A polarity-based strategy for ranking social media reviews

Una strategia basata sulla polarità per ordinare le recensioni sui social media

Simona Balbi, Michelangelo Misuraca and Germana Scepi

Abstract The Opinion Mining methods are widely used to analyse and classify the choices, preferences and behaviours of consumers through the opinions gathered on the Web. On social media like TripAdvisor such opinions are usually expressed with a score and a short text. This paper proposes a strategy for ranking reviews using a scale based jointly on the rating and on the text of the reviews.

Abstract *I metodi di Opinion Mining sono oggi ampiamente utilizzati per analizzare e classificare le scelte, le preferenze e il comportamento dei consumatori attraverso opinioni raccolte sul web. Sui social media come TripAdvisor tali opinioni vengono solitamente espresse con un punteggio e con un breve testo. In questo lavoro si propone una strategia per ordinare le diverse recensioni con una scala di misura basata sia sul punteggio sia sul testo scritto.*

Key words: Textual Data, Opinion Mining, Ranking

1 Introduction

With the rapid expansion of social media, it is more and more widespread the practice of sharing opinions on the Web. The ways for expressing those opinions are many: numbers, texts, emoticons, images, videos, audios. There are often a joint use of these communication tools. It is becoming a habit for users to evaluate the products/services they buy/use, by describing their personal feelings and judgments.

We can find online websites specialised in one or more topics, where people can give their opinion using an evaluation scale (e.g., from “terrible”=1 to “excellent”=5), visualised by bullets or stars, and combined with a written description.

Simona Balbi, Germana Scepi
Università Federico II di Napoli e-mail: simona.balbi@unina.it, germana.scepi@unina.it

Michelangelo Misuraca
Università della Calabria, Arcavacata di Rende e-mail: michelangelo.misuraca@unical.it

As this practice is nowadays considered the core of many marketing strategies, there is a large interest on how to extract knowledge from such a kind of information.

Opinion mining procedures have been developed with the main goal of understanding the mood in a text, transforming it in a numerical value. The basic idea is identifying positive, negative, or neutral viewpoints. Researchers involved in defining proper methods for mining opinions on the Web are mainly computer scientists and computational linguists. They often claim to use statistical techniques.

The main point we are interested in this paper is that we often see the lack of a statistical perspective. Statisticians are professionally involved into the problem of quantifying something that is not quantitative in itself. Furthermore, the implications in the choice of a scale, or in the choice of a weighting system, or in the choice of the proper method for analysing those unconventional data pertain to statisticians.

Here we focus our attention on the so called rating-inference problem [6], and its implications when we refer to "reviews and ratings" social media like TripAdvisor. In this kind of media we usually find ratings in a 1-to-5 stars system, together with written judgments. The challenge is stimulating for a statistician: on one hand, we have a judgment in a 5-point scale; on the other hand we have a (usually) short text. We propose a two-step strategy for taking into account jointly the two assessments and defining a unique rating.

The paper is organised as follows. Section 2 defines the theoretical framework. Section 3 considers the case study. The proposed strategy is presented in Section 4, while the main results of applying the strategy on TripAdvisor reviews are discussed in Section 5.

2 Theoretical framework

Sentiment analysis (SA), also known as opinion mining (OM), refers to the analysis of people's opinions, attitudes, or emotions, in a written text. Note that SA is generally used in industry, while both SA and OM are used in academia. In the following, we interchangeably use the two terms. Opinions are usually published in specialised websites, devoted to peculiar topics like cinema, e-commerce, and so on.

The main goal of SA is to classify documents on the basis of their "polarity". The term polarity is used in linguistics for distinguishing affirmative and negative forms. For a wide review of the different methods of SA refer to [1] [7]. In literature there are three different steps in determining the polarity:

1. the subjectivity/objectivity of a text (SO-polarity): decide if a text has a factual nature or expresses an opinion on its subjective matter.
2. the positivity/negativity of a text (PN-polarity): decide if a subjective text expresses a positive or negative opinion.
3. the positivity/negativity strength of a text (PN-strength): identify different grades of positive or negative sentiments in opinions.

These steps are sequentially ordered, but it is not mandatory to perform all three.

Focusing on the unit of the analysis, we can consider different levels: a document-level, a sentence-level, an aspect-level. The first two levels are usually considered in the so called polarity-based SA, while the latter one is used in a topic-based perspective. The document-level aims at defining the polarity of each document, i.e. if it expresses a positive or a negative sentiment. In the sentence-level each document is segmented into sentences, and we want to determine the polarity of each sentence. The PN-polarity is quantified by considering a score of -1, 0 and 1 for negative, neutral and positive sentiment, respectively [2]. 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 [5]. The aspect-level SA aims at quantifying specific aspects and it allows to obtain fine-grained results. The aspect-level SA requires a greater computational complexity.

