Combining different evaluation systems on social media for measuring user satisfaction

Simona Balbi[a], Michelangelo Misuraca[*,b], Germana Scepi[a]

[a] Department of Economics and Statistics, University of Naples Federico II, Italy
[b] Department of Business Administration and Law, University of Calabria, Arcavacata di Rende, Cosenza 87036, Italy

A R T I C L E   I N F O

A B S T R A C T

Web 2.0 allows people to express and share their opinions about products and services they buy/use. These opinions can be expressed in various ways: numbers, texts, emoticons, pictures, videos, audios, and so on. There has been great interest in the strategies for extracting, organising and analysing this kind of information. In a social media mining framework, in particular, the use of textual data has been explored in depth and still represents a challenge. On a rating and review website, user satisfaction can be detected both from a rating scale and from the written text. However, in common practice, there is a lack of algorithms able to combine judgments provided with both comments and scores. In this paper we propose a strategy to jointly measure the user evaluations obtained from the two systems. Text polarity is detected with a sentiment-based approach, and then combined with the associated rating score. The new rating scale has a finer granularity. Moreover, also enables the reviews to be ranked. We show the effectiveness of our proposal by analysing a set of reviews about the Uffizi Gallery in Florence (Italy) published on TripAdvisor.

1. Introduction

With the rapid expansion of Web 2.0, sharing personal feelings and judgments with others has become a common habit. People evaluate products and services they buy/use by describing their experiences. There are many websites and social media specialised in one or more topics, where people can publish their "opinion". These opinions can be expressed in various ways: numbers, texts, emoticons, pictures, videos, audios, and so on. Following the idea that online evaluations and electronic word-of-mouth can influence customer behaviour (Hennig-Thurau & Walsh, 2004; Sandes & Urdan, 2013), it is important to analyse users' opinions. There is considerable interest in how knowledge can be extracted from this kind of information, and nowadays this task is considered the core of many marketing and business strategies, and in competitive analysis (e.g. He, Zha, & Li, 2013).

In the *rating and review* social media (e.g. Amazon, Yelp, Imdb), users express their opinions with an evaluation scale visualised by bullets or stars - e.g. from *1star* (terrible) to *5stars* (excellent) - and/or a textual review. In the framework of social media mining, in recent years, great attention has been devoted to the so-called *rating inference*, i.e. translating the text into a given number of bullets/stars. However, it is quite difficult to quantify and evaluate opinions expressed in plain text (Baumgartner & Steenkamp, 2001). The most common way of approaching this problem, sometimes referred to "seeing the stars" (Shimada & Endo, 2008), entails using some sentiment analysis tools. In common practice, when both scores and texts are available, there are a limited number of algorithms able to combine the two evaluation systems.

---

* Corresponding author.
  E-mail address: michelangelo.misuraca@unical.it (M. Misuraca).

Furthermore, recommender platforms are becoming increasingly important not only in scoring products and services, but also in ranking them. As an example, let us consider the world-famous TripAdvisor[1]. TripAdvisor shares user-generated contents about hotels, restaurants and touristic attractions. Travellers' satisfaction is visualised through a 1-to-5 star system, and textual reviews are also published to communicate the user experience. TripAdvisor also ranks businesses and attractions, in a given place. This is perhaps one of the most interesting and debated question. They claim that their ranking algorithm is based on three factors: quality (measured by bullets), quantity (number of reviews), and recency of reviews. In May 2016, TripAdvisor modified the algorithm, but these three basic factors did not change. It is interesting to note that no information is extracted from the reviews.

The main research questions underlying this paper are:

- How to analyse different kind of information available on social media?
- How to increase the usefulness of written reviews in recommendation systems?

Our proposal entails combining the two different kinds of information, the rating and the sentiment of the review, In this way it is possible to produce a reliable score, also useful in ranking procedures. From a statistical viewpoint, the idea is to transform the ordinal variable "satisfaction" associated with the explicit quantification given by the customer, into a quantitative variable, obtained by introducing the score of the sentiment underlying the textual description. Easier solutions, based for example on the length of the text or on other linguistic measures, give poor results in practice. Our new measure of satisfaction is little affected by said circumstances. Moving from an ordinal system to a continuous variable gives a more stable and precise measure of quality, also used in the ranking algorithms.

