UNIVERSITI PUTRA MALAYSIA

COMPUTATIONAL STYLOMETRIC MODEL FOR OATH AND OATH-LIKE EXPRESSIONS IN QURANIC TEXT

AHMAD ALQURNEH

FSKTM 2014 32

COMPUTATIONAL STYLOMETRIC MODEL FOR OATH AND OATH-LIKE EXPRESSIONS IN QURANIC TEXT

By

AHMAD ALQURNEH

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

October 2014

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

COMPUTATIONAL STYLOMETRIC MODEL FOR OATH AND OATH-LIKE EXPRESSIONS IN QURANIC TEXT

By

AHMAD ALQURNEH

October 2014

Chair: Aida Mustapha, Ph.D.

Faculty: Computer Science and Information Technology

Computational stylometry is to analyze a written text based on measurable style markers. The current major challenges in stylometric analysis are to be found in terms of scalability, which concerns mainly in authorship attribution applications as how to handle short texts in terms of identifying author; generalization, which investigates property transfer that happen among different text; and explanation, which focuses on support to increase understanding rather than maximizing performance.

To deal with these issues in the domain of Quran, the definitions of scalability, generalizations, and explanations issues are to reshape to be compatible with Quran area. Hence, we illustrate the problem within the domain of oaths and oath-like expressions in the Quran due to its indirect declaration in the context of its literary style as implicit oaths. Therefore, in this research, the scalability issue concerns on how far stylometric features are scalable to detect implicit oaths that can lead to the identification of the oath taker. Generalization issue concerns on measuring any style properties transfer between implicit and explicit oaths that emerge the difference in style characteristics of oath takers. Finally, explanation concerns whether stylometry is able to achieve better understating for the stylistics knowledge of oath compare to the linguistic studies.

Quranic oath is a swear from Quranic verses used to emphasize the importance or truthfulness of a concept that follows the oath in the verse. Oaths are multi-faceted and rich expressions. Oath expressions in Quranic texts are with implicit and explicit forms. Explicit form of oath expression is based on existing of swearing verb while, implicit form is devoid from the swearing verb. This lead to several

interpretations that comes from a variety of aspects. While the Quranic oath has been widely studied from theoretical base in Islamic Studies, it has not being treated from computational perspective. This research proposes a new computational stylometric model for oath and oath like expression detection for implicit and explicit oaths. The model is to detect explicit oaths as well as implicit oaths, inspect any common stylistics properties between the two forms and, has well explanation of the detected oath aspects.

This work proposes an oath-like expression detection algorithm (OLEDA) to develop a computational stylometric model for oaths based on stylometric application-specific features which are structural and content-specific features. The selection of such features is because they can be defined in certain text domains or languages. The aim of these features is to detect both forms of oaths, i.e implicit as well as explicit. Subsequently, the application-specific features are evaluated in terms of their activity towards oath detection through a series of machine learning experiments using various classifiers such as the Bayesian network.

To improve oath detection by OLEDA with scalable features in stylometry, character features are added, particularly character n-gram features, bigram and trigram. Such features are used to select any special characters that commence the oath statement to handle implicit oath in achieving scalability.

To differentiate the oath takers of the implicit and explicit oaths, we examine any common or uncommon properties transfer between them. For this, we performed additional stylometric analysis using syntactic and lexical features. Syntactic features enable us to extract a new feature based on the rewrite rule frequencies. This feature is used to split the oath statements into chunks, where each chunk will be assigned to its corresponding morphological meaning with reference to a standard syntactic Treebank. Next, is the lexical features bag, which include token-based features, short words features, word n-gram features, vocabulary richness functions, Hapax Legomena, Hapax Dislegomena, and frequent function words that discriminate oath-takers in achieving generalization.

Finally, in achieving better explanation on oaths, this research performed two-fold validation of the proposed stylometric model to oath and oath-like expressions. One, because oaths have not being studied from the computational perspective, the stylometric model is compared against an existing linguistic model of oaths. Two, various experimental results from the proposed model are compared against expert evaluation. The results showed that the proposed stylometric model of oaths has scalability in detecting implicit oaths, obtains dissimilar generalization levels in discriminating oath takers, and better adding in oath expressions explanation.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MODEL PENGKOMPUTAN STAILOMETRIK BAGI SUMPAH DAN EXPRESI SUMPAH DI DALAM QURAN

Oleh

AHMAD ALQURNEH

Oktober 2014

Pengerusi: Aida Mustapha, Ph.D.

Fakulti: Sains Komputer dan Teknologi Maklumat

Pengkomputan stailometrik adalah untuk menganalisis teks bertulis berdasarkan penanda gaya yang boleh diukur. Cabaran semasa yang utama di dalam analisis stailometrik adalah dari segi penskalaan yang dititikberatkan di dalam aplikasi pengiktirafan pengarang iaitu bagaimana untuk mengendalikan teks-teks yang pendek bagi mengenal pasti penulis; penyeluruhan yang menyiasat pemindahan ciri yang berlaku di antara teks yang berbeza; dan penjelasan yang menumpukan kepada sokongan bagi meningkatkan kefahaman berbanding dengan memaksimumkan prestasi.

