Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Author gender identification from Arabic text

Kholoud Alsmearat[a], Mahmoud Al-Ayyoub[a,*], Riyad Al-Shalabi[b], Ghassan Kanaan[b]

[a] *Jordan University of Science and Technology, Irbid, Jordan*
[b] *Amman Arab University, Amman, Jordan*

ARTICLE INFO

Article history:

Keywords:
Arabic text processing
Gender identification
Stylometric features
Bag-Of-Words

ABSTRACT

The Gender Identification (GI) problem is concerned with determining the gender of a given text's author. It has a wide range of academic/commercial applications in various fields including literature, security, forensics, electronic markets and trading, etc. To address this problem, researchers have proposed that the writing styles of authors of the same gender share certain aspects, which can be captured by certain stylometric features (SF). Another approach to address this problem focuses mainly on keywords occurrences in each document. This is known as the Bag-Of-Words (BOW) approach. In this work, we study and compare both approaches and focus on the Arabic language for which this problem is still largely understudied despite its importance. To the best of our knowledge, no previous work has considered these approaches for the GI problem of Arabic text. The comparison is carried out under different settings and the results show that the SF approach, which is much cheaper to train, can generate more accurate results under most settings. In fact, the best accuracy levels obtained by the SF and BOW approaches on our in-house dataset are 80.4% and 73.9%, respectively.

1. Introduction

Textual data posted by Internet users (social media content, articles, emails, blogs, literature, web pages, etc.) form a large percentage of the world's data. Mining such data gave rise to many interesting tasks in natural language processing (NLP) such as text categorization (TC), text clustering, concept/entity extraction, plagiarism detection, authorship analysis, production of granular taxonomies and ontologies, extracting useful knowledge/trends, sentiment analysis, document summarization, entity relation modeling, etc. [22,38,59].

One of the challenges related to online textual data is identifying its source by relying only on the data itself. Linguistically, this is a difficult problem. However, the Internet's anonymity and loose accountability add to the complexity of the problem as they make it easy for anyone to post anything online and make any false claims about its authorship [33]. The wide-spread of the social networks and chat services further amplifies this problem since such platforms gave perpetrators new chances and different motivations to spread their false claims. Thus, the need for an automatic system to effectively and efficiently detect such false claims is eminent.

The previous paragraph motivates what is generally known as the authorship analysis (AA) problem. Given, a certain piece of text, the AA problem is concerned with determining the identity of its author (in what is known as the *authorship authentication* or *attribution* problem) or some of his/her characteristics such as gender, age, background, native language, etc. (in what is known as the *authorship characterization* or *profiling* problem). This is achieved solely based on the content of the text [2,10,54,71].

Numerous applications exist for the AA problem with its different variations (attribution and profiling) [16,25,69]. One example is related to security applications where deception and fraud detection are vital for some businesses to survive. Another example is related to targeted advertisement which is one of the main revenue generator for many Internet based companies. The last example is literature. There are many classical texts with doubtful attribution to well-known authors such as Shakespeare. The AA problem is so important that the prestigious Conference and Labs of the Evaluation Forum (formerly known as the Cross-Language Evaluation Forum, or CLEF) has had author profiling tasks since 2013 [53–56].[1]

* Corresponding author.
E-mail addresses: khlood1.smearat@live.com (K. Alsmearat), maalshbool@just.edu.jo, malayyoub@gmail.com (M. Al-Ayyoub), shalabi@aau.edu.jo (R. Al-Shalabi), gkanaan@aau.edu.jo (G. Kanaan).

[1] http://www.uni-weimar.de/medien/webis/events/pan-13/pan13-web/author-profiling.html

http://dx.doi.org/10.1016/j.jisa.2017.06.003
2214-2126/© 2017 Elsevier Ltd. All rights reserved.

Due to its historical roots, the AA problem (or at least some variations of it) has been investigated thoroughly by linguists and literary scholars. The classical approach is to focus on studying the style of each author (or author group) in the given text. The characteristics of the style are captured through stylometric features (SF). The fundamental assumption here is that each author (or group of authors with common characteristics such as gender or age group) has unique writing styles that is relatively fixed and barely changes with time, which means that it can be used to uniquely identify the author (or his/her characteristics) [32]. Another approach to the AA problem uses keywords occurrences/frequencies in each document to capture the patterns used by each author (or group of authors). This approach is known as the Bag-Of-Words (BOW) approach. Unlike the first one, this approach is more language-independent.

In this work, we are interested in one of the authorship profiling problems where the author characteristic of interest is the gender. This is known as the Gender Identification (GI) problem. Similar to other AA problems, the GI problem has its obvious applications in various fields from marketing to security [14,15]. We are interested in addressing the GI problem for Arabic articles using the two approaches mentioned in the previous paragraph. To the best of our knowledge, none of the prior works [7,8,20] have accomplished this.

