previous sections, that all Arabic AA researches collected or built their own corpora. Additionally, both the number of authors and the number of works per author are limited. To build a benchmarking corpus for Arabic AA, we selected newspapers' articles published in Alriyadh, Alhayat and Shorouk newspapers during the period from 2011 to 2013 written by 20 authors. These 20 authors are selected from many authors in order to have regular and well-known authors. The selected authors are (1) عبدالجليل زيد المرهون (Abduljalel Zaid AlMarhon) (2) عبدالله بن عبدالمحسن الفرج (Abdullah AlFaraj), (3) عبدالله القفازي (Abdullah Algafazy) (4) عبدالله الناصر (Abdullah Annasser), (5) عبدالرحمن عبدالعزيز آل الشيخ (Abdurrahman Abdulaziz Aal Ashikh) (6) عبد الرحمن القرني (Abdurrahman Algarni) (7) أحمد الواصل (Ahmed Alwassel) , (8) علي حسن الشاطر (Ali Hassen Ashater), (9) علي ناجي الرعوي

(Ali Naji Alraawi), (10) أنوار عبدالله ابوخالد (Anwaar AboKhaled), (11)أنور أبو العلا (Anwar Abo Alalaa), (12) فهد الدوس (Fahd Addoss), (13) فهد الثنيان (Fahd Athynian), (14) هيا عبدالعزيز المنيع (Hyaa Almanee), (15) خالد الحربي (Khaled Alharbi), (16) منح الصلح (Manah Alselh), (17) محمد محفوظ (Mohamed Mahfodh), (18) اميمة كمال (Omema Kamal), (19) راغدة درغام (Ragedah Dergham), (20) شريفة الشملان (Sharefah Asshamlan). In order to capture features that characterize or reflect the style of authors, we have to provide sufficient works for each author, so we selected 50 articles per author. The texts are in politics, economics, socials and sports. The average length of texts per author ranged between 411 and 1242 words, 2452 and 8132 characters and the average size ranged from 3 to 8 KB.

Extracting the articles from newspapers is preferred than other types of texts (viz. books, novels and messages) for several reasons. Firstly, such type represents the characteristics of the author in one document. This is also true regarding the messages and books when extracted as chapters. Besides, we are not aware of any serious study discussing this type of writings in Arabic. Moreover, considering books as type of writings poses two issues the first one is that the authorships per writer are limited while the second issue is that dividing a few books into many documents (with the same or different size) may affect the writing quality of authors and may not express the style of the authors (Ebrahimpour et al., 2013; E Stamatatos, 2008). Moreover, it may be easier to detect the author as the training and testing data are extracted from the same resource (Ouamour & Sayoud, 2012). On the other hand, most Arabic authorship documents are religious in nature, and hence such authorship texts contain many citations from the main Islamic sources such as The Holy Quran, The Sunnah and/or the Companions and/or the sayings of scholars. The need to meticulously review the text may greatly increase the time build the corpus (dealing with such concerns will be considered as future work). Regarding the other types of writings (messages), articles may be considered as a special case of messages, i.e., the techniques applied for articles can be utilized for messages. Such genres are discussed by (Abbasi & Chen, 2005) while the genre under consideration necessitates advanced research.

We avoided the articles written by more than one author. Another criterion, we considered, when building our corpus, is that it can be used for other authorship analysis tasks such as authorship characterization, since it includes five female authors out of 20 authors. It can also be used for authorship verification and plagiarism detection.

We did our best to collect text with sufficient authors (20 authors) and sufficient works for each author (more than or equal to 50 articles per author). Moreover, we considered selecting regular authors (columnists), carefully, for possible extension of the corpus in the future. Table 3-2 summarizes the statistics of the corpus.

The column ‘Num’ in Table 3-2 shows the number of articles collected per author. However, we used the top 50 articles for each author. The ‘Avg. (words)’ and ‘Avg. (characters)’ columns in Table 3-2 list the average word length and the average character length of articles per author, respectively (as shown in Table 3-2). The average length of texts per author is between 411 to 1242 words, 2452 to 8132 characters and the average size ranged from 3 to 8 KB. The minimum, mean and maximum articles’ lengths in words are 129, 671 and 1688 respectively. The median and the standard deviation are 565 and 304words. The minimum, mean and maximum articles’ lengths in characters are 829, 4017, and 10740. The median and the standard deviation are 3330 and 1862 characters.

Table 3-2 The summary and statistics of the Arabic AA corpus

Author's Name	subject	Num	Avg. (words)	Avg. (Character)	Period(From: To)	Sex *	Source **
Abduljalel Zaid AlMarhon	Politics	55	1242	7680	27-04-2012 : 24-5-2013	M	R
Abdullah AlFaraj	Economic	55	431	2497	20-04-2012 : 17-05-2013	M	R
Abdullah Algafazy	Politics	56	990	5927	05-03-2012 : 27-05-2013	M	R
Abdullah Annasser	Socials & Politics	55	543	3161	03-02-2012: 24-05-2013	M	R
Abdulrahman Abdulaziz Aal Ashikh	Socials & Politics & Economic	55	620	3628	2012-03-23 : 24-05-2013	M	R
Abdulrrahman Algarni	sport	50	422	2452	13-03-2012 : 26-05-2013	M	R
Ahmed Alwassel	Art	52	411	2533	23-11-2011 : 22-05-2013	M	R
AliHassenAssater	Politics	50	687	4413	13-03-2012 : 21-05-2013	M	R
AliNaji Aleaawi	Politics	54	626	3743	22-02-2012 : 22-05-2013	M	R
Anwaar AboKhaled	Socials	55	517	2897	04-11-2011 : 17-05-2013	F	R
Anwar AboAlalaa	Economic	56	458	2716	31-03-2012 : 25-05-2013	M	R
Fahd Addoss	Sports	52	447	2668	17-08-2012 : 12-07-2013	M	R
Fahd Athynian	Economic	55	482	3003	07-04-2013 : 19-05-2013	M	R
Hyaa Almanee	Social	56	455	2670	27-10-2012 : 25-05-2013	F	R
Khaled Alharbi	Sports	51	479	2949	4-03-2013 : 17-05-2013	M	R
Manah Alselh	politics	55	892	5240	13-08-2011 : 24-05-2013	M	R
Mohamed Mahfodh	Social	50	872	5345	30-11-2011 : 21-05-2013	M	R
Omema Kamal	Social & Economic	55	1010	5829	11-08-2011 : 26-05-2013	F	SH
Ragedah Dergham	politics	55	1345	8132	05-04-2012 : 27-05-2013	F	H
Sharefah Asshamlan	Social	53	495	2860	27-09-2012 : 23-05-2013	F	R

* M: Male/ F:Femal

** R: Alriyadh/ SH: Shorouk/ H: Alhayat

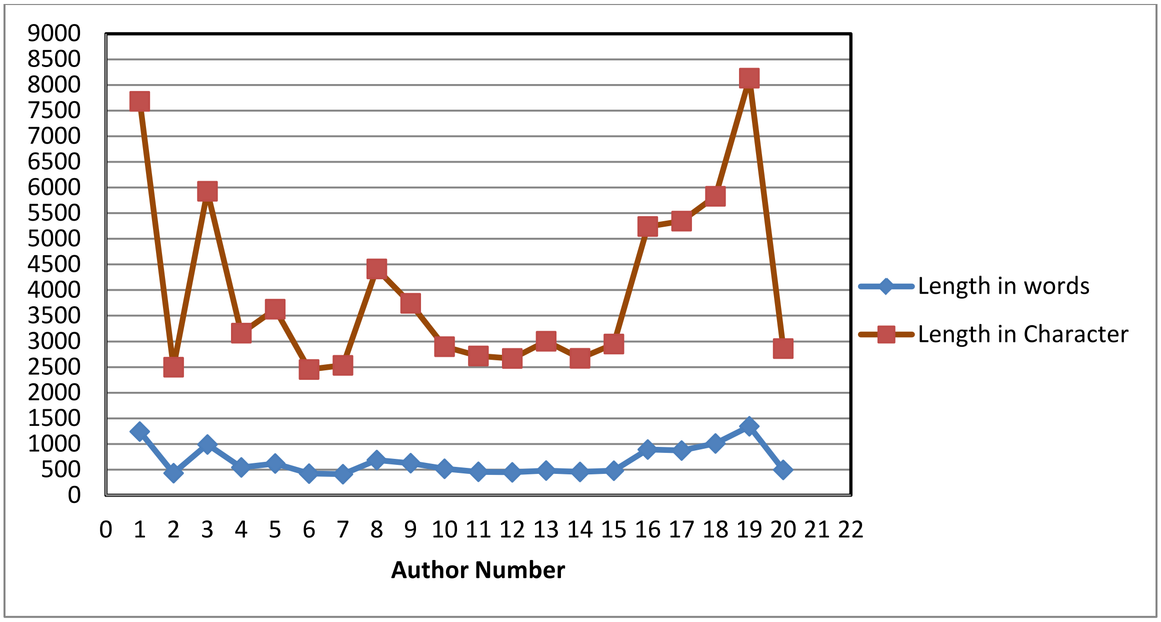


Figure 3-1 The average word and character lengths of articles per author

Table 3-3 shows the counts of both word types and tokens n-grams (n=1-4) for the training set, the testing set and the whole corpus with taking 70% (700 articles) for the training set and 30% (300 articles) for the testing set.

Table 3-3 The word n-grams (n=1-4) statistics of the corpus

	Uni-grams		Bi-grams		Tri-grams		Quad-grams	
	Type	Token	Type	Token	Type	Token	Type	Token
Training	67,058	463,324	319,958	463,323	413,736	463,322	413,331	463,321
Testing	41,750	196,383	153,071	196,382	185,692	196,381	191,068	196,380
The whole corpus	81,543	659,707	436,736	659,706	581,575	659,705	610,016	659,704

3.5 Preprocessing

To determine the probable author of a given text, it is important to do some preprocessing operations on the corpus. We just considered the main body of the text (excluding titles, author names, dates, etc.) for each article manually. This is followed by eliminating the diacritics including (ـٍ ، ـٌّ، ـً ، ـّ ،ـْ ، ـُ ،ـِ، ـَ) and non-Arabic terms and symbols or alphabets. This is in addition to removing the special symbols as they are considered as noisy symbols. We consider every symbol except the punctuation marks (،؛؟!.:-" ([{) as noisy symbols. At the word level preprocessing, we do not take into account the punctuation marks. Considering such features at the word level leads to increases in the size of the feature vectors as the same word is counted as two or more words. For example, the words “الكاتب. *(author.)*” and “الكاتب *(author)*” are considered as different types and hence they are counted two features instead of one if punctuation marks are considered.

While eliminating the titles, author names, dates of articles is done manually, the rest of preprocessing operations is done automatically through our developed system.

3.6 Conclusions

In this chapter, we presented the state of the art in the copra used in the most recent research of AA for non-Arabic languages and the corpora used in all reported AA of Arabic. The previously used corpora in the field of AA are tabulated indicating the number of authors, samples, languages, genre, etc.

We have collected a corpus for Arabic AA which compares favorably with the reported corpora. We think that conducting an advanced research in AA of Arabic texts is hard due to the lack of benchmarking corpus. Our built corpus is a solution to this problem. Increasing the number of authors and their works is possible in the future. The corpus will be made publicly available as a benchmarking corpus for other researchers.

CHAPTER 4

FEATURES EXTRACTION AND SELECTION

We have considered five different types of stylometric features that are applied on AA (viz. lexical, character, application specific, syntactic and semantic features). Lexical, character and syntactic features are the most commonly applied features on AA while application specific ones are more special measures which rely on the type of corpus. Semantic features are applied in rare works since such features require advanced methods. This chapter is divided to two main parts: feature extraction and feature selection.

4.1 Dataset Representation Methods

It is clear that the main objects in AA problem are: a set of authors, a set of attributed documents or training set, a set of unattributed texts or testing set. The goal is to identity who wrote the disputed text. There are different approaches to represent the attributed documents or the training corpus, such as profile-based method and instance-based method. In profile-based method, the known samples of the authors concatenated into a single big document per author and then the information that characterize the author are extracted from this concatenated file as shown in Figure 4-1.

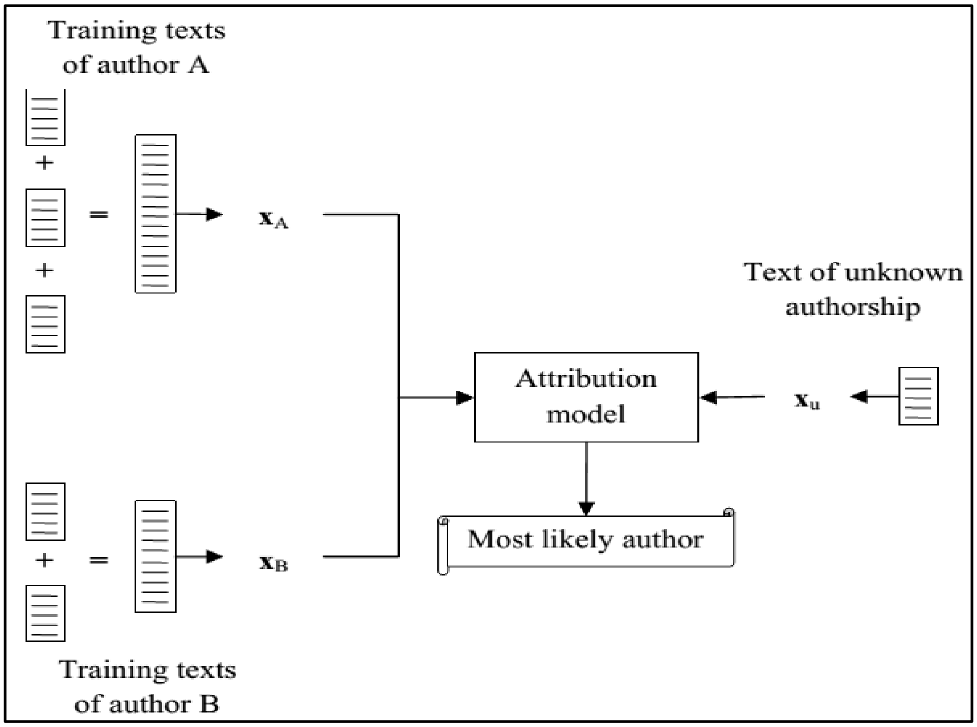


Figure 4-1 The architecture of profile-based approaches (Efstathios Stamatatos, 2009)

While in the instance-based method, each training sample is represented separately to reflect the style of author as shown in Figure 4-2.