This work is organised as follows. In Section 2 we present a brief overview about the research in this field. In Section 3 our proposal for jointly measuring user evaluations with review polarities and ratings is described. In Section 4 we show the effectiveness of the strategy by analysing a dataset of TripAdvisor reviews about the Uffizi Gallery in Florence (Italy). Finally, in Section 5 we conclude with some remarks and the future directions of the research.

2. Background and related work

Nowadays most of the people share their opinions on social media and Web sites, devoted to specific topics such as e-commerce, tourism, points of interest, and so on. Consequently, the amount of available Web data is growing rapidly. This huge and varied set of data cannot be processed manually. Nevertheless, automatic processing also requires a huge computational effort. It is difficult to extract the related information from opinions, and then to understand, summarise and organise them into usable forms (Balahur & Jacquet, 2015). At the same time, it is very important to process the information for making decisions, both for companies as well as for potential users/customers. Due to the huge differences of social media channels as well as their unique characteristics, not all approaches are suitable for each source, i.e. there is no "one-size-fits-all" approach (Petz et al., 2013).

Analysing opinions written in natural language is a very interesting research domain, known as opinion mining (OM) or sentiment analysis (SA) (Petz et al., 2014). According to Pang and Lee (2008):

Opinion mining is a recent discipline at the crossroads of information retrieval, text mining and computational linguistics which tries to detect the opinions expressed in natural language texts.

A systematic literature survey regarding the computational techniques, models and algorithms for mining opinions can be found in Khairullah, Baharum, Aurnagzeb, and Ashraf (2014). These authors share the idea of Tang, Tan, and Cheng (2009) that OM should be deemed as a subarea of SA. Doaa (2016) proposes an interesting comparison of forty-one papers concerning the new challenges in SA. This author consider OM and SA as synonyms, referring exactly to the same research area. Liu (2015) underlines in his book - where all aspects of SA are described - that even if the term SA is generally used in industry, while both SA and OM are used in academia, in a broader sense they refer to the same topic. It is not our aim to review the entire body of literature concerning SA (see Medhat, Hassan, & Korashy, 2014; Qazi, Raj, Hardaker, & Standing, 2017; Ravi & Ravi, 2015).

A large number of papers mention SA in the context of the so-called *polarity classification* (e.g. Cambria, Schuller, Xia, & Havasi, 2013; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). The main goal is to classify documents written in natural language on the basis of their semantic *polarity*. This term is commonly used in linguistics to distinguish affirmative and negative forms. The calculation of the positivity/negativity of a document (PN-polarity) entails deciding if the textual content expresses a positive or negative sentiment. If the document is fractioned into sentences, it is possible to first calculate the polarity of each sentence and then the polarity of the whole document (Tan, Na, Theng, & Chang, 2011). The polarity score of each sentence depends on the lexicon of polarised terms used, while the polarity of the whole document depends on the polarities of its sentences. The PN-polarity is usually quantified by considering a score of -1, 0 and $+1$ for negative, neutral and positive polarity, respectively (Liu, Hu, & Cheng, 2005). Some authors have proposed different scoring systems by defining the polarity not only in terms of sign but also taking into account the PN-strength of the sentiment (Nielsen, 2011). In recent years, research has focused on more efficient term weighting methods in order to improve the performance of SA (Deng, Luo, & Yu, 2014). Nguyen, Chang, and Hui (2011), for example, proposed a supervised term weighting scheme based on the Kullback–Leibler divergence. Lin, Zhang, Wang, and Zhou (2012) and Khan, Qamar, and Bashir (2016) proposed the use of mutual information. Gann, Day, and Zhou (2014) introduced a *total sentiment index* to score the

---

[1] https://www.tripadvisor.com.