Bagi menangani isu-isu ini dalam domain Al-Quran, definisi penskalaan, penyeluruhan, dan penjelasan adalah dibentuk agar bersesuaian dengan bidang Al-Quran. Oleh yang demikian, kami menggambarkan permasalahan di dalam bidang sumpah dan ungkapan sumpah di dalam Al-Quran kerana pengisytiharannya yang tidak langsung dalam dalam konteks gaya kesusasteraan sebagai sumpah yang tersirat. Oleh itu, di dalam penyelidikan ini, isu penskalaan menitikberatkan sejauh mana ciri-ciri stailometrik boleh diskalakan bagi mengesan sumpah tersirat yang membawa kepada pengenalpastian pembawa sumpah. Isu penyeluruhan pula menitikberatkan pengukuran pemindahan ciri-ciri gaya di antara sumpah tersirat dan tersurat yang akan membezakan ciri-ciri gaya pengambil sumpah. Akhir sekali, penjelasan menitikberatkan keupayaan stailometrik dalam memantau kefahaman yang lebih baik bagi pengetahuan gaya sumpah berbanding the pengajian linguistik.

Sumpah Al-Quran adalah sumpah daripada ayat-ayat Al-Quran yang digunakan untuk menekankan kepentingan atau kebenaran konsep menurut sumpah dalam sesuatu ayat. Sumpah adalah ungkapan yang kaya dengan kepelbagaian rupa.

Ungkapan sumpah dalam teks-teks Al-Quran adalah dalam bentuk tersirat dan tersurat. Ekspresi sumpah tersurat adalah berdasarkan kewujudan kata kerja sumpah, manakala bentuk tersirat adalah ketiadaan kata kerja tersebut. Ini Ini membawa kepada beberapa tafsiran yang datang dari pelbagai aspek. Walaupun sumpah Al-Quran telah dikaji secara meluas dari segi teori dalam Pengajian Islam, ianya tidak dikaji dari segi pengkomputan. Kajian ini mencadangkan model pengiraan stailometrik yang baharu untuk mengesan sumpah dan ungkapan sumpah yang tersirat dan tersurat. Model ini juga memeriksan ciri-ciri gaya yang setara antara kedua-dua bentuk sumpah memberi penjelasan kepada aspek pengesanan sumpah tersebut.

Penyelidikan ini mencadangkan satu algoritma pengesan ungkapan sumpah (OLEDA) dalam membangunkan suatu model pengkomputan stailometrik bagi sumpah berdasarkan ciri-ciri khusus aplikasi iaitu ciri-ciri struktur dan kandungan khusus. Pemilihan ciri-ciri adalah kerana ciri-ciri tersebut boleh ditakrifkan di dalam domain teks atau bahasa. Ciri-ciri ini adalah bertujuan untuk mengesan kedua-dua bentuk sumpah, iaitu yang tersirat dan juga tersurat. Seterusnya, ciri-ciri khusus aplikasi akan dinilai dari segi aktiviti terhadap pengesanan sumpah melalui satu siri eksperimen berasaskan pembelajaran mesin eksperimen dengan menggunakan pelbagai pengelas seperti rangkaian Bayesian.

Untuk memperbaiki pengesanan sumpah oleh OLEDA dengan ciri-ciri penskalaan dalam stailometi, ciri-ciri aksara telah ditambah, terutamanya ciri aksara n-gram, bigram dan trigram. Ciri-ciri tersebut digunakan untuk memilih sebarang aksara khas yang memulakan penyataan sumpah dalam mengendalikan sumpah yang tersirat dan seterusnya mencapai penskalaan.

Untuk membezakan pembawa sumpah tersirat dan tersurat, kita memeriksa pemindahan ciri yang biasa atau luar biasa antara kedua-dua jenis sumpah. Untuk ini, kami menjalankan analisis stailometik tambahan dengan menggunakan ciri-ciri sintaktik dan dan leksikal. Ciri-ciri sintaktik membolehkan kami mengeluarkan satu ciri baharu berdasarkan frekuensi petua penulisan semula. Ciri ini digunakan untuk memisahkan penyataan sumpah menjadi bahagian kecil, yang mana setiap bahagian tersebut diperuntukkan kepada maksudnya morfologi masing-masing dengan merujuk Treebank sintaktik yang piawai. Seterusnya, beg ciri leksikal merangkumi ciri-ciri berasaskan token, ciri-ciri kata-kata pendek, ciri-ciri n-gram, fungsi kekayaan kosa kata, Hapax Legomena, Hapax Dislegomena, dan kata-kata fungsi kerap yang mendiskriminasi pengambil sumpah lalu mencapai penyeluruhan.

