The Role of Linguistic Feature Categories in Authorship Verification

Ahmed, H.I.A.A.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 214–221

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates

The Role of Linguistic Feature Categories in Authorship Verification

Hossam Ahmed[a]

[a]*Leiden University, Witte Singel 25, 2311 BZ Leiden, The Netherlands*

Abstract

Authorship verification is a type of authorship analysis that addresses the following problem: given a set of documents known to be written by an author, and a document of doubtful attribution to that author, the task is to decide whether that document is truly written by that author. A combination of a similarity-based method and relevant linguistic features is used to achieve high accuracy authorship verification. The method is an author-profiling approach that dispenses with negative-evidence training data, and a number of lexical, morphological, and syntactic features and feature ensembles are used to determine optimal feature use. The method-feature combination is applied to a test corpus of 31 Classical Arabic books and substantially outperforms best available baselines (with 87.1% accuracy). The varying performance of different features and feature ensembles indicate that Classical Arabic authors are less free to individualize their style lexically or morphologically than when involving syntactic structures.

Keywords: Authorship Verification ; Stylometry; Document Similarity; One-class Classification;

1. Introduction

This paper attempts at identifying what kind of linguistic information is most salient in creating an author's style by examining how well different types of linguistic feature perform in an Authorship Verification (AV) problem. AV problems are a type of problem where it is doubtful whether a known author is the writer of a questionable document. If it is possible to develop a high-accuracy AV system that is based on linguistic features, it may be argued that linguistic features that perform better within this system are good descriptors of an individual user of the language, rather than a characteristic of the language or genre in general. Accordingly, this research serves a double-purpose; describe a high-accuracy AV system for Classical Arabic, and provide evidence as to the underlying variation between authors of Classical Arabic texts.

AV is often compared to Authorship Attribution (AA), where a questionable document is known to be written by one author within a group of candidates. As will be seen in the next section, AA problems are typically addressed as classification problems. The questionable document is compared to known works of the various candidate authors, and the author whose work is most similar to the document is considered the winner. AV, on the other hand, is

* Corresponding author. Tel.: +31(0)71-527-4417.
E-mail address: h.i.a.a.ahmed@hum.leidenuniv.nl

1877-0509
10.1016/j.procs.2018.10.478

more complex due to the fact that there is only one candidate author. Many Machine Learning algorithms convert an AV problem into an AA problem by supplementing negative evidence - examples known to be not written by the candidate author (the impostor method). If the questionable document is more similar to the distractors than to the known documents of the candidate author, it is classified as unauthentic. Although this approach simplifies the AV task, its accuracy depends greatly on the quality and choice of the negative evidence supplemented by the algorithm.

Recent developments have allowed for AV tasks to be addressed without the need for negative evidence [1], relying only on properties of sample texts written by the candidate author. Such developments open the door to addressing more general questions about the nature of language variation on the individual level. If indeed a document can be judged with reasonable certainty to be so different that it is unlikely to be written by the same person as other documents, what are the linguistic characteristics that lead to such distinction? Answering this question does not only contribute to developing better feature-based AV systems, but also helps our understanding of how individuals differentiate themselves using Language.

This paper examines the extent to which lexical, morphological, and syntactic features of Classical Arabic contribute to AV in a single-candidate problem. Token, stem and root frequencies are used as indicators of lexical influence. Diacritics offer morphological information about word patterns, and part-of-speech tags are indicators of word derivation as well. Syntactic properties of phrases and sentences can be extracted from n-gram frequencies of lexical and morphological features. Section 3.1 details feature selection, rationale, and how the interaction between language modules is interpreted as feature categories. To examine the role of the various feature types, I use the algorithm and corpus developed by [1], building on a popular similarity metric developed by [2] and compare the outcome to the baselines of [1] and [3].

Section 2 surveys the literature on AA and AV in Arabic, and describes the contribution of this research. Section 3 describes the training and testing corpora as well as the linguistic features and feature categories used for implementing the algorithm. Section 4 describes the procedures for training, testing, and results. The results of the experiment are evaluated and discussed in section 5. Finally, section 6 describes areas for future research.

