Journal of Computer Science

Original Research Paper

The Impact of Oath Writing Style on Stylometric Features and Machine Learning Classifiers

[1]Ahmad Alqurneh and [2]Aida Mustapha

[1]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Malaysia*
[2]*Faculty of Computer Science and Information Technology, Batu Pahat, Universiti Tun Hussein Onn Malaysia*

Received: 21-05-2014
Revised: 12-06-2014
Accepted: 08-09-2014

Corresponding Author:
Ahmad Alqurneh
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia,
Serdang, Malaysia
Email: ahmad.alqurneh@outlook.com

Abstract: Computational stylometry is the field that studies the distinctive style of a written text using computational tasks. The first task is how to define quantifiable measures in a text and the second is to classify the text into a predefined category. This study propose a stylometric features selection approach evaluated by machine learning algorithms to find the finest of the features and to study the impact of the features selection on the classifiers performance in the domain of oath statement in the Quranic text. The results show that better classifiers performance is highly affected by the best feature selection which is associated to an explicit oath style.

Keywords: Stylometry, Feature Selection, Classifiers Performance, Oath Styles

Introduction

Stylometry is the field that studies the writing style of a text. The computational stylometry tasks are authorship attribution and genre text detection, where authorship attribution is an approach concerned about analyzing texts in text mining (Tareef *et al*., 2010) and aim to find the author of unknown text while, text genre detection identifies the kind of the text (Stamatatos *et al*., 2000). Mainly, the computational tasks are feature extraction that defines a quantifiable measures and classification procedure that classify the text into a predefined category. The stylometry features are categorized as lexical, character, syntactic, semantic and application-specific feature (Stamatatos, 2009). Classification procedure could follow statistical or machine learning approaches. In general, the two primary applications in stylometry are authorial studies and chronology problems (Holmes, 1998). These applications are also reflected into the Quran studies. The Quran is an eloquent religious text written by God (Allah) in an adorable literary style, structurally looks as poetic language (Zaghouani *et al*., 2012) and it has its unique genre recognized through two elements, rhetorical and cohesive (Abdul-Raof, 2001). The contents of Quran knowledge representation is becoming one of the recent research areas for the richness of many existing patterns that can be detected in order to increase the understanding for such patters. Patterns like passages starts with qul (say), similes and oath-like expressions (Saad *et al*., 2010a). Oath-like expressions are one of the interested patterns as it serves both law and human day life. In Arabic, Oath-like expressions expression either start with character '*w* ' or phrase '*la uqsim*' (Saad *et al*., 2010b). The detection of the oath-like expression presented in the form of complete phrases such as the phrase '*la uqsim*' could be applicable to the human as it consists of the swearing verb '*uqsim*'. However, in the case of the oath being dependent on character '*w*', the oath exists without a swearing verb (Issa, 2009; Ibrahim, 2009). Moreover, '*w*' character acts like a conjunction character (Sayoud, 2012) and has several uses not only an oath character (Hassan, 2003). This make it difficult to identify the oath-like expression started with '*w*'. To the best of our knowledge, oath expressions were mainly studied from linguistic prospective (Issa, 2009; Ibrahim, 2009; Hassan, 2003).

Therefore, in this study we propose a stylometric application-specific features selection approach to detect two kind of oath, apparent and narrative. Apparent oath is the oath sworn by God directly such

as the oath-like expression begins with '*w*' or '*la uqsim*', this expression might use the objects such as God features or essence (*rabb*) or creatures. Narrative oath is the oath which is narrated as the oath taker is a human. Narrative oath has many cases in Quran (Hassan, 2003) but in this study we include the most common two cases. First case is the '*t*' character base oath, second is the general swearing verb case. Both cases use common oath object which is term 'Allah'.

The selection of mix oaths is to enrich our datasets with different oath types from apparent and narrative oaths. The objectives of the proposed approach are first, to investigate the performance of the application-specific features selection in different dataset sizes consists of oath statements and second to examine the affect of these features on the machine language algorithms performance.