Akhirnya, dalam mencapai penjelasan yang lebih baik berkenaan sumpah, kajian ini menjalankan dua segi pengesahan ke atas model stailometrik sumpah dan ungkapan sumpah yang dicadangkan. Pertama, oleh kerana sumpah tidak dikaji dari perspektif pengkomputan, model stailometrik ini akan dibandingkan dengan model linguistik sumpah yang sedia ada. Kedua, beberapa keputusan eksperimen yang terhasil daripada model yang dicadangkan akan dibandingkan dengan penilaian pakar. Keputusan yang didapati menunjukkan bahawa model stailometrik

yang dicadangkan mempunyai keupayaan pengskalaan dalam mengesan sumpah tersirat, mencapai paras penyeluruhan dalam membezakan pembawa sumpah, serta penjelasan lebih baik dalam ungkapan-ungkapan sumpah.

ACKNOWLEDGEMENTS

My thanks go firstly, to Allah Almighty, who blessed me with the ability to accomplish this work.

This work could not have been carried out without help and support from my supervisor, Dr. Aida Mustapha. My most heartfelt thanks go to you, for introducing me to the topic of Computational Stylometry and for your continuous support. I am grateful to you for your commitment, the encouragement you provided, and for your patience. Dr. Aida is an excellent supervisor. Albeit many long hours, it has been a pleasure working with you. I am so proud that I am your student.

Great thanks go to my supervisory committee members, Associate Professor Dr. Masrah Azrifah Azmi Murad and and Dr. Nurfadhlina Mohd. Sharef for their valuable comments and discussions.

This research is funded by Fundamental Research Grants Scheme by Ministry of Higher Education Malaysia in collaboration with the Centre of Quranic Research at University of Malaya, Malaysia. Thanks to Universiti Putra Malaysia and the Centre of Quranic Research at University of Malaya for the support.

Finally, I would like to thank my wife and my beloved children for their support and inspiration.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

The members of the Supervisory Committee were as follows:

Aida Mustapha, Ph.D.
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Masrah Azrifah Azmi Murad, Ph.D.
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Nurfadhlina Mohd. Sharef, Ph.D.
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG KIM HUAT, Ph.D.
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by Graduate Student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intelelctual property from the thesis and copyright of thesis are fully-owned by Universiti Putra malsyia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Unievrsiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: _____ Date: _____

Name and Matric No: Ahmad Alqurneh (GS 32676)

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsbilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: _____
Name of
Chairman of
Supervisory
Committee: Aida Mustapha, Ph.D.

Signature: _____
Name of
Member of
Supervisory
Committee: Masrah Azrifah Azmi Murad , Ph.D.

Signature: _____
Name of
Member of
Supervisory
Committee: Nurfadhlina Mohd. Sharef, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BN	Bayesian Networks
DT	Decision Tree
HL	Hapax Legomena
HD	Hapax Dislegomena
IBL	Instance-based Learning
MLP	Multilayer Perceptron
OLE	Oath-like Expressions
OLEDA	Oath-Like Expressions Detection Algorithm
SOD	Stylometric Oath Detection

CHAPTER 1

INTRODUCTION

1.1 Background

Stylometry is defined as a line of research studying quantification of writing style (Sidorov et al., 2014). It is a collection of techniques that allow measurement of styles among authors by the identification of features of style markers which are obtained from textual measurements. Beside, it is also used as a technique to resolve the task of authorship attribution (Lopez-Escobedo et al., 2013). From authorship perspective, it is defined as the processes of examining the writing style of a text to reveal the identity of its author (Holmes, 1998; Stamatatos, 2009).

Stylometry was introduced to complement traditional literary experts work by providing an alternative means of investigating works of doubtful provenance and revealing the identity of works author (Holmes, 1998; Sancho, 2009). The development of stylometry was reflected by the choice of quantifiable features (Holmes, 1998) or style markers (Stamatatos, 2009). In general, stylometry starts with authorship attribution task, uses word length distribution as a feature to discriminate authors, and then more development occurred during the last decades by introducing and testing new features with new techniques.

Recent challenges in computational stylometry concerns on three aspects; scalability, cross-genre stylometry and generalization, and explanation (Daelemans, 2013). In computational stylometry, scalability is a fundamental problem because for a stylometric approach to be scalable, it should be able to handle very short texts. Scalability is defined as the ability to achieve consistent performance under various uncontrolled settings, such as variations in topic, genre, the number of candidate authors, and the amount of textual data available per author (Luyckx, 2010). Generalization is defined as the ability to detect properties travelling from one genre to the other or even within same genre, in distiguishing the author's writing styles. Finally, explanation is defined as the ability to increase understanding among the people who evaluates the texts.

In this research, we intend to address the challenges (scalability, generalization, and explanation) within the domain of Quranic oath and oath-like expressions. To the best of our knowledge, oaths have not being studied from computational or stylometric perspective. The literature has shown that research on oaths in linguistic and Islamic studies are theoretical-based (i.e. Hassan (2003), Hashmi (2008), Issa (2009)) and do not provide any systematic computational methodology to evaluate the components in oath expressions.