2. Related Work

There is a great deal of Machine Learning AA and AV research that makes use of linguistic features of different types. [4] show that statistical ML classifiers (SMO-SVM, and MLP) give better results than purely statistical and distance-based classifiers in short text AA tasks. They indicate that rare words and individual words give better results than word n-grams, with rare words giving best results. [5] and [6] examine Naïve Bayes methods in AA of Classical Arabic texts. [7] examines the usefulness of function words in AA in modern Arabic books using Linear Discriminant Analysis (LDA). [8] uses punctuation, function words and clitics in a variety of modern Arabic texts and use ANOVA to achieve acceptable results. While this research offers insights as to which linguistic features can be manipulated in Arabic stylometry, its experimental design makes it less fit to answering larger questions about language variation or AV. An AA question relies too much on negative evidence, hence raising the question whether a given feature is only adequate given the specific distractors. Furthermore, some of the features used (e.g. punctuation in [8]) do not reflect a systematic linguistic property in Arabic[1].

As far as AV is concerned, a number of recent works investigate the author profiling technique (as opposed to the impostor method). [3] relies on a similarity metric inspired by [2] to conduct a number of experiments using nine feature categories to determine the similarity between a given document and a corpus of documents of a given author, based on a Manhattan Distance measure. The nine feature categories are frequent tokens, and n-grams of punctuation, characters, prefixes, suffixes, and a combination of prefixes and suffixes. Their algorithm predicts verified attribution to a given author if the similarity value exceeds a certain threshold value θ, which is defined as the value where false negatives and false positives in the training set are equal; the point of Equal Error Rate (EER). EER is also used to determine θ in [9] which relies on compression models to calculate distance. While both [3] and [9] do not rely on negative data for training, they still need negative training data to determine θ. To avoid using negative data

[1] In many other languages, punctuation provides syntactic information, and hence can be used as proxy for syntactic features. This is not the case for Arabic. Classical Arabic does not use punctuation marks as we know them. Modern Arabic users vary greatly in their use punctuation, which makes punctuation a suitable behavioral feature for AA, but not an intrinsic syntactic property of language use.

at all, [10] and [1] examine methods for dynamically determining the value for θ in the training phase. [10] uses Common N-gram profiles (of token and character n-grams) with a corpus consisting of the English, Spanish, and Greek portions of the PAN-13 [11] competition training corpus. They use the Area under ROC Curve to determine the acceptability threshold for verifying a question document. [1] uses a simpler Gaussian curve in determining θ using a corpus of Classical Arabic philosophy and religion books. Using bag-of-words token frequencies, [1] shows that an AV system for Arabic can outperform the baseline of [9]. Both [10] and [1] achieve accuracy results that exceed their baselines(88.3% for English and 93.6% for Spanish in [10] and 70.97% in [1] for Arabic). However, both of them suffer a limitation in their choice of feature implementation when it comes to Arabic. In the former, character n-gram is suitable for the languages in question for [10], but not for Arabic (c.f. section 3.1). [1] investigates only one feature category (token frequency), leaving out other linguistically significant features.

This paper has two goals. On the computational linguistics front, it evaluates whether a Classical Arabic AV task can be improved using feature categories other than token frequency. On the purely linguistic side, it explores what type of linguistic features are most salient in defining a language user's thumb-print.

3. Corpus

This section describes the content of the training and testing corpora, selection of features and feature categories in 3.1, and the formatting and preprocessing of the corpus (section 3.2). To allow for a reliable baseline, the same AV task and corpus used by [1] are used for this paper. Using the same corpus and AV problem also mirrors a typical AV situation in Digital Humanities. The corpus consists of 19 works attributed to Al-Ghazali (training corpus). They are also used for testing positive results via the leave-one-out method. The corpus also includes 12 documents for testing negative results: nine classical works of authors belonging to the same time period and genre as the training data; one proven falsely attributed to Al Ghazali using non-computational methods [12]; and two modern documents (one fiction and one non-fiction). Table 1 shows the breakdown of the corpus used.

3.1. Feature Categories

One goal of this paper is to evaluate the role of three modules of language in AV: the lexicon, morphology, and syntax. To do so, a number of textual features is extracted from the corpus. Table 2 shows the five feature categories extracted from the corpus:

- Tokens: tokens are defined as individual words in the corpus, separated by space. A token may include procletics and enclitics.
- Stems: a word stem is a word without inflectional morphology (no case endings, subject or object agreement markers, gender, or number agreement morphology).
- Roots: The three letter roots from which a stem is derived.
- Diacritics: each token is vocalized, then consonants and long vowels are stripped away. What remains are characters for short vowels and gemmination. n-grams of diacritic clusters (one cluster per token) are extracted.
- POS: part-of-speech n-grams (noun, verb, etc.) tagged using MADAMIRA tagset [13]

for each of the feature categories, n-grams are constructed (n = 1, ... 4).