Existing computational works on Quran are like Quran ontology (Saad *et al*., 2010a), authorship attribution (Sayoud, 2012), chapters (surahs) chronology (Nassouro, 2011) and chapters (surahs) categorization (Sharaf and Eric, 2011). The following sections are organized as follow. Section 2 is the application-specific features selection and section 3 is the classification of machine learning algorithms.

Feature Selection for Oath

Feature selection is most important rather than the choice of machine learning method (Daelemans, 2013). However, classifiers selection is highly related to feature extraction and selection. Therefore, the choice of features for oath domain detection requires specifically domain features. Application-specific features are domain specific and can be applied to any language and used to represent the nuances of style in a given text domain (Stamatatos, 2009). The application-specific features selected to detect the oath expression are structural and content-specific features. Structural can quantify the authorial style and significant in very short texts. Also, to better capture the properties of author's style content-specific feature can be used, as definite expressions frequently used within a theme.

Results

In this section, we present a machine learning experiments to oath detection via a series of classification experiments to obtain the main effective factors play an impact role in best feature selection and high classifiers performance. For this we performed two series of classification experiments using the structural and content-specific features with different classifiers, which are Bayesian network, decision tree, instance-based learning and neural network. We used three evaluation measures, which are precision, recall and F-measure as the basis of comparison across the different machine learning algorithms.

Experimental Setup

To run the experiments, we prepared two datasets, the first dataset represents the entire Quran chapters and, the second dataset is the entire Juz' 'Amma (chapter 30) dataset as opposed to global text in the Quran. *Juz' 'Amma* was chosen because oath exists in 40% of its texts (surahs) and it contains 37 surahs (texts), start with the surah number 78 named as The Announcement (*Al-Naba'*) and last with surah number 114 named as People (*Al-Nas*) and its surahs are primarily concerned with oneness of Allah, day of judgment and afterlife (Issa, 2009). The oath types included in these experiments are the apparent oath and narrative oath as explained in section 1.

The proposed stylometric application-specific features selection for oaths is evaluated in four (4) experiments with different datasets as follows:

- Structural feature: Classification of verses (ayat) using structural feature alone, applied on all datasets. Structural feature is used to detect the oath statement occur at the head of the text (surah) whether it is a character '*w*' oath or a keyword oath
- Content-specific feature: Classification of verses (ayat) using content-specific feature alone, applied on all two datasets. Content-specific feature is used to detect only oath statement include specific keyword occur in the text (surah)

In order to control the bias in error rates, we applied 10-fold cross validation (Kohavi, 1995) to split the Quran text into ten approximately equal partitions of training and testing set, each being used in turn for testing while the remainder consolidated for training. The following subsections will present the analysis and discussion in further details.

Structural Feature Performance Evaluation

In the first experiment to analyze structural feature for apparent and narrative oaths detection, we considered the head verses from all 114 texts (surahs) in the Quran chapters' dataset. Meanwhile, in the second experiment for oath detection in small dataset i.e., *Juz' 'Amma*, we considered all head verses from the 37 texts (surahs).

First Experiment: Oath Detection in Entire Quran

In this experiment, any of the head verses from the 114 texts (surahs) will be classified into structural if apparent or narrative oath is detected at the head of the surah. This experiment detected only structural apparent oaths 'w' and 'la uqsim' occurred at the head of the texts (surahs) with absence of structural keywords from narrative oaths. This experiment employed structural-based features in stylometry with Bayesian Networks, DT, IBL and Multilayer Perceptron.

The classification results are represented in Table 1. The results showed that all classifiers have good performance for the precision and F-measure measurements.

Second Experiment: Oath Detection in Juz Amma

In this experiment, all head verses extracted from the 37 surahs in *Juz' 'Amma* will be classified into structural if oath is detected at the head of the surah. This experiment detected structured ap parent oaths 'w' and 'la uqsim' keyword in *Juz' 'Amma*, with the absence of other structured narrative keywords. This experiment employed structural-based features in stylometry with BN, DT, IBL and MLP. The classification results are represented in Table 2. The results showed that the precision and F-measure in DT, IBL and MLP are giving better results than BN.