The motivation of this research is to utilize stylometric concepts from computing perspective in order to contribute and resolve argumentative issues exist in liter-

ary studies such as the dispute opinion towards the value of oath objects used in the oath statements such as in (Hashmi, 2008). For example, one opinion claims that objects have some glory value, while others simply neglect similar value. In addition, the richness of oath expressions make it difficult for individual to recognize implicit (i.e one word) oaths while reading Quran, explore its undetected style characteristics, or identify its oath taker.

This research is hoped to be able to shed answers on inquiries such as on how to explore the deep stylistic properties of oath styles, how to identify the writing characteristics of oath takers in the Quran, how to identify an oath statement without the oath verb or noun or what is the relationship between the oath takers and the stylistic properties of the oath styles. We believe that only through stylometry that the essence of writing style for oaths in the Quranic text can be captured and quantified, which will lead to a successful investigation that satisfies such inquiries.

1.2 Problem Statement

Oath statement has different styles in Quran (Hassan, 2003). It could be implicit and short in the form of one word, or explicit in a complete sentence. With regard to the oath taker of the oath statement, the oath takers are God and human. In some cases, the oath statement is devoid from its main component that is the swearing verb, where it is difficult to detect such statement as well as difficult to recognize its taker. Therefore, the main concern of this research is how to detect implicit and explicit oaths, extract stylistics knowledge of Quranic oath styles, particularly the implicit one word oath and compare it with explicit oaths, and to identify oath styles potential oath takers that enhance explanation on oaths and increase our understanding. For this purpose, we use the recent challenging issues in stylometry and redefine them in our oath domain. These issues are, scalability, generalization, and explanation (Daelemans, 2013).

Scalability in stylometry is concerns on how to examine short texts compare to longer texts within authorship attribution applications in order to identify the author of the short texts (Luyckx, 2010). With regards to Quranic oath, oath takers are God or human. In addition, oath statements may exist in short statements as implicit one word or explicit statements as a complete sentence with all its components, which is easier to be detected. But in the case of implicit one word oath, it is only one single word, which makes it difficult to detect, especially with the absence of major oath statement component such as the swear verb *uqsim*. In this research, since we are are not dealing with authorship attribution and analysing different texts, but with different statements from one text, we mean by scalability is to find scalable features that can detect the implicit one word oath statement with the absence of oath statement components in order to identify the oath taker.

In addition, because the oaths under study are sourced from the Quranic verses, the unique characteristics of Arabic language are critical especially the length in

short text (Shalhoub et al., 2010). We believe that scalability limit should rise to achieve performance in handling more than a short text, but up to a single word. To achieve scalability in this research, the following inquiries must be solved: What are the best features to increase the detection of the approach towards one word oaths? Is it applicable for one standalone feature to accomplish the task or does it requires combination of robust features? Is it possible to use content-specific keywords without compromising scalability?

Generalization concerns on how stylometry handle the stylistic properties of individual or group of authors transfer from one genre to the other. Generalization is most concerned on investigating the relationship between style and content among the same or different genres for individual or group of authors (Daelemans, 2013). However, stylometric applications become more difficult within single textual genre such in authorship attribution (Kestemont et al., 2012). Generalization also explores style detection rather than topic detection. In this research, we mean by generalization, are there any stylistics properties transfer between implicit oath styles and explicit oath styles, and its relation to the oath takers.

In literature studies, the interest is to observe the oath topic rather than the oath style (Hassan, 2003; Hashmi, 2008). A successful stylometric model should be able to detect any stylistic property transfer between authors among same or different genre. In investigating oath styles, we will investigate different oath statement for the same oath taker as well as for different oath takers. The levels of generalization are important to attach each oath style to its taker (i.e. 'Allah' vs. narrative). To achieve generalization, the following research inquiries must be solved: Are there any travelling properties inside the same or different categories of oath styles? In low level generalization, does that lead to the same oath taker? Is there any relation between generalization and scalability that is consistent with identifying the oath takers?

Explanation concerns on how stylometry increase understanding in analyzing the text rather than maximizing performance of detection (Daelemans, 2013). There is a rare explanation of the stylometric results in previous research. In this research, we mean by explanation is to show how computational stylometry can observe details on oath not found in theoretical studies as well as to increase our understanding on oath multiplicity. In linguistic studies, the works on oath are mainly focusing on arguing the glory in oath objects (see Hashmi (2008) or studying the morphological analysis of the oath statements (see Issa (2009). The only study found that distinguish the oath takers in Quran is by Hassan (2003). Although this work distinguishes oaths based on the oath object used, still it does not tell us if there are any specific values of these oath objects or stylistic properties in the objects themselves. However, in linguistic studies, the gap still exist in ongoing discussion on oath object value whether it has a value or not (Hashmi, 2008).

We believe that computational stylometry would be able to fill such gap and quantify some values to answer and explain such queries. To achieve high ex-

planation, the following inquiries must be solved: Does the proposed stylometric model increase the understanding of the apparent and narrative oaths, and able to show the differences between the two oaths? Does it show the differences of style properties of oath takers? Does it satisfy the conditions of reliability, validity, and do not violate (break) the legitimate (authentic) oath structure?

