Digital Investigation xxx (2018) 1–18

Contents lists available at ScienceDirect

Digital Investigation

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

Automatic categorization of Arabic articles based on their political orientation

Raddad Abooraig [a], Shadi Al-Zu'bi [b, *], Tarek Kanan [b], Bilal Hawashin [c], Mahmoud Al Ayoub [a], Ismail Hmeidi [d]

[a] Computer Science Department, School of Information Technology, Jordan University of Science and Technology Irbid, Jordan
[b] Computer Science Department, School of Science and Information Technology, Zaytoonah University of Jordan Amman, Jordan
[c] Computer Information Systems, School of Science and Information Technology, Zaytoonah University of Jordan Amman, Jordan
[d] Computer Information Systems, School of Information Technology, Jordan University of Science and Technology Irbid, Jordan

ARTICLE INFO

Article history:
Received 15 January 2018
Received in revised form
11 April 2018
Accepted 11 April 2018
Available online xxx

Keywords:
Text mining
Supervised classification
Arabic text
Bag-of-words
N-gram
Authorship analysis
Stylometric features
Social networks
Political orientation
Machine learning

ABSTRACT

The ability to automatically determine the political orientation of an article can be of great benefit in many areas from academia to security. However, this problem has been largely understudied for Arabic texts in the literature. The contribution of this work lies in two aspects. First, collecting and manually labeling a corpus of articles and comments from different political orientations in the Arab world and making different versions of it. Second, studying the performance of various feature reduction methods and various classifiers on these synthesized datasets. The two most popular feature extraction approaches for such a problem were compared, namely the Traditional Text Categorization (TC) approach and the Stylometric Features approach (SF). Although the experimental results show the superiority of the TC approach over the SF approach, the results also indicate that the latter approach can be significantly improved by adding new and more discriminating features. The experimental results also show that the feature selection techniques reduce the accuracies of the considered classifiers under the TC and SF approaches in general. The only exception is the Partition Membership (PM) technique which has an opposite effect. The highest accuracies are obtained when PM feature selection method is used with the Support Vector Machine (SVM) classifier.

Introduction

With the emergence of web 2.0, (such as the social networks, forums, personal blogs, etc.), Internet surfers are more aggressively contributing to the collective contents of the Internet by posting articles or comments voicing their opinions and experiences. Web 2.0 is covering many fields such as social events, economic events, political events, etc. Regarding politics and political events, the Internet surfers post comments and articles based on their beliefs and ideologies. The Internet surfers are especially encouraged by the anonymity inherent the Internet. Consequently, these web pages receive millions of comments and articles daily and the process of analyzing them to extract useful information is a very expensive task in terms of both time and effort.

Political articles (especially in the Arab world) are different from other articles due to their subjectivity. A political article might be heavily influenced by the author's convictions and political affiliation. The ability to automatically determine the political orientation of an article can be of great benefit in many areas, from academia to security (Abbasi and Chen, 2005a; Koppel et al., 2009).

Moreover, this is an example of author profiling problems, which are useful for optimizing search engines, sentiment analysis and marketing intelligence (Estival et al., 2007). This problem can be viewed as a special case of the text categorization (classification) problem with the categories being the major political ideologies in the Arab world such as: Liberal, Islamic Sunni (Brotherhood, Salafi, etc.), Islamic Shia (Hezbollah, Ansarollah, Jeaish Almahdi, etc.), Arab Nationalists (Baathi, Nasri, etc.), Communist, Socialist, etc.

There have been several works on Traditional Arabic Text Categorization (TC) and authorship analysis that is related to our work. Although the typical TC usually focuses on identifying a text's domain (Sports, Politics, Economy, etc.) based on its contents

* Corresponding author. Tel: +962 799100034
E-mail address: smalzubi@zuj.edu.jo (S. Al-Zu'bi).

https://doi.org/10.1016/j.diin.2018.04.003
1742-2876/

(topic-based); the authorship analysis focuses on authorship authentication and authorship characterization. Authentication (attribution) deals with verifying whether a text was written by a certain author or not based on stylometric and statistical similarities with other texts written by the same author. On the other hand, characterization (profiling) tries to detect the characteristics of the authors' group such as gender, age group, level of education, social class, etc (Abbasi and Chen, 2005a). While authentication is not directly related to our study, profiling is relevant. So, we discuss them both below.

In Arabic TC, most of previous studies deal with the categorization of the text document in the domain it belongs to, as example, political articles, categorize to the political class, sport articles, categorize to the sport class and so on (Alsaleem, 2011); the only difference was in the domain. Beside the categorization, most of studies filter and preprocess the text documents by removing stop words, removing punctuations, and normalizing (Alsaleem, 2011). Other studies investigate the effects of the Arabic stemmer (Light, Khoja, etc.) and compare the results with stemming and without it (Harrag et al., 2011; Wahbeh et al., 2011). Features reduction (Chi square, information gain, and others) is used in other studies (Thabtah et al., 2009; Mesleh, 2007).

