Internet Research

Understanding the predictive power of social media

Kalampokis Evangelos, Tambouris Efthimios, Tarabanis Konstantinos,

Received 15 June 2012
Revised 4 February 2013
Accepted 11 February 2013

Understanding the predictive power of social media

Evangelos Kalampokis
Information Systems Laboratory, University of Macedonia, Thessaloniki, Greece and Information Technologies Institute, Centre for Research & Technology – Hellas, Thessaloniki, Greece
Efthimios Tambouris
Department of Technology and Management, University of Macedonia, Naousa, Greece and Information Technologies Institute, Centre for Research & Technology – Hellas, Thessaloniki, Greece, and
Konstantinos Tarabanis
Department of Business Administration, University of Macedonia, Thessaloniki, Greece and Information Technologies Institute, Centre for Research & Technology – Hellas, Thessaloniki, Greece

Abstract

Purpose – The purpose of this paper is to consolidate existing knowledge and provide a deeper understanding of the use of social media (SM) data for predictions in various areas, such as disease outbreaks, product sales, stock market volatility and elections outcome predictions.

Design/methodology/approach – The scientific literature was systematically reviewed to identify relevant empirical studies. These studies were analysed and synthesized in the form of a proposed conceptual framework, which was thereafter applied to further analyse this literature, hence gaining new insights into the field.

Findings – The proposed framework reveals that all relevant studies can be decomposed into a small number of steps, and different approaches can be followed in each step. The application of the framework resulted in interesting findings. For example, most studies support SM predictive power, however, more than one-third of these studies infer predictive power without employing predictive analytics. In addition, analysis suggests that there is a clear need for more advanced sentiment analysis methods as well as methods for identifying search terms for collection and filtering of raw SM data.

Originality/value – The proposed framework enables researchers to classify and evaluate existing studies, to design scientifically rigorous new studies and to identify the field's weaknesses, hence proposing future research directions.

Keywords Social networks, Data analysis, Open data, World Wide Web

Paper type Research paper

1. Introduction

In the past years, the use of social media (SM) has dramatically increased with millions of users creating massive amounts of data every day. As of September 2012, the online social networking application Facebook reached one billion monthly active users, while

The authors would like to thank the anonymous reviewers for their valuable comments that have enabled the improvement of manuscript's quality. They would also like to acknowledge that the work presented in this paper has been partially funded by the European Union through the "Linked2Media – An Open Linked Data Platform for Semantically-Interconnecting Online, Social Media Leveraging Corporate Brand and Market Sector Reputation Analysis, FP7-SME-2011 No 286714" project.

Internet Research
Vol. 23 No. 5, 2013
pp. 544-559
© Emerald Group Publishing Limited
1066-2243
DOI 10.1108/IntR-06-2012-0114

the microblogging service Twitter reported more than 140 million active users. SM data are typically in the form of textual content (e.g. in blogs, reviews and status updates), rating scores in Likert scales or stars (e.g. review ratings), like or dislike indications (e.g. reviews helpful votes and Facebook's like or Google's " + 1" buttons), web search queries (e.g. Google trends), tags and profile information (e.g. social network graphs).

SM data incorporates personal opinions, thoughts and behaviours making it a vital component of the web and a fertile ground for a variety of business and research endeavours. In this context, the predictive power of SM has been recently explored. For instance, empirical studies have analysed the Yahoo! Finance message board to predict stock market volatility (Antweiler and Frank, 2004), weblog content to predict movies success (Mishne and Glance, 2006), Google search queries to track influenza-like illnesses (Ginsberg *et al.*, 2009), Amazon reviews to predict product sales (Ghose and Ipeirotis, 2011) and Twitter posts (aka tweets) to infer levels of rainfall (Lampos and Cristianini, 2012).

These research efforts require cross-disciplinary skills as they involve both the transformation of noisy raw SM data into high-quality data suitable for statistical analysis as well as the employment of predictive analytics, which comprise "predictive models designed for predicting new/future observations or scenarios as well as methods for evaluating the predictive power of a model" (Shmueli and Koppius, 2011, p. 555). In this setting, a number of researchers have recently challenged the methods employed and the results reported by empirical studies in the area. For instance, Jungherr *et al.* (2012) repeated the study conducted by Tumasjan *et al.* (2010) and reported controversial results. In addition, Gayo-Avello (2011) and Metaxas *et al.* (2011) conducted a number of experiments and criticized generalizations regarding the predictive power of SM.

This paper aims at consolidating the knowledge created by empirical studies in recent years that exploit SM for predictions, thus enabling an in-depth understanding of SM predictive power. More specific objectives are: to identify steps that characterize all relevant studies as well as approaches that can be followed in each step, and to understand how different steps and approaches are related to SM predictive power.

We anticipate that the proposed framework will enable researchers to classify and evaluate existing studies, to design scientifically rigorous new studies and to identify the field's weaknesses, hence proposing future research directions.

The rest of the paper is structured as follows. Section 2 presents the research approach taken, while Section 3 describes the proposed framework in detail. Section 4 presents the results of employing the framework to further analyse this literature, hence providing interesting results. Finally, Section 5 draws conclusions.

2. Research approach

In order to achieve the objectives of the paper we capitalize on the method proposed by Webster and Watson (2002) for conducting systematic literature reviews in the field of information systems. Initially, we performed a systematic search in order to accumulate a relatively complete body of relevant scientific literature. Towards this end, we started with Google Scholar using the key words predict OR forecast AND social media and we collected an initial pool of papers. Thereafter, we went backward by reviewing citations in the identified papers and forward by using Google Scholar's functionality to identify papers citing the previously identified papers. We thereafter studied and filtered these initially identified papers in order to come up with the final set that was included in our research. For this purpose, we used the following inclusion and exclusion criteria:

- We excluded qualitative or purely theoretical papers (e.g. Louis and Zorlou, 2012).
- We included only studies aiming at making predictions. As a result, we have excluded empirical studies that aim at studying the relationship between SM data and phenomena outcome following an explanatory approach (e.g. Corley *et al.*, 2010; Chen *et al.*, 2011; Chevalier and Mayzlin, 2006; Chunara *et al.*, 2012; Duan *et al.*, 2008; Morales-Arroyo and Pandey, 2010; Reinstein and Snyder, 2005; Ye *et al.*, 2009).
- We included only studies that attempt to predict real world outcomes. Thus, we excluded studies that predict online features such as tie strength (Gilbert and Karahalios, 2009), volume of comments on online news (Tsagkias *et al.*, 2010) or movie rating on IMDB (Oghina *et al.*, 2012).

This approach resulted in a set of 52 papers. For the sake of clarity, the list of these papers is presented at the end of this paper in the Literature Review References section.

In order to synthesize the accumulated knowledge we performed a concept-centric analysis. The main steps and most important aspects composing the whole prediction analysis process were extracted and combined in a conceptual SM data analysis framework for predictions that structures and depicts the area. Finally, the framework was employed to further analyse the literature and to extract insights into the predictive power of SM.

3. The SM data analysis framework

The proposed framework comprises two discrete phases, namely the data conditioning phase and the predictive analysis phase. The former refers to the transformation of noisy raw SM data into high-quality data that is structured based on some predictor variables. The latter phase refers to the creation and evaluation of a predictive model that enables estimating outcome from a new set of observations.

Each of these phases can be further divided into a sequence of stages and each stage into a number of steps. Finally, different approaches can be followed in each step. Figure 1 presents our framework with the two phases, the respective stages along with their steps.