1.3 Objectives of Research

The main objective of this research is to propose and develop a new computational stylometric model for oath and oath-like expressions that is capable of detecting implicit one word oaths that are devoid from main oath component as well as explicit oaths on the basis of proposing a stylometric scalable features that lead to the identification of the oath taker, to quantify any transfer of stylistics properties between implicit oaths and explicit oaths, and computationally attempt to explain argumentative issues on oath object value found in linguistic studies. To achieve this, the following tasks are to require:

- To detect implicit and explicit oaths by analyzing stylometric domain features such as application-specific features and, to improve oath detection by analyzing scalable stylometric features such as character n-gram features.
- To disclose any common or uncommon styles properties transfer between implicit oaths and explicit oaths by analyzing syntactic feature and lexical features that is able to disclose the diversity of oath styles properties.
- To achieve better understanding on oath by corresponding the proposed stylometric approach results to the foundation of the linguistic studies in terms of oath styles and oath takers.

1.4 Scope of Research

The interest of this research is to improve the understanding for oath statement in literary works based on stylistics knowledge extracted for each other taker. The works include original resources, but avoids stylistics description of oath styles in terms of oath takers. The works on oath are chosen from recognized Islamic books and articles for different Islamic scholars.

The dataset of Quran are sourced from the web found in social network sites in MS Access 2007. Because it is a religion book, the verification of the content to match the hard copy of Quran is done manually before any testing is made. The type of oaths investigated in this research in scoped to apparent oaths and two cases of narrative oaths. It is also limited to the analysis of oath and oath-like expressions that appear in the beginning of Quranic text. Narrative oath may be generalized to Arabic literature that may consist similar oath structures used by human. Stylometry features (e.g. structural and character n-grams) may be generalized to classical Arabic texts that consist similar oath structure such as classical poetry that starts with oath as in *Jahili's* litreture (pre-Islamic period).

Since there is no previous computational model to compare with, the comparison is with linguistic model. The attributes used in comparing with a linguistic model are chosen to evaluate the value of oath object such as token-based features, vocabulary features, and function words compare with descriptive values of oath components.

Implementation-wise, the language of choice for datasets is limited to Arabic for the reason of using original text. Meanwhile, all programs are written in Visual Basic 6.0 with MS Access database.

1.5 Research Methodology

The research methodology adopted for this research compromises of collecting the theory background on oath and oath-like expressions, proposing a new stylometric model to detect oath and oath-like expressions which are not detected by the human reading such as implicit oaths, to identify the oath taker of different oath styles, and to output a reliable explanation on oaths. Finally, we use model analysis approaches such as machine learning, statistical analysis, and expert evaluation techniques for validating the proposed stylometric model.

In literature review, we consider extensive literature base on two main areas: computational stylometry since the seminal work of Holmes (1992) and oaths in linguistic and Islamic studies such as Hassan (2003), Issa (2009), Hashmi (2008) as well as the adjacent areas, which include computational studies on Quranic texts such as Sharaf and Atwell (2011) and Sayoud (2012).

In stylometric analysis for the proposed model, we design four investigation levels. The first level uses application-specific features, which are structural and content-specific features. Its responsibility is to detect oath and oath-like expressions. The second level uses the syntactic feature of rewrite rule frequencies. Its responsibility is to assign oath statement components to its morphological meaning. The third level uses the character n-gram features; trigram and bigram. Its responsibility is to handle one-word oath statements to demonstrate scalability. The last level uses lexical features set, which include word n-gram, token-based features, vocabulary richness, and frequent function words. Its responsibility is to show the properties transfer among oath styles to show levels of generalization.

Finally, the proposed stylometric model is then validated in two manners; by comparing the stylometric model against existing linguistic model as well as via expert evaluation.

1.6 Contribution of the Research

Our main focus in this research is to address current challenging issues in computational stylometry within the domain of Quranic oaths in order to detect different oath styles particularly the implicit oaths and to discriminate their oath taker on a computational stylometic basis that evolve and reveal undetected stylistics

knowledge of oaths. The contributions would be in three-fold; scalability, generalization, and explanation of oaths and oath-like expressions. To achieve the three issues, this research introduces a new computational stylometric model based on a proposed oath and oath-like expression detection algorithm (OLEDA). The main contributions are summarized as follow:

- Enhanced OLEDA is achieving oath detection for implicit and explicit oaths based on combining scalable features with domain features.
- Computational stylometric functions such as vocabulary richness and function words are successfuly differentiating oath takers in achieving generalization issue.
- The stylometric validation techniques via the selected questions by experts approved the better understanding in achieving explanation.

1.7 Organization of this Thesis

This thesis is organized in accordance to the standard structure of thesis dissertations for Universiti Putra Malaysia. The thesis is divided into six chapters.

Chapter 1 – Introduction. This chapter introduces the problems in stylometry, which are scalability, generalization, and explanation. These inherent problems in stylometry forms the problem statement and sets to solve the problems within the domain of oaths and oath-like expressions.