The selection of these features is comparable to features used in section 2, given the special characteristics of Arabic orthography and morphology. Feature-based AV tasks ([3] is case in point) often use features that reflect linguistic behavior. For example, prefix and suffix n-grams in English are a reflection of morphological information (derived verbs, nouns, or adjectives). Used correctly, token frequencies and character n-grams can be computationally efficient indicators of lexical choices. Punctuation n-grams and sentence length are indicators of syntactic characteristics.

This connection between textual features and linguistic behavior is often implicit in the literature, but is crucial in dealing with languages morphologically as rich and syntactically as flexible as Arabic. In Arabic, the smallest lexical component of a word is the triliteral root. The morphological component interacts with the lexicon by providing word patterns ("awzaan"). The resulting stems enter syntactic derivations predetermined minimally for part of speech, with properties similar to the interaction between English stems and affixes. Diacritic unigrams capture word patterns for

Table 1. Corpus used

Corpus	Work	Size (1000 tokens)
Al-Gazaly	faḍā'iH al-bāṭiniyya	47
	aṣanāf al Maghrurīn	63.4
	mizān al-'amal	32.7
	al-tibr al-masbūk fī naṣīḥat al-mulūk	31.2
	Bidāyat al-hidāya	14.3
	Tahāfut al-falāsifa	49
	al-Wasīṭ fī al-madhab	400.7
	jawāhir al-Qur'ān	30.3
	iḥyā' 'ulūm al-dīn	831
	al-mustaṣfā min 'ilm al-uṣūl	181.7
	ma'ārij al-Quds fī madārij ma'rifat al-nafs	39.2
	al-Manḳūl min ta'līqāt al-uṣūl	53.4
	miškāt al-anwār	10.3
	miḥak al-naṭr fī al-mantiq	26.5
	mi'yār al-'ilm fī fann al-manṭiq	48.6
	qawā'id al-'aqā'id	18.7
	al-munqidh min al ḍalāl	11
	al-maqṣad al-'asnā fī šarḥ ma'ānī asmā' Allāh al-ḥusnā	34.2
	al-iqtiṣād fī al-i'tiqād	43.4
Others		
Falsely attributed to Ghazali	sirr il'ālimīn	22.3
ḳaṭīb al-bagdādi	šaraf aṣḥāb al-ḥadīth	23.1
	iqtiḍā' al-'ilm wa-al-'amal	13.3
ibn ḥazm al-andalusī	risalat al-radd 'ala al-kindi al-failasūf	10.1
	kitab al-imama wa al-mufadala (al-fiṣal fī al-milal wa-al-ahwā' wa-al-niḥal)	34
Ibn sīnā	al-Qānūn fī al-ṭibb	103.3
	kitāb al-siyāsa	46.1
ibn Rushd	Kitāb Faṣl al-maqāl wa-taqrīr mā bayna al-šarī'a wa-al-ḥikma min al-ittiṣāl	7.3
	Bidāyat al-mugtahid wa-nihāyat al-muqtaṣid	19.8
al-Qurtubi	al-I'lam bima fi din al-nasar ā min al-fasad wa-l-awham	4.8
Modern Texts		
Said Salem	Kaf Maryam (novel)	33.2
Afaf Abdel Moaty	Al mar'a wa al-sulta fi misr	43.4

Table 2. Feature Categories

Feature Category	Description	Example (n= 2)
F1: Token n-grams	Most frequent sequence of consecutive tokens in D.	{أعطيناك مثالا، مثالا في ، في الجملة}
F2: stem n-grams	Most frequent sequence of consecutive stems.	{اعطى مثال، مثال في، في جملة}
F3: root n-grams	Most frequent sequence of consecutive trilateral roots of tokens.	{عطى مثل، مثل في، في جمل}
F4: diacritics n-grams	Most frequent sequence of consecutive diacritics of tokens	{◌ْ◌َ◌ِ◌ِ، ◌َ◌ِ، ◌َ◌ِ، ◌ُ◌ْ◌َ◌ِ}
F5: Part-of-speech tag n-grams	Most frequent sequence of consecutive POS tags	{verb nominal, nominal particle, particle noun}

