Technological Forecasting & Social Change 99 (2015) 92–105

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

#iamhappybecause: Gross National Happiness through Twitter analysis and big data

Ahmet Onur Durahim [1], Mustafa Coşkun *

Management Information Systems Department, Bogazici University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 4 September 2014
Received in revised form 24 June 2015
Accepted 28 June 2015
Available online 9 July 2015

Keywords:
Twitter
Big data
Sentiment analysis
Gross National Happiness
Emotional well being

ABSTRACT

The prominence of social media has contributed to open information access for the researchers. With voluntary information sharing structure of Twitter, user disposition and sentiment analyses can be performed for determining the emotional well-being of the citizens. In this respect, we adopted a sentiment analysis model to calculate the Gross National Happiness (GNH) of a Middle East country, Turkey. For this purpose, over 35 million tweets, published in 2013 and in the first quarter of 2014, of over 20 thousand users were collected and analyzed. In the proposed model, prior to calculating the GNH by considering the polarities of tweets, first convergent and face validities of sentiment analysis and reliability of dataset were tested. After obtaining satisfactory results, the GNH by province survey results of Turkish Statistical Institute was compared to results of sentiment analysis for 2013 in order to state the difference between the surveying method and the proposed social media analysis method. Also, GNH by province in the first quarter of 2014 was analyzed. Additionally, relationships between users' account properties and happiness levels were investigated. Results showed that two GNH calculation approaches give similar results for the country-wide GNH levels. As a conclusion, GNH levels in the first quarter of 2014 were calculated as 47.4% happy, 28.4% neutral and 24.2% unhappy. Besides, strong correlations were found between users' happiness levels and Twitter characteristics.

1. Introduction

Twitter® is a massive social networking web portal tuned toward fast communication and it is designed not only for individual usage but also for business, media and developers (Twitter, Inc., 2014). According to Twitter Activity Monitor (twopcharts, Inc., 2014), over 1 billion currently estimated number of existing Twitter users publish over 400 million tweets every day.

Twitter's popularity as an information source has led to the development of applications and research in various domains. Humanitarian Assistance and Disaster Relief is one domain where information from Twitter is used to provide situational awareness to a crisis situation Kumar et al. (2014). For instance, in their studies Arias et al. (2013) and Asur and Huberman (2010) mined Twitter data for forecasting market trends and possible revenues for movie box office and stock market. In addition, Bouktif and Awad (2013) tried to predict stock market movement from mood collected on Twitter via ant colony based approach. Also, most popular social media analyses using Twitter data were done for the electoral predictions with two common methods. One method was proposed by Tumasjan et al. (2010) inferring votes by simply counting tweets mentioning a given candidate or party, and the other was used by O'Connor et al. (2010) in which sentiment analysis was applied to infer voting intentions from tweets.

In this perspective, Twitter provides APIs that allow developers, researchers and practitioners to collect data relevant to their studies at no cost. Twitter allows programmers to utilize those APIs which can be classified into two in terms of their objectives: a) REST API, which is popularly used for designing

* Corresponding author. Tel.: +90 505 7417058; fax: +90 212 2873297.
E-mail addresses: onur.durahim@boun.edu.tr (A.O. Durahim), mustafa.coskun@boun.edu.tr (M. Coşkun).

[1] Tel.: +90 212 3594644; fax: +90 212 2873297.

http://dx.doi.org/10.1016/j.techfore.2015.06.035
0040-1625/© 2015 Elsevier Inc. All rights reserved.

web APIs to use pull strategy for data retrieval; b) streaming API, which is used for continuous stream of public data with a push strategy. At this point, REST API method was used to collect data that was relevant to this study.

1.1. Sentiment analysis

It is stated by Pang and Lee (2008) that although the area of text analysis and mining has recently enjoyed a huge burst of research activity, there has been a steady undercurrent of interest for quite a while. There are several terms used for the procedure of the kind of text analysis performed also in this study such as "sentiment analysis", "sentiment classification", and "opinion mining". Sentiment analysis is defined by Whitelaw et al. (2005) as "labeling a document as a positive or negative evaluation of a target object". Additionally, Nasukawa and Yi (2003) stated that the essential issues in sentiment analysis are to identify how sentiments are expressed in texts, and whether these expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. Moreover, in order to improve the accuracy of the sentiment analysis, it is important to identify properly the semantic relationships between the sentiment expressions and the subject.