Chapter 2 – Review on Stylometry. This chapter reviews related studies in computational stylometry, and further in depth review on related stylometric works in language studies, Arabic language, and in particular, Quranic text. Most importantly, this chapter summarizes different stylometric features to be used in modeling the oaths.

Chapter 3 – Methodology. This chapter introduces the main methodology structure used in this research, starting from gathering the dataset until the final evaluation step. The backbone for this chapter is the linguistic model derived from previous theoritical studies, which is apparent oath in Quran, and the computational stylometry features and methods used in the previous Literary works, such as vocabulry richness functions, frequent function words, and analysis methods such as machine learning and statistical analysis approachs, as well as validation process performed on the model.

Chapter 4 – Stylometric Model for Oaths. This chapter introduces oaths from linguistic and Islamic studies. Then it presents the proposed stylometric model for oaths and oath-like expression.

Chapter 5 – Analysis and Validation. This chapter presents two parts of stylometric analysis, which is at application-specific level, as well as linguistic level

(e.g. syntactic, character, lexical). Following the analysis, this chapter compares the propsoed stylometric model with existing linguistic model from the literature as well with expert evaluation.

Chapter 6 – Conclusion and Future Works. Finally, this chapter concludes the research with some recommendations for future development.

BIBLIOGRAPHY

Abdel Haleem, M. A. S. 2005. *The Qur'an: A New Translation*. New York: Oxford University Press.

Abdul-Raof, H. 2001. *Qur'an Translation: Discourse, Texture and Exegesis*. Curzon Press.

Al-Kasani, A. 1986. *Badaai us Sanaai*. 2nd edn. Beirut: Daru Kotob Elmiyya.

Al-Muhtaseb, H., Mahmoud, S. A. and Qahwahi, R. S. 2009. A Novel Minimal Arabic Script for Preparing Databases and Benchmarks for Arabic Text Recognition Research. In *Proceedings of Wavelets Theory and Applications in Applied Mathematics, Signal Processing and Modern Science*.

Al-Taani, A. and Al-Gharaibeh, A. 2011. Searching Concepts and Keywords in the Holy Quran. In *Prooceedings of the International Arab Conference on Information Technology*. Sudan.

Aldhaln, K. A. and Zeki, A. M. 2012. Knowledge Extraction in Hadith using Data Mining Technique. In *Proceedings of the 2nd International Conference on E-Learning and Knowledge Management Technologies (ICEKMT 2012)*. Malaysia.

Alkhatib, M. 2010. Classification of Al-Hadith Al-Shareef using Data Mining Algorithm. In *Proceedings of the European Mediterranean and Middle Eastern Conference on Information Systems (EMCIS 2010)*. UAE.

Alotaiby, F., Foda, S. and Alkharashi, I. 2011. Clitics in Arabic Language: A Statistical Study. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 595–601. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Altantawy, M., Habash, N., Rambow, O. and Saleh, I. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Malta.

Argamon, S. 2007. Stylistic Text Classification using Functional Lexical Features. *Journal of the American Society for Information Science and Technology* 58 (6): 802–822.

Atwell, E., Brierley, C., Dukes, K., Sawalha, M. and Sharaf, A. 2011. An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet. In *National Information Technology Symposium (NITS)*. Riyadh, Saudi Arabia.

Bellaachia, A. and Jimenez, E. 2009. Exploring Performance-based Music Attributes for the Stylometric Analysis. *World Academy of Science, Engineering and Technology* 55.

Bergsma, S., Post, M. and Yarowsky, D. 2012. Stylometric Analysis of Scientific Articles. In *Proceedings of HLT-NAACLThe Association for Computational Linguistics*, 327–337.

Boldrini, E., Fernandez, J., Gomez, J. M. and Martinez-Barco, P. 2009. Machine Learning Techniques for Automatic Opinion Detection in Non-Traditional Textual Genres. In *Proceedings of WOMSA*. Seville, Spain.

Chiang, D., Diab, M., Habash, N., Rambow, O. and Sharif, S. 2006. Parsing Arabic Dialects. In *Proceeding of the European Chapter of the Association for Computatinal Linguistics (ACL)*. Trento, Italy.

Daelemans, W. 2013. Explanation in Computational Stylomery. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing*, 451–462.

Dukes, K., Atwell, E. and Habash, N. 2013. Supervised Collaboration for Syntactic Annotation of Quranic Arabic. *Language Resources and Evaluation* 47 (1): 33–62.

Dukes, K., Atwell, E. and Sharaf, A. 2010. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. In *Proceedings of the Language Resources and Evaluation Conference*. Valletta, Malta.

Foster, D. 1999. The Claremont Shakespeare Authorship Clinic: How Severe are the Problems. *Computers and the Humanities* 32 (6): 491–510.

Gill, P. and Swart, T. 2011. Stylometric Analyses using Dirichlet Process Mixture Models. *Journal of Statistical Planning and Inference* 14: 3665–3674.

Golcher, F. 2007. A New Text Statistical Measure and its Application to Stylometry. In *Proceedings of the Corpus Linguistics Conference* (eds. S. H. In Matthew Davies, Paul Rayson and P. D. (eds.)). University of Birmingham.