3.1 Phase 1: data conditioning

The main purpose of the data conditioning phase is the transformation of noisy raw SM data into high-quality data that will enable the computation of predictor variables. In order to define data quality we adopt and adapt a model proposed by Strong *et al.* (1997). In particular, we employ three data quality dimensions from Strong's model that we consider important in the SM data analysis realm, namely objectivity, completeness and amount of data.

Data objectivity are related to the accuracy of data production or the accuracy of the interpretation process, and specifies whether data are what it claims to be and measures what is supposed to measure. For instance, the data produced by interpreting text's sentiment could be of questionable objectivity in the case of non-rigorous sentiment analysis. The same holds when irrelevant data is interpreted as relevant. Data completeness deals with missing values from a data analysis perspective. It specifies whether or not collected data cover all aspects of a phenomenon in terms of, e.g. entities characterizing it and/or predictor variables. Finally, amount of data (or sufficiency) specifies whether or not collected data are sufficient for predictive analysis.

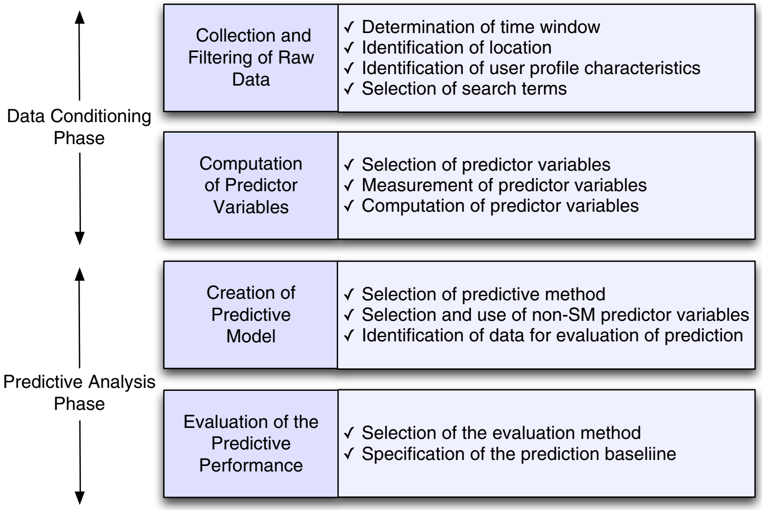
Data Conditioning Phase
Collection and Filtering of Raw Data
✓ Determination of time window
✓ Identification of location
✓ Identification of user profile characteristics
✓ Selection of search terms
Computation of Predictor Variables
✓ Selection of predictor variables
✓ Measurement of predictor variables
✓ Computation of predictor variables
Predictive Analysis Phase
Creation of Predictive Model
✓ Selection of predictive method
✓ Selection and use of non-SM predictor variables
✓ Identification of data for evaluation of prediction
Evaluation of the Predictive Performance
✓ Selection of the evaluation method
✓ Specification of the prediction baseliine

Figure 1.
The two phases and the four stages of the social media data analysis framework for predictions along with the steps that compose each stage

The stages included in this phase along with the steps in each stage are described as follows.

3.1.1 Stage 1.1: collection and filtering of raw data. This stage deals with both raw SM data collection from various sources and filtering of data in order to determine those relevant. After its completion, the final data set that will be further analysed during the next stage is produced. In order to determine the relevant raw data, the when, where, who and what questions should be answered. For example, it can be inferred that a tweet mentioning the Conservative Party one week before the UK elections of 2010 is related to these elections. The same holds for a tweet posted by David Cameron in the same period. The information used to determine relevance is extracted from the actual SM data or their metadata.

The effort required for this stage depends on both the SM and the application area. For example, data filtering in Twitter is challenging because of its noisy nature, while in Amazon it is straightforward as the reviews are aggregated in the product's web page. Detailed steps that are involved in this stage are described as follows.

Determination of time window. The time window is related to the when question as it specifies the duration of the collection activity as well as its relation to the characteristic period of the phenomenon. The characteristic period for product sales could be related to the new-product lifecycle (Liu *et al.*, 2010), while for a disease outbreak to duration of pandemic stages (Ritterman *et al.*, 2009). Clearly, the time window affects both the completeness and the sufficiency of the data.

Identification of location. The identification of location characterizing data are related to the where question. It is crucial in some phenomena (e.g. determination of natural phenomena occurrence) and thus accurate extraction of location is very important. The location characterizing SM data can be extracted from metadata (e.g. Lampos and Cristianini, 2012; Achrekar *et al.*, 2011) or inferred from actual data.

Identification of user profile characteristics. The information related to the online profile of a user answers the who question. In a number of empirical studies (e.g. Forman *et al.*, 2008; Skoric *et al.*, 2012) it is suggested that this information is very important. For instance, Achrekar *et al.* (2011) filter tweets from the same user within a certain syndrome elapsed time in order to avoid duplication from multiple encounters associated with a single episode of the illness.

Selection of search terms. The search terms selection step deals with the what question. In complex phenomena the identification of both the complete and correct set of search terms can be challenging. For example, Da *et al.* (2011) measured the search volume for 3,606 stocks through Google trends based on both "stock ticker" and "company name" and they, interestingly, identified that their correlation was only 9 per cent. The inadequate completion of this task could result in poor quality data regarding its completeness and objectivity.

The different approaches for this step can fall into two broad categories: manual approaches where researchers set search terms (e.g. Polgreen *et al.*, 2008) and dynamic approaches where search terms are derived through a computational process (e.g. Ginsberg *et al.*, 2009). We should note that we consider the use of Google trends' as a dynamic selection approach since the resulting categories are determined based on Google's natural language classification engine.

3.1.2 Stage 1.2: computation of predictor variables. This stage deals with analysis of the raw data resulting from the previous stage in order to compute the values of predictor variables. In this stage, only variables related to SM are considered despite the fact that more variables (e.g. product price) can be finally employed in the predictive analysis stage. The steps composing this stage are the following.

Selection of predictor variables. Although a number of different variables have been used in the literature, we classify them into the following categories:

- Volume-related variables: these measure the amount of SM data in the form of number of tweets, number of reviews, number of queries, etc.
- Sentiment-related variables: these measure the sentiment expressed through the data. The sentiment has been measured in the literature with the bullishness index (Oh and Sheng, 2011), review valence (Forman *et al.*, 2008), review rating (Ghose and Ipeirotis, 2011), etc.
- Profile characteristics of online users such as Facebook friends (Franch, 2013), number of followers of users that posted a tweet (Rui and Whinston, 2011), total posts (Oh and Sheng, 2011), the location of the reviewer (Forman *et al.*, 2008) and in-degree (Livne *et al.*, 2011).

The proper selection of the variables that are employed in the analysis can influence the completeness of the data.

Measurement of predictor variables. The majority of variables are usually measured at successive time instants separated by uniform time intervals and are thus expressed as time series. The time intervals that have been used in the literature vary from hours to months. However, in some cases variables are measured just once, hence resulting in one value per variable (e.g. Tumasjan *et al.*, 2010)

Careful selection of measurement time intervals allows predictor variables to be comparable to the actual outcome. For instance, Forman *et al.* (2008) aggregated data by month because the evaluation data of the outcome was formed in monthly reports. However, in some cases the measurement of variables follows different time intervals than the actual outcome data (e.g. Tumasjan *et al.*, 2010).