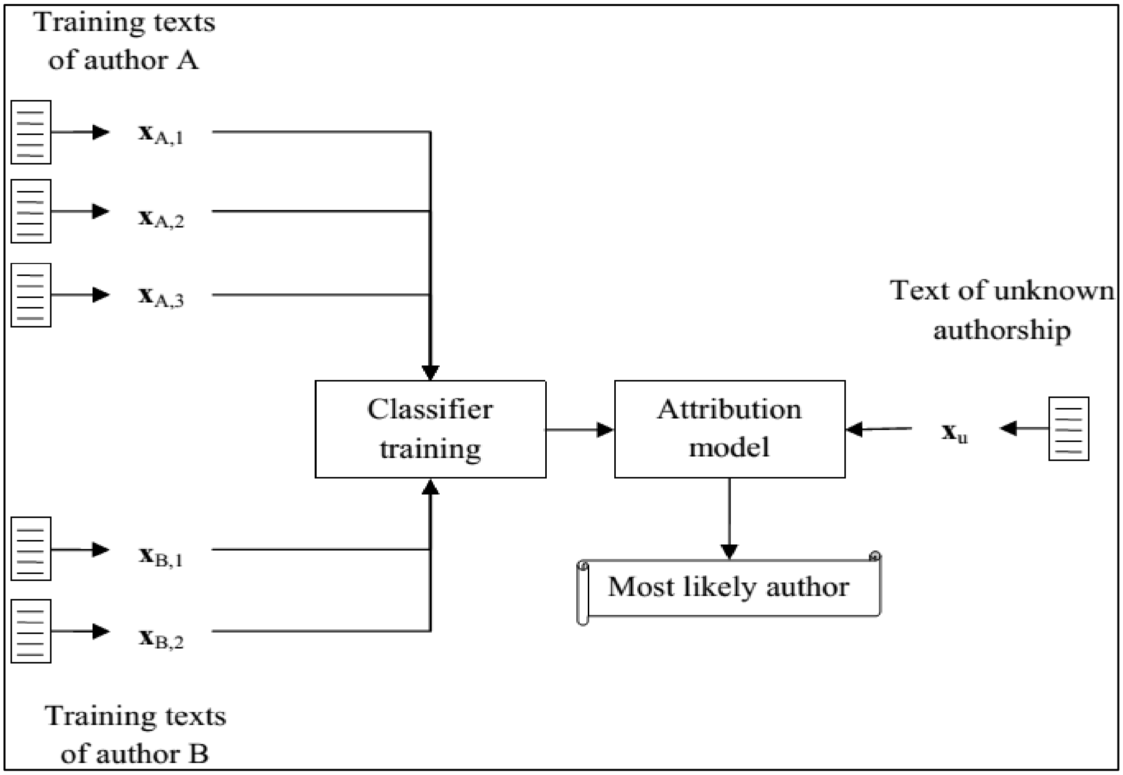


Figure 4-2 **The architecture of instance-based approaches** (Efstathios Stamatatos, 2009)

In our work, we used both methods with emphasis on the instance-based method. We used the profile-based method when applying the Delta rule as a classification method where such classification method necessitates creating the profile of author.

4.2 Feature Extraction Phase

All documents in the dataset were processed to produce numeric feature vectors using the following feature sets.

4.2.1 Lexical features

As lexical features we considered vocabulary richness features and word level n-grams. Vocabulary richness features include type-to-token ratio, Hapax legomena and Hapax dislegomena. They are used in previous works based on vocabulary while we, here, use them with variant word level n-grams (n=1-4); we call them ***word n-grams richness*** features. Therefore, we concatenated word n-grams richness features and we obtained 12-dimentional feature vector called word n-gram richness feature vector (*WR*).

As mentioned above, another type of lexical features is word level n-grams where the value of n depends heavily on the size of dataset. The dependency here means that the bigger value of n is more effective when the size of corpus is large. In other words, the word level n-grams perform better with huge dataset. For example, in our training set the occurrences of word unigrams, and bi-grams that are greater than or equal to 50 occurrences are 1181 and 141 respectively; while the chance of occurring 50 times or more is insignificant in the case of tri-grams and impossible regarding quad-grams. Table 4-1 shows the frequencies of word level n-grams (n=1-4) in the training set.

Table 4-1 Statistics of word n-grams frequencies (n=1-4) in the training set

	>0	>9	>19	>29	>39	>49	>99
Uni-gram	67058	6630	3287	2122	1502	1181	473
Bi-grams	319,958	2479	781	365	222	141	35
Tri-grams	413,736	275	61	25	11	3	0
Quad-grams	431,331	42	11	3	0	0	0

In this thesis, we used the words that occur in the training set more than 49, 99 and 149 times because of the large size of the word level uni-grams feature vector.

We used different types of features normalization, the absolute frequencies, the relative frequencies and Z-scores. The relative frequencies are computed using the following equation:

$$p(w_i) = \frac{count(w_i)}{\sum count(w)}$$

The second standardized score is Z-score of the absolute frequencies for each word type using the following equation:

$$Z\text{-}score(t_{ij}) = \frac{tfr_{ij} - \mu_i}{\sigma_i}$$

Where:

tfr_{ij} *:* the frequency of $term_i$ in a particular document D_j

μ_i : the mean of $term_i$ in the corpus.

σ_i *:* the standard deviation of $term_i$ in the corpus.

Therefore, as shown in Table 4-2, the number of features for words that occur >=50 (word1GTH49fv), the number of features for words that occur >=100 (word1GTH99fv) and the number of features for words that occur >=150 (word1GTH149fv) are 1181, 473 and 287 respectively.

Specific Words per Authors

Specific words are words that are used only by the author. To extract such information, we first concatenated all training documents into a big file (TF). After that, we created the authors' profiles of the training set (TFA_i). The word w_j is specific for the i^{th} author if its count in TF is equal to its count in TFA_i.

We investigated threshold values of two and three for the number of occurring of specific words per author. A threshold value less than three generates large feature vectors including typos while using threshold greater than four leads to authors with no specific words. The numbers of specific words that occur more than two (*specifcwGTH2fv)* and three (*specifcwGTH3fv)* are 1193 and 713, respectively.

Combined words and specific words features

We combined the *word1GTH149fv* along with *specifcwGTH3fv* to obtain combined word and specific words feature vector (*CoSpW*) with size of 1000 words.

It is noteworthy that when combining feature vectors, we have to take into account compatibility for the feature vectors. That is feature vectors should be in the same representation or normalization, for example it is unreasonable to combine relative frequencies with a feature vector normalized with Z-scores (the mean of zero and variance of 1).

4.2.2 Character-based features

Character-based features have been considered as the most discriminate stylometric features to identify the author of disputed text. Unlike word level n-grams, the character level n-grams work well with limited data. This is in addition to the ability of the character n-grams to capture nuances in lexical, syntactical, and structural level. We consider two types of character features: punctuation marks and character-level n-grams. According to Arabic Orthography, we considered eleven punctuation marks (،؛؟!.:-" ([{). As a result, the dimension of the punctuation marks feature vector (PM) is 11.

At the character level n-grams features, we aim to analyze the various character n-grams (n=1-4) effects in identification of the author of unattributed documents in Arabic. We used the following equation to compute uni-gram model:

$$P(ch_i) = \frac{Count(ch_i)}{\sum_{i=1}^{n} Count(ch_i)}$$

While we used the following equation to compute character n-grams for n=2-4:

$$P(ch_n \mid ch_{n-N+1}^{n-1}) = \frac{Count(ch_{n-N+1}^{n-1} ch_n)}{Count(ch_{n-N+1}^{n-1})}$$

We ignored character n-grams features that occur less than 75 in the whole training set. The features below the threshold are considered as noisy or typos. As a result, the dimensions of the character uni-grams (ch1gram), character bi-gram (ch2grams), character tri-grams (ch3gram) and character quad-grams (ch4gram) feature vectors are 61, 969, 4922 and 7178, respectively. The ch1grm, ch2grm, ch3gram, and ch4grams are considered as basic line feature sets. As extended feature set, we concatenated ch1gram and ch2gram to obtain combined character uni-grams and bi-grams feature vector (ch12gram) with the size of 1030 and combined ch2gram and ch3gram to obtain combined bi-grams and tri-grams (ch23gram) feature vector with the size of 5891. We also concatenated ch12grams with ch3grams to obtain combined uni-grams bi-grams and tri-grams (ch123gram) feature vector with the size of 5952. We did not consider combining the ch4grams with the other n-gram features because of the huge size of the generated vector.

4.2.3 Syntactic features

As syntactic features, we consider function words. Such features do not contain information about the documents content and serve to express the grammatical relationships with other words within a sentence. Prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles are examples of function words. The advantage of using function word features lies in that they are topic-independent (i.e. they are writing style markers). However, there is no agreement on a general list of function words for AA purpose. We proposed a general set of 309 function words (FW) including conjunctions (أو، أم، ثم، ...), pronouns (أنا، نحن، أنت، ...), Nedaa (يا، أيا، أي), question words (ماذا، لماذا، متى، ...), time adverbials (يوما، غدا، مؤخرا، ...), place adverbials s (نحو، بين، أمام، ...), prepositions (من، إلى، عن، ...), particles (إن، أن، كأن،...), etc. However, we do not distinguish between homographs, such as, (مَنْ who) and (مِنْ from) as distinguishing such words requires also advanced NLP tools such as part of speech tags which is still immature for Arabic. The proposed function words are listed in Appendix A. Each function word is represented as follows.

$$\frac{count(w)}{\sum_{w'\in FW} count(w')}$$

4.2.4 General feature vector

We combined WR as lexical features, ch1grams as character features and FW as syntactic features to obtain general feature vector (*GFV*) with size of 382 features.

Table 4-2 shows the feature vectors.

Table 4-2 The descriptions of features vectors

	Type of features	Name of vector	Num. of features	Explanation
1	**Lexical**	WR	12	Vocabulary Richness
2		*Word1GTH49fv*	1181	Words with >= 50 occurrences
3		*Word1GTH99fv*	473	Words with >= 100 occurrences
4		*Word1GTH149fv*	287	Words with >= 150 occurrences
5		*SpecifcwGTH2fv*	1193	Specific words occur >2 in the author's profile
6		*SpecifcwGTH3fv*	713	Specific words occur >3 in the author's profile
7		*CoSpW*	1000	Combined 4 & 6
8	**Character Features**	PM	11	Punctuation Marks
9		Ch1gram	61	Character uni-grams
10		Ch2gram	969	Character bi-grams
11		Ch3 gram	4922	Character tri-grams
12		Ch4gram	7178	Character quad-grams
13		Ch12gram	1030	Combined uni-grams bi-grams (9 & 10)
14		Ch23gram	5891	Combined bi-grams quad-grams (10 & 11)
15		Ch123gram	5952	Combined uni-grams bi-grams tri-grams (9-11)
16	Syntactic	FW	309	Function Words
17	General	GFV	382	General feature vector (1 & 9 & 16)

4.3 Feature Selection Phase

Applying stylometric features such as lexical, character and syntactic features generate a huge number of features especially when those features are concatenated to generate combined feature vectors. Extracting the most discriminating features require a suitable feature selection techniques to select the most distinguishing features among the writing styles of authors.

Feature selection is conducted by searching the space of attribute subset, evaluating each one through combining a feature evaluator with a search method. To come up with the best combination of attribute subset evaluators and search methods, we conducted comparative analysis for variety of feature selection techniques through combining attribute subset evaluators and search methods. As feature evaluator, we used Correlation-based Feature Subset Selection, Consistency-based Subset evaluation, Principal Components-based Subset evaluation and Chi-Square Attribute evaluation along with four search methods (viz. Best First, Genetic Search, Rank Search and Ranker)

Two modes can be applied to conduct the feature selection either using the full training set or by cross-validation. In our work, we used the full training set in the feature selection mode.

4.3.1 Evaluators

- Correlation-based Feature Subset Selection (CFSS) is a filter method which assesses the predictive ability of each attribute individually and the degree of redundancy among them. It prefers sets of attributes that are highly correlated within the class while have low inter-correlation (Hall, 1999).

- Consistency-based Subset evaluation (CBSE) is a filter method which evaluates attribute sets by the level of consistency in class values when the training instances are projected onto the set (H. Liu & Setiono, 1996).

- While the previously mentioned evaluators are attribute subset evaluators, the remaining ones are single-attribute evaluators.

- Principal Components-based Subset evaluation (***PCA***) performs principle components analysis and transforms the set of attributes into a reduced set (Binongo, 2003; J. F. Burrows, 1987, 1992; Savoy, 2013b).

- Chi-Square Attribute evaluation (ChiSAE) evaluates attributes with respect to the class through computing the Chi-square statistic(Witten & Frank, 2005).

4.3.2 Search methods

- Best First (BF) performs greedy hill climbing with backtracking facility. The parameters of BestFirst search method include search dirction. The seach direction may be forward, backword or bi-dirctional. forward starts from the empty set of attributes, backword starts from the full set, and bi-dirctional search in both directions by considering all possible single-attribute additions and deletions. We used forward diection as defult parameter.

- Genetic Search (GS) uses a simple genetic algorithm to perform the search (Goldberg, 1989). The parameters of genetic algorithms include crossover probability, max generations, mutation probability, and population size. We used the default parameters such that crossover probability is 0.6, the max generations is 20, the mutation probability is 0.033, and the population size is 20.

- Rank Search (RS) sorts features using an evaluator of single-attribute and then ranks promising subsets using an attribute subset evaluator(Witten & Frank, 2005). To fasten the selection procedure, we used, simple single-attribute evaluator, GainRatioAttributeEval.

- Both Best First and Greedy Stepwise apply greedy hill-climbing but the first method is with backtracking while the second method is without backtracking. However, they select the same features, as we observed, so we used Best First search method.

- The previous methods are used with attribute subset evaluators while the following method is not a search method and it is used with single-attribute evaluator.

- Ranker (R) is a ranking scheme for single attributes and it sorts attributes based on their individual evaluations. In addition to ranking attributes, it performs attribute selection through removing the attributes with lower ranks through setting a cutoff threshold below which attributes are discarded, or specifying the number of attributes to retain. As cutoff threshold we used the default value (-1.8) with PCA. Regarding ChiSAE, we adopt the cutoff threshold to zero since the default threshold does not make any reduction in the size of the features, it just ranks the features. That means 216 features have the ranks of zero (subtracting the number of selected features with default threshold and the selected features with the defined threshold 309-93=216). In other words, the default threshold causes no features being discarded.

4.3.3 Feature selection evaluation

To evaluate the selection feature techniques we employed them to select the most discriminate features from a collection of 309 function words. To evaluate the selected feature set, we used five different machine learning classifiers, namely, Logistic, voting feature intervals (VFI), MLP, SMO and LS-SVM.

The number of features obtained when applying the combination of CFSS and BF is 65 features, while the number of features selected when combining CFSS with GS and RS are 63 and 93, respectively. When applying the combination of CBSE and BF the number of selected features is 16 while when combining CBSE with GS and RS the number of selected features are 82 and 58, respectively. Applying PCA yields 220 selected features while the selected features when applying ChiSAE is 93 features. Figure 4-3 shows the number of selected features obtained using the feature selection techniques.