Habash, N., Rambow, O. and Kiraz, G. 2005. Morphological Analysis and Generation for Arabic Dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 17–24.

Hashmi, T. M. 2008. *A Study of the Quraanic Oaths: An English Translation of Im'an Fi Aqsam Al-Qur'an by Hamid Al-Din Farahi*. Lahore: Almawrid.

Hassan, S. A. 2003. Style of Apparent Oath in the Holy Quran: Eloquence and Purposes. *Journal of Sharia and Islamic Studies* 18 (53).

Hattab, A. 2012. Arabic Content Classification System using Statistical Bayes classifier with Words Detection and Correction. *World of Computer Science and Information Technology Journal* 2 (6): 193–196.

Hilali, M. T. and Khan, M. M. 2007. *Translation of the Meaning of the Noble Qur'an in the English Language*. Madinah, Saudi Arabia: King Fahd Printing Complex.

Hirst, G. and Feiguina, O. 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing* 22 (4): 405–417.

Holmes, D. I. 1992. A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society* 155 (1): 91–120.

Holmes, D. I. 1994. Authorship Attribution. *Computers and the Humanities* 28: 87–106.

Holmes, D. I. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13 (3): 111–117.

Holmes, D. I. and Forsyth, R. S. 1995. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing* 10: 111–127.

Hoover, D, L. 2003. Another Perspective on Vocabulary Richness. *Computers and the Humanities* 37: 151–178.

Houvardas, J. and Stamatatos, E. 2006. N-gram Feature Selection for Authorship Identification. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, 77–86. Varna, Bulgaria: Springer.

Ibn Ya'eesh, M. 2001. *Sharhul Mofassal.* , vol. 8 & 9. Beirut: Darualkotob Al-Elmeyya.

Ibrahim, M. Z. 2009. Oaths in the Qur'an: Bint al-Shati's Literary Contribution. *Islamic Studies* 48 (4): 475–III.

Iqbal, F., Binsalleeh, H., Fung, B. C. M. and Debbabi, M. 2010. Mining Writeprints from Anonymous E-mails for Forensic Investigation. *Digital Investigation* 7 (1): 56–64.

Issa, A. 2009. *Oath in Juz Amma*. PhD thesis, Darululum College, Cairo University.

Kanaris, I. 2009. Learning to Recognize Webpage Genres. *Information Procesing and Managment Journal* 45 (5): 499–512.

Kao, J. and Jurafsky, D. Retrieved July, 2013, A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry, http://www.stanford.edu/ jurafsky/kaojurafsky12.pdf.

Kaplan, D. and Blei, D. 2007. A Computational Approach to Style in American Poetry. In *Proceedings of IEEE Conference on Data Mining*.

Karlgren, J. 2000. *Stylistic Experiments for Information Retrieval*. PhD thesis, University of Stockholm.

Kessler, B., Nunberg, G. and Schutze, H. 1997. Automatic Detection of Text Genre. In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL'97)*, 32–38.

Kestemont, M., Luyckx, K., Daelemans, W. and Crombez, T. 2012. Cross-Genre Authorship Verification using Unmasking. *English Studies* 93 (3): 340–356.

Khreisat, L. 2006. Arabic Text Classification Using N-Gram Frequency Statistics: A Comparative Study. In *Proceedings of DMIN*, 78–82.

Kjell, B. 1994. Discrimination of Authorship using Visualization. *Information Processing and Management* 30 (1): 141–150.

Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the Internation Joint Conference on Artificial Intelligence*, 1137–1145.

Koppel, M., Argamon, S. and Shimoni, A. R. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 14 (4): 401–412.

Lee, Y. and Myaeng, S. 2004. Automatic Identification of Text Genres and Their Roles in Subject-based Categorization. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*.

Lopez-Escobedo, F., Mendez-Cruz, C., Sierra, G. and Solorzano-Soto, J. 2013. Analysis of Stylometric Variables in Long and Short Texts. In *Proceedings of the 5th International Conference on Corpus Linguistics (CILC2013)*.

Luyckx, K. 2010. *Scalability Issues in Authorship Attribution*. PhD thesis, University of Alicante. University of Antwerp, Belgium.

Muhammad, J. 1997. *Lisan Al-Arab*. Beirut: Darul Sadir.

Nassourou, M. 2011. *A Knowledge-based Hybrid Statistical Classifier for Reconstructing the Chronology of the Quran*. University of Wurzburg, Germany.

Nassourou, M. 2012. Using Machine Learning Algorithms for Categorizing Quranic Chapters by Major Phases of Prophet Mohammad's Messengership. *International Journal of Information and Communication Technology Research* 2 (11).

Nguyen, D., Trieschnigg, D., Meder, T. and Theune, M. 2012. Automatic Classification of Folk Narrative Genres. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, 378–382.

Oakes, M. 2004. Ant Colony Optimisation for Stylometry: The Federalist Papers. In *Proceedings of the 5th International Conference on Recent Advances in Soft Computing*, 86–91.