Computation of predictor variables. Although the computation of volume-related variables is straightforward and provides accurate results, the computation of sentiment expressed in text can be cumbersome and may provide poor results. Literature reveals that many research efforts have come up with poor sentiment analysis results (e.g. Gayo-Avello, 2011; Metaxas *et al.*, 2011), mainly because of the informal and noisy nature of SM that creates problems to widely used NLP tools. The poor performance of sentiment analysis is a major source of weakness in the quality of data objectivity as the interpreted sentiment is different than that actually expressed.

In general the approaches used for sentiment computation can be categorized as follows: lexicon-based, where sentiment is defined by the occurrence in the text of words included in a pre-defined lexicon (e.g. Metaxas *et al.*, 2011; O'Connor *et al.*, 2010) and machine learning, where sentiment is computed by language model classifiers (e.g. Asur and Huberman, 2010).

3.2 Phase 2: predictive analysis

The aim of this phase is the creation and evaluation of a predictive model that will enable accurate prediction of phenomenon outcomes based on a new set of observations, where new can be interpreted as observations in future or observations that were not included in the original data sample.

Statisticians recognize that analyses aimed at prediction are different from those aimed at explanation (Konishi and Kitagawa, 2007). Predictive power refers to the ability of predicting new observations accurately, while explanatory power to the strength of association indicated by a statistical model. "A statistically significant effect or relationship does not guarantee high predictive power, because the precision or magnitude of the causal effect might not be sufficient for obtaining levels of predictive accuracy that are practically meaningful" (Shmueli and Koppius, 2011, p. 561). Although statistically significant effects or relationships do not guarantee high predictive power, empirical studies that make predictive claims often infer predictive power from explanatory power without employing predictive analytics (Shmueli and Koppius, 2011).

3.2.1 Stage 2.1: creation of predictive model. In this stage the actual model is created based on statistical or data mining methods. The steps that compose this stage are described as follows.

Selection of predictive method. The actual model of the predictive analysis is built based on different statistical or data mining methods. The most common method in literature is linear regression but many others have been also employed such as logistic regression (Livne *et al.*, 2011), Markov models (Gruhl *et al.*, 2005), neural networks (Bollen *et al.*, 2011), support vector machine (Ritterman *et al.*, 2009) and Granger causality (Gilbert and Karahalios, 2010)

Selection and use of non-SM predictor variables. Apart from the predictor variables computed through SM data, other predictor variables are also used in the predictive model. These usually express objective facts, such as past values of phenomenon outcomes and demographics. For instance, Forman *et al.* (2008) studied the relation between both the average valence of a review and the percentage of reviews disclosing real name or location, and product sales on Amazon. Towards this end, they also employed product price as a control variable in order to reduce the possibility that results reflect differences in average unobserved product quality rather that aspects of the reviews *per se*. In addition, Rui and Whinston (2011) employed non-SM predictor variables such as budget of a movie or the fact that a movie is a sequel in order to enhance the accuracy of the model and Da *et al.* (2011) employed the number of news data from the *Wall Street Journal* in order to predict stock prices.

Identification of data for evaluation of prediction. The data referred to here represent the actual phenomenon outcome. This data are taken from official sources such as governmental documents and web sites (e.g. Lampos and Cristianini, 2012; Sakaki *et al.*, 2010; Ettredge *et al.*, 2005), other trustworthy web sites (e.g. Bollen *et al.*, 2011), international organizations (e.g. O'Connor *et al.*, 2010), etc. The accuracy and timely collection of this data are important for the creation of the predictive model.

3.2.2 Stage 2.2: evaluation of the predictive performance. In this stage prediction accuracy is evaluated against the actual outcome. The steps that comprise this stage are described as follows.

Selection of the evaluation method. The evaluation of predictive performance is very important as it provides the actual result of the study as a whole. In the literature two different approaches are mainly employed: explanatory analytics and predictive analytics. The former assesses the statistical significance of the model using metrics such as p-values or R^2 (e.g. Asur and Huberman, 2010). The latter usually obtains out-of-sample data to be used for actual evaluation based on metrics such as out-of-sample error rate and statistics such as predicted residual sums of squares (e.g. Bordino *et al.*, 2012), root mean square error (e.g. Achrekar *et al.*, 2011), mean absolute percentage error (e.g. Bollen *et al.*, 2011; Liu *et al.*, 2007) and cross-validation summaries.

In general, the criteria that specify whether a study follows a predictive evaluation method or not are the following (Shmueli and Koppius, 2011):

- Was predictive accuracy based on out-of-sample assessment?
- Was predictive accuracy assessed with adequate predictive measures?

Specification of the prediction baseline. The baseline for prediction is an important element in the literature as it provides an extra metric for evaluating predictive power. The predictive power of an SM data-based model is often judged in relation to statistical models fit with traditional data sources (e.g. Goel *et al.*, 2010; Rui and Whinston, 2011) or past values (e.g. Bollen *et al.*, 2011; Ritterman *et al.*, 2009; Wu and Brynjolfsson, 2009). In addition, the results of prediction are sometimes also evaluated against prior models and approaches (e.g. Ghose and Ipeirotis, 2011).

4. Understanding the predictive power of SM

We now employ our framework in order to gain insight into the predictive power of SM. We initially categorize the identified papers based on the application area studied (Table I) and the type of SM employed (Table II).

Table III presents classification of the literature according to the approach employed for selecting search terms, which is vital in Stage 1.1 of the framework. The table suggests that the vast majority of the studies employs manual selection methods.

In Table IV the studies that involve sentiment analysis are aggregated and categorized according to the method they have employed. Selecting such a method is important in Stage 1.2 of the proposed framework. In Table IV we do not include studies that express the sentiment as review ratings since its measurement is straightforward.

Based on the criteria employed by Shmueli and Koppius (2011) we also classify (Table V) literature according to the approach used to infer SM predictive power.

Finally, Table VI categorizes literature according to their final outcome with regard to the predictive power of SM. Some studies provide evidence for both outcomes. These are included in both categories.

Disease outbreaks	Achrekar *et al.* (2011), Althouse *et al.* (2011), Culotta (2010), Ginsberg *et al.* (2009), Hulth *et al.* (2009), Polgreen *et al.* (2008), Ritterman *et al.* (2009), Signorini *et al.* (2011), Wilson and Brownstein (2009)
Elections	Franch (2013), Gayo-Avello (2011), He *et al.* (2012), Jin *et al.* (2010), Jungherr *et al.* (2012), Livne *et al.* (2011), Lui *et al.* (2011), Metaxas *et al.* (2011), Skoric *et al.* (2012), Tjong *et al.* (2012), Tumasjan *et al.* (2010, 2012)
Macroeconomics	Choi and Varian (2012), Ettredge *et al.* (2005), Guzman (2011), O'Connor *et al.* (2010), Vosen and Schmidt (2011, 2012), Wang *et al.* (2012), Wu and Brynjolfsson (2009)
Movies	Asur and Huberman (2010), Bothos *et al.* (2010), Goel *et al.* (2010), Krauss *et al.* (2008), Liu *et al.* (2007, 2010), Mishne and Glance (2006), Rui and Whinston (2011)
Natural phenomena	Earle *et al.* (2011), Lampos and Cristianini (2012), Sakaki *et al.* (2010)
Product sales	Choi and Varian (2012), Forman *et al.* (2008), Ghose and Ipeirotis (2011), Goel *et al.* (2010), Gruhl *et al.* (2005), Jin *et al.* (2010)
Stock market	Antweiler and Frank (2004), Bollen *et al.* (2011), Bordino *et al.* (2012), Da *et al.* (2011), De Choudhury *et al.* (2008), Gilbert and Karahalios (2010), Oh and Sheng (2011), Zhang *et al.* (2011), Zhang *et al.* (2012)