In this paper, we aim at determining the PN-polarity of a document, by considering a sentence-level approach. This is the first step of a mixed strategy that uses both textual and numerical information.

3 The Uffizi Gallery on TripAdvisor

In the last decades several private and public institutions operating in the field of cultural heritage, like museums, have looked at the visitors from a visitor satisfaction perspective. The so called museum audience is became strategically central, because it has a major connection to museums' sustainability. In this framework, it is more and more important to collect and analyse data coming from different sources. Together with classical sample surveys, carried out on a limited number of visitors, it is possible to use secondary data available on the Web. This huge amount of online data can be seen in a big data frame, as they have different natures and are available in real-time. In this paper, we study the audience of the Uffizi Gallery by analysing a set of reviews published on TripAdvisor.

TripAdvisor is a social media specialised in tourism reviews about both businesses and attractions. According to the most general classification of social media, it can be defined as a "reviews and ratings" media. It has been founded in U.S. in February 2000. Since mid-2010 is both an online service on the Web and a mobile application on portable devices. It has been one of the first websites to implement user generated content.

We use a scraping approach by launching a custom crawler on February 11^{th} 2017. In this way we retrieved 9639 reviews written in English and posted on TripAdvisor from February 27^{th} 2003 up to February 10^{th} 2017. The crawler has also provided some meta-information about the author of each review (e.g., location, contribution level on TripAdvisor, number of submitted reviews) and about the review itself (e.g., date, rating, device used for publishing the review). Here in the following we only focus our attention on the reviews and the corresponding ratings. We decide not to perform any lexical pre-treatment on the reviews. Only the parts not in English have been deleted, because some reviews also contained sentences in the mother-tongue language of the author.

4 A two-step strategy for a polarised rating/ranking

The rating scale used by TripAdvisor is an ordinal scale. In details, the ratings from 1 to 5 are associated with the terms *terrible*, *poor*, *average*, *very good* and *excellent*, respectively, and a corresponding number of bullets. In Fig. 1 it is possible to see the rating distribution of the Uffizi Gallery updated at April 5^{th} 2017. The ordinal rating can be seen as a global and comparable measure of the experience, while the textual description is an evaluation highlighting which aspects are positive and negative. Therefore, we propose a two-step strategy for the computation of a polarised rating of a review by combining the rate and the sentiment in the text.

Fig. 1 Visitor rating distribution at April 5^{th} 2017

Step 1: Computing the reviews' polarities

In order to compute the polarity of the reviews, we follow an SA sentence-level approach. This level seems to be more suitable, because in these texts each sentence includes an opinion of the contributor on the different aspects of the offered service.

The polarity scores have been calculated by using the R package *sentiment r*. The equation used in this package is based on the concept of valence shifters [8]. It is a procedure allowing to capture the polarity of a sentence by considering the context of use of its terms. The polarity of each term is weighted by taking into account negators (e.g., "never", "none"), amplifier and de-amplifier (e.g., "very", "few"), adversative and contrasting conjunctions (e.g., "but", "however"). This weighting system allows to emphasise or dampen the positivity and negativity of the terms, and obtain a more proper measure of the sentence sentiment.

Each review d_i (with $i = 1, \ldots, n$) is segmented into a set S_{d_i} of q_i sentences $\{s_{i1}, \ldots, s_{ij}, \ldots, s_{iq_i}\}$, by considering as separators only full stops, question marks and exclamation points. Each sentence j is represented as a sequence of its p_j terms $\{w_{ij1}, \ldots, w_{ijk}, \ldots, w_{ijp_j}\}$. Each term w_{ijk} in the sentence s_{ij} is compared with a lexicon of polarised terms, with a score $r_{w_{ijk}}$ of -1 for negative terms and 1 for positive terms, respectively. The terms not included into the lexicon are assumed to be neutral, with a score $r_{w_{ijk}}$ equal to 0.

The polarity score of each sentence depends on the dictionary of polarised terms used into the analysis, while the PN-polarity of the whole document depends on the polarities of its sentences. Different dictionaries are available. It is possible to consider manually created resources or automatically and partially automatically created resources. There are many papers in literature dealing with the problem of choosing one dictionary [4]. We use the Jockers dictionary, a lexicon of more than 10000 terms developed by the Nebraska Literary Lab for the R package *syuzeht* [3].