Pavelec, D., Oliveria, L. S., Justino, E., Nobre, F. D. and Batista, L. V. 2009. Compression and Stylometry for Author Identification. In *Proceeding of the International Joint Conference on Neural Networks*, 669–674.

Philips, A. A. B. 2005. *Usool At-Tafseer: The Methodology of Quranic Representation*. 2nd edn. IIPH.

Popescu, I. 2011. Vocabulary Richness in Slovak Poetry. *Glottometric* 22: 62–72.

Roper, M., Field, J. and Schaalje, B. 2012. Stylometric Analyses of the Book of Mormon: A Short History. *Journal of the Book of Mormon and Other Restoration Scripture* 21 (1): 28–45.

Saad, S., Salim, N., Ismail, Z. and Zainal, H. 2010. Towards Context-Sensitive Domain of Islamic Knowledge Ontology Extraction. *International Journal for Infonomics* 3 (1).

Sadeghi, B. 2011. The Chronology of the Qur'an: A Stylometric Research Program. *Arabica* 58: 210–299.

Sanan, M., Rammal, M. and Zreik, K. 2008. Arabic Supervised Learning Method using N-Gram. In *Proceedings of Interactive Technology Smart Education*, 157–169.

Sancho, C. 2009. *Stochastic Language Models for Music Information Retrieval*. PhD thesis, University of Alicante. Alicante, Spain.

Sanderson, C. and Guenter, S. 2006. Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. In *Proceedings of the EMNLP*, 482–491.

Santini, M. 2004. A Shallow Approach To Syntactic Feature Extraction For Genre Classification. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 04)*. University of Birmingham, UK.

Sayoud, H. 2012. Author Discrimination between the Holy Quran and Prophet's Statements. *Literary and Linguistic Computing* 27 (4): 427–444.

Shalhoub, G., Simon, R. and Iyer, R. 2010. Stylometry System – Use Cases and Feasibility Study. In *Proceedings of the Student-Faculty Research Day*, *CSIS* , A3.1–A3.8. Pace University, NY, USA.

Sharaf, A. and Atwell, E. 2009. A Corpus-based Computational Model for Knowledge Representation of the Qur'an. In *Proceedings of the 5th Corpus Linguistics Conference*. Liverpool.

Sharaf, A. and Atwell, E. 2011. Automatic Categorization of the Qur'anic Chapters. In *Proceedings of the 7th International Computing Conference in Arabic (ICCA11)*. Imam Mohammed Ibn Saud University, Riyadh, KSA.

Sharoff, S. 2007. Classifying Web Corpora into Domain and Gusing Automatic Feature Identification. In *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve, Belgium.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A. and Chanona-Hernandez, L. 2014. Syntactic N-grams as Machine Learning Features for Natural Language Processing. *Expert Systems with Applications* 41: 853–860.

Simaki, V., Stamou, S. and Kirtsis, N. 2012. Empirical Text Mining for Genre Detection. In *Proceedings of the 8th International Conference on Web Information Systems and Technologies*. Salvador, Brazil.

Snider, N. and Diab, M. 2006. Unsupervised Induction of Modern Standard Arabic Verb Classes using Syntactic Frames and LSA. In *Proceedings of COLING-ACL'06*, 795–802.

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

Stamatatos, E. 2013. On the Robustness of Authorship Attribution based on Character n-gram Features. *Journal of Law and Policy* 21 (2): 421–439.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. 2000a. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics* 26 (4): 461–485.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. 2000b. Text Genre Detection Using Common Word Frequencies. In *Proceedings of the18th International Conference on Computational Linguistics*, 808–814.

Stein, B. and Eissen, S. 2006. Distinguishing Topic from Genre. In *Proceedings of I-KNOW*. Graz, Austria.

Thabet, N. 2005. Understanding the Thematic Structure of the Qur'an: An Exploratory Multivariate Approach. In *Proceedings of the Annual Meeting of Association of Computational Linguistics*, 7–12.

Tweedie, F. J., Singh, S. and Holmes, D. 1996. Neural Network Applications in Stylometry: The Federalist Papers. *Computers and Humanities* 30: 1–10.

Uzuner, O. and Katz, B. 2005. Style versus Expression in Literary Narratives. In *Proceedings of the 28th Annual International ACM SIGIR Conference Workshop on Using Stylistic Analysis of Text for Information Access*. Salvador, Brazil.

Whitelaw, C. and Argamon, S. 2004. Systemic Functional Features in Stylistic Text Classification. In *Proceedings of AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*.

Wolters, M. and Mathias, K. M. 1999. Exploring the Use of Linguistic Features in Domain and Genre Classification. In *Proceedings of EACL*, 142–149.

Zaghouani, W., Hawwari, A. and Diab, M. 2012. A Pilot PropBank Annotation for Quranic Arabic. In *Proceedings of the First Workshop on Computational Linguistics for Literature (NAACL-HLT'2012)* (eds. S. H. In Matthew Davies, Paul Rayson and P. D. (eds.)). Montreal, Canada.