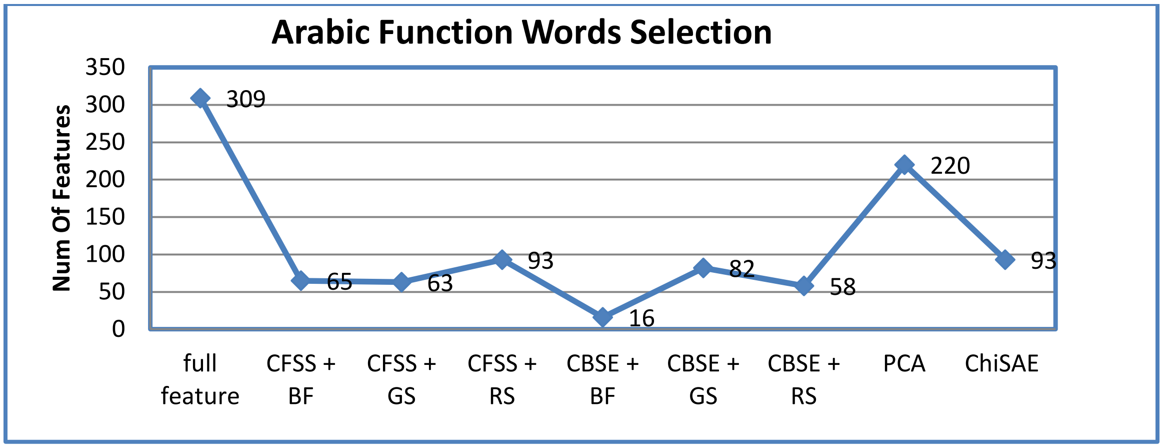


Figure 4-3 The number of selected function words using the feature selection techniques

After conducting word level preprocessing operations (i.e. anything other than Arabic Alphabet are eliminated including punctuations and the diacritical marks), we used 700 documents (70%) for training and 300 documents (30%) for testing.

As shown in Table 4-3, best accuracy of 90.33% is achieved using 93 selected features using the combination of CFSS and RS with MLP classifier. These selected function words are listed in Appendix B. In the case of Logistic classifier, best accuracy of 84.33% is obtained using 65 selected features (the combination of CFSS and BF) compared with 70% when using the full features. Using MLP and SMO, best accuracy rates of 90.33% and 88.67% respectively, are achieved with 93 features that are selected using a combination of CFSS and RS while the accuracy of the more the full features with MLP and SMO are 89.33% and 87.67%, respectively. VFI has the poorest performances over all; however, the highest accuracy of 66.67% (using VFI) is achieved using both the combination of CFSS with RS and ChiSAE compared with 64.33% using full features. With LS-SVM, best performance of 86% is achieved using the full features compared with 81.67% using the selected features (using a combination of CFSS and

RS). Figure 4-4 shows the achieved results using the selected classifiers with the feature selection techniques.

Table 4-3 The classification results

Feature Selection Techniques	Logistic	MLP	SMO	VFI	LS-SVM	Avg
CFSS + BF	**84.33**	85.33	85.67	64.67	77.67	79.53
CFSS + GS	61.33	70.33	71.67	52.67	64.00	64
CFSS + RS	79	**90.33**	**88.67**	**66.67**	81.67	**81.27**
CBSE + BF	58.67	48.67	53.67	43.67	50.33	51.00
CBSE + GS	48.67	58.67	63.67	45	57.00	54.60
CBSE + RS	60.67	82.33	79.33	61.67	72.67	71.32
PCA	75.33	84.33	81.67	62.6	80.00	76.79
ChiSAE	79	89.33	88.33	**66.67**	80.67	80.8
Full Features	70	89.33	87.67	64.33	**86.00**	79.47

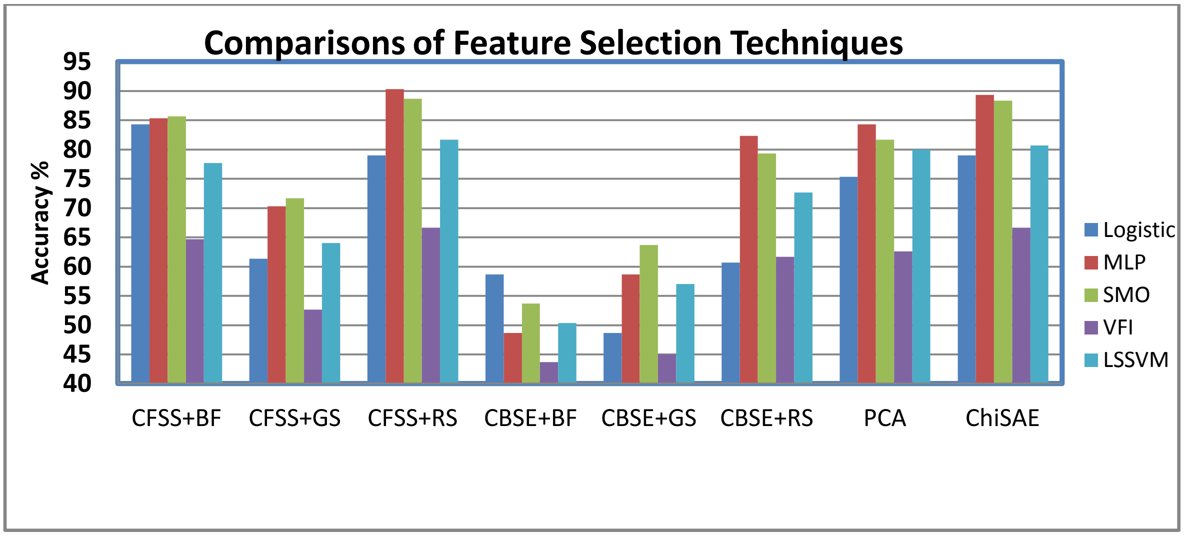
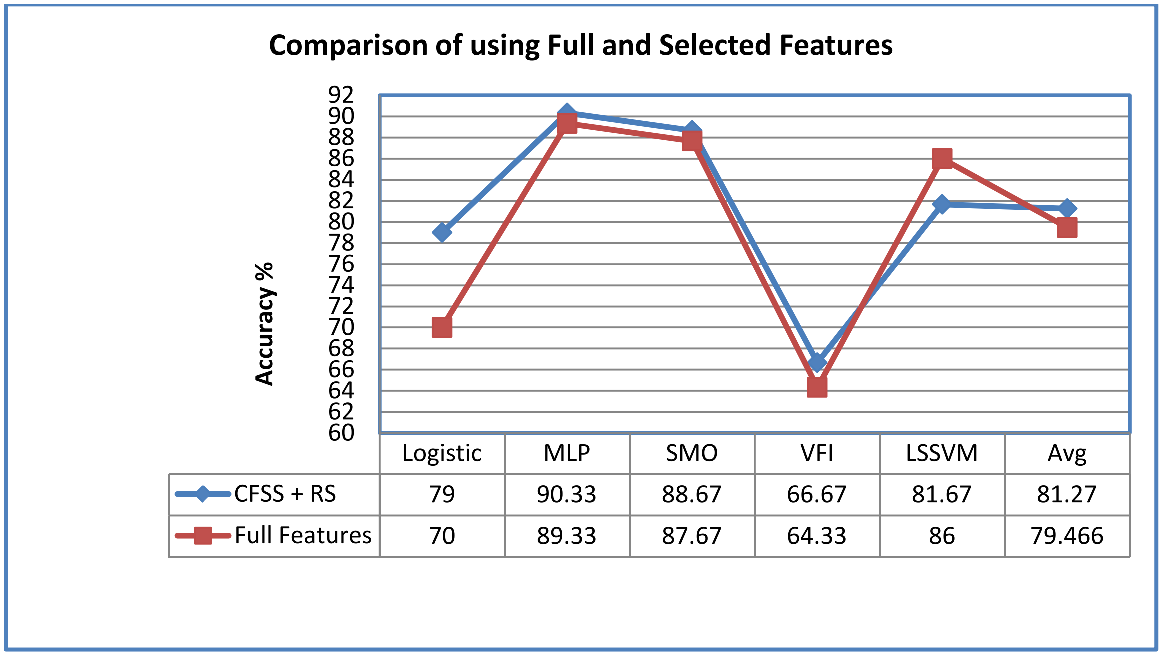


Figure 4-4 Comparisons of feature selection techniques using five machine learning classifiers

It can be observed that, generally, the combination of CFSS and RS outperforms other selection feature techniques under consideration. It also achieves better accuracy than using full features in most cases. The classification ratios of using the full features are 70% with logistic, 89.33% with MLP, 87.67% with SMO, 64.33% with VFI and 86% with LS-SVM while a total of 93 features selected using a combination of CFSS and RS has the accuracy of 79% with logistic, 90.33% with MLP, 88.67% with SMO, 66.33% with VFI and 81.27% with LS-SVM, as shown in Figure 4-5.



	Logistic	MLP	SMO	VFI	LSSVM	Avg
CFSS + RS	79	90.33	88.67	66.67	81.67	81.27
Full Features	70	89.33	87.67	64.33	86	79.466

Figure 4-5 Comparison of using full and the selected features using CFSS and RS with five machine learning classifiers

4.3.4 Selected feature vectors

Based on this comparative analysis, we will use both CFSS with RS and ChiSAE as feature selection techniques on stylometric features in this work. In addition, we use CFSS with BF as a base line to compare with the work of (Türkoğlu et al., 2007).

When applying the feature selection techniques on ch2gram feature vector the number of selected features dropped from 969 to 137 features using a combination of CFSS and BF while it decreased to 231 and 269 features using a combination of CFSS with RS and ChiSAE, respectively.

Regarding ch3gram feature vector, it went down dramatically from 4922 features to 248 using a combination of CFSS and BF. This is also true when applying a combination of CFSS with RS and ChiSAE such that the ch3gram feature vector decreased significantly to 831 and 877 features, respectively. The ch4gram feature vector is also reduced sufficiently from 7178 features to 237 using a combination of CFSS and BF; it also dropped to 1361 features using both a combination of CFSS with RS and ChiSAE.

Table 4-4 The number of full and selected features

	Full features	BF	RS	ChiSAE
FW	309	65	93	93
Ch1gram	61	31	44	44
Ch2gram	969	137	231	269
Ch3gram	4922	248	831	877
Ch4gram	7178	237	1361	1361
Ch12gram	1030	145	272	313
Ch23gram	5891	279	1035	1146
Ch123gram	5952	285	937	1190
GFV	382	99	148	149

Therefore, we have 27 different reduced feature vectors as shown in Table 4-5.

Table 4-5 Names of full and selected feature vectors

Description	Full feature vector	Selected features using CFSS+BF	Selected features using CFSS+RS	Selected features using ChiSAE
Function Words	FW	SFWBF	SFWRS	SFWChi
Character uni-grams	Ch1gram	SCh1BF	SCh1RS	SCh1Chi
Character bi-grams	Ch2gram	SCh2BF	SCh2RS	SCh2RS
Character tri-grams	Ch3gram	SCh3BF	SCh3RS	SCh3RS
Character quad-grams	Ch4gram	SCh4BF	SCh4RS	SCh4RS
Combined uni-grams bi-grams	Ch12gram	SCh12BF	SCh12RS	SCh12RS
Combined bi-grams tri-grams	Ch23gram	SCh23BF	SCh23RS	SCh23RS
Combined uni-grams bi-grams tri-grams	Ch123gram	SCh123BF	SCh123RS	SCh123RS
General feature vector	GFV	SGFVBF	SGFVRS	SGFVRS

4.4 Conclusions

Several types of stylometric features including lexical, character and syntactic features are extracted as basic features. The extracted features are concatenated to form several concatenated feature vectors. Applying lexical, character and syntactic features generates high dimensionality feature vectors especially when those features are concatenated to generate the combined feature vectors. To overcome the problem of high dimensionality of feature vectors we applied feature selection techniques to select the most discriminative features and to reduce the size of these feature vectors. To determine which of these techniques perform well, we carried out a case study on these selected features using five different classifiers. Our results show that a combination of CFSS with

RS and ChiSAE techniques tend to outperform other feature selection techniques. As a result, we used these techniques in our experiments as described in Chapter 6. This is in addition to using a combination of CFSS with BF as base line as it is the default technique and is used in previous work.

CHAPTER 5

ARABIC SEMANTIC FEATURES CONSTRUCTION AND EXTRACTION

In this chapter we extract new stylistic features for Arabic based on the usefulness of information about the style of writing. Arabic is rich with its syntax and rhetorical styles that serve to express or provide knowledge in analyzing and understanding the language. In the topic under consideration (AA), such information can express and define the author’s styles at advanced levels including meaning of expressions, purposes, feelings, and rhetorical styles.

Semantic features have seen limited use. Argamon et al. (2007) addressed English AA based on the Systemic Functional Grammar (SFG) theory which is a functional approach to linguistic analysis (Halliday, 1994). Stamatatos (2009) have considered the work of (Argamon et al., 2007) as the most significant method to employ semantic information. We are not aware of any work utilizing semantic information for Arabic AA. The richness of Arabic in grammatical and rhetorical styles can express the author’s styles at the levels of meaning of expressions, purposes and feelings. These can be defined as semantic features.

Natural language processing (NLP) tasks can be divided into three levels (viz. low-, medium- and high-level tasks). Each level requires special NLP tools to extract the information. Low- and medium- level tasks include tokenizing, streaming, orthographic spell checking, sentence splitting, POS tagging, text chunking, and partial parsing. High-level tasks include full syntactic parsing, semantic analysis, or pragmatic analysis which may be still immature for Arabic. There are insignificant works that have been conducted to utilize advanced stylometric features. These studies are surveyed in (Efstathios Stamatatos, 2009). These techniques resulted in poor accuracy rates. To our knowledge, this is the first work to define such Arabic semantic features and to apply them to AA

This chapter is organized as follows, in section 5.1 we construct the lexicon; we define our Arabic semantic features in Section 5.2; Section 5.3 describes semantic features' extraction algorithm, and we finally conclude the Chapter in Section 5.4.

5.1 Lexicon Construction

We classify the content of our lexicon into several groups based on the type of its elements. Roots and particles consist of one token while a phrase is composed of more than one token. We consider the instances of these types of elements in their base and their derived forms. Derived forms include all forms that have the same meaning of the base form (root) even if they differ in their POS-tags (as our focus is on its semantic meaning, not its syntactic form) as shown in Table 5-1.