The sentiment analysis software which is used in this study is called SentiStrength V2.2. It is defined in Thelwall et al. (2010) and retrieved from SentiStrength official web site[2]. Besides, the Turkish word dictionary used with this software is developed by Vural et al. (2013) on the base of the largest open source Turkish natural language processing library called "Zemberek", which is commonly used in Open Office and Libre Office software.

"Well-being" is commonly used to refer the level of happiness for a group of citizens or nations. Nearly in all countries, these life satisfaction levels have been reported by governmental agencies, mostly provided as yearly rates. They are widely calculated as the level of happiness categorized by province, age, sex, occupation, etc. Today, the most common way of measuring GNH employs a self-reporting methodology, where respondents from households are asked for happiness related questions individually. Kramer (2010) states that the proponents of these metrics argue consistently and convincingly that self-reports are appropriate for this context: "because the construct is very subjective, self-reports effectively have no bias due to misperception (unlike personality measures which may have some error, for instance when one's self-perceptions do not correspond to one's behavior). In other words, if I claim to be happy, who can argue that I'm not?"

In contrast to this, in this study GNH was calculated as a standardized difference between the use of positive and negative words, aggregated across days. Along with the tweet polarity calculations, user happiness polarity calculations were performed via finding the daily happiness percentages of the users who has tweeted considerably.

1.2. Gross National Happiness (GNH)

In detail, tweet polarities were calculated daily to derive yearly happiness level graph and to be used in face validity check, whereas user happiness polarities were calculated to examine the convergent validity and to find GNH rates by province.

1.3. GNH by province survey of Turkish Statistical Institute (TSI) — 2013

In Turkey, first Life Satisfaction Survey (LSS) was implemented in 2003 as an additional module of the Household Budget Survey and it has been performed regularly as an independent survey since 2004. And, in Turkey, LSS was carried out at the province level for the first time in 2013 (TSI, 2014a). In the LSS studies, households, who are 18 years or older, were visited and interviewed face-to-face. The results of this satisfaction survey cover the satisfaction from all municipality services in the boundaries of the province.

However, reliability of this kind of survey-based method of calculating the GNH has been questioned and it is strongly believed that it would have a bias. Consider the case that a respondent of that survey would have negative or positive feelings on the day of face-to-face interview, but may have reverse feelings in the rest of the whole year. Therefore, in order to avoid such a bias, analysis of the respondents' emotional well-beings should be performed by either conducting the same survey for several days with the same people or with the data gathered for several days from social media. These approaches would be appropriate and more accurate for this kind of research. However, implementing the same face-to-face surveys for several days with the same people would be much more costly and difficult, and may contain more bias than performing a similar analysis through social media. Mao et al. (2011) compared the predictive power of traditional investor sentiment survey data and online data from Twitter feeds, news headlines, and volumes of Google search queries. They found that survey sentiment indicators were not statistically significant predictors of financial market values, while all the others were. In addition, Ansolabehere and Hersh (2012) studied the concept of relying on surveys to explain political behavior. They concluded that "studies of representation and participation based on survey reports dramatically mis-estimate the differences between voters and non-voters."

2. Related work

The rise of social broadcasting technologies has led to the open data access for the researchers. Most of the popular social media platforms allow researchers to collect valuable public data for free and conduct studies based on those data.

Facebook is the most popular social media platform that allows users share their feelings, events, actions etc., via text, image, and video messages. In this context, most popular social media studies were designed on Facebook data. For instance, Menon (2012) studied the big data structure of Facebook and how they cope with big data. Additionally, Thusoo et al. (2009) examined warehousing problem of big data on Facebook and stated a simple map reduce framework. Besides, Rieder (2013) studied data collection methods regarding Facebook data and suggested a novel method called Netvizz application. Moreover, several researches such as Lankton and McKnight (2011) and Joinson (2008) were done on the idea of motivations and effects of social media usage. Sentiment analysis is also performed using Facebook data where Ahkter and Soria (2010) examined the Facebook status messages to classify users using Neuro-Linguistic Programming (NLP) information mining. Similarly, Cvijikj and Michahelles (2011) analyzed user generated

[2] http://sentistrength.wlv.ac.uk.