It is imperative to note that both experiments obtained the higher results for apparent oath, which conclude that oath exist at head of surahs is rich with apparent oath compare to narrative oath.

In general, results from the structural feature experiments clearly indicate that structural feature is better to quantify oath statements at the head of the surahs in both small and big datasets. Hence the second stylometric feature is required in order to expand oath investigation at other parts of the surahs instead of the head, which is content-specific feature.

Content-Specific Feature Performance Evaluation

In the first experiment to analyze content-based feature for oaths and oath-like expressions, we considered all 6,236 verses from the 114 surahs in the Quran chapters' dataset. In the second experiment, we considered 564 verses from the 37 surahs of Juz' 'Amma dataset.

First Experiment: Oath Detection in Entire Quran

In this experiment, apparent and narrative oath verses (ayat) in texts (surahs) will be classified into content-specific if a keyword from apparent or narrative oath is detected using content-specific features. This experiment detected keywords of apparent oaths '*la uqsim*', '*rabb*' and narrative oaths keywords '*aqsam'u b'i Allah*', '*t- Allah*'. The classification results are represented in Table 3. The measurements of precision, recall and F-measure showed that the IBL and MLP give better results than BN and DT.

Second Experiment: Oath Detection in Juz Amma

This experiment detected none of oath keywords. The classification results are represented in Table 4. The measurements of precision, recall and F-measure showed that none of the chosen classifiers produce any detection results on this stylometric feature.

Results from the content-specific experiments clearly indicate that it performs better to quantify oath statements in Quran chapters dataset compare with a small dataset. This concludes that Juz' 'Amma contains very little or none keywords compare to the Quran chapters dataset.

Table 1. Results for structural-based oath detection

	Precision		Recall		F-Measure	
Classifier	True	False	False	False	True	False
BN	0.992	1	0.992	1	0.992	1
J48	0.992	1	0.992	1	0.992	1
IBK	0.992	1	0.992	1	0.992	1
Multilayer perceptron	0.992	1	0.992	1	0.992	1

Table 2. Results for structural-based oath in Juz Amma

	Precision		Recall		F-Measure	
Classifier	True	False	False	False	True	False
BN	0.774	0.981	0.706	0.987	0.738	0.984
J48	0.829	1.000	1.000	0.987	0.907	0.993
IBK	0.829	1.000	1.000	0.987	0.907	0.993
Multilayer perceptron	0.829	1.000	1.000	0.987	0.907	0.993

Table 3. Results for content-specific -feature in entire quran

	Precision		Recall		F-Measure	
Classifier	True	False	False	False	True	False
BN	1.000	0.996	0.067	1.000	0.125	0.998
J48	0.767	0.996	0.767	0.999	0.767	0.999
IBK	0.781	0.996	0.833	0.999	0.806	0.999
Multilayer perceptron	0.781	0.996	0.833	0.999	0.806	0.999

Table 4. Results for content-specific -feature in Juz Amma

	Precision		Recall		F-Measure	
Classifier	True	False	False	False	True	False
BN	0	0.984	0	0.995	0	0.989
J48	0	0.984	0	1.000	0	0.992
IBK	0	0.984	0	0.998	0	0.991
Multilayer perceptron	0	0.984	0	0.998	0	0.991

Table 5. Classifiers performance versus feature selection

High classifier's performance r	Feature	Oath style detected	Dataset size
BN, DT, IBL, MLP	Structural	*w* and *la uqsim*	Big
DT, IBL, MLP	Structural	*w* and *la uqsim*	small
IBL, MLP	Content-specific	*rabb*, *la uqsim* and narratives	Big
None	Content-specific	None	small

Classifier Performance Evaluation

Comparing the performance of the classifiers from the previous experiments it is clearly indicate that the best feature is the structural feature obtained better machine learning algorithms performance compare to content-specific feature experiments as in Table 5.