Table 5-1 **Examples of base and derived forms as used in this work**

Base form	type	Derived forms			
		Prefixes	Suffixes	Affixes	Stem/lemma
إضافة إلى	phrase	وإضافة إلى/ بالإضافة إلى	-	-	-
إضافة إلى أن	phrase	وإضافة إلى أن	إضافة إلى أنه	وإضافة إلى أنه	-
ظن	root	يظن/ وظن/ فظن/...	ظننت/ ظنوا/ ظننتم/ ...	يظنون/ تظنين	يظنون/ تظنين/...
كان	root	فكان	كانت	فكانت	يكون/ يكونون/ تكونين/...
إن	particle	فإن	إنهم	فإنهم	-

We listed the forms that can be derived from the roots. For example, the derived forms of the root “حَسِبَ /Hassib/ (he thought)” are “ يحسبون، تحسبون، تحسبين، تحسبان، يحسبان، تحسبينهم، حسبتهم، تحسبهم، ...”. The characteristic of particles differ from roots in that they have the ability of being the suffix and/ or the prefix to the base form. For example, “إن” with a suffix “إنه” and with a prefix “فإن”. While in the case of roots, the root form is changed such as “قال /qaal/ (he said)” is changed to “يقول /yaqool/ (he says)”. This is in addition to adding prefixes and suffixes. We collected most of the base phrases and handled their derived forms including prefixes and suffixes.

To capture most of the semantic information of writing styles, we investigated as many expressions as possible that express semantic information. We included most forms of each element (elements here mean root, particle or phrase). Considering the basic form of each element is not enough to capture sufficient styles. Additionally, listing all derived forms of each basic form of elements is a hard task. So, we listed some of the derived forms and processed others based on heuristics and rules. For example, if the element consists of more than two tokens then any derived phrase that contains these tokens is expected to have the same semantic. This is not true regarding most elements with one token. For example, if any term in a document contains the term “سوف /Sawfa/ (will)” and is considered as “affirmation” feature, then this rule will be true for the terms “وسوف”, “فسوف” while it will include other wrong terms semantically like “كسوف /Kussoof/ (occultation)” and “خسوف /Khussoof/ (eclipse)”.

The elements “حبذا” and “لا حبذا” have different meaning such that the first element is praise style (أسلوب المدح) while the second one belongs to vilification style (أسلوب الذم). The phrase “لا حبذا” when analyzed will be considered as three different styles namely praise style (حبذا), negation (لا) and vilification style (لا حبذا). Another example is analyzed in Table 5-2 such that the phrase “على الرغم من أن” should be considered as contrary to reality style and should be counted as one feature. However, this phrase when analyzed will be considered as several different meanings and is counted as nine features.

Table 5-2 An example of noisy semantic features

The main expression	على الرغم من أن							
The correct meaning	**Contrary to reality**							
Counted as	**One feature**							
Sub expressions	على	على الرغم	على الرغم من	على الرغم من أن	من	من	من	أن
The noisy meanings	**proposition**	**Contrary to reality**	**Contrary to reality**	**Contrary to reality**	**proposition**	**Question style**	**Conditional style**	**Affirmation**
Counted as	**Eight features**							

Considering these elements as is may lead to confusion of the author's style. Such elements require special processing to utilize them properly. We first sort the elements of the lexicon based on the word n-grams and then based on the character n-grams in descending order. Some elements need to be eliminated after counting them. Additionally, some elements need to be matched completely as they may contradict with other terms and some need to be matched partially as they have the same meaning. In addition some articles in Arabic have more than one meaning and distinguishing them is not addressed here. These articles are listed as follows.

- مَنْ (who) is question word and conditional particle while مِنْ (from) is a proposition.
- "إنَّ" is an affirmative style and "إنْ" is a conditional style

- “أي” is a Nedda style, conditional style, expository, question words and quantifiers
- “أما” is a additive adversative expressions and prompting and presenting style
- “إذا” is a conditional style and “إذًا” is a result expressions
- “نَعَمْ” is an answering term while “نِعْمَ” is a praise style.

5.2 Semantic Features Set

We define 39 semantic features (SF) as shown in Table 5-3. The semantic features include the most popular syntax and rhetorical styles in Arabic.

Table 5-3 Arabic semantic lexicon

#	Name of feature	Examples of elements that indicate the style	Comments
1	Additive positive	أيضا، كذلك، بالإضافة إلى أن، بالإضافة، إضافة إلى	link two clauses or sentences that share the same idea
2	Additive adversative	لكن، بل، إلا أن، بينما، غير أن، على العكس من ذلك	link two clauses or sentences that have different ideas
3	Contrary to reality (مخالفة الواقع)	رغم أن، برغم ، بالرغم من أن، على الرغم من أن، مع أن	contrast facts or conflict with rules and admitted matters
4	Results	نتيجة لهذا، نتيجة لذلك، نتيجة ذلك، وعلى هذا، ولهذا، إذن، لذلك، بناء عليه	
5	Causes	كي، حتى، لكيلا، لئلا، لذا، بسبب، بفضل، نظرا ل، لأن، وحيث أن	
6	Doubt and likelihoods	ربما، يبدو، من المحتمل أن، من الممكن، على الأرجح، ممكن ، ظن، حسب، خال، زعم	
7	Necessity and requirements	ينبغي ، يقتضي، يجب، يؤدي إلى، يؤدي، يتطلب، من الضروري، يستوجب	

#	Name of feature	Examples of elements that indicate the style	Comments
8	Examples expression	من هذا القبيل، نحو، مثلا، كما، على سبيل المثال، مثلما، مثال،	
9	Determinism and certainty	حتميا، تماما، بالضبط، بدقة، بالفعل، في الحقيقة، في الواقع، في واقع الأمر، طبيعة الحال، لا محالة، شكل محتوم، بالتأكيد، من المؤكد	
10	Conclusions and summaries	أخيرا، ختاما، في الختام، في النهاية، باختصار، بايجاز، خلاصة القول، في الأخير	
11	Expository	أعني، بعبارة أخرى، معنى ذلك، المراد، أقصد، أي أن، أي، هذا يعني	
12	Particularization	بالأخص، خصوصا، على شكل خاص، بشكل خاص	
13	Generalization	عموما، بشكل عام، على العموم، عامة، على وجه الإجمال، في مجملها، على وجه العموم	
14	Undesirable	لسوء الحظ، من المؤسف ، للأسف، من المحزن، يؤسفني، يحزنني	
15	Desirable	لحسن الحظ، من حسن الحظ، يسرني ، يسعدني	
16	Prediction	لا يثير الدهشة، ليس من المدهش، ليس من الغريب، مما لا يثير الدهشة، يمكن التنبؤ به، متوقع	
17	Surprising	من المدهش، من الغريب، من المستغرب، بدهشة، بانذهال، فجأة، من غير المتوقع، بشكل مفاجئ	
18	Approval (الإقرار والإثبات)	من المقرر، من الثابت، من المعروف	
19	Adverbial accusative of cause or reason (المفعول لأجله)	عنوة، طواعية، قسرا، هدرا، شكرا، تمشيا، تحسبا، وفقا، ابتغاء، خشية، لأجل	explain the motivation, reason or purpose of the verb
20	Affirmation (التوكيد)	إن، أن، سوف، لقد، كلاهما، كليهما، كلتيهما، كلتاهما، جميعهم، جميعا، عامتهم، نفسه، عينه، أجمعون	strengthen the expression and discourses

#	Name of feature	Examples of elements that indicate the style	Comments
21	Quantifiers[2]	كل، كلا، كلتا، جميع، معظم، غالب، عامة، جل، بعض، عدة، بضع،	specify and determine other nouns
22	Time adverbials	من حين لآخر، الآن، اليوم، يوما، نهار، غدا، أمس، بالأمس، مؤخرا، مرارا، مساء، الليلة، حينذاك	
23	Place adverbials	نحو، بين، أمام، خلف، وراء، فوق، تحت، عند، أسفل، أعلى، حول، وسط	
24	Questions	ماذا، من، لماذا، متى، أين، كم، كيف، أي، هل	
25	Conditions	إن، من، ما، مهما، متى، أين، أينما، أنى، حيثما، كيفما، أي، إذ، لو، كلما، لولا، إذا	
26	Negation	ما، لم، لن، لا، ليس	
27	Exceptions	إلا، ما عدا، عدا، سوى، غير، ما خلا، خلا، حاشا، باستثناء	
28	Prompting, presenting (العرض والتحضيض)	هلا، لولا، لوما، ألا، أما	particles that drew the attention of the addressee is followed by requested sentences.
29	Vilification style (أسلوب الذم):	لا حبذا، بئس، ساء	to vilify someone or something
30	Praise style (المدح)	حبذا، نعم، حسن	to praise someone or something

[2] Quantifiers: These are specific nouns or terms in Arabic that specify and determine other nouns. These names may express quantities, majorities, partitions or other types of specification; for examples “كل”, “جميع”, “معظم”, etc. We should differentiate between these nouns (e.g. “جميع”, “كل”) and the terms of affirmation (e.g. “جميع”, “كل”). Assume we have two sentences, “جميع الطلاب حضروا” and “ الطلاب حضروا جميعهم”; “جميع” in the first sentence refer to quantifiers while in the second indicates affirmation. This is also true regarding “كل” and “كلا”. Such concerns are taken into account such that according to Arabic syntax (or Grammar) theory, “جميع”, “كل” and their sisters should be suffixed by personal pronouns to be considered as affirmation like “جميعهم”, “كلاهما”, “كلهم”, etc.

#	Name of feature	Examples of elements that indicate the style	Comments
31	Needa style (أسلوب النداء)	يا، أيا، أي، هيا	call and drew the attention of addressee
32	Answering	نعم، بلا، أجل، بلى، إي، لا	
33	Demonstrative pronouns	هذا، ذا، ذاك، ذلك، تلك، هذه، هذان، ذانك، هؤلاء، أولاء	
34	Relative pronouns	الذي، التي، اللذان، اللذين، اللتان، اللتين، الذين، اللائي، ما، من	
35	Exclusivity style[3] (أسلوب الحصر)	إنما	it expresses that a matter or subject solely belongs to particular thing with no sharing
36	Hopefulness and wishful thinking[4] (الرجاء والتمني)	لعل، ليت، أرجو، رجاء، تمنى، عسى، لو	
37	Incomplete verbs[5] (كان وأخواتها)	كان، أمسى، أصبح، أضحى، ظل، بات، صار، ليس، ما زال، ما انفك، ما فتئ، ما برح، ما دام	Called sisters of verb 'to be'.
38	Propositions	من، إلى، في، عن، على، حتى، منذ، رب، خلا، عدا، حاشا	
39	Deceleration verbs[6]	أجاب، صرح، عبر، أفاد، قال، عقب، سأل، روى، رد، حدث	

[3] The difference between the exception style and the exclusivity style is that the exception style is grammatical style while the exclusivity style is Rhetorical style.

[4] Hopefulness and wishful thinking (الرجاء والتمني): The difference between hopefulness and wishful thinking is that when one want something to happen; if this thing is possible then it is called hopefulness while if it is not possible or very difficult then it is called wishful thinking. Arabic has variety of articles, terms and clauses that refer to them such as "لعل" for wishful thinking and "ليت" for hopefulness.

[5] Incomplete verbs (كان وأخواتها): They are also called verbs of being, becoming, remaining and seeming which are similar in the meaning and syntactic effect. These verbs describe states of existence such as being, inception, duration, and continuation

5.3 Arabic Semantic Feature Extraction

There are techniques that have the ability to extract the expressions when they are embedded with suffixes and prefixes. However, these techniques cannot be applied in our study as is. These techniques work well with some cases, for example “غير (except)” can be derived into several forms with the same semantic such as “أغير”, “غيرها”, “بغير”,etc. However, they fail in other cases, for example “يغير (he changes)” that has a completely different meaning. We notice that this issue is more with terms whose size is 2-5 letters and is less with larger sizes. Therefore, we determine the phrases and the terms with large size which do not conflict with other expressions or terms when applying such techniques. Examples of these elements are “بالإضافة إلى أن”, “مما لا يثير الدهشة”, “ ليس من المستغرب”, “بصراحة”, etc. They retain their meanings even with addition of suffixes, prefixes or infixes. Our Arabic semantic feature extraction algorithm is shown in Figure 5-1.

[6] Deceleration verbs: Such information can characterize an author through either the author write his ideas or declare other opinions. In other words, when these verbs appear frequently in a document (the author just reports or rewords speeches or writings of others otherwise the discourse represent the writer’s opinions and ideas).

```
Algorithm Arabic Semantic Feature Extraction
  Input:
          SF: Sorted elements of the semantic features in descending order based on the length of word n-grams and then the length of character n-grams.
          D: preprocessed document after applying word-level preprocessing operations to remove digits, punctuation and diacritical marks, special or noisy symbols and non-Arabic symbols and alphabets
  Output:
          SFV: vector of normalized values of semantic features

  Begin
      SFV(1:39)=InitializeWithZEROES()
      For each Element E in SF
          if IsMatchedCompletly(E) then
              El=E
          else
               El=*E*   // *  means any prefixes and/ or suffixes
          end if
          if IsFound(El, D) then
              styleNo=GetStyleNO(E)
              SFV (styleNo)++;
              if GetWordNGramsLength(El)>=2
                Eliminate(El, D)  // eliminate the element El from document D
              end if
          end if
      end for
      SFV=NormalizedFeatures(SFV)
  end
```

Figure 5-1 The Arabic semantic feature extraction algorithm

For more details, the algorithm can be described as follows.

1. Apply word-level preprocessing operations which eliminate digits, punctuation and diacritical marks, special or noisy symbols and non-Arabic symbols and alphabets
2. Label each expression according to its category or semantic style
3. Sort the list of expressions according to the word n-grams and then the character n-grams in descending order.
4. Determine which of those elements of the lexicon that should be matched completely and which should be eliminated after matching (as heuristic values)
5. Search the elements of lexicon in a document
6. Based on the heuristic, determine whether the element must be matched in the given document completely.
7. Compute the counts of each feature (or style).
8. If the element contradict with the remaining elements then eliminate it.
9. Reaped steps 5 to 8 for all texts.

5.4 Conclusions

In this chapter, we defined a set of the most popular syntax and rhetorical styles of Arabic that have the ability to characterize an author and reflect writing style of the author.

We first look for expressions of Syntax and Rhetorical styles through various references of Arabic (Abdullatif, Omar, & Zahran, 2005; Othaimeen, 2005) and websites. To come up with semantic features that are as effective as possible, we also collect other elements of Arabic and classify them based on their meaning. We discussed the concerns of representation and extraction of these features and handled them effectively. We constructed the system of Arabic semantic features which aids for Arabic language processing.

Improving this system through providing several semantic levels instead of one level is our future work. Additionally, we aim to apply it on other authorship analysis tasks such as authorship characterization and similarity detections.

CHAPTER 6

EXPERIMENTS AND RESULTS

Several techniques have been developed for authorship attribution. These techniques differ in the used features, classification methods, natural language, corpora, and methodologies.