The final polarity score $r_{s_{ij}}$ of each sentence is computed as the sum of its weighted term scores $r*_{w_{ijk}}$ (taking into account the shifters) on the square-root of the sentence length:

$$r_{s_{ij}} = \frac{\sum_{k=1}^{p_j} r^*_{w_{ijk}}}{\sqrt{p_j}} \tag{1}$$

As we are interested in computing a polarity score for the whole review, we compute the score r_{d_i} of each document by a down-weighted zeros average of its sentence polarities. In this averaging function the sentences with neutral sentiment have minor weight:

$$r_{d_i} = \frac{\sum_{j=1}^{q_i} r_{s_{ij}}}{\tilde{q}_i + \sqrt{\log(2 - \tilde{q}_i)}} \tag{2}$$

where $\tilde{q}$ is the number of sentences with a positive or negative polarity. The logic of down-weighting neutral sentences is that they have less emotional impact in the review than the polarised ones.

Step 2: Computing the polarised rating

The new score for each contributor is obtained by summing the original rating with the polarity score of the review. Because of the unboundedness of the polarity scores, we bring all values into a range [0,1]. For each category c_h in the rating system (with $h = 1, \dots, H$), the $\hat{r}_{d_i}$ rescaled scores are computed as:

$$\hat{r}_{d_i} = \frac{r_{d_i} - \min_{d_i \in c_h} r_{d_i}}{\max_{d_i \in c_h} r_{d_i} - \min_{d_i \in c_h} r_{d_i}} \tag{3}$$

The resulting scoring system has a range [1,H+1], where 1 expresses the strongest criticism and H+1 expresses the strongest appreciation. The polarised rating can be interpreted as a ranking, because the new score allows the sorting of the reviews. Users can not only browse and read the reviews by rating, but also with respect to the sentiment.

5 Main results

After segmenting the 9639 reviews, we have obtained 48684 sentences. In Tab. 1 it is possible to see the statistics about the sentences with respect to the PN-polarity.

Table 1 Statistics on sentences by PN-polarity

	NEG	NEU	POS	ALL
sentence	7653	10072	30959	48684
token (N)	131307	113384	517841	762482
type (V)	6975	5597	9719	15524
hapax (V1)	3318	2827	4543	7228
type/token ratio	5.31%	4.94%	1.88%	2.04%
hapax/type ratio	47.57%	50.91%	46.74%	46.56%

We note that the number of positive sentences is much greater than the number of the negative ones. This result is consistent with the distribution of the rating expressed on TripAdvisor (see Fig. 1).

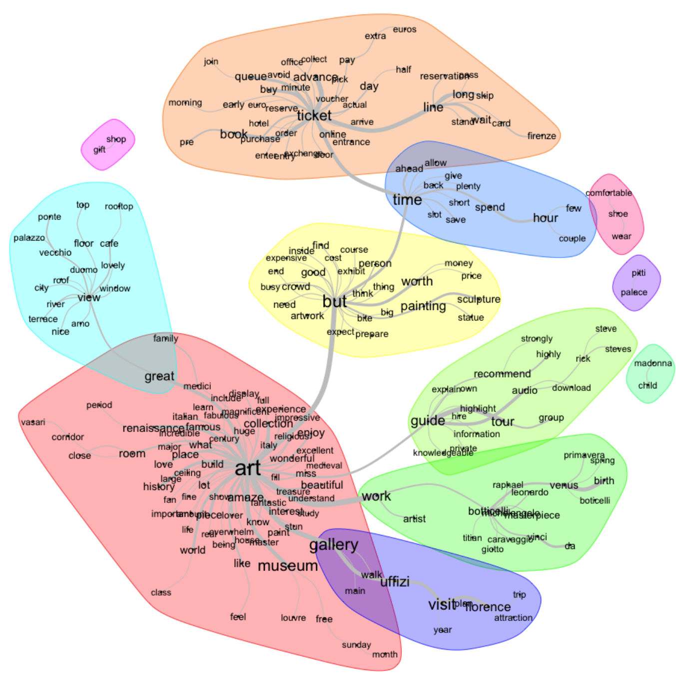

Fig. 2 Community detection on co-occurence network of terms: positive sentences

For visualising the peculiar language associated with positivity and negativity, we explore the *sub-corpora* of positive and negative sentences. After constructing the co-occurrence matrices, the relations among terms are visualised. For identifying a community of terms we consider the edge betweenness (through IRAMUTEQ[1]).

In Fig. 2 the communities related to the positive sentences are highlighted in different colours. We see that each community represents a topic related to the Uffizi experience. The main positive aspects are connected with the way the tickets have been bought, with the possibility of reserving a guided tour, with the different aspects related to the concept of Art, with the most important Masters in the gallery. We note the term "but" in the middle (in terms of betweenness). Its adversative role give, as seen above in Sec. 4, a different weight to the sentence polarities.