Conclusion

The results of all the four (4) experiments have led us to two main conclusions. First, machine learning algorithms performance directly depends on the best feature selection. This can be seen from the structural feature which performed best results in both big and small datasets. Second, best feature performance depends on the specific oath styles, as seen structural feature depends on apparent oath i.e the oath-like expressions '*w*' and '*la uqsim*'. For future work, the oath conjunctions which are neighbors of the main oath statement will be studied based on a syntactic n-gram feature, a recent stylometric feature (Sidorov *et al*., 2014).

Acknowledgement

This study was supported by Fundamental Research Grants Scheme from the Ministry of Higher Education Malaysia in collaboration with the Centre of Quranic Research at Universiti Malaya, Malaysia.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abdul-Raof, H., 2001. Qur'an Translation: Discourse, Texture and Exegesis. 1st Edn., Psychology Press, Curzon Press, ISBN-10: 0700712275, pp: 137.

Daelemans, W., 2013. Explanation in computational stylometry. Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Texr Processing, Mar. 24-30, Samos, Greece, pp: 451-46. DOI: 10.1007/978-3-642-37256-8_37

Hassan, S.A., 2003. Style of apparent oath in the holy quran: Eloquence and purposes. J. Sharia Islamic Studies.

Holmes, D.I., 1998. The evolution of stylometry in humanities scholarship. Literary Linguistic Comput., 13: 111-117. DOI: 10.1093/llc/13.3.111

Ibrahim, M.Z., 2009. Oaths in the Qur'an: Bint al-shati's literary contribution. Islamic Studies, 48: 475-498.

Issa, A.R., 2009. Oath in Juz Amma (morphological study). Ph.D. Thesis, Darululum College, Cairo University.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, (IJCAI' 95), Inc. San Francisco, CA, USA 1137-1145.

Nassouro, M., 2011. A knowledge-based hybrid statistical classifier for reconstructing the chronology of the quran. Department of Computer Philology and Modern German Literature University of Würzburg Am Hubland D-97074 Würzburg.

Saad, S., N. Salim, Z. Ismail and H. Zainal, 2010a. A framework for islamic knowledge via ontology representation. Proceedings of the International Conference on Information Retrieval and Knowledge Management, Mar. 17-18, IEEE Xplore Press, Shah Alam, Selangor, pp: 310-314. DOI: 10.1109/INFRKM.2010.5466897

Saad, S., N. Salim, Z. Ismail and H. Zainal, 2010b. Towards context-sensitive domain of islamic knowledge ontology extraction. Int. J. Inform., 3: 197-206.

Sayoud, H., 2012. Author discrimination between the holy quran and prophet's statements. Literary Linguistic Comput., 27: 427-444. DOI: 10.1093/llc/fqs014

Sharaf, A. and Eric, 2011. A automatic categorization of the qur'anic chapters. Proceedings of the 7th International Computing Conference in Arabic, (ICC '11).

Sidorov, G., F. Velasquez, E. Stamatatos, A. Gelbukh and L. Chanona-Hernandez, 2014. Syntactic N-grams as machine learning features for natural language processing. Expert Syst. Applic., 41: 853-860. DOI: 10.1016/j.eswa.2013.08.015

Stamatatos, E., N. Fakotakis and G. Kokkinakis, 2000. Automatic text categorization in terms of genre and author. Computat. Linguist., 26: 461-485. DOI: 10.1162/089120100750105920

Stamatatos., E., 2009. A survey of modern authorship attribution methods. J. Am. Soci. Inform. Sci. Technol., 60: 538-556. DOI: 10.1002/asi.21001

Tareef, K.M., N. Mustapha, A. Masrah, B.S. Azmi and Nasir, 2010. Dropping down the maximum item set: Improving the stylometric authorship attribution algorithm in the text mining for authorship investigation. J. Comput. Sci., 6: 235-243. DOI: 10.3844/jcssp.2010.235.243

Zaghouani, W., A. Hawwari and M. Diab, 2012. A pilot propbank annotation for quranic arabic. Proceedings of the 1st Workshop on Computational Linguistics for Literature, Jun. 23-24, Portland, Oregon, USA.