We carried out various experiments using our Arabic authorship attribution corpus which is composed of 1000 newspaper articles written by 20 authors in several topics. We investigated several stylometric features including vocabulary richness, word frequency, specific words, character level n-grams, punctuation marks. We investigated the proposed set of function words. We also investigated the combinations of those features. We used several classification methods to evaluate these features namely, Euclidian Distance(ED), K-Nearest Neighbors (K-NN), Delta rule, Multi-Layer Perceptron (MLP), Least Squares Support Vector Machines (LS-SVM) and Sequential Minimum Optimization based Support Vector Machine (SMO). We also investigated the effects of feature normalization methods. We also evaluated training set representation methods including profile-based method and instance-based method. Moreover, we applied several feature selection techniques to discriminate the most effective features that have the ability to identify the author of a given text. We also compared our work with the most related works and we showed that our work compares favorably with published works. We achieved accuracy rates exceeds 95% in many cases on our corpus.

This chapter is organized as follows. Section 6.1 describes the classification methods; the experiments setup is presented in Section 6.2; Section 6.3 shows the experimental results and we summarized the chapter in Section 6.4.

6.1 Classification Methods

Several classification methods have been used for the authorship identification task. In this thesis, we selected five classifiers namely ED, K-NN, Delta rule, LS-SVM, MLP, and SMO.

ED, K-NN and Delta method are distance-based methods that compute the distance between a new pattern with existing instances in the training set. SMO and LS-SVM are support vector machine classifiers. The last one (MLP) is using back propagation artificial Neural Network classifier.

The Euclidian distance is computed using the following equation.

$$D_i = \sqrt{\sum_{j=1}^{n} (x_{ij} - y_j)^2}$$

Where:

- D_i : is the distance between the test sample feature vector and the feature vectors of all models.

- $x_{ij:}$ is the j^{th} feature of the feature vector of model i.

- $y_{j:}$ is the jth feature of the feature vector of the current test sample.

- n is the number of features

K-NN classifies a new pattern based on their similarity to the patterns in the training data. Determining the class of the new instance is by majority vote of its metrically nearest neighbours. For using K-NN classifier, two objects should be set up the value of K (the number of nearest neighbours) and the distance measure such that the default parameters are k=1 and the Euclidean distance. We used K-NN method with value of k=3 as there is no assumptions that are made about the probability distribution of the features and it is suitable for data with complex boundaries between classes (Aha, Kibler, & Albert, 1991). We evaluated both Euclidean and *city block* distances as distance measures of K-NN. *City block* distance is computed using the following equation:

$$D_i = \sum_{i=1}^{n} \left| x_{ij} - y_j \right|$$

The Delta rule have been applied in previous works (Eder & Rybicki, 2013; Eder, 2010, 2013; Savoy, 2012a, 2013a).

- Create author’s profile using the training corpus
- Standardize features using Z-scores such that:

$$Z\text{ - }score(t_{ij}) = \frac{tfr_{ij} - \mu_i}{std_i}$$

- Compute the distance (Delta rule) as follows:

$$\Delta(Q, A_j) = 1/m \sum_{i=1}^{m} \left| Z\text{ - }score(t_{iq}) - Z\text{ - }score(t_{ij}) \right|$$

Where:

- Q is the unattributed document
- A_j is the different authors profiles
- *m* is number of terms

SVM is a powerful classifier and achieves high identification rates in previous works. In our work we used LS-SVM. LS-SVMs are reformulations of the standard SVMs (Suykens & Vandewalle, 1999; Van Gestel et al., 2004) which lead to solving linear Karush-Kuhn-Tucker (KKT) systems (Bradante et al., 2011). LS-SVMs are closely related to regularization networks (Evgeniou, Pontil, & Poggio, 2000) and Gaussian processes (Wahba, 1990) and they also stress primal-dual interpretations. We used LS-SVM with *RBF_kernel* kernel function.

SMO (Platt, 1999) is an algorithm used to speed up the training of SVM through breaking a very large quadratic programming (QP) optimization problem in SVM into a series of smallest possible QP problems. This in turn avoids using a time-consuming numerical QP optimization as an inner loop. The parameters of SMO include the kernel function. We used SMO with *Polykernel* kernel function. SMO have been used in previous works and achieved suitable identification rates (Argamon et al., 2007; Ouamour & Sayoud, 2012; Türkoğlu et al., 2007). Türkoğlu, Diri, & Amasyalı (Türkoğlu et al., 2007) compared several classification methods namely, SMO, Naive Bayes, Random Forest, K-NN, and

MLP that are implemented on WEKA with its default parameters. They reported that SMO achieves higher accuracy rates on Turkish corpora and MLP also achieves good performances. We used both MLP and SMO on WEKA (Witten & Frank, 2005) with its default parameters.

Multilayer Perceptron or MLP is a back propagation neural network. Its parameters include hidden Layers, the learning rate, its momentum, and the number of epochs. The hidden layers present and the number of their nodes are defined by the hidden Layers parameter. In our work we used average of the number of attributes and the number of class values as the value of this parameter. For example the number of features in PM is 11 and the number of classes is 20 then the number of the nodes in the hidden layer is (11+20)/2 =15 as shown in Figure 6-1. The network in Figure 6-1 has three layers: an input layer, hidden layer and output layer. The input layer is on the left in green rectangular box (11 attributes) which is connected to a hidden layer. The hidden layer is represented by red nodes (15 nodes) which are connected to the output layer. The output layer is on the right in orange rectangular boxes (20 classes). For the other parameters we used the value of 0.3 for the learning rate, 0.2 for the momentum parameter and 500 for the number of Epochs parameter.

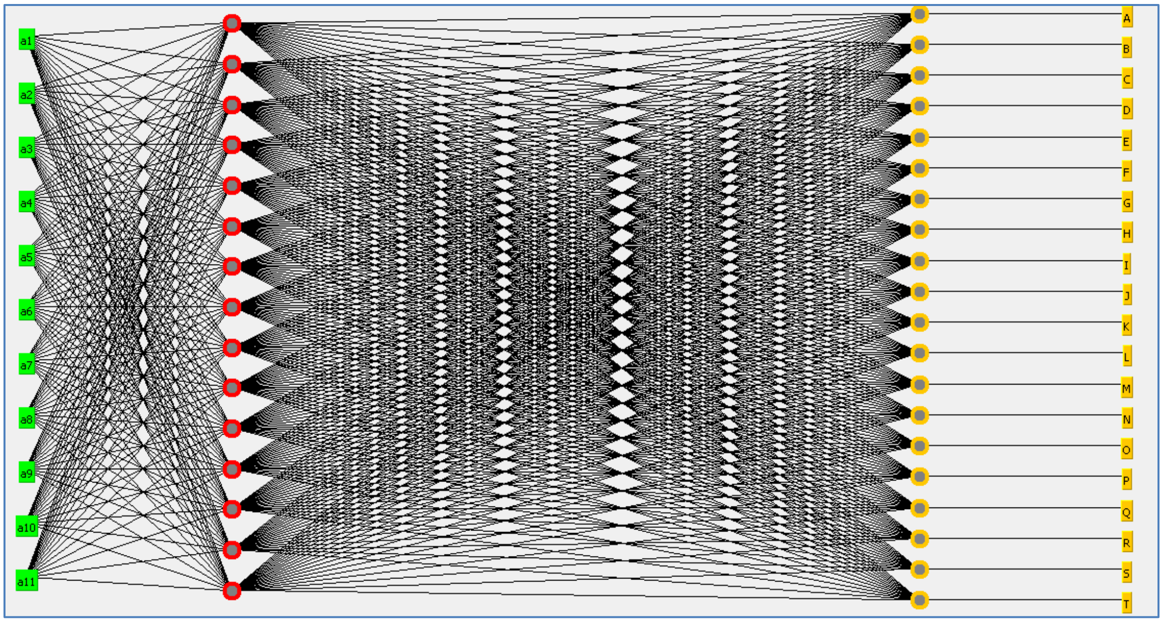


Figure 6-1 Example of the structure of the MLP of PM feature vector

6.2 Experimental Setup

In our experiments, we selected 50 documents per each author (i.e. we selected the first 50 documents for each author). The selected number of documents is almost equal to the available number of some authors. In all our experiments, 700 documents (70%) were used for training and 300 documents (30%) were used for testing.

Some preprocessing operations on the corpus are carried out to determine the probable author of a given text. We considered the main body of the text (excluding titles, author names, dates, etc.). Then we addressed two levels of preprocessing. Character level processing includes some operations: (1) Eliminating the diacritics including (ـَ، ـِ، ـُ، ـْ، ـّ، ـً، ـٌّ، ـٍ), (2) Eliminating none-Arabic terms and symbols or alphabets, (3) Removing the noisy symbols (other than punctuation marks). This level is applied at the character level.

At the word level in the preprocessing phase, we eliminated both punctuation marks and digits. While eliminating the titles, author names, dates of articles is done manually, the rests of preprocessing operations are conducted automatically through our developed system.

6.3 Experimental Results

We carried out many experiments using the selected classifiers to evaluate the performance of the extracted feature vectors that represent several types of stylometric features individually and in combinations. We evaluated the performances of these features before and after applying the feature selection techniques to evaluate the effect of feature selection techniques on the features with regards to Arabic corpus.

6.3.1 Lexical features

We consider three main feature vectors as lexical features namely words frequency, word n-grams richness and specific words. We considered the words that occur in the training corpus more than 49, 99 and 149 (word1GTH49fv, word1GTH99fv, and word1GTH149fv, respectively). We used the different feature normalization methods (absolute frequencies of terms and their standardized scores) to analyze their effects on identifying the possible author of unattributed texts. We used the relative frequencies and Z-scores (with the zero mean and one standard deviation). As a result, we obtained nine feature vectors called: the frequency of terms that occur more than 49 (FWGTH49fv), the frequency of terms that occur more than 99 (FWGTH99fv), the frequency of terms that occur more than 149 (FWGTH149fv), the relative frequencies of terms that occur more than 49 (LWGTH49fv), the relative frequencies of terms that occur more than 99 (LWGTH99fv), the relative frequencies of terms that occur more than 149 (LWGTH149fv), the Z-scores of terms that occur more than 49 (ZWGTH49fv), the Z-scores of terms that occur more than 99 (ZWGTH99fv), and the Z-scores of terms that occur more than 149 (ZWGTH149fv).

As shown in Table 6-1, best accuracy rates of 99% are obtained using the term frequencies that occur more than 49, the relative frequencies of terms that occur more than 49 and the Z-scores of terms that occur more than 49 with SMO. The use of the absolute values or the standardized values does not affect the accuracy of SMO. The average accuracy of 87.78% using Z-scores tends to be the best. The height average accuracy rate of 88.67% is obtained using the Z-scores of terms that occur more than 149 while best average accuracy rate of 98.11% is obtained using SMO. The accuracies achieved using Z-scores outperform those obtained using the absolute frequencies and relative frequencies with distance-based classification. With LS-SVM, however, the highest average of accuracy rates of 93.44% is obtained using the absolute frequencies.

Based on the obtained results shown in Table 6-1 we will use K-NN classifier with *City block* distance in our remaining experiments.

Table 6-1 Accuracy rates of word features that occur more than 49, 99, and 149 using three different feature normalization methods

	Absolute frequencies			Avg.	Relative frequencies			Avg.	Z-scores			Avg.
	FWGTH49	*FWGTH99*	*FWGTH149*		*LWGTH49*	*LWGTH99*	*LWGTH149*		*ZWGTH49*	*ZWGTH99*	*ZWGTH149*	
ED	84.67	83.33	81.33	83.11	88.00	86.00	84.67	86.22	**91.33**	90.33	88.67	90.11
K-NN (CB)	41.67	63.67	71.00	58.78	79.00	83.00	**85.67**	82.56	76.00	82.67	83.33	80.67
K-NN (ED)	75.00	**81.67**	78.33	78.33	72.67	72.67	71.00	72.11	76.00	79.33	79.67	78.33
SMO	99	97	96.67	97.56	99	98	97.67	98.22	99	97.33	98	98.11
LS-SVM	**94.67**	93.33	92.33	93.44	84.67	92.00	93.33	90	90.67	90.67	93.67	91.67
Avg.	79.002	83.8	83.932	82.24	84.67	86.33	86.47	85.82	86.6	88.07	**88.67**	87.78

Specific words

We extracted the terms that are used by each author only or uncommon terms and we called these features *Specific Words*. Because of the complicated inflections of Arabic we evaluated two cases; the specific words that occur more than two in authors' profile (specifcwGTH2fv) and the words that occurs more than three authors' profile (specifcwGTH3fv). For each feature vector we considered the absolute frequency and the relative frequencies. Therefore, we obtained four different feature vectors: the absolute frequencies of specific words that occur more than two (FSpWGTH2fv), the absolute frequencies of specific words that occurs more than three (FSpWGTH3fv), the relative frequencies of specific words that occurs more than two (LSpWGTH2fv) and the relative frequencies of specific words that occurs more than three (LSpWGTH3fv). In general, the accuracy rates obtained using such features are poor where the highest accuracy rate is 66.33% obtained from LSpWGTH2 using SMO as shown in Table 6-2.

Table 6-2 Accuracy rates of specific words per author

	Absolute frequency		Relative frequencies	
	FSpWGTH2	FSpWGTH3	LSpWGTH2	LSpWGTH3
ED	40.00	38.00	43.67	39.33
K-NN (CB)	15.00	21.67	15.00	23.00
K-NN (ED)	23.67	29.00	30.00	34.33
SMO	63	56.67	**66.33**	55.67
LS-SVM	45.00	45.33	40.33	40.67

We tried to improve the accuracy of the specific words by combining them to other vectors such that we combined *word1GTH149fv* and *specifcwGTH3fv* to obtain a new combined feature vector of 1000 features. In addition to the absolute frequency, we considered the relative frequencies to obtain two combined features called FCoSpW and LCoSpW. Unfortunately, the specific words have negative effects on the original feature vectors as shown in the 4th and 5th columns in Table 6-3. This may be attributed to the many zeroes in the vectors of specific words. To improve the results, we suggest using roots of words (or lemmas) instead of words and applying smoothing techniques in the future.

Table 6-3 Accuracy rates of words occurring >=150 and the combined specific words and words occurring >=150

	FWGTH 149	LWGTH 149	FCoSpW	LCoSpW
ED	81.33	84.67	82.00	83.67
K-NN	71.00	85.67	71.67	89.33
K-NN (ED)	78.33	71.00	79.67	72.33
SMO	96.67	97.67	96.33	96.67
LS-SVM	92.33	93.33	92.00	92.33

To our knowledge, the Delta rule has not been used for Arabic AA. Delta rule has been applied to other languages including English Italian, France, German, Hungarian and Greek (Eder & Rybicki, 2013; Eder, 2010, 2013; Savoy, 2012a, 2012b, 2013a). We used it to conduct experiments with words' frequencies. It is reported that the accuracy that was achieved using the most 400 frequency terms (as the best case) is 63.70, and 76.07 for English and Italian corpora respectively using a threshold of 400 (Savoy, 2012a, 2013a). We here need to investigate the best value of n most terms' frequency as parameter for Delta rule. We started with terms that occurs more than or equal to 50, 100, 150, 200, 300 and 400 in the training corpus. Best accuracy rate of 74.33% is obtained with terms that occurs more than or equal to 150 as shown in Table 6-4.

