Chapter 37
Gender Inference for Arabic Language in Social Media

Abdul Rahman I. Al-Ghadir
King Saud University, Saudi Arabia

Abdullatif Alabdullatif
King Saud University, Saudi Arabia

Aqil M. Azmi
King Saud University, Saudi Arabia

ABSTRACT

The widespread usage of social media has attracted a new group of researchers seeking information on who, what and, where the users are. Some of the information retrieval researchers are interested in identifying the gender, age group, and the educational level of the users. The objective of this work is to identify the gender in the Arabic posts in the social media. Most of the works related to gender classification has been for English based content in the social media. Work for other languages, such as Arabic, is almost next to none. Typically people express themselves in the social media using colloquial, so this study is geared towards the identification of genders using the Saudi dialect of the Arabic language. To solve the gender identification problem the authors, a novel method called k-Top Vector (k-TV), which is based on the k-top words based on the words occurrences and the frequency of the stems, was introduced. Part of this work required compiling a dataset of Saudi dialect words. For this, a well-known widely used social site was relied on. To test the system, we compiled 1200 samples equally split between both genders. The authors trained Support Vector Machine (SVM) and k-NN classifiers using different number of samples for training and testing. SVM did a better job and achieved an accuracy of 95% for gender classification.

DOI: 10.4018/978-1-5225-1933-1.ch037

1. INTRODUCTION

Today, we live today in the "information society," in which individuals, governments and organizations alike generate and accumulate enormous amounts of information that can no longer be analyzed manually. Data mining, with its various algorithms and applications, is therefore a natural response to this information barrage. According to the McKinsey Global Institute, the amount of generated data will see the annual increase of 40% worldwide, and already today companies with over 1000 employees have stored at least 200 terabytes of data (Larose 2014). Social media plays a significant role in the daily life of many people and organizations. The growth in the number of users has led to a massive increase in the size of social data. In turn this has led to increase the interest in researching this field. There exists a good deal of researches covering different aspects of social media. This includes, age group estimation, demographic interests, gender, etc. Twitter is considered well-known in Saudi Arabia community where Saudi Arabia ranks second in world's fastest growing countries on Twitter with 90% of the tweets being in Arabic (Jiffry, 2013). Most of the researches in the field of social media have focused on Twitter (Alwagait and Shahzad, 2014a, 2014b). Some of the works that focused on gender classification for English language are (Pennacchiotti and Popescu, 2011; Marquardt et al., 2014; Liu and Ruths, 2013). Few works have addressed Arabic language in the social media (Abdul-Mageed, Diab and Kubler, 2013).

In this paper we focus on gender classification for Arabic social media using the Saudi dialect. Our work differs from other existing methods because it introduces a system to classify gender using a new feature representation of text called k-Top Vector (k-TV). We generated new dataset that covers the social contents for the Saudi dialect. We found the approach in (Liu and Ruths, 2013) to be quite relevant when doing the gender classification for the Arabic language.

The obtained results in this work are comparable to the state-of-art. The rest of this paper is divided as follows. In Section 2, it examined some related works focusing on feature extraction and classification methods. In Section 3, it introduces the feature vector representation (k-TV) for the gender classification for the Saudi dialect. In Section 4, it covers the proposed solution. Results and discussion presented in Sections 5. Finally, conclusion of paper work with future directions in Section 6.

2. RELATED WORK

A literature review is carried out to analyze different approaches for feature extraction for gender classification in social media sites (especially for Twitter). According to the state-of-the-art there are two categorizes for the features: behavior related and statistical related features. In the behavior related features, it focuses on the lexical and behavior of the social users based on the text, images, metadata … etc. In the statistical related features, it focus on some statistical parameters for the words and characters in the text.

2.1. Behavior Related Features

Pennacchiotti and Popescu (2011) introduced a system to infer user attributes such as political orientation or ethnicity by user behavior information network structure and the linguistic content of the Twitter data. They relied on four general feature classes: user profile, user tweeting behavior, linguistic content of user messages and user social network features. They introduced profile-based features (PROF) to identify