Table I.
The application areas studied in the literature

Blogs	De Choudhury *et al.* (2008), Franch (2013), Gilbert and Karahalios (2010), Gruhl *et al.* (2005), Liu *et al.* (2007), Mishne and Glance (2006)
Web search	Althouse *et al.* (2011), Bordino *et al.* (2012), Choi and Varian (2012), Da *et al.* (2011), Ettredge *et al.* (2005), Ginsberg *et al.* (2009), Goel *et al.* (2010), Guzman (2011), Hulth *et al.* (2009), Lui *et al.* (2011), Polgreen *et al.* (2008), Vosen and Schmidt (2011, 2012), Wilson and Brownstein (2009), Wu and Brynjolfsson (2009)
Message boards	Antweiler and Frank (2004), Bothos *et al.* (2010), Krauss *et al.* (2008), Liu *et al.* (2010), Oh and Sheng (2011)
Reviews	Bothos *et al.* (2010), Forman *et al.* (2008), Ghose and Ipeirotis (2011)
Microblogs (Twitter and Facebook updates)	Achrekar *et al.* (2011), Asur and Huberman (2010), Bollen *et al.* (2011), Bothos *et al.* (2010), Culotta (2010), Earle *et al.* (2011), Franch (2013), Gayo-Avello (2011), He *et al.* (2012), Jungherr *et al.* (2012), Lampos and Cristianini (2012), Livne *et al.* (2011), Lui *et al.* (2011), Metaxas *et al.* (2011), O'Connor *et al.* (2010), Oh and Sheng (2011), Ritterman *et al.* (2009), Rui and Whinston (2011), Sakaki *et al.* (2010), Signorini *et al.* (2011), Skoric *et al.* (2012), Tjong *et al.* (2012), Tumasjan *et al.* (2010, 2012), Wang *et al.* (2012), Zhang *et al.* (2011), Zhang *et al.* (2012)
Social multimedia (YouTube, Flickr)	Franch (2013), Jin *et al.* (2010)

Table II.
The social media analyzed in the literature

Manual selection	Achrekar *et al.* (2011), Althouse *et al.* (2011), Asur and Huberman (2010), Bollen *et al.* (2011), Bordino *et al.* (2012), Da *et al.* (2011), De Choudhury *et al.* (2008), Ettredge *et al.* (2005), Franch (2013), Gayo-Avello (2011), Gruhl *et al.* (2005), Guzman (2011), He *et al.* (2012), Jungherr *et al.* (2012), Liu *et al.* (2007, 2011), Metaxas *et al.* (2011), Mishne and Glance (2006), O'Connor *et al.* (2010), Oh and Sheng (2011), Polgreen *et al.* (2008), Rui and Whinston (2011), Signorini *et al.* (2011), Skoric *et al.* (2012), Tjong *et al.* (2012), Tumasjan *et al.* (2010), Wilson and Brownstein (2009), Wu and Brynjolfsson (2009), Zhang *et al.* (2011), Zhang *et al.* (2012)
Dynamic selection	Choi and Varian (2012), Culotta (2010), Ginsberg *et al.* (2009), Goel *et al.* (2010), Hulth *et al.* (2009), Lampos and Cristianini (2012), Ritterman *et al.* (2009), Sakaki *et al.* (2010), Vosen and Schmidt (2011), Wang *et al.* (2012)

Table III.
Classification of literature according to the approach for search term selection

By synthesizing Tables I-VI we can further analyse the empirical studies in the literature and make some interesting observations.

Search term selection

Table III suggests that although dynamic search term selection is used in most application areas (Table I), it only appears in studies that employ web search and microblog data (Table II). Furthermore, all these studies support SM predictive power (Table VI) based on predictive analytics (Table V). In the case of manual search term selection when considering the same two SM categories, the percentage of studies that support SM predictive power falls off to 50 per cent. Hence, we can conclude that search term selection is of vital importance in microblog and web search data, and thus these SM categories call for sophisticated search terms selection methods. For instance,

Table IV.
Classification of literature according to the text's sentiment analysis approach

Lexicon-based	Bollen *et al.* (2011), Gayo-Avello (2011), Liu *et al.* (2010), Metaxas *et al.* (2011), O'Connor *et al.* (2010), Zhang *et al.* (2011), Zhang *et al.* (2012)
Machine learning	Antweiler and Frank (2004), Asur and Huberman (2010), Bothos *et al.* (2010), Gayo-Avello (2011), Gilbert and Karahalios (2010), He *et al.* (2012), Krauss *et al.* (2008), Liu *et al.* (2007), Mishne and Glance (2006), Oh and Sheng (2011), Rui and Whinston (2011)

Table V.
Classification of literature according to the evaluation approach

Explanatory evaluation	Antweiler and Frank (2004), Asur and Huberman (2010), Bordino *et al.* (2012), Da *et al.* (2011), Ettredge *et al.* (2005), Forman *et al.* (2008), Gayo-Avello (2011), He *et al.* (2012), Jin *et al.* (2010), Jungherr *et al.* (2012), Krauss *et al.* (2008), Livne *et al.* (2011), Liu *et al.* (2010, 2011), Metaxas *et al.* (2011), Mishne and Glance (2006), Polgreen *et al.* (2008), Skoric *et al.* (2012), Tjong *et al.* (2012), Tumasjan *et al.* (2010), Wilson and Brownstein (2009), Zhang *et al.* (2011), Zhang *et al.* (2012)
Predictive evaluation	Achrekar *et al.* (2011), Althouse *et al.* (2011), Bollen *et al.* (2011), Bothos *et al.* (2010), Choi and Varian (2012), Culotta (2010), De Choudhury *et al.* (2008), Franch (2013), Ghose and Ipeirotis (2011), Gilbert and Karahalios (2010), Ginsberg *et al.* (2009), Goel *et al.* (2010), Gruhl *et al.* (2005), Guzman (2011), Hulth *et al.* (2009), Lampos and Cristianini (2012), Liu *et al.* (2007), O'Connor *et al.* (2010), Oh and Sheng (2011), Ritterman *et al.* (2009), Rui and Whinston (2011), Sakaki *et al.* (2010), Signorini *et al.* (2011), Vosen and Schmidt (2011, 2012), Wang *et al.* (2012), Wu and Brynjolfsson (2009)