Table 6-4 **Investigating values of n most word frequency**

	n>=50	n>=100	n>=150	n>=200	n>=300	n>=400
Delta rule accuracy%	67.67%	73.00%	**74.33%**	68.33%	58.00%	50.00%

We believe that this threshold is equivalent to that used in previous works (Savoy, 2012a, 2013a) for several reasons; the characteristics of Arabic where it is more inflectional, the number of works per author in these studies seem to be larger than ours and these values achieve the highest accuracies in our work and in (Savoy, 2012a, 2013a). The accuracy rates that are obtained using Delta rule ranged from 63.70 to 76.07 based on three different languages with somewhat equivalent corpora (number of authors and genre of data). We considered these accuracy rates as the baseline in our work. Using ED instead of the Delta rule distance resulted in improved accuracy in all cases as shown in the second row of Table 6-5. Using the Delta rule with instance based method instead of profile-based method to represent the training corpus improved the accuracy rates with most thresholds compared with our baseline performance (except the case of n>=50). Comparing with the ED based profile-based method the accuracy improved dramatically with 200, 300 and 400 thresholds as shown in the third row in Table 6-5. Best accuracy of 92.33% is obtained using the Delta rule with instance based method and with threshold of 200. Using ED with instance-based method to represent the training samples improved the accuracy dramatically with all thresholds as shown in the fourth row of Table 6-5. The accuracy obtained from ED and instance-based method outperforms the other methods for thresholds of 50, 100, and 150. It has less accuracy than the Delta and instance based method with the remaining thresholds of 200, 300 and 400. The ED and instance-based method is preferable as the highest average accuracy of 83.89% is obtained using it. The Delta and instance-based method may be applied with reduced size of features as it outperforms the other cases with thresholds of 200, 300 and 400. It is clear that, the accuracy of authorship attribution is influenced by the training data representation method, the values representation (standardized scores) and the distance measures.

Table 6-5 Comparing the accuracy rates of baseline Delta rule with our modifications

	n>=50	**n>=100**	**n>=150**	**n>=200**	**n>=300**	**n>=400**	**Avg.**
Delta and Profile-based method (Baseline)	67.67	73.00	74.33	68.33	58.00	50.00	65.22
ED based Profile-based method	90.33	86.33	86.33	84.67	75.00	71.33	82.33
Delta and instance-based method	54.00	73.67	83.00	**92.33**	89.00	81.33	78.89
ED and instance-based method	**91.33**	**90.33**	**88.67**	**84.33**	**76.00**	**72.67**	83.89

We compared our work with previous approaches (Savoy, 2012a, 2013a) as shown in Table 6-6 which shows our methods outperform these approaches.

Table 6-6 Comparison of our methods and the approaches of (Savoy, 2012a, 2013a) **using Delta rule.**

Approach	(Savoy, 2012a, 2013a)	(Savoy, 2012a, 2013a)	**Our baseline method**	**Our modified method**
Language	English	Italian	Arabic	Arabic
Features	Frequency terms	Frequency terms	Frequency terms	Frequency terms
Number of authors	20	20	20	20
Size of corpus	5408	4326	1000	1000
Genre	Newspaper articles	Newspaper articles	Newspaper articles	Newspaper articles
Subjects	Several topics	Several topics	Several topics	Several topics
Accuracy	63. 70%	76.07%	74.33%	92%

6.3.2 Character-based features

As mentioned we considered punctuation marks (PM) and character level n-grams (n=1-4) as character features. Intuitively, punctuation marks individually are poor for identifying authors. The obtained accuracies using PM ranged between 40 with SMO to 53% with LS-SVM.

Regarding character level n-grams, a best accuracy rate of 99.67% is obtained using both character tri- and quad-grams with SMO. Best accuracy rate of 97.33% is obtained using Ch4grams using ED. The accuracy rates fluctuate between 80.33% and 95.67% using Ch1gram, Ch2gram and Ch3gram with ED. In general, the lowest accuracy rates are achieved with K-NN where a poorest accuracy rate of 9.67% is obtained using Ch4gram while the other accuracy rates ranged from

71% to 78.67% using Ch1gram, Ch2gram and Ch3gram as shown in Table 6-7. The highest accuracy rate of 96% is achieved using Ch2gram with LS-SVMO whereas accuracy rates of 89.33%, 76% and 83.67% are obtained using Ch1gram, Ch3gram and Ch4gram, respectively.

With respect to variable character level n-grams feature vectors including Ch12gram, Ch23gram and Ch123gram, best accuracy rate of 99.67% is obtained using Ch23gram with SMO. This is followed by an accuracy rate of 99% which is achieved using Ch12gram with SMO. We mean here by variable n-grams that the feature vector contains different length of n for example Ch12grams contains both uni- and bi-grams. In contrast, the fixed length of n-grams contains a fixed value for n such as Ch2gram contains just measures of character bi-grams.

Best average accuracy rate of 99.45% is obtained using variable character level n-grams (or combined character n-grams) including Ch12gram, Ch13gram and Ch123gram with SMO. This is better than an average accuracy rate of 97.75% obtained using fixed character level n-grams including Ch1gram, Ch2gram and Ch2gram as shown in Table 6-7.

An average accuracy rate of 96% is obtained using variable length character level n-grams with ED which is better than an average accuracy rate of 88.92% using fixed character level n-grams. This is true regarding K-NN where an average accuracy rate of 86% is obtained using the variable length character level n-grams. This is compared with an average accuracy rate of 59.25% obtained using the fixed length character level n-grams. Unlike LS-SVM, an average accuracy rate of 86.25% is obtained using fixed length character level n-grams which is better than that obtained using the variable length character level n-grams where the average accuracy rate is 83.44%.

6.3.3 Syntactic features

As reported above, we have proposed a collection of 309 function words (FW). The best accuracy rate of 87.67% is achieved with SMO. The accuracy rates of 63.33%, 71.33% and 86% are obtained with ED, K-NN, and LS-SVM, respectively as shown in Table 6-7.

6.3.4 General feature vector

As described above, we combined FW, Ch1gram and WR to obtain a new feature vector called general feature vector (GFV). Best accuracy rate of 98.67% is obtained with SMO. The accuracy rates of 63.67%, 81.67% and 92.67% are achieved with ED, K-NN, and LS-SVM, respectively.

Table 6-7 Character, syntactic and general feature vectors results

Stylometric	Character features										Syntactic Features	General Feature vector
	PM	Fixed length character n-grams					Variable length character n-grams					
Feature vector/ classifier		Ch1gram	Ch2gram	Ch3gram	Ch4gram	Avg.	Ch12 gram	Ch23 gram	Ch123 gram	Avg.	FW	GFV [FW+ Ch1+WR]
ED	41.33	82.33	80.33	95.67	97.33	88.915	80.33	96.00	96.00	90.78	63.33	63.67
K-NN	52.00	78.67	77.67	71.00	9.67	59.25	78.00	85.67	86.00	83.22	71.33	81.67
LS-SVM	53.00	89.33	**96.00**	76.00	83.67	86.25	95.67	77.33	77.33	83.44	86.00	92.67
SMO	40	92.67	99.00	**99.67**	**99.67**	97.75	99	**99.67**	**99.67**	99.45	87.67	98.67

6.3.5 Semantic features

To our knowledge, this is the first work to define semantic features using syntax and rhetorical Arabic styles and apply them on Arabic AA. One serious work that applied semantic features to English (Argamon et al., 2007). We proposed a set of 39 semantic features (SF) which include the most popular grammatical and rhetorical Arabic styles. The extracted features are evaluated on our built corpus. We tested the SF using different classification methods including ED, K-NN, MLP, LS-SVM and SMO. We also used SMO with 10 fold cross validation in order to compare our proposed approach with the approach of (Argamon et al., 2007). That approach is based on principles of Systemic Functional Grammar (SFG) (Halliday, 1994). As shown in Table 6-8, the highest accuracy of 71.8% is obtained using SMO 10- fold cross validation. Both of LS-SVM and SMO also tend to achieve good accuracy of 70%. We then combined SF with FW to obtain a combined function words and semantic features (FWSF) with size of 348 features. Highest accuracy rate of about 91% is obtained using SMO and MLP. The accuracy of about 90% is obtained using SMO 10-fold-CV. The semantic features improved the performance of FW in all cases.

Table 6-8 Accuracy rates using semantic features

	SF	FW	FWSF
ED	59.33	63.33	66.67
K-NN	56.67	71.33	72.67
LS-SVM	70.00	86.00	86.67
MLP	63.33	90.33	91.33
SMO	69.67	87.67	91.00
SMO 10 fold CV	71.8	88.8	89.6

It is noteworthy that the more advanced features are used the low accuracy rate is obtained with AA. Character features outperform lexical features and lexical features outperform syntactic features which outperform semantic features. The works of (Ouamour & Sayoud, 2012; Türkoğlu et al., 2007) confirm these findings regarding character and lexical and syntactic features while the work of (Argamon et al., 2007) confirms our findings regarding syntactic and semantic features. So the accuracy of our semantic system is suitable for AA since such features deal with and detect hidden writing styles of authors. We compared our work with the work of (Argamon et al., 2007) as shown in Table 6-9. Our semantic features compare favorably with the reported system. Additionally, our system is applied to 20 authors while the compared work is applied to just eight authors. In general, the number of candidate authors is inversely proportional to the accuracy of the features (Luyckx & Daelemans, 2011).

Table 6-9 **Comparisons of our semantic features accuracy with** (Argamon et al., 2007)

	(Argamon et al., 2007)	**Our approach**
Language	English	Arabic
Num of authors	8	20
Genre of corpus	Novel chapters	Newspaper articles
features	FW, SF and SF+FW	FW, SF and SF+FW
Classifier	SMO 10 fold CV	SMO 10 fold CV
Identification rate using FW	85%	88.8%
Identification rate using SF	71.5%	71.8%
Identification rate using FWSF	89%	89.6%

This comparison seems week as we compare English results with Arabic. We are not aware of any similar work on Arabic. This gives us indication of our work.

6.4 Selected Features

We have conducted comparative analysis of different feature selection techniques. Based on our experiments reported in Chapter 4, we decided to use CFSS with RS and ChiSAE as feature selection techniques on stylometric features in this work. We used CFSS with BF as a base line to compare with the feature selection technique used in the work of (Türkoğlu et al., 2007). These techniques are applied to optimize the accuracy rate of both character and syntactic features since with respect to lexical features we used the most frequency terms as described above.

The best accuracy rates for function words are obtained from selected features using both CFSS with RS and ChiSAE where best accuracy rate of 90.33% is obtained using CFSS+RS method as shown in Table 6-10. The accuracy rates using function words in the literature does not exceed 90%. For example, the accuracy rates obtained by (Pavelec et al., 2008) and (Varela et al., 2011) are 83.2% and 74%, respectively in the best cases on Portuguese. To consider the work of (Shaker & Corne, 2010) we should take the average accuracy of all pair of authors since they reported their results based on binary classification. The accuracy rate of 87.64% based on the reported accuracies was obtained using 65 Arabic function words applied on Arabic novels written by six authors. An accuracy of 85% is obtained by (Argamon et al., 2007) using 675 English function words. Other works did not use FW individually but they combined them with other features (Abbasi & Chen, 2005; Argamon et al., 2007; Türkoğlu et al., 2007; Zheng et al., 2005). Our proposed function words can identify the possible author of disputed texts efficiently and compares with the literature favorably.

Table 6-10 Accuracy rates of full and selected feature vectors using function words

	FW	SFWBF	SFWRS	SFWChi
ED	**63.33**	59.33	61.33	61.33
K-NN	**71.33**	65.67	70.00	70.00
LS-SVM	**86.00**	79.00	81.00	81.33
SMO	87.67	85.67	**88.67**	88.33
MLP	89.33	85.33	**90.33**	89.33
Avg.	79.532	75	78.266	78.064

With respect to character level uni-gram, the best accuracy rate of 93.67% is obtained from optimized features using CFSS+RS and ChiSAE, as shown in Table 6-11. In many cases CFSS+RS and ChiSAE tend to outperform the full feature vector and CFSS+BF. The full feature vector and selected features using CFSS+RS are achieved better accuracy rates of 82.33% and 81.00%, respectively using ED. This is also true with MLP, where the full feature vector with an accuracy rate of 92.33% outperforms the selected feature vectors.

Table 6-11 Accuracy rates of full and selected feature vectors using character uni-gram

	Ch1gram	SCh1BF	SCh1RS	SCh1Chi
ED	**82.33**	81.00	79.00	79.00
K-NN	78.67	77.33	79.33	79.33
LS-SVM	89.33	87.33	89.00	89.33
SMO	92.67	92.33	**93.67**	**93.67**
MLP	92.33	91.67	91.67	90.33
Avg.	87.07	85.932	86.53	86.33

Regarding character level bi-grams, best accuracy rate of 99% is obtained using the full features with SMO, as shown in Table 6-12. In cases of ED, SMO and LS-SVM, full feature vectors tend to outperform the selected feature vector. These insignificant drops, however, in the accuracy rates are negligible so selected features still outperform the full features regarding time and space. In case of K-NN, the highest accuracy rate of 84% is obtained from SCh2BF while the best accuracy rate of 98.33% is obtained using SChi2Chi.

Table 6-12 Accuracy rates of full and selected feature vectors using character bi-gram

	Ch2gram	SCh2BF	SCh2RS	SCh2Chi
ED	80.33	62.00	64	66.67
K-NN	77.67	84.00	78.67	80.33
LS-SVM	**96.00**	93.00	93.33	95.33
SMO	99.00	97.67	**98.67**	98
MLP	89.33	98	97.67	98.33
Avg.	88.466	86.93	86.47	87.73

With regard to character tri-grams, as shown in Table 6-13, best accuracy rate of 99.67% is obtained using full features and SMO. In the case of ED and SMO, full features outperform the selected features insignificantly (the differences in accuracy rates are negligible). Therefore, selected features still outperform the full features regarding time and space. In the case of K-NN and LS-SVM, highest accuracy rates of 90.33% and 96% respectively are obtained using SCh2ChiS.