most of the tokens, but not all (as there are many patterns that contain long vowels), POS unigrams help complement missing information from diacritics. The syntactic component of Arabic can be abstracted through n-grams of the aforementioned features. A diacritic or POS n-gram reflects the grammar of Arabic (e.g. the ban on a sequence of verbs or determiners), but it also reflects choices that the author makes (e.g. multi-word idafas, preference for relative

clauses or nominal sentences). Syntax interacts with the lexicon directly, e.g. in compound nouns and selectional restrictions of verbs and subjects/objects, and with morphology in some fixed styles (e.g. the so-called unreal idafa: indefinite adjective + definite noun sequence). Figure 1 summarizes the relation between linguistic modules and the feature categories selected for this experiment.

3.2. Corpus preprocessing, formatting, and feature extraction

For preprocessing, punctuation marks, kashida and numerals are removed. White spaces are normalized to single spaces. Tokens are defined in this experiment as Arabic Character strings separated by white space. Roots are generated using ISRI [14, 15]. Stems, diacritic, and POS features are extracted using MADAMIRA[13], where the top-ranked analysis of each token is selected, and the rest discarded. Hapax legomena are discarded.

4. Verification Method

The AV task is built into a number of problems. Each problem P consists of a question document D and a set of known documents S. S is the entire body of works of Al-Ghazali. In the evaluation phase, if $D \in S$, then $S = S - D$. For each P, training and testing is conducted for the x% most common n-gram of each of the feature categories outlined above; $x \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30\}$ and $n \in \{1, 2, 3, 4\}$. Training and Testing will follow the algorithm outlined by [1]: use Manhattan Distance as a similarity metric, establish a confidence interval of similarity values in the training phase, and use the lower-bound value of the confidence interval as a similarity threshold θ for accepting test documents.

4.1. Training Procedure

Input to the training procedure is a set of documents containing a string of features extracted from a book known to be written by Al-Ghazali: words, stems, diacritics, POS tags, and roots. N-grams are created using NLTK [14], and hapax legomena are removed. Normalized frequencies of n-grams are calculated using NLTK. Output of the training procedure is a set of similarity values $S = S_1, S_2, S_3, ...S_n$, where $0 < S_n < 1$ represents the similarity of a training document n and the rest of the training corpus.

Calculating Similarity. Similarity is calculated using the Manhattan Distance function between a document X and a corpus of known documents Y:

$$dist(X, Y) = \Sigma_{j=1}^{n} |x_j - y_j| \tag{1}$$

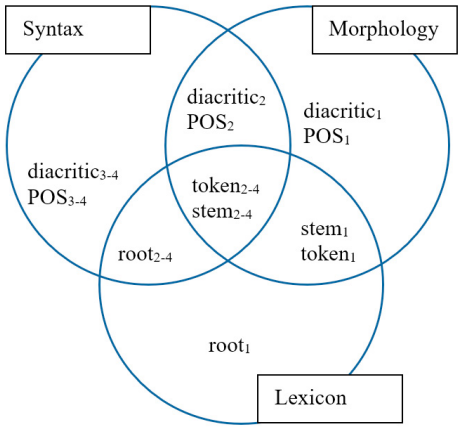

Fig. 1. Feature categories (feature$_{n-gram}$) related to language modules

Table 3. Results

Feature Category Best performing subcategory	**Highest accuracy(%)**
Lexical	
Root unigrams	77.4
Lexical-morphological	
Stem unigrams	77.4
Token unigrams	77.4
Lexical-syntactic	
Root n-grams (n = 2)	83.8
Morphological	
POS unigrams	74.1
Diacretics unigrams	80.6
Morphological-syntactic	
POS bigrams	77.4
Diacretics bigrams	74.1
Syntactic	
POS n-grams (n = 3)	80.6
Diacretics n-grams (n = 4)	70.1
lexical-morphological-syntactic	
Token n-grams (n = 2)	77.4
Stem n-grams (n = 2)	**87.1**
Baseline	**7.97**

Where x_j and y_j are the normalized frequencies of feature j n-gram. Distance is then converted to a similarity score:

$$Sim(X, Y) = \frac{1}{1 + dist(X, Y)} \tag{2}$$

Estimating similarity threshold θ. Having calculated the similarity value S between each document in the training set and the rest of the training set documents, The similarity threshold θ is defined as the lower bound of a confidence interval of the training set similarity values at $p < 0.005$.