Table VI.
Classification of literature based on main outcome

Support SM predictive power	Achrekar *et al.* (2011), Althouse *et al.* (2011), Antweiler and Frank (2004), Asur and Huberman (2010), Bollen *et al.* (2011), Bordino *et al.* (2012), Bothos *et al.* (2010), Choi and Varian (2012), Culotta (2010), Da *et al.* (2011), De Choudhury *et al.* (2008), Ettredge *et al.* (2005), Forman *et al.* (2008), Franch (2013), Ghose and Ipeirotis (2011), Gilbert and Karahalios (2010), Ginsberg *et al.* (2009), Goel *et al.* (2010), Gruhl *et al.* (2005), Guzman (2011), Hulth *et al.* (2009), Jin *et al.* (2010), Krauss *et al.* (2008), Lampos and Cristianini (2012), Liu *et al.* (2007, 2010), Livne *et al.* (2011), Oh and Sheng (2011), Polgreen *et al.* (2008), Ritterman *et al.* (2009), Rui and Whinston (2011), Sakaki *et al.* (2010), Signorini *et al.* (2011), Tjong *et al.* (2012), Tumasjan *et al.* (2010), Vosen and Schmidt (2011, 2012), Wang *et al.* (2012), Wu and Brynjolfsson (2009), Zhang *et al.* (2011, 2012)
Challenge SM predictive power	Bollen *et al.* (2011), Forman *et al.* (2008), Gayo-Avello (2011), Goel *et al.* (2010), He *et al.* (2012), Jungherr *et al.* (2012), Liu *et al.* (2010, 2011), Metaxas *et al.* (2011), Mishne and Glance (2006), O'Connor *et al.* (2010), Skoric *et al.* (2012), Tjong *et al.* (2012), Wilson and Brownstein (2009)

Lampos and Cristianini (2012) successfully estimated daily rainfall rates for five UK cities by identifying relevant tweets through the application of Bolasso (i.e. the bootstrapped version of Least Absolute Shrinkage and Selection Operator) for search term selection.

Sentiment analysis

Table IV suggests that the majority of studies that employ sentiment analysis investigate stock market and movies (Table I). Although sentiment seems to be important in application areas such as elections, product sales and macroeconomics, only six out of 24 studies include a sentiment-related independent variable. Disease outbreaks and natural phenomena-related studies do not employ sentiment, as one might have expected. Interestingly however, 40 per cent of studies that have used sentiment-related variables challenge SM predictive power. This number increases to 65 per cent in the case of lexicon-based approaches, while it falls off to 20 per cent in those of machine learning. Hence, it seems that sentiment analysis in SM requires innovative approaches that could address the noisy and informal nature of SM.

Evaluation method

In general, half of the studies do not use predictive analytics to draw conclusions on the predictive performance of SM. These studies span equally across all SM categories (Table II). With regard to application areas (Table I), the vast majority of election-related cases do not follow a predictive analytics evaluation, while most studies related to macroeconomic indices, natural phenomena and product sales application areas evaluate predictive power based on prediction analytics. The evaluation of a predictive model with out-of-sample data is sometimes challenging. For instance, in the case of election-related studies the outcome is produced once every four or five years. In order to overcome this limitation Franch (2013) used poll data.

Tables V and VI suggest that ten out of 14 studies that challenge SM predictive power have used explanatory evaluation methods. This fact does not imply that these studies do not contribute to the understanding of SM predictive power as lack of a statistically significant relationship indicates low predictive power. In addition, 14 out of 40 studies that support SM predictive power infer predictive power without employing predictive analytics. Here we should also note that if these studies had used predictive evaluation methods, they could have presented high predictive power. However, based on the reported results we cannot assess their predictive power because a statistically significant relationship does not always ensure high predictive power. For example, "low predictive power can result from over-fitting, where an empirical model fits the training data so well that it underperforms in predicting new data" (Breiman, 2001, p. 204).

Application areas

The application area of a study seems to be related to the accuracy of the prediction that the study presents. Some application areas, such as disease outbreak and natural phenomena, do not involve the expression of any kind of opinion or sentiment. The signal that the researcher has to decode in these cases has to do with the occurrence or not of the event. As a result, these studies are expected to provide more accurate predictions than studies requiring extracting opinions or sentiment out of raw data. Moreover, some application areas, such as elections or macroeconomics, can be characterized as complex because they involve multiple and interrelated real-world entities such as political parties and politicians or complex concepts such as consumer confidence or inflation rate. The identification of the complete set of relevant raw SM data in these cases is challenging and hence call for sophisticated methods.

This becomes evident if we elaborate on two of the identified applications areas, namely elections and disease outbreak. The former involves opinion expression and is characterized by multiple and interrelated real-world entities (i.e. political parties, candidates, election constituencies), while the latter does not require opinion extraction. Table I suggests that all 11 election-related studies selected their search terms manually (Table III) and only three of them employed sentiment-related variables (Table IV). These facts could provide an explanation of the unfavourable and controversial results reported in the literature regarding predictability of election results through SM. In addition, half of the disease outbreak-related studies employed sophisticated search unit selection approaches, 80 per cent used predictive analytics evaluation and 90 per cent supported SM predictive power.

5. Conclusions

SM are a vital component of the contemporary web as they enable the production of data that reflects personal opinions, thoughts and behaviour. Since the emergence of blogs and forums, several research efforts have explored the potential of SM data for predictions of outcomes such as disease outbreaks, product sales, stock market volatility and elections. As the field is immature, some studies produce controversial results and doubtful outcomes.

In this paper, we aim at consolidating knowledge created in the past eight years by empirical studies that aim at predicting real world outcomes through SM, thus enabling an in-depth understanding of SM predictive power. Towards this end, we identify and synthesize the literature and we create a SM data analysis conceptual framework for predictions. Using this framework we further analyse the literature and classify studies according to the approaches they follow and the results they report.

The proposed framework suggests that all relevant studies can be decomposed into a small number of steps and that different choices can be made in each step. The application of the framework enabled us to make some interesting observations. The majority of the empirical studies support SM predictive power, however more than one-third of these studies infer predictive power without employing predictive analytics. Sophisticated search term selection is crucial in web search and microblog data. In addition, the use of sentiment-related variables resulted often in controversial outcomes proving that SM data call for sophisticated sentiment analysis approaches.

We anticipate that both the framework and analysis results will enable researchers to design scientifically rigorous new studies and to more easily identify the field's weaknesses, hence proposing new future research directions.

References

Breiman, L. (2001), "Statistical modeling: the two cultures", *Statistical Science*, Vol. 16 No. 3, pp. 199-215.

Chen, Y., Wang, Q. and Xie, J. (2011), "Online social interactions: a natural experiment on word of mouth versus observational learning", *Journal of Marketing Research*, Vol. 48 No. 2, pp. 238-254.

Chevalier, J.A. and Mayzlin, D. (2006), "The effect of word of mouth on sales: online book reviews", *Journal of Marketing Research*, Vol. 43 No. 3, pp. 345-354.

Chunara, R., Andrews, J.R. and Brownstein, J.S. (2012), "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak", *The American Journal of Tropical Medicine and Hygiene*, Vol. 86 No. 1, pp. 39-45.

Corley, C.D., Cook, D.J., Mikler, A.R. and Singh, K.P. (2010), "Text and structural data mining of influenza mentions in web and social media", *International Journal of Environmental Research and Public Health*, Vol. 7 No. 2, pp. 596-615.

Duan, W., Gu, B. and Whinston, A.B. (2008), "Do online reviews matter? – an empirical investigation of panel data", *Decision Support Systems*, Vol. 45 No. 4, pp. 1007-1016.

Gilbert, E. and Karahalios, K. (2009), "Predicting tie strength with social media", *27th International Conference on Human Factors in Computing Systems – CHI*, pp. 211-220.

Konishi, S. and Kitagawa, G. (2007), *Information Criteria and Statistical Modeling*, Springer, New York, NY.

Louis, C. St and Zorlou, G. (2012), "Can twitter predict disease outbreaks?", *British Medical Journal*, Vol. 344, p. e235.

Morales-Arroyo, M. and Pandey, T. (2010), "Identification of critical eWOM dimensions for music albums", *IEEE International Conference on Management of Innovation and Technology, IEEE,* pp. 1230-1235.