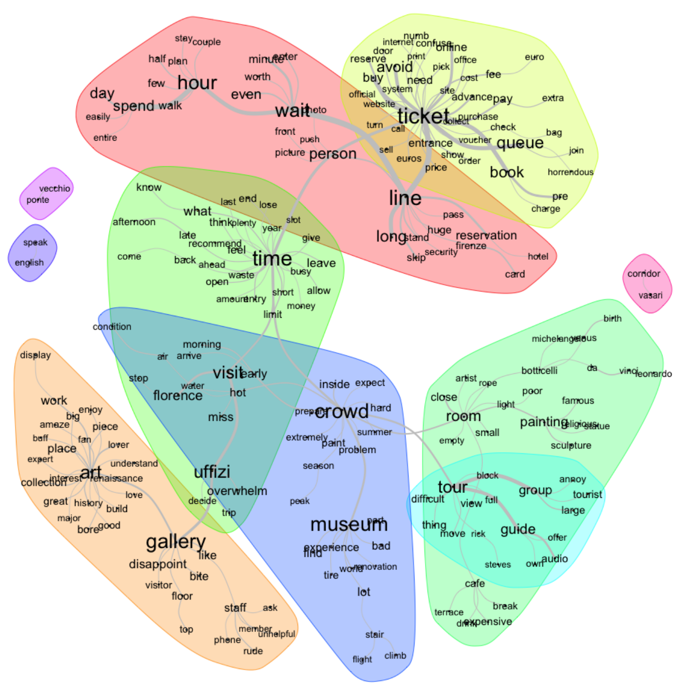

Fig. 3 Community detection on co-occurence network of terms: negative sentences

Analogously, in Fig. 3 the communities related to the negative sentences are highlighted. It is interesting to note that although we find some topics in common in the two networks, we find different paths. For example, "art" and "gallery" in the network of negative sentences are related to the inefficiency of the "staff", while in the network of positive sentences (Fig. 2) the same terms describe the visit experience.

[1] http://www.iramuteq.org/documentation

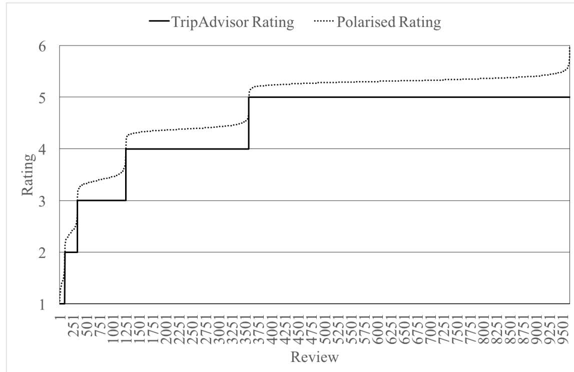

Fig. 4 Distributions of the TripAdvisor ratings and the polarised ratings

In Fig. 4 we show the distribution of the original ratings and the distribution of the polarised ratings. The new scale introduces a useful gradation in the judgments. Here in the following we can see two examples of reviews rated 1 by the contributor, and rated 1.0 and 1.9 by the polarised rating, respectively:

Review #2061: *I'm not sure why this museum is so famous, the truth is: it's extremely boring, full of statues and religious paintings, all the same, not even the building is nice!! The line up is insane, even if you buy tickets in advance, it's ridiculous, lots of people! Worthless!!! Save yourself the trouble, go browse Florence, so much to see outside. Totally waste of time and energy, nothing interesting, we were in and out!! Horrible!!*
Review #1121: *Buy your tickets online beforehand otherwise you will wait a long time in a queue. There is a very good rooftop cafe with reasonably priced food and drinks. Some spectacular photo opportunities through the windows overlooking Florence.*

References

1. Bing, L.: Sentiment Analysis: mining opinions, sentiments, and emotions. Cambridge University Press (2015)
2. Bing, L., Minqing, H., Junsheng C.: Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proceedings of the 14^{th} IW3C2 conference, pp. 342–352 (2005)
3. Jockers, M.L.: Syuzhet: Extract sentiment and plot arcs from text (2017) Available via https://github.com/mjockers/syuzhet
4. Marquez, F.B., Mendoza, M., Poblete, B.: Meta-level sentiment models for big social data analysis, Knowledge-Based Systems. **69**, pp. 86–99 (2014)
5. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: Rowe, M., Stankovic, Dadzie, A., Hardey, M. (eds) Proceedings of the Workshop on Making Sense of Microposts: Big things come in small packages, Vol. 718, pp. 93–98 (2011)
6. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43^{rd} Annual Meeting on Association for Computational Linguistics (ACL05), pp. 115–124 (2005)
7. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. In: Foundations and Trends in Information Retrieval, Vol. 2, pp. 1–135 (2008) doi: 10.1561/1500000011
8. Polanyi, L., Zaenen, A.: Contextual valence shifters. In: Shanahan, J., Qu, Y., Wiebe J. (eds) Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, Vol. 20, pp. 1–9. Springer, Dordrecht (2004)