Testing. For each question document feature n-gram subset, similarity to the training corpus is calculated as shown above. The document is judged fake if its similarity value is lower than θ.

Evaluation Baselines. To evaluate the results returned by the experiment, I use as a baseline the accuracies reported by [1], which reports 70.97% accuracy of using the same algorithm with the most common 3-9% tokens.

4.2. Experiment and Results

This experiment has two goals. The first is to identify what feature (sub)category performs best in Classical Arabic AV. The second is to identify whether features pertaining to a certain linguistic module or ensemble of modules are particularly relevant in distinguishing an author's style. To achieve this, the proposed verification method is applied to each of the documents in the corpus. To evaluate true positives, evaluation is implemented using the leave-one-out method. The experiment is conducted for each document $D_{f,n,p}$, where f is Feature Category, n is n-gram (1 - 4), and p is the percentage of most common instances of a given feature category to be compared. Accuracy of the outcome is calculated as the number of correctly classified documents divided by the total number documents (31).

Table 3 shows the resulting accuracies across the best performing parameters for each feature category. As table 3 shows, the classifier is most accurate when considering similarity at the most common stem bigrams, with accuracy of 87.1%. Appendix A shows the full range of accuracies returned by the experiment.

5. Evaluation and Discussion

Like the research in [1] and [10], this experiment shows that an author profiling method based solely on positive evidence can indeed yield highly accurate AV results. The best performing subcategories outperformed the baseline of [1] in all feature categories, including token unigrams. Higher accuracy than [1] in the token unigram category is unexpected; the current experiment uses the same corpus and the same algorithm, except for removing hapax legmena. The results also show that performance is consistent across most common x% features (1 – 30 %).

The results of this experiment could not be compared to the accuracy of the other single-class classifier discussed in section 2 ([10]) as it calculates accuracy differently. It only reports accuracy calculated as the harmonic mean of precision and recall, while in accuracy in our case is defined in terms of precision only. This is because the experimental design of [10] is different from this experiment. While [10] allows an 'I do not know' answer, our experiment always yields a response. Hence, recall must be taken into account for [10], but not in our case.

Comparing how feature categories fare against each other, it can be seen that features involving syntactic decisions (n-grams) are more powerful than purely lexical or morphological features in distinguishing an author. Best synergies involve syntax and the lexicon (stem bigrams at 87.1% followed by root bigrams at 83.8%). Morphological features perform poorly, whether alone such as in POS and diacretic bigrams, or even in consort with syntactic information such as diacretic n-grams, which are among the lowest accuracy outcomes. Indeed, there is no difference in accuracy between using roots, stems, or tokens, which means that morphological information involved in creating stems from roots, and in creating full-fledged words does not distinguish the style of individual authors.

Although this system is built on top of another ML tool (MADAMIRA), poor perormance of morphological information cannot be due to MADAMIRA extracting diacritics or POS information less efficiently than stemming, rather than to intrinsic properties of language use. This is an unlikely because MADAMIRA has much higher accuracy in POS tagging than diacritization (95.9% and 86.3%, respectively [13]. If the quality of the preprocessing negatively affected feature performance, it would be in the opposite direction (POS performing better than diacritics, especially in unigrams).

Relating this experiment to the bigger picture in AV cross-linguistically, the results can explain why using the same features in an AV task results in lower accuracy in Arabic, but not extremely lower than other languages. The best performing feature subset in this experiment is stem bigrams. In Arabic, this means that a most successful system should ignore inflectional morphology (mostly agreement markers). like many Roman-character languages, this involves removing suffixes and prefixes. Unlike these languages, Arabic also involves infixation as well, in irregular (broken) plural nouns, and middle long vowels in verbs (e.g. the so-called Hollow verbs).

6. Conclusion and Future Work

This paper shows that it is possible to achieve high accuracy Classical Arabic AV using only positive evidence, a simple distance-based learner, and some preprocessing. Using Manhattan Distance, stem n-grams, and an acceptability threshold calculated from the training set, the single-class classifier achieves an accuracy of 87.1%, outperforming a baseline of best known classifier of 71.9%. This paper confirms that using no negative training data can render better AV performance. It additionally shows that careful selection of linguistic features has a significant impact on classifier performance. Stem bigrams are the best performing features under that algorithm, which suggests that syntactic characteristics of an author are a strong predictor, followed by lexical information. Morphological choices of language users seem to matter least.