Oghina, A., Breuss, M., Tsagkias, M. and de Rijke, M. (2012), "Predicting IMDB movie rating using social media", *34th European Conference on Information Retrieval (ECIR)*, *Springer*, pp. 503-507.

Reinstein, D. and Snyder, C.M. (2005), "The influence of expert reviews on consumer demand for experience goods: a case study of movie critics", *Journal of Industrial Economics*, Vol. 53 No. 1, pp. 27-51.

Shmueli, G. and Koppius, O.R. (2011), "Predictive analytics in information systems research", *MIS Quarterly*, Vol. 35 No. 3, pp. 553-572.

Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997), "Data quality in context", *Communications of the ACM*, Vol. 40 No. 5, pp. 103-110.

Tsagkias, E., Weerkamp, W. and de Rijke, M. (2010), "News comments: exploring, modeling, and online predicting", *ECIR*, *Springer*, pp. 191-203.

Webster, J. and Watson, R.T. (2002), "Analyzing the past to prepare for the future: writing a literature review", *MIS Quarterly*, Vol. 26 No. 2, pp. xiii-xxiii.

Ye, Q., Law, R. and Gu, B. (2009), "The impact of online user reviews on hotel room sales", *International Journal of Hospitality Management*, Vol. 28 No. 1, pp. 180-182.

Literature review references

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H. and Liu, B. (2011), "Predicting flu trends using twitter data", *IEEE Conference on Computer Communications Workshops*, *IEEE*, pp. 702-707.

Althouse, B.M., Ng, Y.Y. and Cummings, D.A.T. (2011), "Prediction of dengue incidence using search query surveillance", *Public Library of Science*, Vol. 5 No. 8, pp. 1-7.

Antweiler, W. and Frank, M.Z. (2004), "Is all that talk just noise? The information content of internet stock message boards", *Journal of Finance*, Vol. 59 No. 3, pp. 1259-1294.

Asur, S. and Huberman, B.A. (2010), "Predicting the future with social media", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Press*, pp. 492-499.

Bollen, J., Mao, H. and Zeng, X.J. (2011), "Twitter mood predicts the stock market", *Journal of Computational Science*, Vol. 2 No. 1, pp. 1-8.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A. and Weber, I. (2012), "Web search queries can predict stock market volumes", *PLoS ONE*, Vol. 7 No. 7, p. e40014.

Bothos, E., Apostolou, D. and Mentzas, G. (2010), “Using social media to predict future events with agent-based markets”, *IEEE Intelligent Systems*, Vol. 25 No. 6, pp. 50-58.

Choi, H. and Varian, H. (2012), “Predicting the present with Google trends”, *The Economic Record*, Vol. 88 No. S1, pp. 2-9.

Culotta, A. (2010), “Towards detecting influenza epidemics by analyzing Twitter messages”, First Workshop on Social Media Analytics, 25-28 July, ACM Press, Washington, DC, pp. 115-122.

Da, Z., Engelberg, J. and Gao, P. (2011), “In search of attention”, *The Journal of Finance*, Vol. 66 No. 5, pp. 1461-1499.

De Choudhury, M., Sundaram, H., John, A. and Seligmann, D.D. (2008), “Can blog communication dynamics be correlated with stock market activity?”, *19th ACM Conference on Hypertext and Hypermedia*, *ACM Press*, pp. 55-60.

Earle, P., Bowden, D.C. and Guy, M. (2011), “Twitter earthquake detection: earthquake monitoring in a social world”, *Annals of Geophysics*, Vol. 54 No. 6, pp. 708-715.

Ettredge, M., Gerdes, J. and Karuga, G. (2005), “Using web-based search data to predict macroeconomic statistics”, *Communications of the ACM*, Vol. 48 No. 11, pp. 87-92.

Forman, C., Ghose, A. and Wiesenfeld, B. (2008), “Examining the relationship between reviews and sales: the role of the reviewer identity disclosure in electronic markets”, *Information Systems Research*, Vol. 19 No. 3, pp. 291-313.

Franch, F. (2013), “(Wisdom of the crowds)^2: 2010 UK election prediction with social media”, *Journal of Information Technology & Politics*, Vol. 10 No. 1, pp. 57-71.

Gayo-Avello, D. (2011), “Don’t turn social media into another ‘Literary Digest’ poll”, *Communications of the ACM*, Vol. 54 No. 10, pp. 121-128.

Ghose, A. and Ipeirotis, P.G. (2011), “Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics”, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, Vol. 23 No. 10, pp. 1498-1512.

Gilbert, E. and Karahalios, K. (2010), “Widespread worry and the stock market”, *Fourth International Conference on Weblogs and Social Media*, *AAAI Press*, pp. 58-65.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), “Detecting influenza epidemics using search engine query data”, *Nature*, Vol. 457 No. 7232, pp. 1012-1014.

Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M. and Watts, D.J. (2010), “Predicting consumer behaviour with web search”, *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 107 No. 41, pp. 17486-17490.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005), “The predictive power of online chatter”, *Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, *ACM Press*, pp. 78-87.

Guzman, G. (2011), “Internet search behavior as an economic forecasting tool: the case of inflation expectations”, *Journal of Economic and Social Measurement*, Vol. 36 No. 3, pp. 119-167.

He, Y., Saif, H., Wei, Z. and Wong, K. (2012), “Quantising opinions for political tweets analysis”, *Eight International Conference on Language Resources and Evaluation, European Language Resources Association*, pp. 3901-3906.

Hulth, A., Rydevik, G. and Linde, A. (2009), “Web queries as a source for syndromic surveillance”, *PLoS ONE*, Vol. 4 No. 2, p. e4378.

Jin, X., Gallagher, A., Cao, L., Luo, J. and Han, J. (2010), “The wisdom of social multimedia: using flickr for prediction and forecast”, *ACM Multimedia 2010*, ACM Press, pp. 1235-1244.

Jungherr, A., Jurgens, P. and Schoen, H. (2012), “Why the pirate party won the German election of 2009 or the trouble with predictions: a response to Tumasjan, A., Sprenger, T.O., Sander, P.G. and Welpe, I.M. ‘predicting elections with twitter: what 140 characters reveal about political and sentiment’”, *Social Science Computer Review*, Vol. 30 No. 2, pp. 229-234.

Krauss, J., Nann, S., Simon, D., Fischbach, K. and Gloor, P. (2008), "Predicting movie success and academy awards through sentiment and social network analysis", *16th European Conference on Information Systems*, pp. 2026-2037.

Lampos, V. and Cristianini, N. (2012), "Nowcasting events from the social web with statistical learning", *ACM Transactions on Intelligent Systems and Technology*, Vol. 3 No. 4, pp. 72:1-72:22.

Liu, Y., Huang, X., An, A. and Yu, X. (2007), "ARSA: a sentiment-aware model for predicting sales performance using blogs", *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press*, pp. 607-614.

Liu, Y., Chen, Y., Lusch, R.F., Chen, H., Zimbra, D. and Zeng, S. (2010), "User-generated content on social media: predicting market success with online word-of-mouth", *IEEE Intelligent Systems*, Vol. 25 No. 1, pp. 75-78.

Livne, A., Simmons, P.S., Adar, E. and Adamic, L.A. (2011), "The party is over here: structure and content in the 2010 election", *Fifth International AAAI Conference on Weblogs and Social Media, AAAI Press*, pp. 201-208.