The rest of this paper is organized as follows. The following section gives a general overview of the current literature on authorship analysis with a focus on the Arabic language. Our work is discussed in Section 3 and the experimental results obtained are discussed in Section 4. Finally, concluding remarks along with a discussion of the future work are presented in Section 5.

2. Background and related works

The problem at hand (the GI problem) is basically a binary text categorization (TC) problem where the classes are simply male/female. Thus, we start our discussion of the related works by giving a brief overview of some recent works on Arabic TC. Unfortunately, to the best of our knowledge, the field of AA and the use of stylometric features are largely understudied topics in the Arabic NLP community.

2.1. Arabic text categorization (TC)

The Arabic language is among the most widely used languages in the world with hundreds of millions of native speakers. Moreover, it is an integral part of the lives of more than 1.5 billion Muslims all over the world who use it for their religious practices and rituals. Another reason that makes Arabic so appealing to study is related to the special and unique challenges associated with the automated handling and understanding of the Arabic language and the relatively young and under-developed field of Arabic NLP. To get an overview of the Arabic language and a comprehensive coverage of the general field of Arabic NLP, the interested reader is referred to [21,26,59].

Among the many characteristics of the Arabic language, there are some characteristics that have significant effect on the general approaches to handle the AA problem. Examples include the language's complex nature (derivational, inflectional, etc.). Another example is the existence of many variations of the language such as the Classical Arabic (CA) (which is mainly used in ancient and theological texts but is still understood due to its use in the Holy Quran), the Modern Standard Arabic (MSA) (which is a modernized and simplified version of CA) and Dialectal/colloquial Arabic (DA) (where each region has its own dialect). The most widely-understood variation among Arabic speakers is MSA due to its wide adoption in education, media, and formal communication across the different Arabic speaking countries [10].

Over the past two decades, a good number of papers have been published addressing the Arabic TC problem in various settings. For more recent and more comprehensive comparative studies on Arabic TC, the interested reader is referred to the works of Said et al. [60], Saad [59] and Khorsheed and Al-Thubaity [34].

2.2. Authorship analysis (AA)

Compared to TC, AA is largely understudied in the Arabic language [9,10]. Below, we discuss some of the works in this field starting with authorship authentication (attribution) before going into the more relevant works on authorship characterization (profiling).

2.2.1. Authorship authentication (attribution)

The authorship authentication problem is a classification problem in which the goal is to determine the author of a certain text given a set of texts written by various authors. Intuitively, the writing style is the most intuitive aspect on which to focus. This is manifested in the computation of SF. On the other hand, the BOW approach has been shown to be very effective [10]. There have been many papers on AA of English text. The interested reader is referred to [58] for a recent comprehensive survey of such papers. Another interesting work worth mentioning is the work of Potthast et al. [51], in which the authors tested the reproducibility of existing AA papers by re-implementing them from scratch.

To the best of our knowledge the AA problem for the Arabic language has not been studied well as the number of published works on AAA is very small. One of the most notable works is that of Abbasi and Chen [1,2], in which the authors used different sets of features including lexical, syntactic, structural and content-specific features for the authorship identification problem. The classification techniques used in their study were Support Vector Machine (SVM) and Decision Tree (DT). In [1], they applied their technique to web forum messages, whereas, in [2], they collected and analyzed messages posted on extremist groups' web forums both in Arabic and English.

Following the same approach of Abbasi and Chen [1,2], Otoom et al. [46] extracted 27 stylometric features that were further reduced to the most discriminating 12 features. They applied two classifiers: Functional Trees (FT) and SVM on a rather small dataset of 456 articles written by seven authors.

In [68], the authors focused on the problem of small and imbalanced datasets, which is a common problem with authorship identification datasets. They represented each document using the bag-of-characters n-gram approach which they claim is better than the BOW approach as it can capture stylistic as well as thematic information more accurately.

Shaker and Corne [67] relied on the usage pattern of function words to identify the author. They exploited a set of 104 function words reflecting the semantics of the English function words of Mosteller and Wallace [44]. As for the classification approach, they used a hybrid of evolutionary algorithms (EA) and linear-discriminant analysis (LDA), which is known to have excellent performance for authorship authentication of English text [66]. They tested their approach on a dataset of 14 Arabic novels by six different writers. Another work relying on function words in addition to punctuation marks is [24].

Ouamour and Sayoud [48,49] considered a dataset of 30 historic texts written by ten different authors. What is special about this dataset is that the authors are all famous well-educated Arab explorers describing their travels and expeditions to different regions of the world. The consistency in the topic and the educational background of the authors adds to the difficulty of distinguishing