Future research should further explore some of the experimental design choices in this paper. Distance vectors in this paper are based on Manhattan Distance, a popular yet rather old measure. Sample size is another area of further research. Many of the texts in the corpus used are very large books, which reflects a real-life situation in Classical Arabic Digital Humanities. On the other hand, using such large samples comes with disadvantages. The results indicate that increasing the number of items involved in calculating distance (optimally stem bigrams) does not improve accuracy - basically that anything more than a 1% portion of the question document is wasted computation. Furthermore, using such large dataset of 'known documents' could cast doubt on the validity of the simple Gaussian algorithm used to calculate acceptance threshold θ. Future research should investigate the minimum usable corpus size that delivers comparable accuracy, and expand to include Modern literary and non-literary Arabic texts. Predictions

made in this paper on the value of syntactic and lexical, compared to morphological, features, should be examined cross-linguistically.

Appendix A. Results

Results of the experiment for most common 1-30% feature n-grams. Highest score for each feature is in bold.

Feature	n-gram	Accuracy	Feature	n-gram	Accuracy
Token	1	**77.4%**	POS	1	74.2%
	2	**77.4%**		2	77.4%
	3	58.1%		3	**80.6%**
	4	54.8%		4	77.4%
Stem	1	74.2%	Diacritics	1	**80.6%**
	2	**87.1%**		2	74.2%
	3	61.3%		3	67.7%
	4	58.1%		4	71.0%
Root	1	77.4%			
	2	**83.9%**			
	3	61.3%			
	4	58.1%			

References

[1] Ahmed H. Dynamic Similarity Threshold in Authorship Verification: Evidence from Classical Arabic.*Procedia Computer Science* 2007;117:145 – 152.

[2] Burrows J. " "Delta": a measure of stylistic difference and a guide to likely authorship." *Literary and Linguistic Computing*2002;17(3);267–287.

[3] Halvani O, Winter C, and Pflug A. "Authorship verification for different languages, genres and topics." *Digital Investigation* 2016;16; S33–S43.

[4] Ouamour S, Sayoud H . "Authorship attribution of short historical Arabic texts based on lexical features." In *Proceedings - 2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2013* ; 144–147.

[5] Howedi F, and Mohd, M. "Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data." *Computer Engineering and Intelligent Systems* 2014; 5(4); 48–56.

[6] Altheneyan A S, Menai M E B. "Naïve Bayes classifiers for authorship attribution of Arabic texts." *Journal of King Saud University - Computer and Information Sciences* 2014; 26(4); 473–484.

[7] Shaker K . *Investigating features and techniques for Arabic authorship attribution.* Heriot-Watt University; 2012.

[8] García-Barrero D, Feria M, Turell M T . "Using function words and punctuation marks in Arabic forensic authorship attribution." In R. Sousa-Silva, R. Faria, N. Gavaldà, & B. Maia (Eds.), Proceedings of the 3rd European Conference of the International Association of Forensic Linguists ; 2013; (pp. 42–56). Porto, Portugal: Faculdade de Letras da Universidade do Porto.

[9] Halvani O, Winter C, Graner L ."Authorship verification based on compression-models". *arXiv preprint*; 2017; 1706.00516.

[10] Jankowska J, Milios E, Keselj V . "Author Verification Using Common N-Gram Profiles of Text Documents". *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*; 2014; 387 –397.

[11] Rangel F, Rosso P, Koppel M, Stamatatos E, Inches G . "Overview of the author profiling task at PAN 2013". In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* 2013; 352-365. CELCT.

[12] Watt W M . "The Authenticity of the Works Attributed to al-Ghazālī." Journal of the Royal Asiatic Society of Great Britain and Ireland1952; 2(1); 2445.

[13] Pasha A, Al-Badrashiny M, Diab M T, El Kholy A, Eskander R, Habash N, Pooleery M, Rambow O, Roth R . "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic". In *LREC* 2014; 14;1094-1101.

[14] Bird S, Klein E, Loper E. "Natural language processing with Python: analyzing text with the natural language toolkit." O'Reilly Media, Inc.; 2009.

[15] Taghva K, Elkhoury R, Coombs J. "Arabic stemming without a root dictionary." In *International Conference on Information Technology: Coding and Computing, 2005. ITCC 2005*. (Vol. 1; 152-157). IEEE.