Lui, C., Metaxas, P.T. and Mustafaraj, E. (2011), "On the predictability of the US Elections through search volume activity", *IADIS International Conference e-Society*, pp. 165-172.

Metaxas, P.T., Mustafaraj, E. and Gayo-Avello, D. (2011), "How (not) to predict election", *IEEE Third International Conference on Social Computing, IEEE*, pp. 165-171.

Mishne, G. and Glance, N. (2006), "Predicting movie sales from blogger sentiment", American Association for Artificial Intelligence 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, Stanford University in Stanford, CA, 27-29 March.

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A. (2010), "From tweets to polls: linking text sentiment to public opinion time series", *Proceedings of the International AAAI Conference on Weblogs and Social Media*, *AAAI Press*, pp. 122-129.

Oh, C. and Sheng, O. (2011), "Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement", *32nd International Conference on Information Systems*, *AIS*, p. 17.

Polgreen, P.M., Chen, Y., Pennock, D.M. and Nelson, F.D. (2008), "Using internet searches for influenza surveillance", *Clinical Infectious Diseases*, Vol. 47 No. 11, pp. 1443-1448.

Ritterman, J., Osborne, M. and Klein, E. (2009), "Using prediction markets and twitter to predict a swine flu pandemic", First International Workshop on Mining Social Media, 9 November, Sevilla, pp. 9-17.

Rui, H. and Whinston, A. (2011), "Designing a social-broadcasting-based business intelligence system", *ACM Transactions on Management Information Systems*, Vol. 2 No. 4, pp. 22:1-22:19.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), "Earthquake shakes twitter users: real-time event detection by social sensors", *19th International Conference on World Wide Web (WWW'10), ACM Press*, pp. 851-860.

Signorini, A., Segre, A.M. and Polgreen, P.M. (2011), "The use of twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic", *PLoS ONE*, Vol. 6 No. 5, p. e19467.

Skoric, M., Poor, N., Achananuparp, P., Lim, E. and Jiang, J. (2012), "Tweets and votes: a study of the 2011 Singapore General Election", *45th Hawaii International Conference on System Sciences*, *IEEE*, pp. 2583-2591.

Tjong, E., Sang, K. and Bos, J. (2012), "Predicting the 2011 Dutch senate election results with twitter", *13th Conference of the European Chapter of the Association for Computational Linguistics*, *Association for Computational Linguistics*, pp. 53-60.

Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M. (2010), "Predicting elections with twitter: what 140 characters reveal about political sentiment", *Fourth International AAAI Conference on Weblogs and Social Media, AAAI Press*, pp. 178-185.

Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M. (2012), "Where there is a sea there are pirates: response to Jungherr, Jurgens, and Schoen", *Social Science Computer Review*, Vol. 30 No. 2, pp. 235-239.

Vosen, S. and Schmidt, T. (2011), "Forecasting private consumption: survey-based indicators vs Google trends", *Journal of Forecasting*, Vol. 30 No. 6, pp. 565-578.

Vosen, S. and Schmidt, T. (2012), "A monthly consumption indicator for Germany based on internet search query data", *Applied Economics Letters*, Vol. 19 No. 7, pp. 683-687.

Wang, X., Gerber, M.S. and Brown, D.E. (2012), "Automatic crime prediction using events extracted from twitter posts", in Yang, S.J., Greengerg, A.M. and Endsley, M. (Eds), *SBP 2012, LNCS 7227*, Springer-Verlag, Berlin Heidelberg, pp. 231-238.

Wilson, K. and Brownstein, J.S. (2009), "Early detection of disease outbreaks using the internet", *Canadian Medical Association Journal*, Vol. 180 No. 8, pp. 829-831.

Wu, L. and Brynjolfsson, E. (2009), "The future of prediction: how google searches foreshadow housing prices and quantities", *30th International Conference on Information Systems, AISE*, available at: http://aisel.aisnet.org/icis2009/147 (accessed February 2012).

Zhang, X., Fuehres, H. and Gloor, P.A. (2011), "Predicting stock market indicators through twitter 'I hope it is not as bad as I fear' ", *Procedia – Social and Behavioral Sciences*, Vol. 26, pp. 55-62.

Zhang, X., Fuehres, H. and Gloor, P.A. (2012), "Predicting asset value through twitter buzz", in Altmann, J. *et al.* (Eds), *Advances in Collective Intelligence*, AISC 113, Springer Berlin Heidelberg, pp. 23-34.

Further reading

Shmueli, G. (2010), "To explain or to predict?", *Statistical Science*, Vol. 25 No. 3, pp. 289-310.

About the authors

Evangelos Kalampokis is a PhD Student at the Information Systems Laboratory at the University of Macedonia, Greece. He is also a Research Assistant with the Information Technologies Institute of the Centre for Research & Technology – Hellas (CERTH/ITI) in Thessaloniki, Greece, while from October 2010 until November 2011 he was a Research Intern at the Digital Enterprise Research Institute (DERI) of the National University of Ireland, Galway (NUIG). His main research interests include Open and Linked Data, Social Media, eGovernment and Data Mining. Evangelos Kalampokis is the corresponding author and can be contacted at: ekal@uom.gr

Dr Efthimios Tambouris is an Assistant Professor of Information Systems at the Technology Management Department at the University of Macedonia, Thessaloniki, Greece. Before that, he served as a Researcher Grade D at the research center CERTH/ITI and at research center NCSR "Demokritos". He was also Founder and Manager of the eGovernment Unit at Archetypon SA, an international IT company. He holds a Diploma in Electrical Engineering from the National Technical University of Athens, Greece, and an MSc and PhD from Brunel University, UK. During the past years he has initiated and managed several research projects (e.g. IST EURO-CITI, IST eGOV, eContent eMate). He has also participated in numerous research projects (FP6/IST, e.g. OneStopGov, DEMO-net, FP5/IST, TAP, ACTS, ESPRIT, SPRITE-S2, etc.), service contracts (e.g. MODINIS Interoperability Study, European eParticipation Study) and standardization activities (CEN/ISSS project on eGovernment metadata, CEN/ISSS eGovernment Focus Group). He has more than 120 publications in eGovernment, eParticipation, eLearning and eHealth.

Professor Konstantinos Tarabanis is a Professor at the Department of Business Administration of the University of Macedonia, Greece, and the Director of the Information Systems Laboratory at the same university. He received an Engineering Diploma in Mechanical Engineering from the

National Technical University of Athens (1983), an MS in both Mechanical Engineering and Computer Science (1984 and 1988, respectively), and a PhD in Computer Science (1991), at Columbia University, New York, NY. He was a Research Staff Member at the IBM T.J. Watson Research Centre, 1991-1994, and was employed by the IBM Corporation as a whole during 1984-1994. In recognition of his research, he was the recipient of the Anton Philips Best Paper Award at the 1991 IEEE International Conference on Robotics and Automation. He has about 200 research publications in the areas of software modeling and development for the domains of eGovernment, eBusiness, eLearning, eManufacturing, etc.