Table 6-13 Accuracy rates of full and selected features using character tri-gram

	Ch3gram	SCh3BF	SCh3RS	SCh3Chi
ED	95.67	90.67	92.67	93.67
K-NN	71.00	86.67	89.67	90.33
LS-SVM	76.00	95.33	91.00	**96.00**
SMO	99.67	97.00	98.67	98.33
Avg.	85.59	92.42	93.00	94.58

With respect to character quad-grams, as shown in Table 6-14, best accuracy rate of 99.67% is obtained using Ch4gram, SCh4RS and SCh4Chi with SMO. In the case of ED and SMO, full feature vectors outperform the selected feature vectors insignificantly (the differences in accuracy rates are negligible). Therefore, selected features still outperform the full features regarding time and space. In the case of K-NN and LS-SVM highest accuracy rates of 90.33% and 96% respectively are obtained using SCh2ChiS. In most cases the selected feature vectors achieved higher accuracy rates than full features. Poorest accuracy rate of 9.67% is obtained using Ch4gram with K-NN. This may be attributed to the many zeroes in the feature vector.

Table 6-14 Accuracy rates of full and selected features using character quad-gram

	Ch4gram	SCh4BF	SCh4RS	SCh4Chi
ED	97.33	95.00	97.00	97.00
K-NN	9.67	85.67	80.00	80.00
LS-SVM	83.67	90.33	88.67	**96.33**
SMO	**99.67**	98.00	**99.67**	**99.67**
Avg.	72.585	92.25	91.335	**93.25**

Best accuracy rate of about 99% is obtained using Ch12gram, SCh12RS and SCh12Chi using SMO, as shown in Table 6-15. Best accuracy rate of 80.33% is obtained using Ch12gram and ED. Using K-NN, the selected features outperform the full features with a best accuracy rate of 85.00% using SCh12BF. Using LS-SVM, best accuracy rate of 95.67 is obtained using both Ch12gram and SCh12Chi.

Table 6-15 **Accuracy rates of full and selected features using combined character uni- and bi-grams**

	Ch12gram	SCh12BF	SCh12RS	SCh12Chi
ED	80.33	64.33	64.00	66.67
K-NN	78.00	85.00	79.67	80.67
LS-SVM	95.67	95.00	94.33	95.67
SMO	**99**	97	**98.67**	**98.67**
Avg.	88.25	85.33	84.17	85.42

Best accuracy rate of about 99.67% is obtained using Ch23gram, SCh23RS and SCh23Chi with SMO, as shown in Table 6-16. Best accuracy rate of 96% is obtained using Ch23gram and ED. The selected features outperform the full features and other selected features using K-NN such best accuracy rate of 91.33% is obtained using SCh23Chi. Best accuracy rate of about 92% is obtained using the selected features (SCh23RS) and LS-SVM.

Table 6-16 **Accuracy rates of full and selected features using combined character bi- and tri-grams**

	Ch23gram	SCh23BF	SCh23RS	SCh23Chi
ED	**96.00**	89.67	92.00	93.33
K-NN	85.67	88.67	90.67	**91.33**
LS-SVM	77.33	91.67	**92.00**	91.33
SMO	**99.67**	98.33	**99.33**	**99.33**
Avg.	89.6675	92.085	93.5	**93.83**

Best accuracy rate of 99.67% is obtained from Ch123gram and SCh123RS with SMO as shown in Table 6-17. Best accuracy rate of 96% is obtained using Ch123gram and ED. The selected features (SCh123Chi) outperform the full features (Ch123gram) using K-NN such that best accuracy rate of 91.33% is obtained using SCh123Chi. Best accuracy rate of 97% is obtained using SCh123Chi and LS-SVM.

Table 6-17 Accuracy rates of full and selected feature vectors using combined character uni-, bi- and tri-grams

	Ch123gram	SCh123BF	SCh123RS	SCh123Chi
ED	**96.00**	87.67	90.33	93.33
K-NN	86.00	88.33	86.00	**91.33**
LS-SVM	77.33	96.33	94.67	**97.00**
SMO	**99.67**	99	**99.67**	99.33
Avg.	89.6675	92.085	**93.5**	**93.83**

The average accuracy rates of selected features are higher than the full feature vectors, as shown in Table 6-18. Highest average accuracy rate of 89.31% is obtained using ChiSAE which is followed by an average accuracy rate of 88.24% obtained using CFSS+RS. These averages outperform the average of the base line feature selection technique (CFSS+BF). Highest average accuracy rate of 95.25% is obtained from the selected features of Ch123gram using ChiSAE.

Table 6-18 The average accuracy rates of full and selected features obtained using different classification methods

	Avg. (full features)	Avg.(CFSS+FB)	Avg. (CFSS+RS)	Avg. (ChiSAE)
FW	**79.53**	75	78.27	78.06
Ch1gram	**87.07**	85.93	86.53	86.33
Ch2gram	**88.47**	86.93	86.47	87.73
Ch3gram	85.59	92.42	93.00	**94.58**
Ch4gram	72.59	92.25	91.34	**93.25**
Ch12gram	**88.25**	85.33	84.17	85.42
Ch23gram	89.67	92.08	93.5	**93.83**
Ch123gram	89.75	92.83	92.67	**95.25**
Avg.	85.11	87.85	88.24	**89.31**

We compared these accuracy rates with the most related work (Türkoğlu et al., 2007). Türkoğlu et al.(Türkoğlu et al., 2007) used CFSS+BF with five WEKA classifiers (viz. Naive Bayes, SVM, Random Forest, k-Nearest Neighbor, and MLP) with many stylometric features extracted from Turkish corpora. The used feature selection technique by Türkoğlu et al.(Türkoğlu et al., 2007) is one of the feature selection techniques that we used as baseline (CFSS+BF). Comparing its accuracy rate using (CFSS+BF) with the other used techniques (CFSS+RS and ChiSAE), we find in most cases (four cases out of five) that the combination of CFSS+RS, and ChiSAE outperform (CFSS+BF) technique. In general, the accuracy rates obtained using CFSS+RS, and ChiSAE are higher than the accuracy rates of the baseline feature selection technique (CFSS+BF).

6.5 Conclusions

We investigated several types of stylometric features including lexical, character, syntactic and semantic features. We obtained high accuracy rates exceeding 99% in many cases on the corpus of 20 authors and 1000 documents. We investigated vocabulary richness and word level as lexical features. To our knowledge, this is the first time to investigate word n-grams richness in word bi-, tri- and quad-grams and it is the first time to use specific words per authors. We investigated punctuation marks and character level n-grams. For character n-grams we used both fixed length character n-grams and variable length n-grams by combining uni-, bi-, and tri-grams feature vectors. We also investigated the proposed collection of Arabic function words. We combined an instance of lexical, an instance of character, and syntactic features to create the general feature vector which achieves higher accuracy rates than the individual ones. We tested our Arabic semantic lexicon and showed how such features aid in gaining insight about the author of a given text and which compare favorably with other works. We used ED, K-NN, Delta rule, MLP, LS-SVM, and SMO classification methods to investigate our features. Character features tend to outperform the other features and SMO achieves the highest accuracy rates. Other classification methods performed well, too. The achieved accuracies of the proposed function words indicate that they might be generalized for AA purpose.

Best accuracy rates of 99.67% are obtained from Ch3gram Ch4gram, Ch23gram, Ch123gram, SCh4RS, SCh4Chi and SCh123RS with SMO. These are followed by the accuracy rates of about 99% which are obtained from word1GTH49fv, Ch2gram, Ch12gram, SCh12RS and SCh12Chi. High accuracy rates are obtained using other techniques. For example, LS-SVM achieves the accuracy rate of 97.00% using SCh123Chi. ED achieved accuracy rates of about 97.33% with Ch4gram, SCh4Chi and SCh4RS. K-NN achieved best accuracy rate of 91.33% using SCh23Chi and SCh123Chi. Other methods and techniques achieved acceptable accuracy rates and they compare favorably with previous works. Both fixed length and variable length character n-grams resulted in high accuracy rates. The fixed length character n-grams outperformed the variable length character n-grams. To our knowledge, it is also the first time to investigate the effects of using different feature normalization methods to AA. We also investigated the effects of training set representation methods including profile-based method and instance-based method. The instance-based method achieved better accuracy rates. We also optimized the extracted features by applying feature selection techniques. The selected features tend to perform better especially in terms of time and memory.

CHAPTER 7

CONCLUSIONS AND FUTURE WORKS

We conducted a comprehensive literature survey for authorship attribution researches. In this survey, the contributions, strengths and limitations of published works are discussed. These publications were classified based on the types of stylometric features, the AA classification methods and techniques, the selection feature techniques and the corpora used. The characteristics of Arabic and its challenges from the point of view AA are presented.

We designed a corpus which consists of selected newspapers' articles published in Alriyadh, Alhayat and Shorouk newspapers during the period from 2011 to 2013 written by a total of 20 well-known regular authors (columnist) to build a benchmarking dataset. In order to capture features that characterize or reflect the style of authors, we collected sufficient works for each author (where we selected 50 articles per author). The texts cover different topics namely, politics, economics, socials and sports. The average length of texts per author ranged between 411 and 1242 words, 2452 and 8132 characters and the average size ranged from 3 to 8 KB.

In order to show how our corpus compares with other used corpora by researchers, we surveyed the corpora used in most related and recent works, in general and those used in Arabic researches, in particular.

We presented different stylometric features and feature selection techniques. We extracted several types of stylometric features and designed new lexical features (viz.

word n-grams richness and specific words per author). We also conducted a case study to select the most appropriate feature selection techniques.

We constructed novel stylometric features (viz. Arabic semantic features) and presented our methodology in extracting them. Our Arabic semantic lexicon may be applied in other Arabic language processing and understanding topics.

Several experiments are conducted to evaluate our techniques. The obtained accuracy rates show that the used techniques can identify the authors successfully. This is in addition to our modifications and designed techniques that significantly improved the obtained accuracies. In many cases our accuracies outperformed published works.

Best accuracy rate of 99.67% is obtained using character tri-grams (ch3gram), character quad-grams (ch4gram), combined character bi-grams and tri-grams features (ch23gram), combined character uni-grams, bi-grams and tri-grams features (ch123gram), selected character quad-grams features using CFSS+RS (SCh4RS), selected character quad-grams features using ChiSAE (SCh4Chi) and selected combined uni-grams bi-grams and trigrams features using CFSS+RS (SCh123RS) with SMO. Accuracy rates of about 99% are obtained using terms that occur more than 49, general feature vector (GFV), character bi-grams (ch2gram), combined character uni-grams and bi-grams features (ch12gram), selected combined character uni-grams and bi-grams features using CFSS+RS (SCh12RS) and selected combined character uni-grams and bi-grams features using ChiSAE (SCh12Chi). This is in addition to high accuracy rates that are obtained using other techniques. For example LS-SVM achieved accuracy rate of 97.00% using selected combined character uni-grams bi-grams and tri-grams features using ChiSAE (SCh123Chi). ED achieved performances of about

97.33% with character quad-grams feature vector (Ch4gram), selected character quad-grams feature vector using ChiSAE (SCh4Chi) and selected character quad-grams feature vector using CFSS+RS (SCh4RS). K-NN achieved best performance of 91.33% using selected combined character bi-grams and tri-grams feature vector using ChiSAE (SCh23Chi) and selected combined character uni-grams bi-grams and tri-grams feature vector using ChiSAE (SCh123Chi). Other methods and techniques achieved accuracies that compare favorably with published works.

The main contributions of this thesis can be listed as follows:

- A literature survey of AA is conducted.
- A corpus of Arabic AA texts is built. The corpus includes more than 1000 documents written by 20 authors. To our knowledge, this is the first benchmarking corpus for Arabic AA. We aim to make the corpus freely available to the research community. This is expected to provide a platform for researchers to compare their results with other researchers.
- Several types of stylometric features are extracted for Arabic AA including lexical, character and syntactic features using our built features extractor. Additionally, we designed and applied new lexical features such as specific words per authors and word n-grams richness.
- We proposed a collection of Arabic function words that we evaluated for AA.
- We constructed a novel Arabic semantic lexicon and used it for AA.
- Feature selection techniques are applied to reduce the high-dimensionality feature vectors which are evaluated on Arabic AA. The selected features compare favorably with the more complex ones.

- We developed a prototype system for automated authorship attribution of Arabic texts which will be able to handle different authors' authorship.
- Two journal publications are submitted.

Future works

We are planning to increase the size of the corpus through:

- Increasing the number of authors
- Increasing the number of works per author

This is in addition to, dealing with the texts which rely on quotations such as religious articles. We will investigate applying smoothing techniques, using lemmas, roots and POS-tags. We are working on improving our semantic system to use several semantic levels.

References

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, 20*(5), 67–75.

Abdullatif, M., Omar, A., & Zahran, M. (2005). *Annaho Al-Assassi (In Arabic)* (p. 495). Qairo: Dar Alfekr Al-Aarabi.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning, 6*(1), 37–66.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features: Research Articles. *Journal of the American Society for Information Science and Technology*, *58*(6), 802–822.

Arun, R., Saradha, R., Suresh, V., Murty, M., & Madhavan, C. (2009). Stopwords and stylometry: a latent Dirichlet allocation approach. In *NIPS workshop on Applications for Topic Models*. Whistler (BC).

Baayen, H., Halteren Van, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. *6th JADT*, *I*, 69–75. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.6139&amp;rep=rep1&amp;type=pdf

Baayen, H., Van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *11*(3), 121–132. doi:10.1093/llc/11.3.121

Bandara, U., & Wijayarathna, G. (2013). Source code author identification with unsupervised feature learning. *Pattern Recognition Letters, 34*(3), 330–334. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167865512003571

Benedetto, D., Caglioti, E., & Loreto, V. (2002). Language trees and zipping. *Physical Review Letters*, *88,* 048702. doi:10.1103/PhysRevLett.88.048702

Binongo, J. N. G. (2003). Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. *Chance, 16,* 9–17. doi:10.1080/09332480.2003.10554843

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Bosch, R. A., & Smith, J. A. (1998). Separating hyperplanes and the authorship of the disputed federalist papers. *The American Mathematical Monthly*, *105*, 601–608.

Bradante, K. De, Karsmakers, P., Ojeda, F., Alzate, C., Brabanter, J. De, Pelckmans, K., … J.A.K., S. (2011). *"LS-SVMlab Toolbox User's Guide version 1.8*.

Burrows, J. (2002). "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, *17*(3), 267–287. doi:10.1093/llc/17.3.267

Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, *2*(2), 61–70.

Burrows, J. F. (1992). Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, *7*(2), 91–109.

Burrows, S. (2012). Comparing techniques for authorship attribution of source code. *Software: Practice and Experience,*, *44*(1), 1–32. doi:10.1002/spe.2146

Chandrasekaran, R., & Manimannan, G. (2013). Use of Generalized Regression Neural Network in Authorship Attribution. *International Journal of Computer Application*, *62*(4), 7–10.

Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, *4*(1), 1–13. Retrieved from https://library.utica.edu/academic/institutes/ecii/publications/articles/B49F9C4A-0362-765C-6A235CB8ABDFACFF.pdf

De Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, *30*(4), 55–64. doi:10.1145/604264.604272

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, *19*(1-2), 109–123. doi:10.1023/A:1023824908771

Ebrahimpour, M., Putniņš, T. J., Berryman, M. J., Allison, A., Ng, B. W.-H., & Abbott, D. (2013). Automated authorship attribution using advanced signal classification techniques. *PloS One*, *8*, 1–12. doi:10.1371/journal.pone.0054998

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. In *Digital Humanities 2010: Conference Abstracts* (pp. 132–35). London: King's College London. doi:10.1093/llc/fqt066

Eder, M. (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, *28*(4), 603–614.