Arabic is a Semitic language written from right to left. It is used by 5% of the world's population. There are 22 countries with more than 400 million who speak Arabic as their first language. Also, the Arabic language is the fastest-growing language on the Web; researchers revealed that the annual growth rate for Arabic-speaking users on the Web is 2501.2% in 2010 (Korayem et al., 2012). Working on Arabic text gives rise to many unique challenges with respect to the language structure and stylistic features, such as inflection, diacritics, word length and elongation, and tri-glossic nature. These challenges make authorship analysis for Arabic more difficult than for other languages (Harrag et al., 2011; Wahbeh et al., 2011; Kanan and Fox, 2016).

The aim of this study is to analyze articles/comments written in Modern Standard Arabic (MSA) to determine their political orientation. The decision will rely on statistical as well as semantic features of the studied texts. The specific political orientations (ideologies) we consider in this work are Liberal, Communist or Socialist, Arab Nationalist (Baathi, Nasri, etc.), Islamic-Sunni (Brotherhood, Salafi, etc.), and Islamic Shia (Hezbollah, Ansarollah, Jeaish Almahdi, etc.). We treat this as a supervised learning problem, in which a manually annotated corpus is collected and used to train and test a classifier.

This study is divided into three main parts. The first part is building a manually annotated corpus of political articles/comments written in Arabic. As customary in the literature, it is beneficial to present and discuss some statistical features of the corpus such as the total numbers of articles, characters/article, characters/sentence, characters/word, punctuation, words/article, unique words/article, words/sentence, and sentences/article. The second part is creating different versions of the corpus for both feature extraction approaches, the TC and SF approaches. The third part is dedicated to doing experimentations to find the best combination of feature extraction, feature selection and classification techniques for the problem at hand.

Related literature

The closest works to the problem at hand are those on TC and authorship analysis. We briefly discuss some of recent works on general Arabic TC. Unfortunately, to the best of our knowledge, the field of authorship analysis is still largely understudied for the Arabic language. Before discussing the TC and authorship analysis, we discuss briefly different classifiers and different types of preprocessing mechanisms (stemming, n-gram, and feature selection) that are used in our study.

Classification

Machine Learning (ML) is the study of computer algorithms that learn automatically through experience (Russell and Norvig, 2010). ML can be usually classified into supervised and unsupervised learning. The supervised method depends on learning from a categorized data set to create a model for later prediction or classification (Russell and Norvig, 2010). The outcome function can be classification (discrete) or regression (continuous). The supervised learning can be classification or regression. The classification predicts the class labels and classifies data based on the training set that learns the classification function. In this review, we will concentrate only on the classification function. We start by discussing some of the most common classifiers.

Naive Bayes (NB) classifier

NB is a classification algorithm that relies on conditional probability. It utilizes Bayes' theorem and assumes a strong independence between features. It computes the occurrences of the features and the relation between them in the corpus considered. NB is good in practice even when the dimensionality of the input is high (Han et al., 2012a). NB is a good classifier for TC in general (Jurafsky and Martin, 2009).

Discriminative Multinomial Naive Bayes (DMNB) classifier

Multinomial Naive Bayes (MNB) is widely used in TC because the computation is very simple, but it is not efficient like other generative classifiers. It maximizes probability rather than conditional probability. DMNB appears as an extension of MNB that considers together the probability and the categorization points through the frequency calculation. DMNB is shown in many cases to give better results than NB and SVM with TC. (Su et al., 2008).

Sparse generative model (SGM) classifier

From its name, SGM is a generative classifier that is based on the idea of using a sparse computing method to reduce its complexity and make it more scalable. The sparse computing method represents documents word counts in two vectors of indexes. SGM was developed on MNB. It proposed a sparse time complexity algorithm for MNB classification (Puurula, 2012).

Support Vector Machine (SVM) classifier

SVM is a classifier that can be utilized for categorizing linear and nonlinear data. SVM constructs an N-dimensional hyper-plane that completely divides the data set into two classes (Han et al., 2012b). It is a very excellent classifier for TC (Puurula, 2012).

The SVM prediction result is usually high, strong (good theoretic assurances about over-fitting), runs even if the training set includes errors, and with an appropriate kernel, it can function well regardless of whether the data is linearly separable or not (Han et al., 2012b).

Random forest (RF) classifier

RF is a grouping of tree predictors, where every predictor of tree relies on the values of a random vector. These values are independently sampled. Every random vector gets equal distribution for other tree predictors in the forests. In classification, the forest makes voting between trees classification and choose the classification with the highest votes overall tree in the forest (Breiman, 2001).

Please cite this article in press as: Abooraig, R., et al., Automatic categorization of Arabic articles based on their political orientation, Digital Investigation (2018), https://doi.org/10.1016/j.diin.2018.04.003