This article has been cited by:

1. Naheed Bashir, K.Nadia Papamichail, Khaleel Malik. 2017. Use of Social Media Applications for Supporting New Product Development Processes in Multinational Corporations. *Technological Forecasting and Social Change* **120**, 176-183. [CrossRef]
2. OhSehwan, Sehwan Oh, BaekHyunmi, Hyunmi Baek, AhnJoongHo, JoongHo Ahn. 2017. Predictive value of video-sharing behavior: sharing of movie trailers and box-office revenue. *Internet Research* **27**:3, 691-708. [Abstract] [Full Text] [PDF]
3. Antonella Angelini, Paola Ferretti, Gabriele Ferrante, Paolo Graziani. 2017. Social Media Development Paths in Banks. *Journal of Promotion Management* **23**:3, 345-358. [CrossRef]
4. Jiayin Pei, Guang Yu, Xianyun Tian, Maureen Renee Donnelley. 2017. A new method for early detection of mass concern about public health issues. *Journal of Risk Research* **20**:4, 516-532. [CrossRef]
5. JunSeung-Pyo, Seung-Pyo Jun, ParkDo-Hyung, Do-Hyung Park. 2017. Visualization of brand positioning based on consumer web search information. *Internet Research* **27**:2, 381-407. [Abstract] [Full Text] [PDF]
6. Md Safiullah, Pramod Pathak, Saumya Singh, Ankita Anshul. 2017. Social media as an upcoming tool for political marketing effectiveness. *Asia Pacific Management Review* **22**:1, 10-15. [CrossRef]
7. R. Piryani, D. Madhavi, V.K. Singh. 2017. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management* **53**:1, 122-150. [CrossRef]
8. Guilherme M. Thomaz, Alexandre A. Biz, Eduardo M. Bettoni, Luiz Mendes-Filho, Dimitrios Buhalis. 2016. CONTENT MINING FRAMEWORK IN SOCIAL MEDIA: A FIFA WORLD CUP 2014 CASE ANALYSIS. *Information & Management* . [CrossRef]
9. Andrea Ceron, Luigi Curini, Stefano Maria Iacus. 2016. iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences* **367-368**, 105-124. [CrossRef]
10. Perez-GarciaLorena, Lorena Perez-Garcia, Jan Broekaert, Jan Broekaert, NoteNicole, Nicole Note. 2016. The temporal evolution of the normalized web distance. *Internet Research* **26**:5, 1269-1290. [Abstract] [Full Text] [PDF]
11. Gálvez-RodríguezMaría del Mar, María del Mar Gálvez-Rodríguez, Caba-PérezCarmen, Carmen Caba-Pérez, López-GodoyManuel, Manuel López-Godoy. 2016. Drivers of Twitter as a strategic communication tool for non-profit organizations. *Internet Research* **26**:5, 1052-1071. [Abstract] [Full Text] [PDF]
12. Andrew Sun, Michael Lachanski, Frank J. Fabozzi. 2016. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis* . [CrossRef]
13. MaizAnder, Ander Maiz, ArranzNieves, Nieves Arranz, Fdez. de ArroyabeJuan Carlos, Juan Carlos Fdez. de Arroyabe. 2016. Factors affecting social interaction on social network sites: the Facebook case. *Journal of Enterprise Information Management* **29**:5, 630-649. [Abstract] [Full Text] [PDF]
14. Panagiotis Andriotis, George Oikonomou, Theo Tryfonas, Shancang Li. 2016. Highlighting Relationships of a Smartphone's Social Ecosystem in Potentially Large Investigations. *IEEE Transactions on Cybernetics* **46**:9, 1974-1985. [CrossRef]
15. Ya-Ling Wu, Eldon Y. Li, Wei-Lun Chang. 2016. Nurturing user creative performance in social media networks. *Internet Research* **26**:4, 869-900. [Abstract] [Full Text] [PDF]
16. Alexander Wieneke, Christiane Lehrer. 2016. Generating and exploiting customer insights from social media data. *Electronic Markets* **26**:3, 245-268. [CrossRef]

17. Yean-Fu Wen, Ko-Yu Hung, Yi-Ting Hwang, Yeong-Sung Frank Lin. 2016. Sports lottery game prediction system development and evaluation on social networks. *Internet Research* **26**:3, 758-788. [Abstract] [Full Text] [PDF]

18. Yulia Sidorova, Michela Arnaboldi, Jacopo Radaelli. 2016. Social media and performance measurement systems: towards a new model?. *International Journal of Productivity and Performance Management* **65**:2, 139-161. [Abstract] [Full Text] [PDF]

19. Jie Qin, Tai-Quan Peng. 2016. Googling environmental issues. *Internet Research* **26**:1, 57-73. [Abstract] [Full Text] [PDF]

20. Yuxian Eugene Liang, Soe-Tsyr Daphne Yuan. 2016. Predicting investor funding behavior using crunchbase social network features. *Internet Research* **26**:1, 74-100. [Abstract] [Full Text] [PDF]

21. Martin Hilbert. 2016. Big Data for Development: A Review of Promises and Challenges. *Development Policy Review* **34**:1, 135-174. [CrossRef]

22. Veronica Liljander, Johanna Gummerus, Magnus Söderlund. 2015. Young consumers' responses to suspected covert and overt blog marketing. *Internet Research* **25**:4, 610-632. [Abstract] [Full Text] [PDF]

23. Rodney Graeme Duffett. 2015. Facebook advertising's influence on intention-to-purchase and purchase amongst Millennials. *Internet Research* **25**:4, 498-526. [Abstract] [Full Text] [PDF]

24. Carlo Lipizzi, Luca Iandoli, José Emmanuel Ramirez Marquez. 2015. Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. *International Journal of Information Management* **35**:4, 490-503. [CrossRef]

25. Yang Yu, Xiao Wang. 2015. World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. *Computers in Human Behavior* **48**, 392-400. [CrossRef]

26. Bongsug (Kevin) Chae. 2015. Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics* **165**, 247-259. [CrossRef]

27. Benedikt Boecking, Margeret Hall, Jeff Schneider. 2015. Event Prediction With Learning Algorithms-A Study of Events Surrounding the Egyptian Revolution of 2011 on the Basis of Micro Blog Data. *Policy & Internet* **7**:2, 159-184. [CrossRef]

28. Pin Luarn, Yu-Ping Chiu. 2015. Key variables to predict tie strength on social network sites. *Internet Research* **25**:2, 218-238. [Abstract] [Full Text] [PDF]

29. Liaquat Hossain, Muhammad Rabiul Hassan, Rolf T. Wigand. 2015. Resilient Information Networks for Coordination of Foodborne Disease Outbreaks. *Disaster Medicine and Public Health Preparedness* **9**:02, 186-198. [CrossRef]

30. Chenyan Xu, Yang Yu. Measuring NBA Players' Mood by Mining Athlete-Generated Content 1706-1713. [CrossRef]

31. Martin Hilbert. ICT4ICTD: Computational Social Science for Digital Development 2145-2157. [CrossRef]

32. Moez Ltifi. 2014. Roles of social media in the retail sector in Tunisia: the case of Facebook. *International Strategic Management Review* **2**:2, 79-88. [CrossRef]

33. Lisa Madlberger, Amai Almansour. Predictions based on Twitter &#x2014; A critical view on the research process 1-6. [CrossRef]

34. Matthew S. Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* **61**, 115-125. [CrossRef]

35. Wu He, Yong Chen. 2014. Using Blog Mining as an Analytical Method to Study the Use of Social Media by Small Businesses. *Journal of Information Technology Case and Application Research* **16**:2, 91-104. [CrossRef]

36. Schoen Harald, Gayo-Avello Daniel, Takis Metaxas Panagiotis, Mustafaraj Eni, Strohmaier Markus, Gloor Peter. 2013. The power of prediction with social media. *Internet Research* **23**:5, 528-543. [Abstract] [Full Text] [PDF]