Eder, M., & Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, *28*(2), 229–236.

Escalante, H. J., Nicol, S., Garza, D. L., & Montes-y-g, M. (2011). Local Histograms of Character N -grams for Authorship Attribution. *Computational Linguistics*, 288–298. Retrieved from http://www.aclweb.org/anthology/P11-1030

Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, *13*(1), 1–50.

Forsyth, R., & Holmes, D. (1996). Feature-Finding for Text Classification. *Literary and Linguistic Computing*, *11*, 163–174. doi:10.1093/llc/11.4.163

Gill, P. S., & Swartz, T. B. (2011). Stylometric analyses using Dirichlet process mixture models. *Journal of Statistical Planning and Inference*, *141*(11), 3674–3665. doi:10.1016/j.jspi.2011.05.020

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. *Addison Wesley* (Vol. Addison-We, p. 432). doi:10.1007/s10589-009-9261-6

Hall, M. a. (1999). Correlation-based Feature Selection for Machine Learning. *Methodology*, *21i195-i20*, 1–5. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.9584&amp;rep=rep1&amp;type=pdf

Halliday, M. A. K. (1994). *Introduction to functional grammar* (Second Edi.).

Holmes, D. I., & Forsyth, R. S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, *10*(2), 111–127.

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. *Artificial Intelligence Methodology Systems and Applications, 4183*, 77–86. doi:10.1007/11861461_10

Jamak, A., Savatić, A., & Can, M. (2012). Principal component analysis for authorship. *Business Systems Research*, *3*, 49–56.

Jockers, M. L., & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, *25*(2), 215–223. doi:10.1093/llc/fqq001

Kaster, A., Siersdorfer, S., & Weikum, G. (2005). Combining text and linguistic document representations for authorship attribution. In *SIGIR Workshop Stylistic Analysis of Text for Information Access STYLE* (Vol. 1, pp. 27–35). Citeseer. doi:10.1.1.84.4407

Khmelev, D. V, & Teahan, W. J. (2003). Comment on “Language trees and zipping”. *Physical Review Letters*. doi:10.1103/PhysRevLett.90.089803

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*, 9–26. doi:10.1002/asi.20961

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship Attribution with Thousands of Candidate Authors. *SIGKDD Explorations*, *pp*, 659–660. Retrieved from http://eprints.pascal-network.org/archive/00002680/

Kukushkina, O. V, Polikarpov, A. A., & Khmelev, D. V. (2001). Using Literal and Grammatical Statistics for Authorship Attribution. *Problems of Information Transmission*, *37*, 172–184. doi:10.1023/A:1010478226705

Li, J., Zheng, R., & Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM*, *49*(4), 76–82. doi:10.1145/1121949.1121951

Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. B. (2004). The similarity metric. *Information Theory, IEEE Transactions on*, *50*(12), 3250–3264.

Liu, H., & Setiono, R. (1996). A Probabilistic Approach to Feature Selection A Filter Solution. *Proc 13th International Conference on Machine Learning, pages*, 319–327. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.1270

Liu, Z., Yang, Z., Liu, S., & Shi, Y. (2013). Semi-random subspace method for writeprint identification. *Neurocomputing*, *108*, 93–102.

Luyckx, K., & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, *26*(1), 35–55.

Malyutov, M. B. (2005). Authorship attribution of texts: a review. *Electronic Notes in Discrete Mathematics*. doi:10.1016/j.endm.2005.07.064

Marton, Y., Wu, N., & Hellerstein, L. (2005). On compression-based text classification. In *European Conference on Information Retrieval* (pp. 300–314).

Mehri, A., Darooneh, A. H., & Shariati, A. (2012). The complex networks approach for authorship attribution of books. *Physica A: Statistical Mechanics and Its Applications*, *391*(7), 2429–2437.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science, IX*, 237–246.

Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The Federalist. Reading, MA: Addison-Wesley*. Addison-Wesley.

Oliveira, W., Justino, E., & Oliveira, L. S. (2013). Comparing compression models for authorship attribution. *Forensic Science International*, *228*, 100–4. doi:10.1016/j.forsciint.2013.02.025

Othaimeen, M. (2005). *Sharh Al-Ojromeah (In Arabic)*. Ryadh: Arroshd Bookstore.

Ouamour, S., & Sayoud, H. (2012). Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier. In *The 2nd International Conference on Communications and Information Technology (ICCIT): Digital Information Management* (pp. 44 –47). Hammamet, Tunisia: IEEE.

Ouamour, S., & Sayoud, H. (2013). Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on* (pp. 144–147). Beijing: IEEE.

Pavelec, D., Oliveira, L., Justino, E., & Batista, L. (2008). Using Conjunctions and Adverbs for Author Verification. *Journal Of Universal Computer Science*, *14*, 2967–2981. Retrieved from http://apps.isiknowledge.com.libproxy.unm.edu/full_record.do?product=WOS&colname=WOS&search_mode=RelatedRecords&qid=189&SID=1BFE94Ekeg2KHDJkJJ8&page=6&doc=53

Pillay, S. R., & Solorio, T. Authorship attribution of web forum posts. , eCrime Researchers Summit eCrime 2010 1–7 (2010). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5706693

Plakias, S., & Stamatatos, E. (2008). Tensor space models for authorship identification. In *Artificial Intelligence: Theories, Models and Applications* (pp. 239–249). Springer Berlin Heidelberg.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods*, *12*, 185–208. doi:10.1109/ISKE.2008.4731075

Savoy, J. (2012a). Authorship Attribution Based on Specific Vocabulary. *ACM Transactions on Information Systems*, *30*(2), 1–30. doi:10.1145/2180868.2180874

Savoy, J. (2012b). Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics*, *19*(2), 132–161.

Savoy, J. (2013a). Authorship attribution based on a probabilistic topic model. *Information Processing & Management*, *49*(1), 341–354.

Savoy, J. (2013b). Feature Selections for Authorship Attribution. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 939–941). Coimbra, Portugal.

Schaalje, G. B., Blades, N. J., & Funai, T. (2013). An open-set size-adjusted Bayesian classifier for authorship attribution. *Journal of the American Society for Information Science and Technology*, *64*(9), 1815–1825.

Seroussi, Y., Zukerman, I., & Bohnert, F. (2011). Authorship Attribution with Latent Dirichlet Allocation. In *the Fifteenth Conference on Computational Natural Language Learning* (pp. 181–189). Portland, Oregon, USA.

Shaker, K., & Corne, D. (2010). Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. In *2010 UK Workshop on Computational Intelligence (UKCI)* (pp. 1–6). IEEE. doi:10.1109/UKCI.2010.5625580

Solorio, T., & Al., E. (2011). Modality specific meta features for authorship attribution in web forum posts. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (p. 164).

Stamatatos, E. (2007). Author Identification Using Imbalanced and Limited Training Texts. *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*. doi:10.1109/DEXA.2007.5

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, *44*(2), 790–799. doi:10.1016/j.ipm.2007.05.012

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556. doi:10.1002/asi.21001

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, *26*(4), 471–495. doi:10.1162/089120100750105920

Suykens, J. A. K., & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, *9*, 293–300. doi:10.1023/A:1018628609742

Türkoğlu, F., Diri, B., & Amasyalı, M. (2007). Author attribution of turkish texts by feature mining. In *Third International Conference on Intelligent Computing, ICIC2007* (pp. 1086–1093). Qingdao, China. doi:10.1007/978-3-540-74171-8_110

Van Gestel, T., Suykens, J., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., … Vandewalle, J. (2004). Benchmarking least squares support vector machine classifiers. *Machine Learning*, *54*, 5–32. Retrieved from https://lirias.kuleuven.be/bitstream/123456789/73836/1/Benchmarking.pdf

Varela, P., Justino, E., & Oliveira, L. S. (2010). Verbs and Pronouns for Authorship Attribution. In *International Conference on Systems, Signals and Image Processing, Rio de Janeiro* (pp. 89–92).

Varela, P., Justino, E., & Oliveira, L. S. (2011). Selecting syntactic attributes for authorship attribution. In *Neural Networks (IJCNN), The 2011 International Joint Conference on Neural Networks,San Jose, California, USA* (pp. 167–172).

Wahba, G. (1990). *Spline models for observational data*. Siam.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. *Machine Learning* (p. 525). Retrieved from http://books.google.com/books?hl=en&amp;lr=&amp;id=QTnOcZJzlUoC&amp;oi=fnd&amp;pg=PR5&amp;dq=Data+Mining:+Practical+machine+learning+tools+and+techniques&amp;ots=3fozatWmTb&amp;sig=zOqH_wgtAn5d_guFVFa4rXYxCaQ

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika*, *30*, 363–390. doi:10.1093/biomet/30.3-4.363

Yule, G. U. (1944). *The statistical study of literary vocabulary*. *Cambridge England* (p. 306). Cambridge University Press. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp;path=ASIN/0208006893

Zhao, Y., & Zobel, J. (2005). Effective and Scalable Authorship Attribution Using Function Words. *INFORMATION RETRIEVAL TECHNOLOGY PROCEEDINGS*, *3689*, 174–189. doi:10.1007/11562382_14

Zhao, Y., & Zobel, J. (2007). Searching with style: authorship attribution in classic literature. In *ACSC 07 Proceedings of the thirtieth Australasian conference on Computer science* (Vol. 62, pp. 59–68). Australian Computer Society, Inc. Retrieved from http://portal.acm.org/citation.cfm?id=1273757

Zheng, R., Li, J., Chen, H., & Huang, Z. (2005). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378–393. doi:10.1002/asi

Appendices

Appendix A Arabic Function Words

آنذاك	أبدا	أثناء	أسفل	أصبح	أعلى
أغلب	أكثر	ألا	ألم	أم	أما
أمام	أمس	أن	أنا	أنت	أنتم
أنتما	أنتن	أو	أولئك	أي	أيا
أيضا	أين	أينما	إبان	إثر	إذ
إذا	إذن	إزاء	إلا	إلى	إما
إن	إنما	إياك	إياكم	إياكما	إياكن
إيانا	إياه	إياها	إياهم	إياهما	إياهن
إياي	الآن	البتة	التي	الذي	الذين
اللات	اللاتي	اللتان	اللتين	اللذان	اللذين
اللهم	اللوات	اللواتي	الليلة	اليوم	بألا
بئس	بئست	باتجاه	بالأخص	بالأمس	بالذات
بالضبط	بالطبع	بالفعل	بالقرب	بالكامل	بتاتا
بجانب	بحسب	بحوالي	بحيث	برمته	بشتى
بضع	بضعة	بعدما	بعض	بغية	بقرب
بل	بلا	بلى	بما	بمفرده	بين
بينما	تاما	تباعا	تبعا	تجاه	تحت
تحديدا	تحسبا	تقريبا	تلك	تلو	تماما
تمشيا	ثم	ثمة	جانب	جاهدا	جدا
جديا	جراء	جميع	جميعا	جنوب	جنوبي
حتما	حتميا	حتى	حسب	حسبما	حوالي
حول	حيال	حيث	حيثما	حين	حينئذ
حينا	حينذاك	حينما	خارج	ختاما	خلال
خلف	دائما	داخل	دوما	دون	دونما
ذاك	ذلك	رغم	ريثما	زهاء	ساعة
سنة	سوف	سوى	سويا	شتى	شرق
شريطة	شكرا	شمال	صبيحة	صوب	ضد
طالما	طبقا	طواعية	طوعا	طيلة	ظل
عادة	عام	عامة	عبر	عدا	عدة
عسى	عشية	عقب	على	عما	عمن
عموما	عن	عند	عندئذ	عندما	عنوة
غالب	غالبا	غدا	غداة	غرب	غير
فجأة	فجر	فحسب	فصاعدا	فضلا	فلا
فور	فورا	فوق	في	فيما	قبالة
قبل	قبيل	قد	قدما	قرابة	قرب
قسرا	قطعيا	قليلا	كأن	كالمعتاد	كان
كثيرا	كذا	كل	كلا	كلتا	كم
كما	كي	كيف	لئلا	لا	لابد
لاحقا	لاسيما	لحظة	لحوالي	لدى	لذا

لعل
لقد
لكن
لكي
للتو
لم
لماذا
لن
لو
لولا
ليت
ليس
ليلة
مؤخرا
مؤقتا
ما
ماذا
مباشرة
متى
مثل
مثلا
مثلما
مجانا
مجددا
مجرد
محض
مرارا
مساء
مطلقا
مع
معا
معظم
مما
ممن
من
منذ
مهما
نادرا
نحن
نحو
نسبيا
نعم
نعمت
نفس
نهار
نهارا
هؤلاء
ها
هاتان
هاتين
هدرا
هذا
هذان
هذه
هذين
هكذا
هل
هم
هما
هن
هنا
هناك
هنالك
هو
هي
وراء
وسط
وفق
وفقا
وقت
يا
يوم
يوما
يوميا

Appendix B Optimized Arabic Function Words using feature selection techniques

أكثر
أمام
أن
أو
أي
أيضا
إذ
إذا
إزاء
إلا
إلى
إن
إنما
الآن
التي
الذي
الذين
اليوم
بالذات
بجانب
بحيث
بعدما
بل
بلا
بينما
تلك
ثم
ثمة
جدا
جنوب
حوالي
حول
حيث
حيثما
حين
حينما
خلال
دائما
دون
ذلك
سنة
سوف
سوى
شمال
ضد
طبقا
طيلة
ظل
عام
عبر
عقب
على
عند
عندما
فحسب
فضلا
في
فيما
قد
كان
كل
كي
لا
لذا
لقد
لكن
لم
لماذا
لن
لو
ليس
ما
مثلما
مساء
معا
معظم
من
نحو
نعم
نفس
هذا
هذه
هكذا
هنا
هناك
هو
هي
وراء
وسط
وفق
وفقا
وقد
يا

Vitae

Name	:Sadam Hussein Mohammed Al-Azani
Nationality	:Yemeni
Date of Birth	:2/1/1982
Email	: g201002580@kfupm.edu.sa; Sadamal-azani@hotmail.com
Academic Background	:

- Received Bachelor of Science (B.S.) in Computer Science from Thamar University, Yemen in 2004 with a GPA of 87.02%.
- Joined the Faculty of Computer Sciences and Information Systems as Demonstrator at Thamar University, Thamar, Yemen, from 2007 to 2010.
- Joined the Information and Computer science department as full time student at King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia in September 2011.
- Completed Master of Science (M.S.) in Computer science from King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia in April 2014, with a GPA of 3.75/4.0.