International Journal of Information Retrieval Research
Volume 10 • Issue 1 • January-March 2020

Twitter Users' Classification Based on Interest:
Case Study on Arabic Tweets

Noura A. AlSomaikhi, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia
Zakarya A. Alzamil, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

Microblogging platforms, such as Twitter, have become a popular interaction media that are used widely for different daily purposes, such as communication and knowledge sharing. Understanding the behaviors and interests of these platforms' users become a challenge that can help in different areas such as recommendation and filtering. In this article, an approach is proposed for classifying Twitter users with respect to their interests based on their Arabic tweets. A Multinomial Naïve Bayes machine learning algorithm is used for such classification. The proposed approach has been developed as a web-based software system that is integrated with Twitter using Twitter API. An experimental study on Arabic tweets has been investigated on the proposed system as a case study.

KEYWORDS

Arabic Tweets, Big Data Analytics, Big Data Classification, Interest-Based Classification, Text Classification, Tweet Classification, Twitter User Classification

INTRODUCTION

The social media applications have become an important part of the daily life of millions of users. Twitter is an example of these famous social media platforms that is used by millions of users for different purposes. For instance, users in Twitter can share their interests, opinions and knowledge, as well as searching for latest news and reviews. This form of multi-usage by millions of users leads to generate massive amount of data in different types and forms. This makes the social media as a form of big data that is difficult to manage and utilize.

Big data concept has been introduced recently, and has been defined as the dataset that could not be handled by traditional software/hardware tools to process and manage within an acceptable time (Chen et al., 2014). Big data refers to the large complicated dataset that is difficult to process with regular data processing systems (Samuel et al., 2015). In addition, big data has been characterized by three properties, volume, velocity and variety (3V model). The volume is the size of the dataset that should be increasingly big; velocity is the speed of data generation, analysis and delivery that must be rapidly and timely conducted; and variety indicates the various types of data from different sources that include unstructured, semi-structured as well as structured data type (Chen et al., 2014;

DOI: 10.4018/IJIRR.2020010101

Tsai et al., 2015). Additional property of big data has been added to extend 3V model, in which 4V model has been introduced to include value property that indicates discovering values, e.g., meaningful information, from the dataset (Chen et al., 2014; Tsai et al., 2015).

Understanding the big data within certain context is a challenge in the social media. As a result, analysis process on this massive amount of big data is needed to better understand and utilize big data for specific purpose such as text classification. Text classification aims to assign pre-defined classes to text documents, such as labeling each news story with a topic like health, economy or sport (Hotho et al., 2005). There have been several different analysis techniques used for the purpose of text classification such as sentiment analysis (Waykar et al., 2016; Cai et al., 2010) which aims to analyze text to extract and classify user opinion either as positive or negative. In addition, classifying users' opinions and interests is very important to understand the users' concerns to provide them with better utilities and recommendations.

Many users in Twitter participate in writing tweets mainly for networking with others, and may not explicitly indicate their interests. Although some information such as name, age, location and short summary of interests may be available in the user's profile; it can be incomplete, users may prefer not to share them, or deceptive users may choose to write fake information. Knowing users' interests in social media is useful for different purposes, such as recommendation systems and marketing systems. Recommendation system may use the users' interests to recommend friends for users that share the same interests, and marketing systems may use them for marketing purposes. In addition, it may be used for detecting abusive accounts.

There are several research studies such as Mangal et al. (2016), Michelson and Macskassy (2010), Lim and Datta (2013), Lee et al. (2011), and Magdy et al. (2015) that focused on Twitter classification for different purposes. Although these research studies have investigated the Twitter users' opinions by understanding their tweets to classify users based on their interests such as Mangal et al. (2016), and Michelson and Macskassy (2010), most of these studies have been applied to English tweets. In spite of the fact that, millions of active Twitter users are Arabic speaking native, there is a lack of research that is conducted on Arabic language in comparison to English language due to the Arabic language's morphological complexity and limited availability of software that is compatible with the Arabic language (Refaee and Rieser, 2014).

This research aims at developing a big data analysis system that classifies Twitter users based on their interests, which is of the interest of many people and associated with their daily life. Additionally, the focus is on Arabic tweets generated by Twitter, which is currently considered as an important source of information and communication for many Arab people in a way that may help in improving several services for their societies and countries. Therefore, in this paper we propose an approach for classifying Twitter users with respect to their interests based on their Arabic tweets using Multinomial Naïve Bayes (Aggarwal and Zhai, 2012). The proposed approach has been developed as a web-based software system that is integrated with Twitter using Twitter API to collect Arabic tweets, and then analyze them to identify users' interests based on their tweets. Such software system enables users to identify certain Twitter user's interest by collecting user's tweets and classifying them based on a predefined interest categories/classes i.e., sport, religion, technology, health, economy and literature.

The rest of the paper is organized as follows; next section presents the related work. After that, the proposed classification approach is introduced, in which the classifier and its major components are described. Then the experimental study is illustrated, and the conclusion is presented in the last section.

RELATED WORK

There have been many research studies which have investigated the data classification for different purposes, such as classification of Arabic text documents by Wahbeh et al. (2011), road accidents data classification by Kumar et al. (2018), textual plagiarism detection by Bouarara